COMILLAS
UNIVERSIDAD PONTIFICIA

Facultad de Ciencias Empresariales

LA CONEXIÓN ENTRE LAS OPORTUNIDADES LABORALES Y EL CRIMEN

¿En qué medida una mejoría en las oportunidades laborales afecta la criminalidad? ¿Qué evidencia hay al respecto?

Autor: Dolores Soubrié Guadalfajara
Director: Riccardo Ciacci

MADRID | Abril 2021

ÍNDICE

RESUMEN

El presente trabajo trata de profundizar en el estudio sobre la reincidencia y el desistimiento de los delincuentes. Concretamente, trata de mostrar las influencia que tiene las tasas de desempleo de un mercado concreto en la reincidencia de las personas que son puestas en libertad tras la comisión de un delito. El estudio se basa en un análisis de datos utilizando los métodos de Chi-cuadrado de Pearson y regresión logística, y concluye con una interesante conclusión: para los 26.020 delincuentes del Estado de Iowa que fueron puestos en libertad entre los años 2010-2015, cuanto mayor era la tasa de desempleo del mercado de Iowa, menor era el número de personas que cometían otro delito durante los 3 años siguientes a su puesta en libertad.

Palabras clave: Reincidencia, desistimiento, actividad criminal, empleo legal, costo de oportunidad, Chi-cuadrado Pearson, regresión logística, estadística, independencia, tasa de desempleo.

ABSTRACT

This paper aims to deepen the study of recidivism and desistance of offenders. Specifically, it tries to show the influence that the unemployment rates in a specific market have on the recidivism of people who are released from prison. It is based on a data analysis using Pearson's Chi-square and logistic regression methods, and concludes with an interesting finding: for the 26,020 offenders in the State of Iowa who were released between years 2010-2015, the higher the unemployment rate of the Iowa market, the lower the number of people who committed another crime during the 3 years following their release.

Key words: Recidivism, desistance, criminal activity, legal employment, opportunity cost, Pearson Chi-square, logistic regression, statistics, independence, unemployment rate.

INTRODUCCIÓN

a) Justificación del interés del tema

Durante estos últimos años he participado en varios proyectos con personas conflictivas, que han pasado tiempo de su vida encarceladas. Al tratar con estas personas de manera directa y al escuchar sus testimonios, empecé a tener gran curiosidad por profundizar más en estos temas, por estudiar sus patrones de comportamiento y aquellas variables que les llevan a reincidir, o, por el contrario, los ayudan a alejarse gradualmente de la vida conflictiva. A mi modo de ver, todavía queda un largo recorrido para alcanzar niveles de reincidencia aceptables en nuestra sociedad, y ello pasa por seguir estudiando e investigando sobre la mejor estrategia a implantar para que estos individuos queden completamente reinsertados en la sociedad. Además, estudio un doble grado en Derecho y Business Analytics, y estoy segura que, aunque el TFG esté enfocado en el análisis de datos, haber cursado asignaturas como Derecho Penal me permitirán tener una perspectiva diferente que complementará mi desarrollo sobre el tema. Realizar este TFG me aportará mucho como persona y como futura profesional, puesto que me servirá para corroborar, una vez más, que la exploración de datos y su interpretación es la herramienta más eficaz de la que disponemos para diseñar y proponer estrategias futuras, tanto a nivel empresarial como a nivel social.

b) Objetivos del TFG

El trabajo de investigación que se realiza pretende contestar a la siguiente pregunta: ¿*En qué medida una mejoría en las oportunidades laborales afecta la criminalidad*? ¿*Qué evidencia hay al respecto*?. Sin embargo, las "oportunidades laborales" es un concepto muy amplio que debe ser concretado, por lo que la primera parte del trabajo tiene como objetivo la definición de una pregunta de investigación más concreta a través de un estudio cualitativo del campo. Para este estudio se acudirán a estudios realizados durante los últimos años en relación al tema objeto de estudio. Una vez determinada la pregunta de investigación, se trata de dar respuesta a la misma a través de un análisis de datos llevado a cabo con el software R y con distintas herramientas estadísticas que se eligen en base a las características del dataset escogido.

c) Metodología

El presente trabajo se divide en tres fases, que se detallan a continuación:

La primera fase comienza con una breve introducción sobre la actividad criminal. Se expone la situación actual del fenómeno de la reincidencia de reclusos, así como de la importancia de la función del costo de oportunidad en cuanto a la actividad criminal se refiere.

La segunda fase reside en una revisión de la literatura, fundada en diferentes artículos publicados y de gran interés sobre el tema, y centrada en el efecto que puede provocar un aumento en el acceso a un empleo legal sobre el desistimiento de este colectivo. Esta segunda fase es la que ayuda y guía a formular la pregunta concreta de investigación relacionada con el título del trabajo.

La tercera y más relevante fase consiste en el análisis de una base de datos con R, una herramienta de programación que posee un importante enfoque al análisis estadístico. Se realiza un breve resumen sobre la metodología utilizada para llevar a cabo este análisis de datos y, asimismo, se presentan los resultados obtenidos, tratando de dar una respuesta completa a la pregunta formulada. Para ello, se hace uso de dos técnicas estadísticas, la primera la prueba de Chi- cuadrado de Pearson, y la segunda, para superar algunas de la limitaciones de la primera prueba, la regresión logística simple. Este análisis de datos incluye, tanto la ejecución del código como la interpretación de resultados, siendo esta última la parte más relevante de todo el trabajo ya que nos permitirá concluir con una respuesta clara a la pregunta de investigación.

CAPÍTULO 1: LA ACTIVIDAD CRIMINAL

1. LA REINCIDENCIA

La reincidencia tras el encarcelamiento es un problema grave y extendido mundialmente. Durante estos últimos años, el Bureau of Justice Statistics (BJS), organismo encargado de recoger informes estadísticos sobre la justicia estadounidense, ha publicado varios informes sobre la reincidencia de los ex reclusos en las cárceles y prisiones de Estados Unidos. Uno de estos informes estudia los datos sobre presos estadounidenses liberados en 2005 en 30 estados. Según dicho informe, el 67.8% de los 404.638 reclusos puestos en libertad en 2005 volvieron a ser arrestados por cometer un crimen durante los tres años siguientes a su liberación, y el 76,6% fueron arrestados dentro de los cinco años de la puesta en libertad. (Durose, Cooper, & Snyder, 2014)

Los datos arrojados sobre la reincidencia son preocupantes. Durante las últimas décadas se ha ido generando un gran interés público dirigido a frenar el círculo de criminalidad y reincidencia, y muchas han sido las personas dedicadas a estudiar esta cuestión. Numerosos estudios tratan de evidenciar las razones por las que los reclusos reinciden, sin embargo, no es trabajo sencillo. Todo lo que ocurre antes y después de cometer un crimen es desconocido, lo que hace muy complicada dicha labor. Igualmente, a medida que se va investigando más y se van realizando estudios, se entiende en mejor medida las factores que llevan a estas personas a reincidir, pero las conclusiones de los estudios nos siempre coinciden. Esto hace que la implantación de estrategias para hacer frente a la reincidencia sea una ardua tarea.

En esta misma línea, las investigaciones muestran que las políticas de justicia para combatir la delincuencia han fracasado en muchos países, ya que, generalmente, las tasas de reincidencia se caracterizan por ser altas. Entre las principales dificultades se encuentran el elevado coste que se necesita para llevar a cabo políticas de reinserción o los prejuicios que rodean a estas personas en el entorno social en el que viven. En este sentido, Noruega es de los pocos países que ha conseguido lograr que sus esfuerzos dirigidos a transformar a sus criminales en no reincidentes se materialicen. Hace 20 años, cuando Noruega presentaba altas tasas de reincidencia, decidió distanciarse del enfoque "punitivo" de los encarcelamientos y someterse a serias reformas para bajar dichos porcentajes. Los funcionarios de las cárceles adoptaron el papel de mentores y modelos a seguir para los prisioneros y se ofreció a los prisioneros programas diarios de educación. Con estas reformas, la reincidencia en el país se redujo enormemente, hasta alcanzar valores apenas de un 20% después de dos años (Kirby, 2019). Sin embargo, como se ha dicho al principio, esta exitosa estrategia no es común entre los demás países, sino que más bien se trata de una excepción.

Dicho esto, el interés por encontrar acciones dirigidas a fomentar la reinserción social y desistimiento delictivo no cesa. Aunque no sea tarea fácil, las altas tasas de reincidencia hacen necesario y útil profundizar en la realidad de los ex reclusos y en sus trayectorias una vez salen de su encarcelamiento. A continuación, se detalla la cuestión relevante al costo de oportunidad del delito, una cuestión clave en el fenómeno de la reincidencia, pues es una de las maneras más sencillas de entender los comportamientos de los reclusos una vez son puestos en libertad.

1.1.Costo de oportunidad del delito

En cuanto a la forma de alentar el desistimiento criminal, encontramos numerosos enfoques, entre los cuales destacan los siguientes cinco: 1) modificar las penas; 2) modificar las probabilidades de ser arrestado; 3) modificar el costo de oportunidad del delito; 4) asistir a los individuos a evitar influencias criminógenas; o 5) modificar las preferencias sobre el comportamiento legal frente al ilegal (Doleac J. , 2019)

Centrándonos en la modificación del costo de oportunidad del delito como forma de prevención de la reincidencia, podemos definir el costo de oportunidad del delito como el beneficio neto de la actividad jurídica a la que se renuncia mientras se planifica, ejecuta y oculta el acto delictivo. Cuanto más bajo es el nivel de ingresos de un individuo, más bajo es su costo de oportunidad de participar en una actividad ilegal. Esto es porque, un individuo con pocos ingresos que comete una actividad criminal, no renuncia a importantes ingresos legales. Consecuentemente, el beneficio neto de su actividad criminal será mayor que el de un individuo con mayores ingresos (Becker, 1968). Normalmente, cuanto menor es el costo de oportunidad de delito de un individuo, mayor reincidencia encontramos, pues el individuo renuncia prácticamente a nada cometiendo un segundo delito. Sin embargo, si conseguimos como sociedad que el costo de oportunidad del delito sea alto para la mayoría de los individuos, la reincidencia se vería disminuida considerablemente. Esto es porque los individuos verían mucho más atractivos otros comportamientos legales antes que la comisión de un delito, ya que tendrían algo que perder. Para ello, es necesario entender los componentes de esta función.

1.2.Función costo de oportunidad del delito

Dados los beneficios particulares que se esperan del comportamiento criminal, podríamos disuadir futuros delitos aumentando la función de costo de oportunidad del delito (U_{nc}). Definimos dicha función de la siguiente forma:

$$U_{nc} = f(v, w, x, z)$$

Siendo *v*, los costes materiales necesarios para participar en una actividad no delictiva (ej. medios de transporte para acudir a un puesto de trabajo legal); *w*, salarios

legales; *x*, la inutilidad generada por realizar el trabajo requerido para ganar dichos salarios; and *z*, cualquier beneficio psíquico derivado de una actividad no delictiva (ej. orgullo y satisfacción de los contribuyentes a la sociedad).

Disuadir el crimen a través de la función de coste de oportunidad criminal supone incrementar U_{nc}, es decir, aumentar las variables *w* and *z*, y disminuir las variables *v* y *x*. Tal y como expone en el artículo "*Encouraging desistance from crime"*, hay diferentes maneras de incrementar el coste de oportunidad del delito. Éstas son las siguientes: facilitar e incentivar la inversión en capital humano, aumentar los salarios legales, aumentar el acceso a un empleo legal, ofrecer complementos de apoyo al empleo e incrementar el acceso a la asistencia pública. (Doleac J. , 2019)

Por ejemplo, si aumentan los salarios legales los individuos que salen de prisión tendrán mayor facilidad de reinserción en la sociedad que les rodea y menos necesidad de cometer otro delito, pues disponen de mayores salarios que les permiten mejorar su calidad de vida. Sobre esta cuestión podemos encontrar abundante literatura, pero el presente trabajo tiene limitaciones y sería imposible ofrecer un estudio de la literatura completo sobre la función del costo de oportunidad del delito. Por ello, este trabajo se enfoca en estudiar concretamente una de los medios a través del cual se puede incrementar el coste de oportunidad: el aumento del acceso al empleo legal. Como se ha expuesto anteriormente, el análisis cualitativo de esta parte será clave para precisar la pregunta de investigación más concreta que se tratará de responder posteriormente.

CAPÍTULO 2. ESTUDIO DE LA LITERATURA: AUMENTO DEL ACCESO A UN EMPLEO LEGAL

A pesar de que un grupo de trabajadores presente las mismas cualidades para desempeñar un trabajo concreto, varios estudios se han encargado de evidenciar que los trabajadores con algún tipo de antecedente penal suelen ser discriminados a la hora de ser elegidos para puestos de trabajos. Esto ocurre, no solo cuando los trabajadores presentan las mismas cualidades, sino incluso en supuestos en los que los trabajadores con algún tipo de antecedente están mejor cualificados que otros trabajadores sin antecedentes criminales (Agan & Starr, 2019). El estudio llevado a cabo por (Pager, 2003), revela un mecanismo de estratificación importante y muy poco reconocido, confirmando que los

antecedentes penales suponen un gran obstáculo para el empleo, con importantes implicaciones para las disparidades raciales.

Según (Doleac J. L., 2016), una de la principales razones por las que los empleadores se muestran reacios a contratar a este grupo de personas es la cuestión concerniente a la responsabilidad legal. En el caso de que alguno de estos trabajadores cometa otro delito mientras está en su puesto de trabajo, los empleadores que han tomado la decisión sobre su incorporación pueden enfrentarse a cargos contra la diligencia debida en procesos de selección, o simplemente ganarse una mala reputación en el sector. En este sentido, atender a las inquietudes de los reclutadores es un paso crucial para diseñar intervenciones que aumentan el empleo de este grupo de personas. Tal y como confirma (Doleac J. L., 2016), es imprescindible adoptar un enfoque múltiple que incluya no solo estrategias destinada a desarrollar las habilidades de los trabajadores poco cualificados, sino también políticas destinadas a la comprensión de la interacción de los empleadores con los solicitantes a puestos de empleo con antecedentes penales.

Una de las campañas estadounidenses más destacadas para hacer frente a esta cuestión es la política "Ban the Box", una política que surgió en Hawái a finales de los años 90 y cobró fuerza en otros estados de Estados Unidos tras la recesión de 2007-2009. Esta política restringe a todos los reclutadores la posibilidad de preguntar sobre los antecedentes penales de los solicitantes a los puestos de trabajos. Sin embargo, recientes investigaciones sobre su impacto muestran algunas de las deficiencias de esta política. En primer lugar, (Agan & Starr, 2019) y (Doleac & Hansen, 2020) destacan que esta política no aborda concretamente la cuestión relativa las preocupaciones de los reclutadores sobre la contratación de este colectivo, lo que puede conducir a una discriminación de aquellos grupos que contienen un mayor número de personas con antecedentes penales. De esta manera, ocultar la información sobre los antecedentes penales podría fomentar la discriminación racial: los reclutadores pueden hacer suposiciones sobre la criminalidad basadas en la raza del solicitante. Otros estudios, como aquellos impulsados por (Jackson & Zhao, 2017) revelaron que la política "Ban the Box" no aumenta el empleo entre las personas con antecedentes penales.

Por otro lado, algunas jurisdicciones ofrecen certificados de rehabilitación emitidos por los tribunales que ratifican la capacidad de las personas con antecedentes penales para asumir la responsabilidad que un puesto de trabajo conlleva. Además, estos certificados aseguran una protección frente a la responsabilidad legal de los reclutadores

para el caso de que el contratado vuelva a cometer un delito. En este sentido, podríamos decir que estos certificados se encargan de trasladar el riesgo de los reclutadores a los tribunales que emiten dichos certificados. El estudio dirigido por (Leasure & Andersen, 2016) demuestra que los candidatos que poseen certificados de rehabilitación tienen las mismas probabilidades de recibir una llamada del reclutador que los candidatos sin ningún tipo de antecedentes penales.

Por último, resulta interesante hacer referencia al efecto que tiene un aumento de la visibilidad de los puestos de trabajos disponibles en el mercado laboral. En esta línea, (Galbiati, Ouss, & Philippe, 2021) estudian cómo las condiciones del mercado laboral y la información sobre la disponibilidad de puestos de trabajo afectan a la reincidencia tras el encarcelamiento, manteniendo constante el número real de puestos de trabajo disponibles en ese momento. Para su estudio combinaron datos administrativos a nivel individual sobre individuos ex reclusos en Francia con datos diarios a nivel condado sobre nuevas ofertas de puestos de trabajo y sobre su cobertura periodística. Los autores alcanzaron con dos principales conclusiones: en primer lugar, que la cobertura mediática de la creación mediática del empleo reduce la reincidencia, por lo que aquellas políticas que promueven el acceso a la información sobre las oportunidades laborales pueden contribuir a reducir la reincidencia de los recién salidos de un encarcelamiento. En segundo lugar, también encontraron que en Francia las mejores oportunidades en los puestos de trabajo de la industria manufacturera en el momento de la puesta en libertad de estos individuos reduce la reincidencia. Dicho esto, los autores afirman que existe una heterogeneidad en cuanto a los tipos de empleo que afectan a la reincidencia. Otros estudios llevados a cabo por autores como (Schnepel, 2018) y (Yang, 2017) refuerzan esta última tesis confirmada por Gabialti, Ouss y Philippe.

CAPÍTULO 3. PREGUNTA DE INVESTIGACIÓN

Habiendo analizado el estado de la cuestión relativo a los efectos que tiene el acceso a un empleo legal sobre la reincidencia, resulta interesante estudiar en qué medida afecta el número de puestos de trabajos disponibles que existen en el mercado en el momento de la puesta en libertad. En la línea de lo expuesto en el epígrafe anterior, parece lógico pensar que si un aumento de la visibilidad de los puestos de trabajo afecta positivamente a la reincidencia, la existencia de una mayor oferta laboral también afectaría de manera positiva a este fenómeno.

Para poder estudiar esta relación, en el presente trabajo se hace uso de la tasa de desempleo, indicador que expresa la proporción de personas que se encuentran sin un puesto de trabajo respecto del total de personas activas[1] de la población. Así, trataremos de evidenciar que, cuanta mayor es la tasa de desempleo, más dificultades tienen los ex reclusos de acceder a un puesto de trabajo legal porque menor es la oferta laboral en el mercado y, por ende, mayor tasa de reincidencia habrá. Consecuentemente y en base a lo anterior, cuanto menor sea la tasa de desempleo en un momento dado, menor será la dificultad para acceder a un puesto de empleo legal pues la competencia es menor, y, por tanto, el fenómeno de la reincidencia se observará en menor medida.

Dicho esto, podemos concluir que la pregunta de investigación concretada para llevar a cabo el análisis de datos es la siguiente:

¿En qué medida una mayor tasa de desempleo afecta a la criminalidad de los recién salidos de prisión?

CAPÍTULO 4. ANÁLISIS DE DATOS

1. CONJUNTO DE DATOS

Para el análisis de datos realizado en el presente trabajo, se ha obtenido una base de datos del portal de datos del Estado de Iowa, Estados Unidos[2], a través de una búsqueda en internet. La nube es una herramienta muy útil para encontrar bases de datos. A mi disposición tenía varios conjuntos de datos sobre la reincidencia de los ex reclusos de diferentes países. Sin embargo, no todas disponían de datos concretos que me facilitasen la ejecución del trabajo, pues no todos los datos sirven para hacer un análisis que pretende responder a una pregunta en concreto. Por este motivo, escogí este dataset que era el que mejor se adaptaba para la elaboración del trabajo.

El conjunto de datos informa sobre la reincidencia de los delincuentes durante los tres años siguientes a su puesta en libertad y recoge datos de los delincuentes que son puestos en libertad en el Estado de Iowa entre 2010 y 2015, con el seguimiento de la reincidencia entre 2013 y 2018. Las variables en el conjunto de datos son las siguientes:

[1] Entendemos por personas activas todas aquellas personas en edad de trabajar que bien trabajan en un empleo remunerado o bien se encuentran en búsqueda del mismo

[2] Disponible en: https://data.iowa.gov/Correctional-System/3-Year-Recidivism-for-Offenders-Released-from-Pris/mw8r-vqy4

- "*Fiscal year released*": año en el que el recluso fue puesto en libertad. Como se ha mencionado en el párrafo anterior, los años van de 2010 a 2015.

- "*Recidivism Reporting Year*": año que marca el final del periodo de seguimiento de 3 años. Por ejemplo, los delincuentes que salieron de prisión en el año 2012, se encuentran en el año de informe de reincidencia 2015. Es preciso resaltar que, por la configuración de los datos, esta variable y la anterior nos proporciona la misma información.

- "*Race-Ethnicity*": raza/etnia del delincuente, entre las que encontramos: American Indian or Alaska Native (Hispanic y Non-Hispanic), Asiano r Pacific Islander (Hispanic y Non-Hispanic), Black (Hispanic y Non-Hispanic) y White (Hispanic y Non-Hispanic)

- "*Age At Release*": edad del delincuente en el momento de su puesta en libertad. Clasifica a los delincuentes en 5 grupos de edad: menores de 25 años, entre 25-34 años, entre 35-44 años, entre 45-54 años y mayores de 55 años

- "*Convicting Offense Classification*": penas máximas a las que tiene que hacer frente el delincuente, entre las que se incluyen penas de cadena perpetua, de 25-50 años, de 10 años, de 5 años, de 2 años, de 1 año y de 30 días.

- "*Convicting Offense Type*": categoría general del delito más grave por el que el delincuente fue encarcelado (delitos de drogas, propiedad, violencia, orden público u otros).

- "*Convicting Offense Subtype*": clasificación adicional del delito más grave por el que el delincuente fue encarcelado. Esta variable tiene a su vez 26 categorías de delitos entre los que se encuentran asalto, vandalismo, delitos sexuales, tráfico de drogas, robo, etc.

- "*Main Supervising District*": distrito judicial que supervisa al delincuente durante el mayor tiempo del periodo de seguimiento.

- "*Release Type*": motivo por el que el delincuente salió de prisión. Entre las categorías de esta variable encontramos el cumplimiento de la sentencia,

libertad condicional, sentencia especial y distintos tipos de libertad condiciona a detención en función de la geografía.

- "*Release Type: Paroled to Detainder united*": misma información que la variable anterior, pero aglomerando el tipo de libertad condicional a detención en una única categoría.

- "*Part of Target Population*": variable que recoge si los delincuentes forman parte de la población objetivo de las estrategias específicas para reducir las tasas de reincidencia de los presos que están en libertad condicional, dirigidas por el Departamento de Instituciones Penitenciarias .

- "*Recidivism – Return to Prison numeric*": variable que recoge si hay ingreso en prisión por cualquier motivo dentro del periodo de seguimiento de 3 años desde la puesta en libertad del delincuente. Es una variable binaria que recoge la información con valores 0 o 1, siendo 0 indicativo de que el individuo no cometió ningún delito tras su puesta en libertad, y siendo 1 indicativo de que el individuo reincidió en los 3 años siguientes a su puesta en libertad.

Este dataset no recoge ninguna información relativa a la tasa de desempleo del estado de Iowa durante estos años. Por este motivo, para poder llevar a cabo el análisis, se ha escogido otro conjunto de datos disponibles en la página web del *Federal Reserve Bank of St. Louis*, que recoge la tasa de desempleo en Iowa desde 1976[3]. Para el presente trabajo, únicamente se ha descargado los datos de los años objeto del otro conjunto de datos para poder trabajar con ambos datasets. Sin embargo, la frecuencia de los datos de este último dataset es mensual, por lo que, previa la unión de ambos datasets, se ha procedido a calcular la media anual de las tasas mensuales de desempleo. Una vez realizado obtenidos los datos de reincidencia en frecuencia anual, ambos datasets se han fusionado para tener toda la información en un único conjunto de datos y comenzar con el análisis en R. Por tanto, el dataset con el que se va a trabajar, además de todas las variables mencionadas, tiene una última variable que recoge la tasa de desempleo del mercado de Iowa en términos porcentuales para el año en el que los individuos fueron puestos en libertad.

[3] Disponible en: https://fred.stlouisfed.org/series/IAUR

2. METODOLOGÍA

Como se ha expuesto en las primeras líneas del trabajo, la metodología que se ha elegido para proceder con este análisis es una metodología del tipo cuantitativa, concretamente de métodos estadísticos. Este método resulta idóneo ya que el propósito del trabajo es la comprobación, en el grupo de reclusos recién salidos de prisión, de la consecuencia deducida de una hipótesis general de investigación en relación con la tasa de desempleo del mercado local. Concretamente, la hipótesis para el análisis de datos es la siguiente: el fenómeno de la reincidencia es independiente de la tasa de desempleo que haya en el mercado laboral local en el momento de cometer el delito.

Las características que adoptará el procedimiento propio del método estadístico dependen del diseño de la investigación y de los datos que tenemos a nuestra disposición. Para el presente trabajo, y teniendo en cuenta las características de nuestros datos y de las variables con las que queremos trabajar, se llevará a cabo a través del método estadístico de Chi-cuadrado de Pearson y de la regresión logística. Estos métodos se introducirán en sus epígrafes correspondientes y de igual manera se explicaran los conceptos que los respaldan y las explicaciones matemáticas en las que se sustentan. A través de esta explicación, se configurará la revisión y profundización de estas dos técnicas esbozadas durante las asignaturas de grado de Business Analytics.

Habiendo introducido el conjunto de datos con el que se va a trabajar y la metodología a utilizar, a continuación se procede con el análisis de datos ayudándonos de la herramienta R. Esta herramienta es un software apto para el análisis estadístico y gráfico que nos ayudará a dar respuesta a la pregunta de investigación. Resulta muy útil ya que es una herramienta gratis, de código abierto y muy flexible que permite ser ampliada a través de la instalación de paquetes, librerías o la definición de funciones propias. Además, permite la creación de gráficos de alta calidad que nos ayudaran a entender mejor los datos con los que trabajamos y a presentar los resultados obtenidos del análisis. Para el presente trabajo, se ha creado un R script con un código específico a través de conocimientos que he adquirido en el grado de Business Analytics y de otros conocimientos adquiridos para la realización del presente trabajo a partir de la lectura de manuales interesantes sobre el tema. A continuación, se irá desgranando el código para poder explicar las técnicas estadísticas utilizadas y las conclusiones alcanzadas.

3. ANÁLISIS

3.1.Estudio y limpieza del dataset

Para comenzar el estudio de datos, lo primero que se debe realizar es cargar los datos en el directorio y guardarlos para poder trabajar con ello. En este caso, se ha guardado bajo el nombre "dataset". Una vez realizado esto, podemos utilizar la función nrow(), ncol() y dim() para conocer las dimensiones (número de columnas y número de filas) del dataset con el que vamos a trabajar.

```
dataset <- read_delim("3 Year_Recidivism_for_Offenders_Released_from_Prison_in_Iowa_elaborated.csv",";", escape_double = FALSE, trim_ws = TRUE)

nrow(dataset)

ncol(dataset)

dim(dataset)
```

Al ejecutar esta parte del código, obtenemos que nuestro dataset está compuesto por 26.020 filas, es decir, 26.020 observaciones. Tenemos información sobre 26.020 ex reclusos que nos ayudará a sacar las conclusiones oportunas. Este número es considerable, sin embargo, cuanto mayor número de observaciones se tiene para un análisis, mejor resultados se obtienen. Esto es porque se recogen mayor número de supuestos y es más difícil que la muestra este sesgada por alguna razón. De igual manera, hallamos que existen 13 columnas que se corresponden a las 13 variables mencionadas en el apartado anterior. Por tanto, el dataset es una matriz de datos de una dimensión 26060x13.

Para poder trabajar de una manera más sencilla, resulta conveniente crear un sub dataset que incluya únicamente las variables que nos sean de interés y que nos ofrezcan datos interesantes para entender el comportamiento de los reclusos y los resultados de nuestro análisis. Así, tener un dataset más pequeño y limpio nos permitirá trabajar mejor. A través de la función select(), nos hemos desprendido de las variables `*Main Supervising District*`, `*Part of Target Population*`, `*Release Type*`, `*Convicting Offense Subtype*` y `*Convicting Offense Classification*`.

```
dataset<- select(dataset, -`Main Supervising District`, -`Part of Target Population`,-`Release Type`,-`Convicting Offense Subtype`,-`Convicting Offense Classification`)

ncol(dataset)

str(dataset)
```

Tras la ejecución de esta parte del código, se crea un sub dataset formado únicamente por ocho variables. Estas ocho variables son las siguientes: `Fiscal Year Released`, `Recidivism Reporting Year`, `Race - Ethnicity`, `Age At Release`, `Convicting Offense Type`, `Release type: Paroled to Detainder united`, `Recidivism - Return to Prison numeric`, `Unemployment rate`. Gracias a la función str(), podemos analizar la estructura de las variables mencionadas, y observamos que hay 4 variables numéricas y 4 variables categóricas. Las variables numéricas que representan números y por tanto, con ellas se pueden llevar a cabo operaciones aritméticas. Las variables numéricas a su vez pueden ser continuas o discretas. Por ejemplo, la medida de una persona sería una variable continua, pues puede adoptar cualquier medida en el marco del intervalo en cuestión. Las variables discretas no adoptan cualquier medida, sino que se dan separaciones entre los valores que puede tomar la variable. Un ejemplo de esta última sería el número de hijos que tiene un individuo. Por otro lado, son variables categóricas aquellas que se enmarcan dentro de un número finito de categorías o grupos. Por ejemplo, si la variable representa el color de un objeto, los valores pueden ser asociados a las diferentes clases de colores que existan.

Dicho esto, para nuestro estudio de datos es necesario convertir algunas de las variables, pues aunque R las clasifique como numéricas o categóricas, muchas veces pueden estar mal denominadas. Por ejemplo, a pesar de que la variable `Recidivism - Return to Prison numeric` tome valores numéricos (1 y 0), esta variable recoge información cualitativa, pues hace referencia a si los reclusos han reincidido o no han reincidido en los 3 años siguientes a su puesta en libertad. Entonces, se trata en realidad de una variable categórica que divide los datos en reincidentes o no reincidentes, y por ello debe ser transformada. Para ello, acudimos a la función as.factor() que permite convertir las variables en factor, un tipo de dato estadístico en R que almacena variables categóricas.

```
str(dataset)
dataset$`Fiscal Year Released`<-as.factor(dataset$`Fiscal Year Released`)
dataset$`Recidivism Reporting Year`<-as.factor(dataset$`Recidivism Reporting Year`)
dataset$`Race - Ethnicity`<-as.factor(dataset$`Race - Ethnicity`)
dataset$`Age At Release`<-as.factor(dataset$`Age At Release`)
dataset$`Convicting Offense Type`<-as.factor(dataset$`Convicting Offense Type`)
```

```
dataset$`Release type: Paroled to Detainder united`<-as.factor(dataset$`Release type: Paroled to Detainder united`)

dataset$`Recidivism - Return to Prison numeric`<-as.factor(dataset$`Recidivism - Return to Prison numeric`)
```

La única variable que no hemos convertido a factor es la variable `Unemployment rate`, ya que recoge la información sobre la tasa de desempleo; un valor numérico que no precisa ser convertido a categórico.

El siguiente paso que se ha llevado a cabo para la preparación de la base de datos es la comprobación y la eliminación de los missing values. Los valores perdidos, que son muy habituales en los conjuntos de datos, existen en aquellos casos en los que no se almacena ningún valor para la variable de una observación. Estos valores perdidos aparecen por diversas razones, entre las que se encuentran la introducción manual de datos, los errores de los equipos o las mediciones incorrectas. Es importante trabajarlos, pues pueden tener un efecto significativo a la hora de trabajar con los datos ya que modifican los análisis y las estadísticas, dificultando la labor de interpretación de los analistas. Por este motivo, resulta importante identificar cuántos hay en nuestro conjunto de datos para así eliminarnos y quedarnos con un dataset limpio. Para ello, llamamos a la función is.na() de R, que al ejecutarla nos devuelve TRUE o FALSE dependiendo si efectivamente existen missing values o no para cada observación.

```
is.na(dataset) #vemos que tenemos valores perdidos

mean(is.na(dataset)) #concretamente un 0.8%

dataset<-na.omit(dataset) #borramos los valores perdidos

mean(is.na(dataset)) #los datos limpios ya no tienen NAs
```

Podemos observar que, concretamente, un 0.8% de los datos del dataset está formado por valores perdidos. No son muchos, pero teniendo en cuenta lo que se ha dicho en el párrafo anterior, es importante deshacernos de ellos. Como se ha dicho, para que los resultados no queden influenciados por estos datos, se ha procedido a su eliminación a través de la función na.omit(). Tras su eliminación, la media del valor de missing values en el conjunto de datos es 0 y podemos afirmar que nuestros datos están limpios y no se verán afectados por valores perdidos.

El siguiente paso dentro de nuestro análisis es tratar de conocer los datos con los que estamos trabajando. Tal y como he aprendido durante mi grado, una de las cosas más

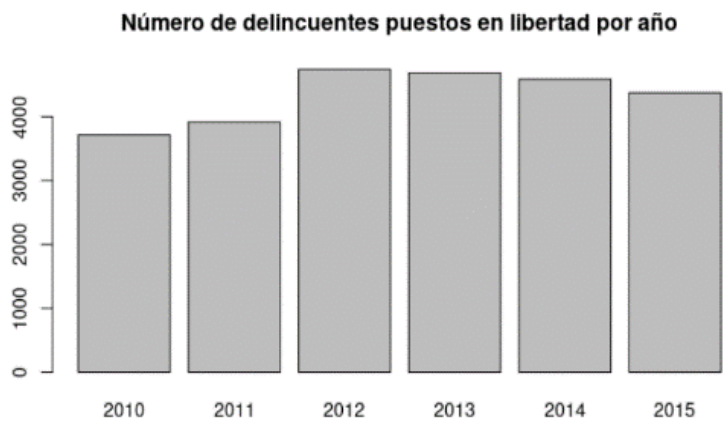
importante es tener contexto de los datos con los que se trabaja, pues puede ser de gran utilidad en la parte final de las conclusiones. Ejecutando la función describe() obtenemos el número de observaciones que entran dentro de cada una de las categorías de cada variable, así como la proporción de las mismas. Sin embargo, resulta conveniente analizar una por una y obtener así un conocimiento más exhaustivo del conjunto de datos en cuestión.

En primer lugar, se crea una tabla ("tabla 1") para ver el número de delincuentes que han sido puestos en libertad cada año. A través de técnicas de visualización, se obtiene un gráfico de barras que nos permite apreciar la proporción de delincuentes puestos en libertad cada año. Para el trabajo, haremos uso de diferentes tipos de visualizaciones, que como aprendido en mi grado de Business Analytics, siempre proporcionan una información más clara y visual que ayuda al entendimiento del análisis.

```
tabla1<- table(dataset$`Fiscal Year Released`)
barplot(tabla1, beside=TRUE)
title(main='Número de delincuentes puestos en libertad por año')
```

Obtenemos que el conjunto de datos recoge para el año 2010 un total de 3.716 observaciones, para el año 2011 3.917 observaciones, para el 2012 4.740 observaciones, para el año 2013 4687 observaciones, para el año 2014 4.586 observaciones y, finalmente, para el año 2015, 4.374 observaciones. Por tanto, el año que mayor número de los delincuentes fueron puestos en libertad en el año 2012. Podemos observar en el siguiente gráfico como el número de delincuentes puestos en libertad durante los últimos 4 años que disponemos no varía considerablemente:

Figura 1: Número de delincuentes puestos en libertad por año



Para la variable `Race - Ethnicity` se ha procedido de igual manera, mediante la creación de una tabla que recoge el número de observaciones correspondientes a cada categoría. El resultado de ejecutar la tabla 2 nos muestra que el dataset está formado principalmente por personas que encajan dentro de la categoría "White-Non Hispanic", ya que es la categoría que representa el 67,57% de las observaciones. Le sigue los delincuentes pertenecientes a la categoría "Black-Non Hispanic" que representan el 23.5% de la muestra. Resulta interesante esta variable porque como mencionábamos en el epígrafe del estudio de la literatura, políticas como la de "Ban the Box" en Estados Unidos pueden fomentar la discriminación racial, pues los reclutadores pueden, al no poder preguntar sobre los antecedentes criminales de los candidatos, hacer suposiciones sobre su criminalidad basadas en la raza del solicitante. Tal y como expresan (Agan & Starr, 2019) en su estudio, antes de la política "Ban the Box", los solicitantes de raza blanca recibían un 7% más de devoluciones de llamadas por parte de los reclutadores que los solicitantes similares de raza negra. Sin embargo, tras la política "Ban the Box", la diferencia aumentó hasta el 43%, corroborando así la creencia de que esta política aviva la discriminación racial.

De acuerdo con esta política y los resultados que hemos estudiado en el epígrafe anterior sobre la influencia del acceso a un empleo legal en la reincidencia, podríamos suponer a priori que los delincuentes de raza negra que salen de prisión tienen mayor probabilidad de reincidir pues tienen una menor probabilidad de acceder al mercado laboral por temas de discriminación racial. Esta sería cuestión de otro estudio adicional, sin embargo, resulta interesante ver la proporción de delincuentes que reinciden según su raza/etnia. Para poder obtener dichos resultados, creamos una segunda tabla llamada "tabla2.2" que recoja las frecuencias por cada categoría.

```
tabla2.2<- table(dataset$`Recidivism - Return to Prison numeric`,dataset$`Race - Ethnicity`)
```

Al ejecutar el código adjuntado, podemos observar cómo dentro de la categoría "Black" que incluye tanto "Black", como "Black-Hispanic" y "Black-Non-Hispanic", un 33.39% de un total de 6.148 obversaciones son reincidentes. Por otro lado, dentro de la categoría "White", que engloba "White", "White-Hispanic" y "White-Non-Hipanic", un 33.35% de un total de 19.118 observaciones son reincidentes. De acuerdo con nuestros datos, y sin realizar un estudio detenido de los datos, comprobamos que para nuestro

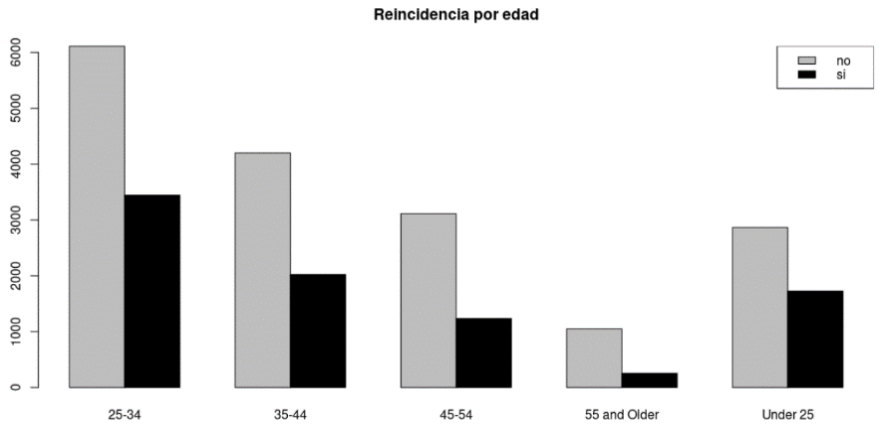
dataset, la raza/etnia no influye en el fenómeno de la reincidencia de los delincuentes puestos en libertad, pues la proporción es prácticamente la misma.

La siguiente variable interesante de estudiar es la variable edad. Es una variable categórica, que categoriza a los delincuentes en 5 clases: menores de 25 años, entre 25-34 años, entre 35-44 años, entre 45-54 años y mayores de 55 años. De la misma manera que con la anterior, es de especial interés analizarla junto con la reincidencia, para poder sacar mejores conclusiones en relación a este fenómeno.

```
tabla3<- table(dataset$`Recidivism - Return to Prison numeric`,dataset$`Age At Release`)
tabla33<-addmargins(tabla3)
par(mfrow=c(1,2))
colores <- c("gray", "black")
barplot(tabla3, col=colores, beside=TRUE)
legend("topright", legend=c("no", "si"), fill=colores)
title(main='Reincidencia por edad')
par(mfrow=c(1,1))
```

En el código de R, creamos una variable "tabla 3" que recoja las frecuencias relativas de las observaciones según su reincidencia y la edad. Además, mediante la función barplot(), R nos permite crear un gráfico donde se detallan estas frecuencias. En el siguiente gráfico observamos como la categoría que prevalece entre los delincuentes es la categoría de entre 25-34 años (concretamente, hay 9.554 observaciones, es decir, casi un 40% de la muestra). Si nos fijamos en el grupo de edad que proporcionalmente reincide en mayor medida, los menores de 25 tienen una tasa de reincidencia del 38% seguidos de los delincuentes de entre 25-34 años con una tasa del 36%. Por el contrario, el grupo de edad que menor tasa de reincidencia presenta son los mayores de 55 años, que del total de 1.303 delincuentes, únicamente el 19% cometió otro delito en el periodo de los 3 años siguientes a su puesta en libertad.

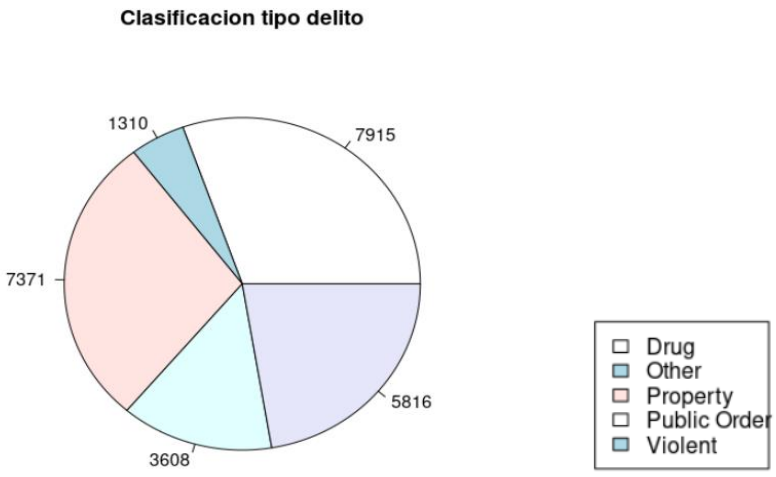
Figura 2: Número de delincuentes en función de la edad y de su comportamiento



La variable `Convicting Offense Type` recoge las distintas categorías de delito más grave por el que los delincuentes fueron encarcelados. En un primer momento y a través de la función pie(), se visualiza un gráfico de sectores donde se muestra las proporciones del número de delincuentes que cometieron cada delito.

```
describe(dataset$`Convicting Offense Type`)
tabla4<-table(dataset$`Convicting Offense Type`)
pie(tabla4,labels=paste0(tabla4),main="Clasificacion tipo delito")
legend("topleft", legend = c("Drug", "Other", "Property", "Public Order", "Violent"),
       fill =  c("white", "lightblue", "mistyrose"))
tabla44<- table(dataset$`Recidivism - Return to Prison numeric`,dataset$`Convicting Offense Type`)
tabla44<-addmargins(tabla44)
```

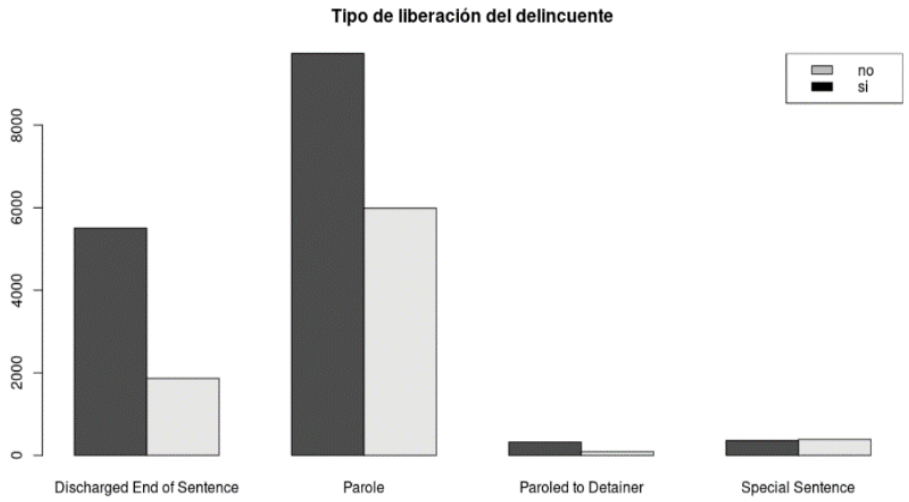
Figura 3: Clasificación de delincuentes en función del tipo de delito más grave cometido



Tras la ejecución del código en R, el gráfico de sectores nos muestra que los delitos que se engloban bajo la categoría "Drug" y "Property", son los más comunes entre los delincuentes recogidos en nuestra base de datos. Más del 50% de los individuos han cometido delitos asociados a las drogas o a la propiedad, dato muy llamativo ya que ordenamientos jurídicos suelen recoger multitud de categorías bajo las que se engloban las actuaciones penadas. Asimismo, para analizar la relación que tiene el tipo de delito por el que los delincuentes fueron encarcelados con la posterior de reincidencia de este grupo de personas, obtenemos una segunda tabla guardada como "tabla44" en la que guardamos las frecuencias relativas de los delincuentes en función del delito cometido y de su reincidencia. El resultado de esta tabla nos muestra que proporcionalmente, los delincuentes que mayor tasa de reincidencia presentan son aquellos que han sido encarcelados por cometer un delito que entra dentro de la categoría de "Others". Contrariamente, los delincuentes que menos reinciden proporcionalmente según el delito por los que son encarcelados son aquellos que cometen delitos pertenecientes a la categoría "Violent", concretamente con una tasa de reincidencia del 26.75%.

Para la variable `Release type: Paroled to Detainder united`, realizamos lo mismo que para la variable que nos proporcionaba la información sobre la edad de los delincuentes: a través de la creación de una tabla de frecuencias que representamos en una visualización de barras, podemos observar cómo se comportan los delincuentes en función del tipo de medio por el cual se ha llevado a cabo su puesta en libertad. En este sentido, encontramos que el mayor número de delincuentes bien son puestos en libertad condicional bien cumplen su sentencia hasta el final. De estos dos grupos de delincuentes, el grupo que mayor tasa de reincidencia presenta es el grupo de delincuentes que son puestos en libertad condicional, ya que presenta un 38% frente a un 25% de los que son liberados después de cumplir su pena. Desde un punto de vista cualitativo, es coherente que las personas que cumplan la pena se sientan en menor medida impulsadas a cometer otro delito, pues han pasado por la experiencia del encarcelamiento y, a la hora de cometer otro delito, debe influir. En cambio, para aquellas personas que han cometido delitos pero sin embargo no han pasado tiempo entre rejas porque han sido puestos en libertad condicional, resulta más fácil cometer otro con vistas a volver a evitar el paso por prisión.

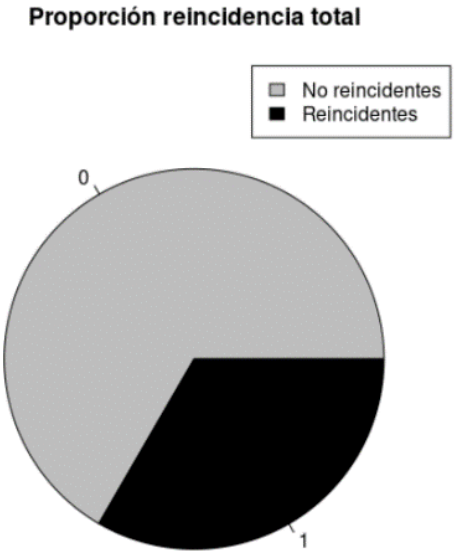
Figura 4: Número de delincuentes en función de su modo de puesta en libertad y de su comportamiento



Por último, resulta interesante estudiar la variable reincidencia independientemente del resto, ya que es la variable que mediante 1 y 0, nos proporciona información sobre si los delincuentes han vuelto a cometer un delito en el periodo de 3 años siguientes a su puesta en libertad. Para ello, utilizamos el siguiente código en R:

```
describe(dataset$`Recidivism - Return to Prison numeric`)
tabla6<-table(dataset$`Recidivism - Return to Prison numeric`)
colores <- c("gray", "black")
pie(tabla6, col=colores, main="Proporción reincidencia total")
legend("topright", legend=c("No reincidentes", "Reincidentes"), fill=colores)
```

Figura 5: Proporción de delincuentes reincidentes en los 3 años siguientes a su puesta en libertad



A través de otro gráfico de sectores, podemos observar cómo el 66.6% de los delincuentes que forman parte de nuestro dataset no cometió otro delito en los tres años siguientes a su puesta en libertad, frente a un 33.36% que efectivamente sí que reincidió. Esta proporción no respalda las conclusiones alcanzadas por (Durose, Cooper, & Snyder, 2014) en su estudio sobre la criminalidad que se ha mencionado *supra.* Según estos investigadores, el 68% de los ex reclusos volvieron a cometer un crimen durante los tres años siguientes a su puesta en libertad. Podemos apreciar que las proporciones son bien diferentes, sin embargo, es preciso recalcar que este último estudio recogía datos de alrededor de 405.000 reclusos que fueron puestos en libertad en cárceles de todo el territorio estadounidense durante el año 2005, mientras que nuestro dataset se centra únicamente en alrededor de 26.000 reclusos en el Estado de Iowa y durante años posteriores. Es una de las limitaciones con la que se encuentra nuestro dataset, el pequeño número de observaciones que tenemos a nuestro alcance, ya que como se ha expuesto en párrafos anteriores, cuanto mayor número de observaciones, más objetivas son las conclusiones de un análisis.

A partir de este análisis inicial de las variables que forman el dataset hemos podido entender en mayor medida las características de estos 26.020 reclusos de los que tenemos información: sus edades, su raza, el tipo de delito que cometieron, cuándo fueron puestos en libertad, etc. Considero Teniendo en cuenta esta información de referencia, a continuación se procederá al estudio de las dos variables principales a través de las cuales se pretende dar respuesta a la pregunta de investigación: `Recidivism - Return to Prison numeric` y `Unemployment rate`.

3.2.Contrastes no paramétricos: prueba de chi-cuadrado de Pearson

En las ciencias estadísticas existen una serie de procedimientos diseñados para estudiar variables cuantitativas, que entran dentro de un grupo de técnicas de análisis denominadas contrastes paramétricos. Estas técnicas se caracterizan por: i) permitir contrastar hipótesis referidas a algún parámetro concreto; ii) exigir el cumplimiento de supuestos concretos sobre las poblaciones originales de las que se extraen los datos; iii) analizar los datos obtenidos con una escala de medida de intervalo o razón. Sin embargo, cuando el dataset con el que trabajamos no cumplen todos los supuestos necesarios para poder aplicar estas técnicas, debemos acudir a otras técnicas estadísticas que nos permitan poner a prueba la hipótesis para dar respuesta a la pregunta de investigación. Este otro

grupo de técnicas estadísticas se conoce como contrastes no paramétricos, y tal y como se ha expuesto, permiten referirse a hipótesis que analizan propiedades nominales de los datos.

En el presente trabajo se quiere dar respuesta al posible efecto que tienen las tasas de desempleo en el Estado de Iowa sobre la reincidencia de los reclusos, por lo que al estar tratando con variables categóricas, se procederá a la prueba de Chi-cuadrado de Pearson para averiguar si las tasas de desempleo tienen influencia en la reincidencia o el desistimiento de los reclusos. Para estudiar esta relación, en primer lugar es preciso introducir el concepto de tabla de contingencia. Esta tabla es una herramienta que se utiliza en las técnicas estadísticas a través de la cual se muestran datos categóricos en términos de conteos de frecuencia. Así, cada celda de la tabla recoge el total de observaciones que efectivamente cumplen ambas categorías que se cruzan.

Antes de ejecutar la tabla de contingencia en R, resulta interesante crear las tablas de frecuencia de cada variable, por un lado la variable `Recidivism - Return to Prison numeric` y, por otro, `Unemployment rate`. Con estas tablas de frecuencias tenemos una idea de cómo se distribuyen los datos en ambas variables, aunque en un principio no nos aporte nada sobre la relación entre ambas.

```
#Tabla de frecuencias relativas reincidencia

table(dataset$`Recidivism - Return to Prison numeric`)
```

```
    0     1
17339  8681
```

```
#Tabla de frecuencias relativas unemployment

table(dataset$`Unemployment rate`)
```

```
 2.59  3.07  3.57  3.69  4.22 4.722
 4374  4586  4687  4740  3917  3716
```

```
#Tabla de contingencia (frec relativa): variable x seria atributo (si reincide o no), y la
Y otro (tasa de desempleo)

tabla<- table(dataset$`Recidivism - Return to Prison numeric`, dataset$`Unemployment
rate`)

t1<-addmargins(tabla) #añadimos addmargins para mostrar la frec relativa acumulada

pander(t1)
```

Figura 6: Tabla de contingencia para las variables Recidivism - Return to Prison numeric` y `Unemployment rate`

	2.59	3.07	3.57	3.69	4.22	4.722	Sum
0	2720	2963	3082	3230	2755	2589	17339
1	1654	1623	1605	1510	1162	1127	8681
Sum	4374	4586	4687	4740	3917	3716	26020

Tras la ejecución del código, obtenemos la generación de tres tablas diferentes: la primera, la tabla de frecuencias para la variable que recoge datos sobre la reincidencia; la segunda, la tabla de frecuencias que recoge el número de observaciones que hay para cada tasa de desempleo y la tercera, la fusión de ambas tablas en una tabla de contingencia que nos permite comparar el número de delincuentes reincidentes por cada tasa de desempleo. Como se ha expuesto anteriormente, esta tabla de contingencia está compuesta por celdas que recogen las combinaciones de filas por columnas. Por ejemplo, para el año en el que la tasa de desempleo alcanzó los 3.69 puntos porcentuales, un total de 1.510 reclusos de nuestro dataset cometieron otro delito en el periodo de los 3 años siguientes a su liberación. Para poder interpretar esta tabla mejor, en R podemos hacer uso de la función prop.table(), que nos devuelve la tabla de contingencia pero en valores absolutos, lo que nos permite comparar más fácilmente entre distintas categorías. Además, si añadimos a la función "margin=2", nos devuelve las frecuencias absolutas calculadas por columnas, es decir la proporción de reincidentes y no reincidentes por cada valor de tasa de desempleo. Lo vemos en la siguiente tabla:

Figura 7: Tabla de frecuencias absolutas calculadas por columnas para las variables Recidivism - Return to Prison numeric` y `Unemployment rate`

	2.59	3.07	3.57	3.69	4.22	4.722
0	0.6218564	0.6460968	0.6575635	0.6814346	0.7033444	0.6967169
1	0.3781436	0.3539032	0.3424365	0.3185654	0.2966556	0.3032831

Así, para el momento en el que la tasa de desempleo en el Estado de Iowa alcanzaba un valor de 3.69%, un total de 31% delincuentes volvieron a cometer un delito. En un primer momento, podemos apreciar que la tabla de contingencia nos aporta ya alguna de las conclusiones a las que queremos llegar, pues por ejemplo para las tasas de desempleo más bajas encontramos las tasas de reincidencia más altas. Adicionalmente, podemos crear un gráfico de barras que nos permite visualizar la tabla de contingencia utilizando la función barplot() que ya han sido expuesta en el epígrafe anterior al tratar de entender las variables que forman el dataset. Asimismo, podemos realizar esta

representación gráfica a través de una representación de mosaicos, en la que el área de los cuadros resultantes pueden ser tomados como una señal de independencia: cuanto más similar sea el área de los mosaicos, más señal de independencia entre ambas variables. El resultado de ejecutar las funciones barplot () y el mosaicplot () en R para estas variables es el siguiente:

Figura 8: Gráfico de barras representando la reincidencia por tasa de desempleo

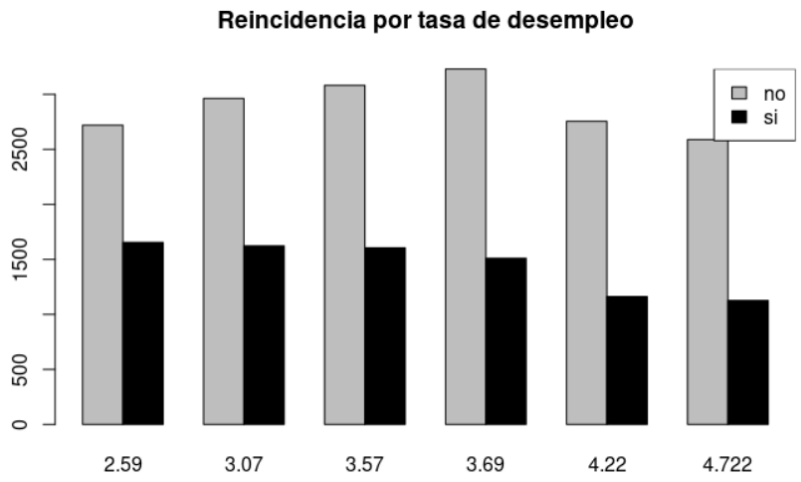
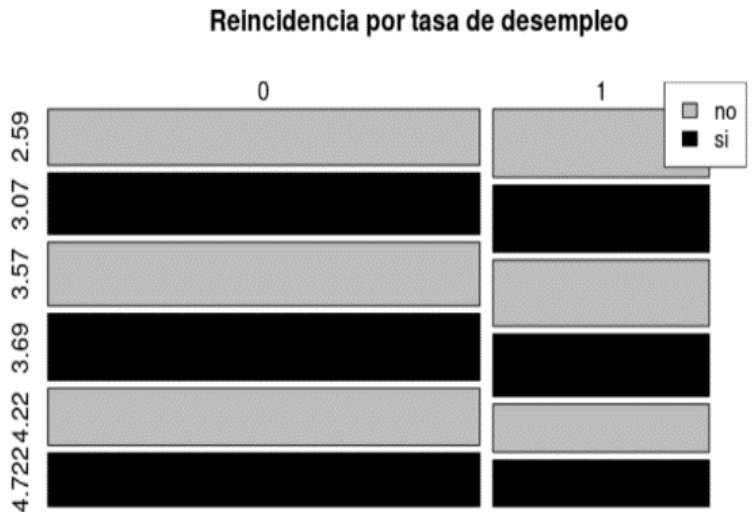


Figura 9: Gráfico de mosaicos representando la reincidencia por tasa de desempleo



Tanto la tabla de contingencia como ambos gráficos proyectan algo de claridad a nuestro estudio, sin embargo, esto no basta para alcanzar conclusiones. Para ello,

debemos cuantificar la relación entre ambas variables a través de una expresión numérica que indique el grado de relación que existe.

Para cuantificar estadísticamente esta relación, hacemos uso de la técnica de Chi cuadrado de Pearson, una estrategia que nos permite medir la distancia que existe entre lo que ocurre en nuestros datos y lo que ocurriría en el caso de que ambas variables fueran completamente independientes. En este último caso, en el cual no hubiese ninguna relación entre las variables `Recidivism - Return to Prison numeric` y `Unemployment rate`, el índice estadístico nos devolverá un valor igual a 0. Por el contrario, cuanta más intensa sea la relación entre ambas variables, el índice nos devolverá valores más alejados de 0. Por tanto, la prueba de independencia de chi cuadrado contrasta la hipótesis nula (en este caso, la reincidencia es independiente de la tasa de desempleo) frente a la hipótesis alternativa (en este caso, la reincidencia se distribuye de manera diferente para cada tasa de desempleo).

Matemáticamente, a través del coeficiente Chi-cuadrado de Pearson, se trata de identificar lo que ocurriría en la tabla de contingencia para el supuesto en el que no existiera relación entre ambas variables, es decir, para el caso de la hipótesis nula. Para ello, lo relevante no son tanto los valores de cada celda, sino los valores marginales de la tabla de contingencia, es decir, los totales. Partiendo de nuestra tabla de contingencia, el objetivo del método estadístico es deducir las llamadas frecuencias esperadas, que son aquellos valores que tomaría la tabla en el caso de una independencia total entre ambas variables. Así, tomando la tabla de contingencia representada en la figura 6, las puntuaciones totales de la variable `Recidivism - Return to Prison numeric` nos muestran que hay 17.339 delincuentes que no hay cometido otro delito los 3 años siguientes a su puesta en libertad, frente a 8.681 que sí que han reincidido. En el caso de que no existiese ninguna relación entre ambas variables, deberíamos observar debidamente la misma proporción de reincidencia (66,6% y 33,4%) para cada una de las tasas de empleo que quedan recogidas en la otra variable. Por ejemplo, en el caso de la primera tasa de empleo, 2.59 donde el total de observaciones es 4.374, tendría que haber un total de 2.913 personas no reincidentes y un total de 1.460 personas reincidentes. Así, construimos una nueva tabla de contingencia con las frecuencias esperadas para cada una de las celdas. Una vez realizado esto, la función de Chi-cuadrado eleva las diferencias entre lo observado y lo esperado al cuadrado y posteriormente se divide cada diferencia cuadrática entre la frecuencia esperada, para así poder expresar la distancia en la escala de los valores que

se estén utilizando. Este método estadístico fue definido por Karl Pearson, por eso recibe el nombre de Chi cuadrado de Pearson. Dicho esto, la expresión del cálculo de este coeficiente es la siguiente:

$$X^2 = \sum_{i,j} \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$$

Antes de proceder a ejecución de esta prueba de Chi cuadrado, es preciso comentar que, aunque sea una técnica que se enfrenta a pocas limitaciones, existen algunas que deben ser comentadas para tenerlas en cuenta en nuestro análisis. En primer lugar, como el cálculo de Chi-cuadrado se basa en el cálculo de las diferencias entre las observaciones observadas y las esperadas, para que se pueda considerar correcto el grado de significación entre ambas, debe cumplirse la condición de que las frecuencias esperadas no tomen valores muy pequeños en más del 20% de las celdas de la tabla de contingencia. Se entiende por valores pequeños aquellos valores inferiores a 5. Adicionalmente, si las muestras son muy grandes, es necesario además de la prueba de chi-cuadrado, una inspección visual que permita comprobar que las diferencias observadas son clínicamente interesantes. Por otro lado, si las variables tienen un alto número de categorías, la prueba estadística probablemente no aporte interés suficiente, ya que cuanto mayor sea el número de categorías, mayor grado de diferencias se espera encontrar. Por último, el contraste de chi cuadrado únicamente ofrece datos sobre la independencia entre ambas variables y no se puede considerar como una medida de asociación entre ambas (Barón López & Téllez Montiel, 2004). Por este motivo, en el caso de que, tras el análisis de chi cuadrado, concluyamos que ambas variables son dependientes, se llevará a cabo el método de la regresión logística para poder estudiar la asociación existente entre estas dos variables y ampliar así nuestras conclusiones sobre la pregunta de investigación.

Habiendo expuesto el razonamiento matemático que hay detrás del método de chi-cuadrado y sus limitaciones, a continuación se procede a ejecutar la herramienta estadística en R. La herramienta R tiene una función denominada chisq.test() para poder realizar este contraste, así que para realizar el método chi-cuadrado de Pearson basta con ejecutar dicha función. Para el presente estudio, se ha ejecutado el siguiente código:

```
tabla<-rbind(Reincidentes= c("2.59"=2720,"3.07"=1551, "3.57"=3082, "3.69"=3197,
"4.22"=2755, "4.722"=2589), NoReincidentes=c("2.59"=1654,"3.07"=1273,
"3.57"=1605, "3.69"=1510, "4.22"=1162, "4.722"=1127))
```

test<-chisq.test(tabla, correct = FALSE)

test

```
        Pearson's Chi-squared test

data:  tabla
X-squared = 242.8, df = 5, p-value < 2.2e-16
```

En cuanto a la interpretación de los resultados, el suizo Crammer afirmó que el máximo valor que puede tomar el coeficiente de independencia es *n(k-1)*, siendo n el número de datos y k el número de categorías que tiene la variable que menos valores recoge. Para este caso, el valor máximo que puede tomar X^2 es 26.020, pues n=26.020 y k=2. Los resultados de este método de contraste resultan en un valor de X^2 igual a 242.8 con 5 grados de libertad, lo que permite concluir que no se cumple la hipótesis nula, y por tanto, las variables no son independientes entre sí. Además, se ha obtenido un p-value muy pequeño, concretamente 2.22e-16. Este valor es considerablemente inferior a 0.01, por lo que también de acuerdo con esto, rechazamos la hipótesis nula en la que se afirmaba que la reincidencia es independiente de la tasa de desempleo que exista en el mercado laboral, y por tanto, concluimos que se cumple la hipótesis alternativa.

Además, en el código, hemos se ha creado la variable "test" para almacenar el resultado del contraste porque así, utilizando $ podemos acceder a otros aspectos del contraste que R no muestra por defecto en el output de la función chisq.test().

test$expected

```
                2.59      3.07     3.57     3.69    4.22    4.722
Reincidentes   2869.777 1852.8238 3075.136 3088.258 2569.94 2438.064
NoReincidentes 1504.223  971.1762 1611.864 1618.742 1347.06 1277.936
```

test$statistic

```
X-squared
 242.7979
```

De esta manera, podemos hacer el mismo análisis pero si acudir a la función R, detallando matemáticamente como se calcula el coeficiente de independencia. Ejecutando los siguientes comandos, obtenemos el coeficiente X2 pero creando nosotros mismos la función, para comprobar que obtenemos los mismos resultados que utilizando la función predeterminada de R, chisq.test:

estChi<-sum((tabla-test$expected)^2/test$expected)

```
> estChi
[1] 242.7979
```

df<-(nrow(tabla)-1)(ncol(tabla)-1)*

```
> df
[1] 5
```

Comprobamos que los resultados obtenidos de manera manual como los obtenidos a través de la función predeterminada de R son los mismos, y, que por tanto, ambas variables son dependientes. La reincidencia de los reclusos durante los 3 próximos años a su puesta en libertad depende en cierta medida de la tasa de desempleo que haya en el mercado laboral. Sin embargo, como se ha adelantando en las limitaciones de este método estadístico, que a través de este contraste no paramétrico afirmemos que existe una dependencia, no podemos concluir sobre la fuerza de dicha dependencia. Autores como el ya mencionado Crammer han propuesto soluciones al problema de interpretación sobre este índice de independencia. Concretamente, Crammer propuso el índice V de Crammer, calculado como lo siguiente:

$$V = \sqrt{\frac{X^2}{n(k-1)}}$$

El resultado del índice en este caso sería igual a aproximadamente 0.10. No obstante, vuelve a aparecer el problema de cómo interpretar este número y cómo concluir sobre la fuerza de la relación entre ambas variables. Para ello, otros autores como Jacob Cohen han detallado listas de valores que permiten guiar el resultado de la V de Crammer. Así, para Cohen, un resultado de 0-0.10 se traduce en una relación de dependencia entre ambas variables casi despreciable. Sin embargo, a día de hoy todavía no hay una solución completa para interpretar la fuerza entre las variables y por ello sigue siendo una limitación fuerte de este método estadístico.

Como consecuencia de todo lo anteriormente expuesto, en el siguiente epígrafe se estudiará dicha relación desde un punto de vista de la regresión logística, para tratar de alcanzar conclusiones sobre como de fuerte es la relación entre las variables `Recidivism - Return to Prison numeric` y `Unemployment rate`.

3.3.Regresión logística

Tal y como se introducía en el apartado anterior, para poder superar la limitación sobre la interpretación del coeficiente de independencia, es preciso aplicar otros métodos estadísticos como la regresión logística que nos permitan alcanzar conclusiones sobre la asociación o fuerza de dependencia que existe entre las dos variables en cuestión.

La regresión logística simple es un método de regresión desarrollada por David Cox a finales de los años 50, que permite estimar la probabilidad de una variable categórica en función de otras variables, que pueden ser tanto cuantitativas como cualitativas. Como principal aplicación de la regresión logística, encontramos la clasificación binaria, es decir, la clasificación de observaciones en un grupo u otro dependiendo de los valores que tomen el resto de variables que se utilizan como predictores. En la dataset que estamos utilizando, la clasificación de binaria consistiría en clasificar a los reclusos en reincidentes o en no reincidentes, en función de la tasa de desempleo que haya en el mercado en ese momento determinado. (Amat Rodrigo, 2016)

A partir de esta método, se logra modelar la probabilidad de que una observación pertenezca al grupo de reincidentes o al grupo de no reincidentes P(Y=1). En el presente trabajo, la variable `Recidivism - Return to Prison numeric` únicamente tiene dos clases, 0 y 1 que representan si efectivamente se ha cometido otro delito o no, pero la regresión logística se puede igualmente aplicar para el caso de variables que tengan más de dos categorías. Dicho esto, también es preciso recalcar que las variables predictoras pueden ser más de una y estaríamos ante una regresión logística múltiple. La regresión logística múltiple se considera una extensión de la regresión logística simple, sin embargo, como queremos estudiar el impacto que tiene la tasa de desempleo en la reincidencia, nuestra regresión logística será simple.

Este tipo de regresión se diferencia de la regresión lineal en que no se expresa como una función lineal de los predictores en la forma $p = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ ... + \beta_q X_q$ porque de esta manera no se garantiza que los β generen valores dentro del intervalo [0,1]. Por ello, se utiliza la función no lineal denominada *logit*: $p = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+ \ ...+\beta_q X_q)}}$. Las interpretaciones de ambas regresiones difieren, pues en el caso de la regresión logística los β ya no representan los efectos marginales y su estimación no se hace a partir de la técnica de mínimos cuadrados. (Moral Peláez, 2016)

También es importante aclarar el concepto de *odds* o razón de probabilidad y el logaritmo de *odds*. En este tipo de regresión, se modela la probabilidad de una observación pertenezca al grupo de reincidentes o no reincidentes a través del uso del logaritmo de *odds*. El concepto de *odds* de pertenecer a la clase 1 (delincuentes que han reincidido en los 3 años siguientes a su puesta en libertad) se define como la relación entre la probabilidad de pertenecer a la clase en cuestión y la probabilidad de pertenecer a la otra clase, es decir, a la clase 0. Así, el *odds* de pertenecer a la clase 1 queda definido como:

$$Odds(\mathrm{Y}=1)=\frac{P}{1-P}$$

Si la probabilidad de pertenecer a la clase de reincidentes es del 60%, entonces la probabilidad de que el individuo no reincida es del 40%. Así, los *odds* de verdadero se definen como el ratio entre la hipótesis verdadera y la hipótesis falsa, que en este caso tomaría un valor de 1.5, lo que es lo mismo que afirmar que se esperan 1.5 individuos que reincidan por cada uno que no lo haga. Vemos como el rango de valores que puede tomar los *odds* es de $[0,\infty]$. Como las probabilidades únicamente pueden tomar valores entre $[0,1]$, resulta necesario la trasnformacion *logit*, consistente en el logaritmo de los *odds*. Así, el logit toma valores desde -∞ (*odds* muy bajos) hasta +∞ (*odds* muy altos). Dicho esto, la interpretación de los coeficientes resultantes de la regresión logística se realiza de la siguiente manera: para un cambio en 1 unidad de x_1, *Logit* se modifica en β_1 unidades. (Chitarroni, 2002)

Habiendo introducido los conceptos básicos de la regresión logística, procedemos a la realización de este método en R para nuestro dataset. La función que se utiliza en R es glm(formula, family=binomial(logit), data), siendo "formula" la descripción simbólica del modelo que se va a ajustar, "family=binomial(logit)" la función a utilizar en el modelo, que en nuestro caso es la clasificación regresión logística) y "data", que en este caso es el dataset que contiene las variables del modelo. Así, el código en R queda definido de la siguiente manera:

```
model1 <- glm(formula=`Recidivism - Return to Prison numeric` ~ `Unemployment rate`, family=binomial(logit), data = dataset)

summary(model1)
```

El output que obtenemos tras ejecutar esta parte del código es el siguiente, mostrándose en la figura número 10:

Figura 10: Resultado del modelo logístico ejecutada en R

```
Call:
glm(formula = `Recidivism - Return to Prison numeric` ~ `Unemployment rate`,
    family = binomial(logit), data = dataset)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.9680  -0.9022  -0.8601   1.4405   1.5724

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)         -0.05560    0.07092  -0.784    0.433
`Unemployment rate` -0.17733    0.01950  -9.096   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 33135  on 26019  degrees of freedom
Residual deviance: 33052  on 26018  degrees of freedom
AIC: 33056
```

En la salida de R que se muestra en la figura 10, lo primero que obtenemos es la llamada donde R nos recuerda el modelo que se ha ejecutado con las opciones especificadas. Seguidamente, obtenemos los residuos de desviación que son una medida del ajuste del modelo de regresión que muestra la distribución de los residuos de desviación para los casos individuales utilizados para el modelo. Para el presente trabajo, estos residuos de desviación no son de interés pues no tratamos de realizar un modelo de predicción, sino que a través de la regresión tratamos de dar respuesta a la relación entre ambas variables. La siguiente parte de la salida ofrece los valores para los coeficientes, con sus errores estándar, el estadístico z (conocido también como el estadístico de Wald) y los valores p. En cuanto a los asteriscos, estos nos informan sobre el nivel de significación de cada variable, siendo *** el máximo y ningún asterisco el mínimo. Vemos como el modelo establece que la variable `Unemployment rate` tiene un alto nivel de significación.

Si analizamos los coeficientes resultantes del modelo, vemos que, según la figura 10, el resultado de la función logit para este modelo es el siguiente:

$$logit(probabilidad\ de\ reincidencia) = -0.055 - 0.177 * Tasa\ de\ desemeplo$$

Como se ha especificado antes, estos coeficientes muestran el cambio en las probabilidades logarítmicas de reincidencia para un aumento de una unidad en la variable de tasa de desempleo. Según nuestros datos, por cada cambio de una unidad de la tasa de desempleo, las probabilidades de que un individuo reincida en los 3 años siguientes a su puesta en libertad se reducen un 0.177. Por tanto, de acuerdo con este resultado, podemos afirmar que efectivamente el fenómeno de la reincidencia no es independiente de la tasa de desempleo del mercado en ese momento concreto ya que el resultado de la regresión nos afirma que es una variable muy significativa, y además tiene una relación negativa.

De igual manera, resulta interesante comentar las ultimas filas del output del modelo de regresión. Así, debajo de los coeficientes encontramos los índices de ajuste del modelo, que incluyen los residuos nulos, los residuos de desviación y el AIC. Estos valores ayudan al procedimiento de evaluación del modelo de predicción basado en la regresión *logit*, sin embargo, no es cuestión objeto de nuestro estudio.

Para terminar, podemos hacer uso de la función confint() para obtener los intervalos de confianza para las estimaciones de los coeficientes. De esta manera tendremos una medida que nos proporciona mayor información sobre la fiabilidad de los coeficientes obtenidos. Se debe tener en cuenta que para los modelos de regresión logística, como el que se ha realizado en este trabajo, los intervalos de confianza se basan en la función de log-verosimilitud. Asimismo, podemos obtener intervalos de confianza basados únicamente en los errores estándar a través del método por defecto. A continuación se recoge el código ejecutado para obtener estos intervalos de confianza así como la salida que ofrece R:

```
#Intervalos de confianza utilizando la función log-verosimilitud
confint(model1)
```

```
                        2.5 %      97.5 %
(Intercept)         -0.1946104  0.08338404
`Unemployment rate` -0.2155730 -0.13914531
```

```
#intervalos de confianza utilizando los errores estándares
confint.default(model1)
```

```
                        2.5 %      97.5 %
(Intercept)         -0.1945914  0.08339514
`Unemployment rate` -0.2155453 -0.13911982
```

Vemos que los intervalos de confianza obtenidos de ambas maneras nos proporcionan prácticamente la misma información. La confianza que se establece por defecto en R es del 95%, sin embargo, se puede ampliar o reducir en nuestro código en función de nuestros intereses. Para este trabajo, hemos dejado una confianza del 95%, por lo que podemos decir que estamos seguros en un 95% que, como el intervalo no contiene el 0, la diferencia es negativa y el coeficiente se encuentra entre un -0.21 y un -0.13. Encontramos que es muy poco probable que el coeficiente sea 0 y, por tanto, que las variables sean independientes.

CAPÍTULO 4. RESULTADOS

Tras la explicación y ejecución de los métodos estadísticos de chi-cuadrado de Pearson y la regresión logística simple en los apartados anteriores, podemos comenzar la recopilación de los resultados finales del estudio. Cabe recordar que la pregunta de investigación a la cual se intentaba dar respuesta a través de este análisis de datos es la siguiente: *¿en qué medida una mayor tasa de desempleo afecta a la criminalidad de los recién salidos de prisión?*

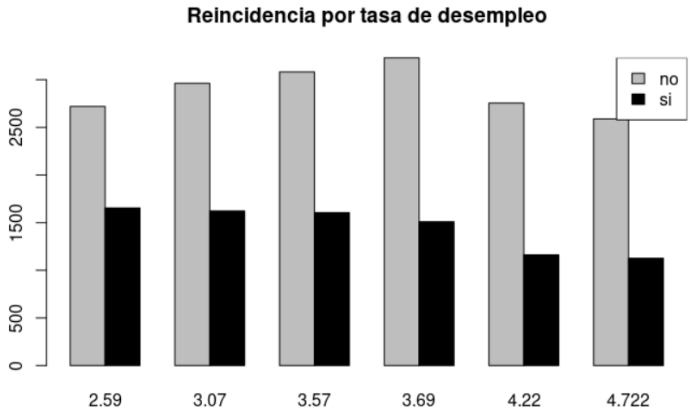
En un primer análisis, hemos obtenido un índice de chi-cuadrado igual a 242.79, lo que nos permite negar la hipótesis nula consistente en que el fenómeno de la reincidencia es independiente a la tasa de empleo del mercado. Por tanto, confirmábamos que la variable de reincidencia queda influida por la variable tasa de desempleo en el mercado. Sin embargo, como se ha expuesto anteriormente, una de las limitaciones principales de este método estadístico es la interpretación de este índice de independencia, ya que no nos permite ver la fuerza de la relación que existe entre ambas variables. Aunque se ha expuesto alguna otra técnica para interpretar el índice de chi-cuadrado obtenido, se ha preferido hacer uso de otra técnica estadística para obtener un índice que nos proporcione mayor información.

Por tanto, para poder superar estas limitaciones, hemos llevado a cabo una regresión logística que nos ha permitido obtener nuevos datos sobre la relación entre ambas variables. Tal y como se ha expuesto y razonado en el apartado anterior, el coeficiente obtenido nos informaba que por cada cambio de una unidad de la tasa de desempleo, las probabilidades de que un individuo reincida en los 3 años siguientes a su puesta en libertad se reducen un 0.177.

Dicho esto, se deduce que, tal y como el índice de Chi-cuadrado mostraba, estas variables son dependientes, sin embargo, esta dependencia se caracteriza por ser negativa, tal y como confirmaba en un segundo momento la regresión logística. Esto quiere decir, que la respuesta a nuestra pregunta, aunque en un primer momento parezca sorprendente, es que una mayor tasa de desempleo afecta negativamente a la reincidencia. Digo sorprendente porque, de acuerdo al estudio de la literatura, parecía lógico afirmar que cuantas más dificultades tenían los delincuentes que salían de prisión para acceder a un empleo legal, mayor probabilidad de que volviesen a cometer un delito.

Para comprobar los resultados que hemos obtenido, resulta interesante volver a analizar el histograma que mostraba las proporciones entre el número de delincuentes que vuelven a reincidir o no, en función de la tasa de empleo que había en ese momento en el mercado. Es importante recalcar que las conclusiones no se pueden tomar en base a un juicio visual sobre un histograma, y han de basarse en índices de modelos estadísticos. Sin embargo, a modo de comprobación sí que resulta útil para apoyar los resultados así como conclusiones que hemos obtenido.

Figura 10 (idem figura 8): Gráfico de barras representando la reincidencia por tasa de desempleo



Si prestamos atención al número de reincidentes en función de la evolución de la tasa de desempleo, podemos confirmar lo que los resultados de la regresión logística nos mostraban. Para la tasa de desempleo 4.7%, es decir, la tasa más alta registrada en nuestro base de datos, el número de personas que reinciden es el más bajo. Por el contrario, para la tasa de desempleo más baja (2.59%), el número de personas que cometen otro delito durante los 3 años siguientes a su puesta en libertad es el más alto, mostrando así la

dependencia negativa que tiene la tasa de desempleo del mercado en la tasa de reincidencia de los individuos recogidos en nuestro dataset.

CAPÍTULO 5. CONCLUSIONES

Para concluir, es preciso recordar lo analizado previo el análisis de datos en relación al estado de la cuestión del fenómeno de la reincidencia. Tal y como se exponía en el epígrafe correspondiente, una de las maneras de alentar el desistimiento es incrementar el acceso a empleos legales para las personas que salen de prisión. Se ha demostrado como, a partir de políticas encauzadas para conseguir esto, las personas que salen de prisión pueden obtener un trabajo legal que les permita irse alejando de la criminalidad que en etapas anteriores de su vida les rodeaba. Centrándonos en este tema, como pregunta de investigación concreta para estudiar la conexión entre las oportunidades laborales y el crimen, se planteaba en qué medida una mayor tasa de desempleo afectaba a la criminalidad de los ex reclusos, pues a priori, los momentos en los que mayores tasas de desempleo hay son los momentos en los que menores oportunidades laborales hay en el mercado.

Siguiendo con esta línea, parecía lógico que afirmar que, si aumentar el acceso a empleo legal es una forma de aumentar la función de coste de oportunidad del delito, momentos temporales en los que hubiese mayor oportunidad laboral y por ende, la tasa de desempleo fuese menor, menor criminalidad habría. Sin embargo, tal y como se ha expuesto en los resultados, tras el análisis de datos realizado para la base de datos del Estado de Iowa, esto no ocurre. Concretamente, ocurre lo contrario, siendo las tasas de reincidencia mayores cuanto menor es la tasa de desempleo.

De igual manera es preciso señalar que el presente estudio se enfrenta a limitaciones, entre la que destaca el número de observaciones de las que se disponían. A pesar de que 26.020 observaciones parezcan en un principio suficientes, estas observaciones únicamente muestran una pequeña porción de los delincuentes que son puestos en libertad. Esto supone que, aunque las resultados alcanzados sean concluyentes para la muestra utilizada en el trabajo, no podemos afirmar que los resultados sean representativos para la sociedad en general y se ajusten a los patrones de comportamiento de los presos que son puestos en libertad a más amplios niveles geográficos. Para poder llegar a confirmar esto último, sería necesario trabajar con una muestra mucho más amplia

que cumpla con los requisitos de aleatoriedad. Además, la muestra con la que se ha trabajo recoge únicamente información sobre reclusos en el mercado de Iowa. Esto es importante porque pueden existir otros factores que no entran dentro del ámbito del trabajo que estén influenciando el comportamiento del mercado o de los delincuentes. Si, por el contrario, tuviésemos una muestra que no se limitase a un territorio en concreto, habría menos arbitrariedad y los datos podrían ser considerados como más objetivos.

Por último, me gustaría comentar unas breves líneas sobre la exploración de datos y el análisis de datos, y sobre su importancia para llevar a cabo políticas que ayuden a crecer como sociedad. El tema de la reincidencia es un tema relevante, que, a través de políticas adecuadas, puede llegar a reducirse en gran medida, como hemos podido comprobar en algunos países. Por eso, la extracción de conclusiones apoyadas en datos resulta crucial, pues muchas veces lo que creemos en un primer momento dista de los informaciones que nos aportan los datos. Como ha ocurrido en el presente trabajo, al comienzo del mismo exponía como en línea con la literatura estudiada, lo más lógico era pensar que la reincidencia se redujese cuanto mayor número de oportunidades laborales existiesen en el mercado. Sin embargo, los datos han arrojado una conclusión diferente, que probablemente, sin su exploración, no hubiésemos alcanzado. La exploración e interpretación de datos permite reconocer oportunidades son las más idóneas en las que invertir o dejar de invertir. Por tanto, a la hora de poner en marcha actuaciones para intentar hacer frente a una cuestión tan importante como es la reincidencia, resulta crucial apoyarse siempre en datos objetivos y técnicas que nos permiten extraer la mejor conclusión posible.

BIBLIOGRAFÍA

Agan, A., & Makowsky, M. D. (2018). The Minimum Wage, EITC, and Criminal Recidivism.

Agan, A., & Starr, S. (2019). Ban the Box, Criminal Records, and Racial Discrimination: a field experiment. *The Quarterly Journal of Economics, Volume 133, Issue 1*, 191-235.

Amat Rodrigo, J. (2016). *Regresión logística simple y múltiple.*

Barón López, F., & Téllez Montiel, F. (2004). Apuntes de bioestadística. *Tercer Ciclo en Ciencias* .

Beauchamp, A., & Chan, S. (2014). The Minimum Wage and Crime. *The B.E. Journal of Economic Analysis & Policy, vol. 14(3)*, 1-23.

Becker, G. (1968). Crime and Punishment: An Economic Approach. *76 The Journal of Political Economy 169*, 176-177.

Charte Ojeda, F. (2014). *Análisis exploratorio y visualización de datos con R.*

Chitarroni, H. (2002). *La regresión logística.*

Corso, J. A. (2005). *Estadística no paramétrica: Métodos basados en rangos.* Editorial UN.

Cunningham, S. (2021). Causal inference: The mixtape. *Yale University Press*.

Doleac, J. (2019). Encouraging desistance from crime.

Doleac, J. L. (2016). "Increasing Employment for Individuals with Criminal Records". Hamilton Project policy proposal.

Doleac, J., & Hansen, B. (2020). The unintended consequences of "ban the box": Statistical discrimination and employment outcomes when criminal histories are hideen. *Journal of Labor Economics, Volumen 38(2)*, 321-374.

Durose, M., Cooper, A., & Snyder, H. (2014). Recidivism of prisoners released in 30 states in 2005: Patters from 2005-2015 (Special Report NCJ244205). *Bureau of Justice Statistics*.

Galbiati, R., Ouss, A., & Philippe, A. (2021). News and Reoffending after Incarcelation. *The Economic Journal, Volume 131, Issue 633*.

Gomez Villegas, M. (2005). *Inferencia estadística* . Ediciones Díaz de Santos.

Hernandez, F., & Usuga, O. (2020). *Model: Model Package Contains Useful Functions for Modelling Regresion.*

Jackson, O., & Zhao, B. (2017). *The effect of changing employer's access to criminal histories on ex-offenders' labor market outcomes: evidence from the 2010-2012 Massachusetts CORI Reform.*

Kirby, E. J. (9 de Julio de 2019). La exitosa estrategia de Noruega para transformar a sus criminales en "buenas personas". *BBC*.

Kling, J. R. (2006). Incarcelation length, employment, and earnings. *American Economic Review, 96*, 863-876.

Leasure, P., & Andersen, T. (2016). The effectiveness of certificates of relief as collateral consequence relief mechanisms: An experimental study. *Yale L. & Pol'y Rev. Inter Alia, vol 35*, 11.

Moral Peláez, I. (2016). Modelos de regresión: lineal simple y regresión logística. *Revista Seden, 14*, 195-214.

Pager, D. (2003). The Mark of a Criminal Record. *American Journal of Sociology, 108*, 937-975.

Schnepel, K. (2018). Good Jobs and Recidivism. *The Economic Journal*.

Yang, C. (2017). Local labor markets and criminal recidivism. *Journal of Public Economics, vol. 147*, 16-29.

ANEXO: Código R Script

```
#Trabajo Fin de Grado Analytics
#Dolores Soubrie
#Abril 2021

#-----------------------------------------------------
#INSTALACIÓN DE PAQUETES PARA PODER TRABAJAR
install.packages("readr") #instalacion de paquetes
library(readr) #descarga de la librería correspondiente
install.packages("dplyr")
library(dplyr)
install.packages("tidyr")
library(tidyr)
install.packages("pander")
library(pander)
install.packages("ggplot2")
library(ggplot2)
install.packages("DescTools")
library(DescTools)
install.packages("vcd")
library(vcd)
install.packages("Hmisc")
library(Hmisc)
install.packages("dendextend")
library(dendextend)
```

```
#-------------------------------------------------------

#CARGA DE DATOS

dataset <- read_delim("3-
Year_Recidivism_for_Offenders_Released_from_Prison_in_Iowa_elaborated.csv",";",
escape_double = FALSE, trim_ws = TRUE)

View(dataset) #función que permite ver el dataset completo


#-------------------------------------------------------

#ESTUDIO Y PREPARACIÓN DE DATOS

nrow(dataset) #número de filas/observaciones

ncol(dataset) #número de columnas/variables

dim(dataset) #dimensiones del dataset

str(dataset) #estructura del dataset y tipo de variables


#Borrar las variables que no nos interesan para el análisis y unicamente hace que el dataset
sea más pesado

dataset<- select(dataset, -`Main Supervising District`, -`Part of Target Population`,-
`Release Type`,-`Convicting Offense Subtype`,-`Convicting Offense Classification`)

ncol(dataset) #tenemos un dataset con ocho variables

#necesario transformación previas de los datos antes de hacer los análisis

#4 variables numericas, el resto categóricas --> hay que transformarlas a categóricas,
porque a pesar de ser números, corresponden a categorías

#utilizamos funcion as.factor

str(dataset)

dataset$`Fiscal Year Released`<-as.factor(dataset$`Fiscal Year Released`)

dataset$`Recidivism Reporting Year`<-as.factor(dataset$`Recidivism Reporting Year`)

dataset$`Race - Ethnicity`<-as.factor(dataset$`Race - Ethnicity`)

dataset$`Age At Release`<-as.factor(dataset$`Age At Release`)

dataset$`Convicting Offense Type`<-as.factor(dataset$`Convicting Offense Type`)
```

```
dataset$`Release type: Paroled to Detainder united`<-as.factor(dataset$`Release type: Paroled to Detainder united`)

dataset$`Recidivism - Return to Prison numeric`<-as.factor(dataset$`Recidivism - Return to Prison numeric`)

summary(dataset)


#Comprobar y borrar los missing values

is.na(dataset) #vemos que tenemos valores perdidos

mean(is.na(dataset)) #concretamente un 0,8%

dataset<-na.omit(dataset) #borramos los valores perdidos

mean(is.na(dataset)) #los datos limpios ya no tienen NAs


#---------------------------------------------------------------

#CONOCER LOS DATOS

describe(dataset)#nos dice el numero y prop de variables

#1. variable años (para ver la proporcion de presos cada año)

tabla1<- table(dataset$`Fiscal Year Released`)

barplot(tabla1, beside=TRUE)

title(main='Número de delincuentes puestos en libertad por año')

#2. variable ethnicity

tabla2<- table(dataset$`Race - Ethnicity`)

tabla22<- table(dataset$`Recidivism - Return to Prison numeric`,dataset$`Race - Ethnicity`)

tabla22

#3. variable edad

describe(dataset$`Age At Release`)

tabla3<- table(dataset$`Recidivism - Return to Prison numeric`,dataset$`Age At Release`)

tabla33<-addmargins(tabla3)
```

```
par(mfrow=c(1,2))

colores <- c("gray", "black")

barplot(tabla3, col=colores, beside=TRUE)

legend("topright", legend=c("no", "si"), fill=colores)

title(main='Reincidencia por edad')

par(mfrow=c(1,1))

#4. variable tipo de delito

describe(dataset$`Convicting Offense Type`)

tabla4<-table(dataset$`Convicting Offense Type`)

pie(tabla4,labels=paste0(tabla4),main="Clasificacion tipo delito")

legend("topleft", legend = c("Drug", "Other", "Property", "Public Order", "Violent"),

       fill =  c("white", "lightblue", "mistyrose"))

tabla44<- table(dataset$`Recidivism - Return to Prison numeric`,dataset$`Convicting Offense Type`)

tabla44<-addmargins(tabla44)

#5. variable tipo de release

describe(dataset$`Release type: Paroled to Detainder united`)

tabla5<-table(dataset$`Recidivism - Return to Prison numeric`,dataset$`Release type: Paroled to Detainder united`)

tabla55<-addmargins(tabla5)

barplot(tabla5, beside=TRUE)

title(main='Tipo de liberación del delincuente')

legend("topright", legend=c("no", "si"), fill=colores)

#6. variable reincidencia

describe(dataset$`Recidivism - Return to Prison numeric`)

tabla6<-table(dataset$`Recidivism - Return to Prison numeric`)

colores <- c("gray", "black")

pie(tabla6, col=colores, main="Proporción reincidencia total")
```

```
legend("topright", legend=c("No reincidentes", "Reincidentes"), fill=colores)


#---------------------------------------------------------

#DESCRIPTIVA DOS-DIMENSIONAL: CALCULO DE CHI-CUADRADRO

dataset$`Unemployment rate` <- dataset$`Unemployment rate`

#Tabla de frecuencias reincidencia (relativa)

table(dataset$`Recidivism - Return to Prison numeric`)

#Tabla de frecuencias unemployment (relativa)

table(dataset$`Unemployment rate`)

#Tabla de contingencia (frec relativa): variable x seria atributo (si reincide o no), y la Y otro (tasa de desempleo)

tabla<- table(dataset$`Recidivism - Return to Prison numeric`, dataset$`Unemployment rate`)

t1<-addmargins(tabla) #añadimos addmargins para mostrar la frec relativa acumulada

pander(t1)

#frecuencia absoluta

tabla_prop<-prop.table(tabla)

t2<-addmargins(tabla_prop)

pander(t2)

#frecuencias absolutas calculadas por columnas, por lo que cada columna suma 1

prop.table(tabla, margin=2)

#Representación gráfica --> grafico de barras

par(mfrow=c(1,2))

colores <- c("gray", "black")

barplot(tabla, col=colores, beside=TRUE)

legend("topright", legend=c("no", "si"), fill=colores)

title(main='Reincidencia por tasa de desempleo')

par(mfrow=c(1,1))
```

```
#representacion gráfica --> mosaicos

#Si todos los cuadros en todas las categorías tienen la misma área estamos ante una señal de independencia --> ¿poca independencia? veremos a continuacion

par(mfrow = c(1, 2))

mosaicplot(tabla, cex = 1.1, col=colores)

legend("topright", legend=c("no", "si"), fill=colores)

title(main='Reincidencia por tasa de desempleo')

par(mfrow = c(1, 1))


#-----------------------------------------------------------

#Pearson's Chi-squared test

#Función a utilizar: chisq.test

#HO: la reincidencia es independiente de la tasa de desemepleo

#H1: la reincidencia es dependiente de la tasa de desemepleo

# La fucnion chisq.test nos permite obtener el estadístico del contraste, los grados de libertad y el p-valor de forma muy sencilla. En nuestro ejemplo bastaría con hacer:

tabla<-rbind(Reincidentes= c("2.59"=2720,"3.07"=1551, "3.57"=3082, "3.69"=3197, "4.22"=2755, "4.722"=2589), NoReincidentes=c("2.59"=1654,"3.07"=1273, "3.57"=1605, "3.69"=1510, "4.22"=1162, "4.722"=1127))

test<-chisq.test(tabla, correct = FALSE)

test

test$expected

test$statistic

test$p.value

#como p-valor es <0.01 rechazamos la hipotesis nula: rechazamos qeu reincidencia sea independiente de la tasa de desemeplo

#Hay evidencia en la muestra de que ambas variables son dependientes

estChi<-sum((tabla-test$expected)^2/test$expected)
```

```
df<-(nrow(tabla)-1)*(ncol(tabla)-1)

#Pero que exista una dependencia no nos dice gran cosa sobre la fuerza de esa relación. Para medir la fuerza Podemos calcular el V-crammer es un índice que nos ayuda a la interpretacion del índice de chi-cuadrado



#-----------------------------------------------------

#REGRESION LOGISTICA - VARIABLE CATEGORICA

#Logit=Log (Odds)=β_0+β_1 X_1+β_2 X_2+ …+β_q X_q

install.packages("GGally")

library(GGally)

#la columna de color debe ser categorica, convertimos la variable en categorica, que en este caso es una variable binaria. El resto de variables predictoras X1, X2, .... , Xn pueden ser variables categóricas, variables continuas o una mezcla de estos dos tipos. En nuestro caso es una regression simple, así que unicamente utilizaremos una variable predictora

data2<-select(dataset, -'Convicting Offense Subtype')

data2$`Recidivism - Return to Prison numeric`<-as.factor(data2$`Recidivism - Return to Prison numeric`)

ggpairs(data2,aes(color=`Recidivism - Return to Prison numeric`, alpha=0.3))

#reincidencia es nuestra variable independiente

#regresion logistica simple

model1 <- glm(formula=`Recidivism - Return to Prison numeric` ~ `Unemployment rate`, family=binomial(logit), data = dataset)

summary(model1)

#Intervalos de confianza utilizando la funcion log-verosimilitud

confint(model1)

#intervalos de confianza utilizando los errores estándares

confint.default(model1)
```