Bitácoras Grupo #5, CA-204 (II-2024)

Análisis de las variables cualitativas en relación al riesgo crediticio

Jeikel Navarro Solis, Gabriel Valverde, Erick Venegas

2024-11-14

Tabla de contenidos

Introducción

Este estudio utiliza una base de datos de riesgos financieros en el cual se toman en cuanta variables como lo son la edad, genero, pais en el que vive y estado marital de las personas, esto con el fin de realizar un perfil descriptivo de los prestatario, asi como tambien se toman variables cuantitativas como los ingresos, activos y deuda de la persona en cuestion. De manera que la base contiene un total de 15000 casos y 20 variables. El objetivo es generar empiricamente un metodo que basado en las variables permita crear una calificación de riesgo, tal que se puedan relacionar las variables cuantitativas y cualitativas de modo tal se logre llegar a algún resultado satisfactorio a partir de estos datos. Por estas razones, la investigación se basa en un marco teorico y empirico donde los autores buscan salir de la norma y explorar mas factores que solo la parte economica de los prestatarios.

1 Bitacora 1

1.1 Parte de planificación

1.1.1 Definión de la idea

Cuando una persona quiere acceder a un préstamo bancario, es común que los bancos requieran información personal y profesional de la persona, por ejemplo su profesión, ingresos mensuales, edad, gastos mensuales, etc. Esto se debe a que los bancos necesitan saber si esta persona es apta para pagar el préstamo en un tiempo conveniente. Generalmente se piensa que, mientras más ingresos tenga la persona, es más probable que le acepten el crédito a financiar. Añadido a esto, también es frecuente escuchar que mientras la persona tenga un mayor grado académico, ésta tendrá un mejor salario o tendrá más formas de ingresar dinero a sus cuentas. Ante esta situación, nos surgió la idea de comprobar si estos estereotipos son ciertos ó no. Por lo que nuestra meta es lograr determinar si realmente el nivel de educación y el nivel de ingresos influye en la capacidad de pago de una persona en un préstamo, así como analizar si los factores cualitativos tienen cierta relevancia al momento de que haya un impago.

1.1.2 Conceptualización de la idea

"Verificar una relación entre la calificación de riesgo con las variables Nivel de educación, Ingresos, Monto del préstamo". De la pregunta anterior, nos interesa conceptualizar dicha idea, por lo que, buscando la definición de las palabras que conforman la idea, en la RAE, encontramos lo siguiente:

- Relación: Conexión, correspondencia.
- Nivel: Medida de una cantidad con referencia a una escala determinada.
- Educación: Acción y efecto de educar, instrucción por medio de la acción docente.
- Ingresos: Caudal que entra en poder de alguien, y que le es de cargo en las cuentas.
- Calificación: Puntuación obtenida en un examen o en cualquier tipo de prueba.
- Riesgo: Contingencia o proximidad de un daño.

1.1.3 Identificación de tensiones

Como mencionamos en el apartado anterior, el trabajo se centra más en el estudio de las herramientas utilizadas, que en la información del trabajo per se. Sin embargo, vamos a desarrollar la teoría de manera satisfactoria con el hecho de crear un trabajo bien estructurado. Dicho lo anterior, es claro que la calificación de riesgo es un fenómeno que depende de muchas más variables, un ejemplo de ello puede ser la cultura de la sociedad en la cual se ve inmersa la persona que solicita el préstamo, como la entidad que lo desembolsa. Por ello, el origen de los datos de esta tabla de datos es de suma importancia, pues la cultura de las personas que arrojaron estos datos puede influir de manera sustancial en la salida de los datos.

Otro factor a tomar en cuenta es por supuesto, la tabla de datos, pues podría contener información errónea, o datos que no estén correctamente digitados, además de que la tabla de datos tiene que ser convertida a información más trabajable, es decir, muchas de las variables son categóricas, por lo que una conversión a datos de diferente especie, podría provocar que haya errores en la nueva tabla que se va a manipular.

1.1.4 Reformulación de la idea en modo preguntas

La idea principal de este trabajo es: "Verificar una relación entre la calificación de riesgo con las variables Nivel de educación, Ingresos, Monto del préstamo", por lo que para etsa parte del trabajo, vamos a formular la idea de diferentes formas, esto con el fin de formular una pregunta de investigación que sea clara.

¿Cuáles son las variables que influyen más en la calificación de riesgo? ¿Existe una correlación positiva entre las variables de interés y la calificación de riesgo? *¿Es la calificación de riesgo un buen calificador de las variables que la determinan?* ¿Cómo se mide la calificación de riesgo dadas las variables consideradas?

1.1.5 Argumentación de las preguntas

Estas preguntas deben poseer una argumentación detrás de ellas, pues es necesario tener una idea de qué es lo que queremos realizar con estas preguntas o dicho de otra forma, cuál es el problema que potencialmente podemos resolver o la incógnita a contestar.8

1.1.5.1 ¿Cuáles son las variables que influyen más en la calificación de riesgo?

Contraargumentos

Las variables que se toman son realmente las correctas, cuál es el método que se utiliza para determinar que las variables que se están tomando son mejores para un resultado que otras. Al ser una calificación de riesgo, que facilita el préstamo a personas o no, las personas podrían

tener intenciones perversas a siempre manipular su información para obtener buenas calificaciones. Por otro lado, son muchas las variables contempladas en la base de datos, hacer uso de un par de variables, podría influir de manera considerable en buscar una relación entre las variables.

Argumentos

Al realizar un análisis previo de la tabla de datos, podríamos determinar cuáles son las variables con mayor impacto en esta calificación, y así simplificar el modelo y darle un mayor énfasis en estas variables que afectan de manera considerable en la calificación.

Concluya

Al ser una calificación de riesgo, las empresas deben considerar varios factores, para ver si sus clientes son ideales para el préstamo o no. Por otro lado, concentrarnos en las variables que tienen mayor peso, podría ayudar a un mejor entendimiento de la calificación de riesgo.

1.1.5.2 ¿Existe una correlación positiva entre las variables de interés y la calificación de riesgo?

Contraargumentos Una relación positiva no implica que haya una mejor calificación, pues pueden existir variables con relación negativa, que tengan un efecto positivo sobre la calificación de riesgo, por pura intuición, podemos pensar en las personas que tienen un récord crediticio limpio, esto implicaría que esta variable debería tener un valor nulo para mejorar la calificación de riesgo. Otra variable que podría afectar es las veces que la persona ha caído en impago, pues entre más suba este valor, la calificación de riesgo, debería ser menor.

Argumentos

Las relaciones que buscamos no tienen por qué ser positivas, desde un tipo de vista de coeficiente de correlación, sino desde un tipo de vista de mejoría, es decir, variables que en ausencia contribuyan a una mejor calificación, nos son de interés para el trabajo. Como hemos mencionado encontrar las relaciones más contribuyentes, son de utilidad, pues favorecen a simplificar el modelo y obtener a cambio una mejor interpretación del estudio.

Concluya

Todas las relaciones son de importancia, hasta donde no existe relación, pues sirven para delimitar el modelo y determinar cuáles son las verdaderas variables que si influyen.

1.1.5.3 ¿Es la calificación de riesgo un buen calificador de las variables que la determinan?

Contraargumentos

Esta pregunta, no es del todo objetiva, pues dependerá de lo que la empresa quiera detectar en estas evaluaciones, es decir, un factor que determina en gran medida a la calificación de riesgo, en otra empresa no tiene por qué ser así, pues dependerá del público objetivo de la empresa. Un buen calificador de riesgo depende tanto de las variables que se toman en cuenta, como del contexto de la empresa.

Argumentos

Desde el punto de vista de la empresa, la calificación de riesgo es una herramienta que sirve para determinar en gran medida si un préstamo se realiza o no, por ello, saber si su calificación de riesgo logra captar la información deseada de las variables utilizadas, entonces se podría considerar un buen factor, pues, aunque las empresas deban tener cuidado a quiénes otorgan los préstamos, también es un hecho, que si no lo hacen, se quedan sin negocio

Concluya

Si la calificación de riesgo logra simplificar y consolidar la información, entonces podríamos conjeturar que se comporta como una buena calificación.

1.1.5.4 ¿Cómo se mide la calificación de riesgo dadas las variables consideradas?

Contraargumentos

La pregunta es más compleja que las hechas anteriormente, pues estamos entrando a un método de calificación, es decir, hacer un análisis de la medición, como mencionamos como un contraargumento en la pregunta anterior, esto no tiene por qué ser universal en las empresas, esto puede variar, por lo que dependerá del contexto en el cual se realice la pregunta.

Argumentos

En realidad que sea una pregunta que depende de sus variables, podría ser beneficioso, pues si existen varias metodologías, podríamos tener una mejor cartera de préstamos, es decir, dado cierto tipo de cliente, se podrían realizar cierto tipo de préstamos.

Concluya

La subjetividad de esta pregunta, no es ni buena ni mala, esta depende su contexto, si es aplicada de una buena forma, podría ser beneficioso tanto para la empresa, por ampliar su mercado, como para el cliente, al recibir el préstamo deseado.

1.1.6 Argumentación a través de datos

La base de datos utilizada en este trabajo se toma de la base de datos de Kaggle. Sin embargo, el autor de la base de datos no publica la información de cuándo es sacada la información. La información fue de fácil acceso pues está disponible en la de datos de Kaggle. La muestra observada son personas entre las edades de 18 a 69 años, de donde toman muchas variables, las cuales veremos más adelante. La unidad estadística estudiada para este trabajo son individuos que buscan obtener un préstamo.

Para la siguiente vamos a tomar el nombre las variables, las cuales dejaremos en el idioma original y vamos a dar la descripción de ellas, las cuales viene a su vez con la tabla de datos.

- Age: La edad del individuo, una variable continua que influye en la estabilidad financiera.
- Gender: Género del individuo, categorizado en Masculino, Femenino y No binario.
- Education Level: Nivel de educación alcanzado, que varía desde la Secundaria hasta el Doctorado.
- Marital Status: Estado civil actual, categorizado como Soltero, Casado, Divorciado o Viudo.
- Income: Ingreso anual en USD, que representa la capacidad de ganancia del individuo.
- Credit Score: Valor numérico que indica la solvencia crediticia, que varía de 600 a 800.
- Loan Amount: La cantidad de préstamo solicitada por el individuo, que representa las necesidades financieras.
- Loan Purpose: El propósito del préstamo, categorizado en Vivienda, Auto, Personal o Negocios.
- Employment Status: Situación laboral del individuo, incluyendo Empleado, Desempleado o Autónomo.
- Years at Current Job: Duración del empleo en el trabajo actual, que refleja la estabilidad laboral.
- Payment History: Desempeño histórico de pagos, categorizado como Excelente, Bueno, Regular o Malo.
- Debt-to-Income Ratio: Relación entre deuda e ingreso, que indica el apalancamiento financiero y el riesgo.
- Assets Value: Valor total de los activos poseídos por el individuo.
- Number of Dependents: Número de dependientes a cargo del individuo, que afecta las responsabilidades financieras.
- City: Ciudad en la que reside el individuo, proporcionando contexto geográfico.

- State: Estado en el que reside el individuo, proporcionando más detalles geográficos.
- Country: País de residencia, añadiendo una perspectiva global.
- Previous Defaults: Número de incumplimientos de préstamos anteriores, indicando el riesgo financiero histórico.
- Marital Status: Número de cambios en el estado civil, reflejando cambios en la vida personal.
- Risk Rating: Columna objetivo que categoriza el riesgo financiero en Bajo, Medio o Alto.

1.2 Revisión bibliográfica

1.2.1 Búsqueda de bibliografía

Entre las posibles combinaciones de palabras clave se se pueden encontrar:

- Ingreso + situación laboral + solvencia crediticia
- Prestamo + incumplimiento de préstamos + riesgo financiero
- Ciudad + propósito del prestamo + ingreso
- Situacion laboral + duración de empleo + historico de pagos
- Historico de pagos + incumplimiento de prestamos + categoria de riesgo
- Valor de activos + prestamo + relación entre deuda e ingreso

1.2.2 Fichas de literatura

Título: La valoración del riesgo financiero.

- Autor: Dorina Chicu.
- Año: 2020.
- Nombre del tema: Métodos para medir el riesgo.
- Cronología: 2020.
- Metodología: Recolección de datos.
- Temática: Estudios económicos.
- Teórica: Valoración de riesgos.
- Resumen en una oración: Distintos riesgos existentes y algunas formas de medirlos.

- Argumento central: Analizar algunos de los distintos métodos existentes para medir los riesgos financieros.
- Problema con el argumento o el tema: Aunque el tema principal gire en torno a la valoración de riesgos financieros, el trabajo queda falente de varios detalles que, si pudieran ser notorios a la hora de hacer un análisis más exhaustivo, además de la falta de ejemplos u aplicaciones de estos, quedan solo como algo teórico.
- Resumen en un párrafo: El estudio busca centrarse en uno de los tres componentes de la inversión el cual es el riesgo financiero, de manera que se sabe que el objetivo principal de una empresa es maximizar sus beneficios, tal que llegue a asegurar la máxima rentabilidad posible. Por tanto, lo que se quiere brindar son los distintos métodos existentes que sirven para medir o valorar el riesgo, lo cual lleva a que sea posible generar una estrategia que permita mitigar los mismos.

Título: La evaluación del riesgo de crédito en las instituciones de microfinanzas: estado del arte.

- Autor: María Seijas, Milagro Vivel, Rubén Lado, Sara Fernández.
- Año: 2017.
- Nombre del tema: Riesgos en los microcréditos.
- Cronología: 2015 - 2017.
- Metodología: Recolección y comparación de datos.
- Temática: Estudios económicos.
- Teórica: Valoración de riesgos.
- Resumen en una oración: medición del riesgo de los microcréditos y análisis de los clientes.
- Argumento central: Explorar la diversa teoría existente al riesgo de crédito de las instituciones financieras.
- Problema con el argumento o el tema: Aunque el tema principal gire en torno a la valoración de riesgos financieros, el trabajo queda falente de varios detalles que, si pudieran ser notorios a la hora de hacer un análisis más exhaustivo, además de la falta de ejemplos u aplicaciones de estos, quedan solo como algo teórico.
- Resumen en un párrafo: Este trabajo busca exponer a través de las investigaciones que se han centrado en la evaluación del riesgo de crédito en las Instituciones de Microfinanzas, aquella teoría relacionada con el riesgo presente en los microcréditos, además busca analizar todos aquellos factores que llegan a ser determinantes en el riesgo de que haya algún tipo de impago., por lo que este estudio también ofrece ciertas técnicas que generan una mayor consistencia y trasparencia en la evaluación y seguimiento de los clientes y su perfil.

Título: Modelos para otorgamiento y seguimiento en la gestión del riesgo de crédito.

- Autor: Millán Solarte, Julio César; Cerezo, Edinson Caicedo.
- Año: 2018.
- Nombre del tema: Riesgo de crédito.
- Cronología: 2018.
- Metodología: Análisis cuantitativa.
- Temática: Estudios económicos.
- Teórica: Valoración de riesgos.
- Resumen en una oración: Riesgos financieros y análisis en la gestión del riesgo de crédito.
- Argumento central: Explorar maneras de gestionar el riesgo financiero a través del análisis de las solicitudes.
- Problema con el argumento o el tema: Uno de los posibles problemas que se destacan en este estudio es que no se llega a profundizar en las posibles limitaciones que se pueden llegar a presentar a la hora de usar las técnicas que se narran y exploran, así como no se toma en cuenta que varias de las variables llegan a ser muy volátiles a lo largo del tiempo, además de los comportamientos de mercado que puedan incidir en el incumplimiento de pago.
- Resumen en un párrafo: Este trabajo habla acerca del riesgo financiero, más específicamente en el riesgo de crédito el cual se refiere a las perdidas derivadas del incumplimiento de obligaciones financieras, de manera que las instituciones financieras buscan gestionar este tipo de riesgos a través del análisis de las solicitudes por medio del sistema de scoring de crédito, dónde se evalúan variables como la situación financiera e historial de pagos del solicitante para diferenciar entre los buenos y malos clientes.

Título: Variables determinantes de la probabilidad de incumplimiento de un microcrédito en una entidad microfinanciera del Perú, una aproximación bajo el modelo de regresión logística binaria.

- Autor: María Calixto, Luis Casaverde
- Año: 2011.
- Nombre del tema: Incumplimiento de un microcrédito.
- Cronología: 2011.
- Metodología: Recolección y comparación de datos.
- Temática: Estudios económicos.

- Teórica: Valoración de riesgos.
- Resumen en una oración: Medición del riesgo de los microcréditos y el incumplimiento de pago.
- Argumento central: Explorar la probabilidad de incumplimiento de pago de los clientes a través del modelo de la regresión logística binaria.
- Problema con el argumento o el tema: La mayor problemática es que solo se enfoca en el modelo predictivo, de manera que no se llegan a abordar las posibles causas estructurales que al final terminan por inducir que hayan morosidades, de tal forma que los modelos como lo son la regresión logística binaria se va a ver muy limitada para lograr el propósito de reducir las morosidades.
- Resumen en un párrafo: Esta investigación gira entorno a los microcréditos y el incumplimiento de pago en estos, de manera que se busca identificar los factores que influyen en que exista mayor probabilidad de incumplimiento de los clientes, donde se utilizan modelos predictivos que permiten evaluar tanto variables cuantitativas como cualitativas relacionadas con las personas prestatarias, para ello se toman en cuenta factores socio-demográficos, económicos y financieros que lleguen a ser determinantes para el incumplimiento de pago.

Título: Credit Scoring en Costa Rica y la probabilidad de clasificación de créditos personales basados en un modelo estadístico-matemático para aprobar o rechazar.

- Autor: Patricia Hernández, Pablo Montoya, Allan Villareal
- Año: 2013.
- Nombre del tema: Incumplimiento de un microcrédito.
- Cronología: 2012-2013.
- Metodología: Recolección y comparación de datos.
- Temática: Estudios económicos.
- Teórica: Valoración de riesgos.
- Resumen en una oración: Analizar los riesgos en la designación de créditos a través de los modelos de credit scoring, tomando en cuenta variables cualitativas.
- Argumento central: Explorar la probabilidad de incumplimiento de pago de los clientes a través del modelo de la regresión logística binaria.

- Problema con el argumento o el tema: La mayor problemática es que solo se basa en los créditos personales para el consumo, y no toma en cuenta otros propósitos que se les puedan dar a los mismos, además de que en el estudio se menciona que no pudieron tener acceso a ciertas políticas que toman en cuenta los bancos para aceptar o rechazar un crédito.
- Resumen en un párrafo: Esta investigación se basa en el análisis de riesgos en la designación de créditos y el incumplimiento de pago a través de los modelos de credit scoring , de manera que analizar como a partir de estos se puede tratar de llegar a optimizar la asignación de los créditos y minimizar los riesgos, de manera que se termina analizando no solo el nivel de ingreso de una persona, sino también otros factores cualitativos como son la edad, el estado civil, el género, la escolaridad, entre otros.

1.3 Construcción de la UVE de Gowin

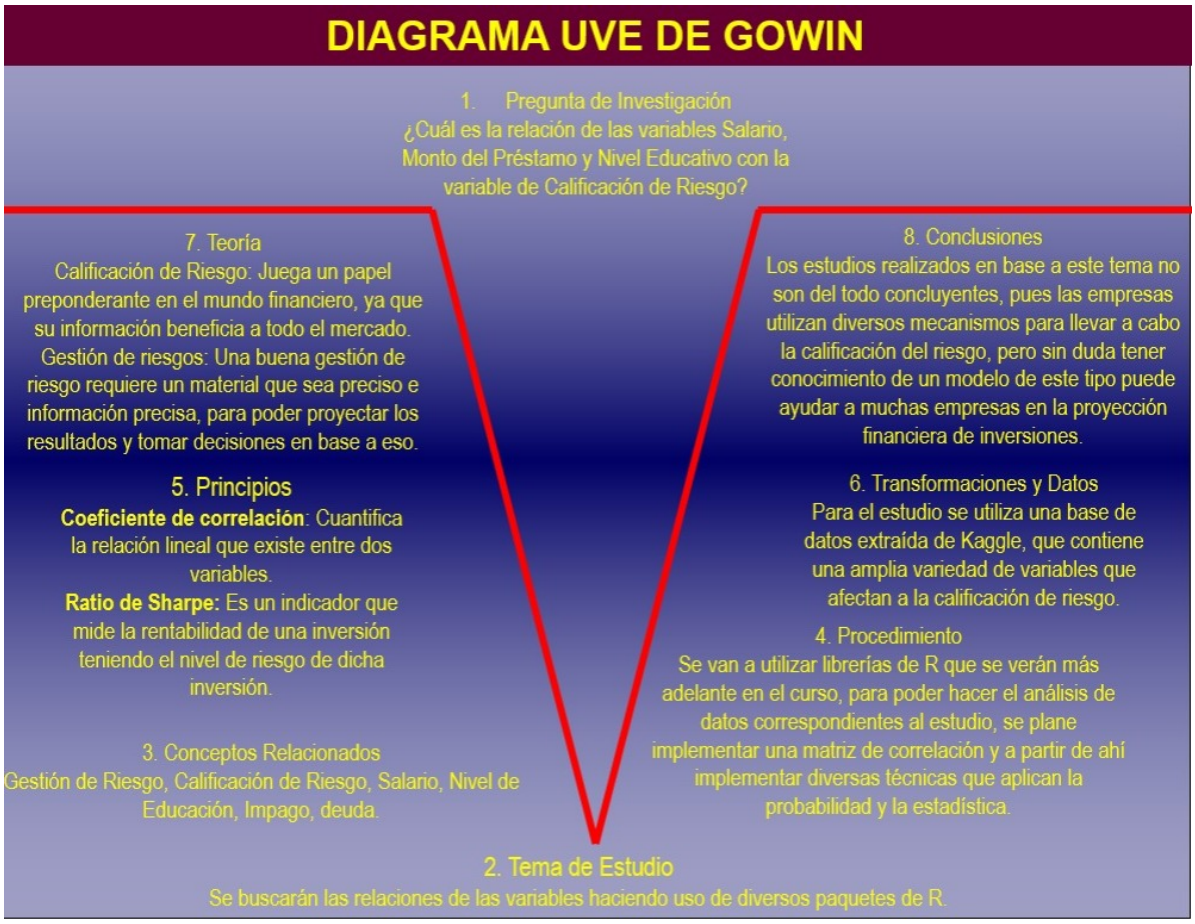


Figura 1.1: V de Gowin

1.3.1 Conceptos básicos

Como hemos estado hablando a lo largo del trabajo, la calificación de riesgo es de suma importancia, por lo que el estudio posee como objetivo determinar una relación fuerte de las variables, para aproximar un método empírico de una calificación de riesgo, pues ya vimos que las empresas por lo general no comparten esta información.

Por otro lado, la gestión de riesgo y la incertidumbre se vuelven esenciales a lo largo del trabajo, pues son conceptos que son base para el estudio posterior. El riesgo de impago también se vuelve un concepto a tener en cuenta, el cual vamos a entender como la incapacidad de cumplir una obligación financiera.

1.3.2 Principios y teorías

Para llevar a cabo la investigación en estas etapas preliminares hemos hecho estudio bibliográfico de lo que son los análisis de calificación financiera, que es justo el tema y variable objetivo de nuestro estudio. También se incorpora la teoría de la gestión de riesgos, que a su vez hace uso de técnicas de probabilidad pues dentro de su corpus la idea principal es maximizar una ganancia en eventos que presentan incertidumbre, de ahí la estrecha relación que tiene con Probabilidad.

Por otro lado, utilizar los conceptos de la valoración de riesgos financieros, sirve de base para entender cómo funciona el mercado financiero, y cuáles son los impactos directos de poseer una calificación de riesgos, y que el impacto es de hecho, para todos los participantes.

1.4 Parte de escritura

El problema que se va a tratar en el presente trabajo es de determinar la relación existente entre las variables de ingresos, nivel de educación y Monto del préstamo con la calificación de riesgo. Desde el punto de vista teórico, el autor (Palacios 2012), menciona que "La principal función que radica en las calificaciones crediticias es la evaluación de la mayor o menor probabilidad de pago de la deuda y los intereses, proporcionando indicadores que sirvan de referencia a los inversores con el fin de que puedan tener conocimiento del riesgo crediticio de una forma simple y accesible". Desde este punto de vista, hay un apoyo en la investigación que tratamos de realizar, pues la calificación de riesgo es de suma importancia en el mundo financiero. Este mismo autor menciona que "Su importancia deriva de su implantación dentro de la regulación, lo que afecta a todo el entramado institucional, y sectores clave de la sociedad como son el bancario y las agencias de seguros y reaseguros", podemos ver entonces que la teoría respalda la importancia que hemos estado conjeturando en esta presente bitácora (Entiéndase conjeturando, porque aún no hemos realizado el análisis de la tabla de datos).

El estudio del análisis financiero es de suma importancia en la actualidad, ya que las transacciones de los flujos de dinero cada vez son mayores, es decir, vender deuda para obtener financiamiento en el corto plazo es una de las estrategias más aplicadas, por ello tanto inversores como prestatarios, según (Palacios 2012), "Los inversores hacen uso de las calificaciones crediticias como un indicador de la probabilidad de recuperar su dinero. Adicionalmente, los prestatarios pueden beneficiarse de tener calificada su deuda, con el objetivo de colocarla con mayor facilidad y eliminar las dudas que haya relación a ellos." Por ello, ambas parten obtienen beneficio de que exista este rating en el mundo de la información financiera. Y desde el punto de vista del inversor, como menciona la autora (Chicu 2020), "…a la hora de analizar una inversión, debemos valorar la rentabilidad esperada, así como la liquidez que perdemos y el riesgo que estamos dispuestos a sumir". Por lo tanto, poseer la información de rating es de suma utilidad, pues ayuda a los inversores a realizar mejores proyecciones. En adición, haciendo referencia a esta misma autora "…la gestión de riesgos tiene un lugar cada vez que un

inversor analiza e intenta cuantificar las pérdidas potenciales en una inversión y luego toma las medidas apropiadas, considerando sus objetivos de inversión y su tolerancia al riesgo." Esto último viene de la mano con lo que son las proyecciones, pues le permite al inversionista hacer un mejor análisis y una gestión de riesgos adecuada, que podemos definir según Chicu como "El proceso de identificación, análisis e incorporación de la incertidumbre en las decisiones de inversión" (Chicu 2020). Reforzando lo que menciona Chicu, la autora (Maria de los Ángeles Herrera 2024), menciona en adición a la gestión de riesgos "el contexto de incertidumbre genera inevitablemente un riesgo, y es ahí cuando la institución financiera debe preservar su valor económico y la integridad de los recursos confiados por los depositantes y socios." Y añadiendo la definición de esta misma autora tenemos que la gestión de riesgos es "la denominación que se utiliza para el conjunto de técnicas y herramientas que pretenden maximizar el valor económico de la institución financiera, en un contexto de incertidumbre". Concluyendo, la gestión de riesgos depende íntimamente de la calificación de riesgo, pues permite tener un parámetro ante la incertidumbre que representa invertir.

2 Bitácora 2

2.1 Parte de Planificación

2.1.1 Ordenamiento de la Literatura

Tabla de Organización y Literatura

Tipo	**Tema General**	**Tema Específico**	**Título**	**Año**	**Autor(es)**
Metodológica	Correlación y Análisis de datos	Análisis del perfil de los prestatarios	La evaluación del riesgo de crédito en las instituciones de microfinanzas: estado del arte	2017	María Seijas, Milagro Vivel, Rubén Lado, Sara Fernández
Metodológica	Correlación y Análisis de datos	Relevancia del scoring de crédito	Modelos para el otorgamiento y seguimiento en la gestión del riesgo crediticio	2018	Millan Solarte, Julio Cesar, Edinson Caicedo
Metodológica	Correlación y Análisis de datos	Relación variables cuantitativas y cualitativas en el incumplimiento de pago	Variables determinantes de la probabilidad de incumplimiento de un microcrédito	2011	María Calixto, Luis Casaverde
Metodológica	Correlación y Análisis de datos	Clasificación de perfiles basándose en variables cualitativas	Credit Scoring en Costa Rica y la probabilidad de clasificación de créditos personales	2013	Patricia Hernández, José Montoya, Allan Villareal

```r
if (!requireNamespace("kableExtra", quietly = TRUE)) {
  install.packages("kableExtra")
}
library(kableExtra)

data <- data.frame(
  Tipo = rep("Metodológica", 4),
  Tema_General = rep("Correlación y Análisis de datos", 4),
  Tema_Especifico = c("Análisis del perfil de los prestatarios",
                      "Relevancia del scoring de crédito",
                      "Relación variables cuantitativas y cualitativas en el incumplimiento c
                      "Clasificación de perfiles basándose en variables cualitativas"),
  Título = c("La evaluación del riesgo de crédito en las instituciones de microfinanzas: esta
             "Modelos para el otorgamiento y seguimiento en la gestión del riesgo crediticio"
             "Variables determinantes de la probabilidad de incumplimiento de un microcrédito
             "Credit Scoring en Costa Rica y la probabilidad de clasificación de créditos per
  Año = c(2017, 2018, 2011, 2013),
  Autor = c("María Seijas, Milagro Vivel, Rubén Lado, Sara Fernández",
            "Millan Solarte, Julio Cesar, Edinson Caicedo",
            "María Calixto, Luis Casaverde",
            "Patricia Hernández, José Montoya, Allan Villareal")
)


kable(data, format = "html", escape = FALSE,
      col.names = c("Tipo", "Tema General", "Tema Específico", "Título", "Año", "Autor(es)")
  kable_styling(full_width = FALSE, position = "left")
```

2.2 Enlaces de la Literatura

Título: *La valoración del riesgo financiero.*

Resumen: El riesgo financiero es uno de los tipos de riesgo más importantes, por eso es de suma importancia entender su significado y qué herramientas se pueden utilizar para su debida gestión. En este caso se enfocan en tres métodos distintos: la desviación estándar, la beta del mercado y el valor en riesgo; así como comprender la rentabilidad ajustada con el ratio de Sharpe. Al utilizarlos en estrategias de cobertura de riesgo financiero se logra una reducción o mitigación del mismo en diferentes intrumentos financieros como las carteras de inversión.

Contraste: Este trabajo, a comparación del artículo de Solarte J. y Caicedo E. titulado Modelos para otorgamiento y seguimiento en la gestión del riesgo de crédito, logra explicar de

manera más amigable y general los conceptos del riesgo financiero, lo que hace que el artículo de Chicu sea más comprensible para todas las personas. Sin embargo, esta falta de rigurosidad puede generar fallas en los procesos más complejos del tema de investigación, ya que el mismo artículo de Solarte y Caicedo utiliza fórmulas y métodos más complejos que si bien no son tan fáciles de entender, logran una mejor recopilación de resultados.

Comentario Propio: Si bien nos parece importante que una explicación sea comprensible para la mayor cantidad de personas, no tener la suficiente precisión en un tema tan complicado como lo es el riesgo financiero puede causar ignorancia en los conceptos avanzados, como por ejemplo los métodos paramétricos y no paramétricos y su metodología. Sin embargo, sentimos que esta investigación nos ayudó con la afinidad a la hora de seleccionar las palabras adecuadas en el desarrollo de nuestra investigación.

Título: *La evaluación del riesgo de crédito en las instituciones de microfinanzas: estado del arte.*

Resumen: El análisis de los modelos de credit scoring muestra que estos pueden ser utilizados en diversas dimensiones, aunque hay una preferencia por predecir el riesgo de retrasos costosos en los microcréditos los cuales están bajo control de las instituciones financieras para mitigarlos. La literatura teórica destaca la importancia de la información estadística cualitativa en estos modelos, además de incluir variables relacionadas con el prestatario, su negocio y el préstamo. Existe una preferencia por las téncicas paramétricas como por ejemplo la regresión logística. Aún así, en los últimos años se registra un aumento en el uso de técnicas no paramétricas que logran un mayor poder predictivo en la detección de incumplimientos de microcréditos. Con esto, las instituciones están incorporando estas herramientas para convertirlas en una práctica estándar, como en las instituciones bancarias; las cuales mejoran la consistencia, transparencia, control de calidad y optimización de los procesos.

Contraste: Éste trabajo, a diferencia del artículo de Chicu D. titulado La valoración del riesgo financiero, logra realizar no sólo un análisis del credit scoring por medio de teoría estadística, sino que también logra plantear un análisis macroeconómico y toma en cuenta las repercusiones que las fallas en los procesos de gestion de riesgos podría generar en el sistema financiero. Sin embargo, no es tan específico en la parte matemática a la hora de estimar probabilidades de que algo pase.

Comentario Propio: Para nosotros es importante entender también la teoría desde la perspectiva economista ya que nos ayuda a entender a qué nos enfrentamos y cómo podemos aplicar diferentes metodologías de investigación.

Título: *Modelos para otorgamiento y seguimiento en la gestión del riesgo de crédito.*

Resumen: El credit scoring es un método estadístico para estimar la probabilidad de inclumpimiento de un prestatario, usando su información histórica y estadística para obtener un indicador que permita distinguir la calidad de un deudor. Los modelos de scoring son muy importantes para los procesos de gestión del crédito, los cuales buscan explicar la composición y operatividad de estos modelos utilizando grandes bases de datos. La información que resulta

de estos modelos permite el análisis de la toma de decisión de si se otorga ó no un crédito a una persona. Por medio de cuatro modelos distintos se logra un procedimiento multicapa para obtener información precisa para calificar a un cliente como bueno o malo para la empresa financiera.

Contraste: Este trabajo logra ser más específico que los demás en cuanto a los métodos de gestión de riesgo se refiere. Además coincide con varios de los artículos recopilados respecto a la utilización del método de regresión logística como herramienta de calificación de crédito.

Comentario Propio: Consideramos que este artículo es el más completo en cuanto a materia matemática se refiere, ya que logra explicar los métodos de manera clara y además los muestra con las gráficas y resultados obtenidos. De hecho, es de gran ayuda que las variables a trabajar sean similares a las que la investigación utiliza.

Título: *Variables determinantes de la probabilidad de incumplimiento de un microcrédito en una entidad microfinanciera del Perú, una aproximación bajo el modelo de regresión logística binaria.*

Resumen: Después de realizar una estimación probabilística del incumplimiento de un microcrédito por medio de la aproximación de una función logística binaria, se determina que son las variables cualitativas como el estado civil, edad y tipo de vivienda junto con las variables de plazo, número de créditos con la entidad y el saldo deudor; las que generan un modelo correctamente ajustado bajo el modelo de regresión logística. La cual logra una capacidad predictiva aceptable medida por la curva ROC.

Contraste: Este trabajo, a diferencia de "La evaluación del riesgo de crédito en las instituciones de microfinanzas: estado del arte", determina las variables necesarias para un estudio de microfinanzas. Sin embargo, coinciden en todo lo referente a las ventajas y deventajas del uso de la calificación crediticia, aunque si cabe mencionar que esta investigación tambien logra hacer conclusiones de los inversores y reguladores como parte de los responsables de la ineficiencia que ha dado el uso excesivo de ésta calificación.

Comentario Propio: Fue de mucha ayuda entender la noción de qué variables utilizar y porqué son importantes para un estudio de calificación de crédito.

Título: *Calificación de riesgo: definición e influencia en la última década.*

Resumen: Las agencias de calificación crediticia han crecido en las últimas décadas por su capacidad para reducir las asimetrías de información en los mercados, facilitando la liquidez y aumentando los participantes. Sin embargo, no han cumplido con los efectos positivos esperados en la última década, se han expuesto fallas en su funcionamiento, como su papel en la burbuja de deuda y la crisis económica, lo que ha generado inestabilidad y ralentizado la recuperación. La importancia que los inversores le dieron a las calificaciones crediticias para la toma de desiciones fue desmedida al no tomar en cuenta el nivel de riesgo, lo cual se relaciona a las causas de las últimas crisis financieras e hipotecarias.

Contraste: En comparación de la investigación titulada "La evaluación del riesgo de crédito en las instituciones de microfinanzas: estado del arte", ambos logran concluir cosas muy similares en cuanto a la importancia del credit score, sus ventajas y desventajas. Sin embargo, logra aportar más en cuanto a las ineficiencias que ha tenido sobrevalorarlo y los momentos en los que ha generado una crisis.

Comentario Propio: Este trabajo, de igual forma, es de suma importancia para entender la manera en que se realizan calificaciones crediticias y de qué formas analizarlas. Para así utilizarlas en nuestros estudios estadísticos y entender qué otras variables ó conceptos tomar en cuenta para los futuros resultados a obtener.

Título: Conceptualización del riesgo de los mercados financieros.

Resumen: El riesgo está inmerso en todas las actividades humanas y es entendido como la probabilidad de ocurrencia de un evento que podría inducir un perjuicio. Cuando se habla de riesgo financiero, se habla de una eventual pérdida de dinero que signifique una afectación al sistema financiero ó a alguna institución que sea parte del mismo. Este trabajo es una recopilación algunos de los riesgos en los mercados financieros y presenta algunos métodos válidos para su valoración.

Contraste: Esta investigación, a diferencia de las demás investigaciones, logra tener un glosario completo de definiciones relacionadas al riesgo financiero, sin embargo no va más allá de ser sólamente un artículo de definiciones. Por lo que en materia teórica no tiene nada que aportar.

Comentario Propio: Gracias a estas definiciones, hemos logrado definir mejor qué queremos estudiar de manera más específica y cómo utilizar ciertos conceptos de mejor manera.

2.3 Análisis Estadístico

Como base para realizar este análisis estadístico, nos estamos guíando con la guía del curso de Herramientas de Ciencias de Datos, el cual adjuntamos el link a dicha guía (Solis 2024) y también estamos utilizando el libro escrito por Wickham, el cual también adjuntamos el link (Hadley Wickham 2019).

A modo de introducción, el análisis estadístico consiste en un conjunto de herramientas o técnias que se utilizan para la recolección, el análisis e interpretación de datos. Para este trabajo es imprescindible contar con este set de herramientas.

2.3.1 Análisis Descriptivo

La base de datos ya se encuentra en formato en tidy, recordemos que el formato tidy fue popularizado por el autor Hadley Wickham, donde indican que cada variable debe tener su

propia columna y cada observación su propia fila. Nuestra base de datos cumple con estar en formato tidy.

Vamos a llamar a nuestra base de datos, la cual vamos a utilizar durante el trabajo.

```
Rows: 15000 Columns: 20
-- Column specification --------------------------------------------------------
Delimiter: ";"
chr (10): Gender, Education Level, Marital Status, Loan Purpose, Employment ...
dbl (10): Age, Income, Credit Score, Loan Amount, Years at Current Job, Debt...

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
head(base_financial_risk_assessment)
```

```
# A tibble: 6 x 20
    Age Gender     `Education Level` `Marital Status` Income `Credit Score`
  <dbl> <chr>      <chr>             <chr>             <dbl>          <dbl>
1    49 Male       PhD               Divorced          72799            688
2    57 Female     Bachelor's        Widowed              NA            690
3    21 Non-binary Master's          Single            55687            600
4    59 Male       Bachelor's        Single            26508            622
5    25 Non-binary Bachelor's        Widowed           49427            766
6    30 Non-binary PhD               Divorced             NA            717
# i 14 more variables: `Loan Amount` <dbl>, `Loan Purpose` <chr>,
#   `Employment Status` <chr>, `Years at Current Job` <dbl>,
#   `Payment History` <chr>, `Debt-to-Income Ratio` <dbl>,
#   `Assets Value` <dbl>, `Number of Dependents` <dbl>, City <chr>,
#   State <chr>, Country <chr>, `Previous Defaults` <dbl>,
#   `Marital Status Change` <dbl>, `Risk Rating` <chr>
```

Antes de aplicar cualquier gráfico o análisis de datos a nuestra base de datos, es importante eliminar las variables que no aportan al estudio, por ello, vamos a eliminar los valores NA que vengan en nuestra base de datos.

```
Base_limpia <- na.omit(base_financial_risk_assessment)
head(Base_limpia)
```

```
# A tibble: 6 x 20
    Age Gender     `Education Level` `Marital Status` Income `Credit Score`
```

```
  <dbl> <chr>      <chr>             <chr>             <dbl>          <dbl>
1    49 Male       PhD               Divorced          72799            688
2    21 Non-binary Master's          Single            55687            600
3    59 Male       Bachelor's        Single            26508            622
4    42 Non-binary Master's          Single           116212            707
5    55 Male       High School       Married           70978            706
6    56 Non-binary PhD               Married           21084            702
# i 14 more variables: `Loan Amount` <dbl>, `Loan Purpose` <chr>,
#   `Employment Status` <chr>, `Years at Current Job` <dbl>,
#   `Payment History` <chr>, `Debt-to-Income Ratio` <dbl>,
#   `Assets Value` <dbl>, `Number of Dependents` <dbl>, City <chr>,
#   State <chr>, Country <chr>, `Previous Defaults` <dbl>,
#   `Marital Status Change` <dbl>, `Risk Rating` <chr>
```

Ahora hacemos un análisis estadístico de nuestra base de datos, de todas las variables.

```r
# Instalar kableExtra si no lo tienes
# install.packages("kableExtra")

library(dplyr)
library(tidyr)
library(knitr)
library(kableExtra)

# Como nuestras variables no son del todo numéricas, hay que hacerlo para las variables que s

# Primero hacemos las variables numéricas.
resumen_numericas <- Base_limpia %>%
  summarise(
    Edad_Media = mean(Age, na.rm = TRUE),
    Edad_Minima = min(Age, na.rm = TRUE),
    Edad_Maxima = max(Age, na.rm = TRUE),
    Ingreso_Medio = mean(Income, na.rm = TRUE),
    Ingreso_Varianza = var(Income, na.rm = TRUE),
    Ingreso_Minimo = min(Income, na.rm = TRUE),
    Ingreso_Maximo = max(Income, na.rm = TRUE),
    Prestamo_Medio = mean(`Loan Amount`, na.rm = TRUE),
    Prestamo_Varianza = var(`Loan Amount`, na.rm = TRUE),
    Prestamo_Minimo = min(`Loan Amount`, na.rm = TRUE),
    Prestamo_Maximo = max(`Loan Amount`, na.rm = TRUE)
  )
```

```r
# Variables Categóricas
resumen_categoricas <- tibble(
  Variable = c("Nivel de Educación", "Género", "Estado Civil"),
  Frecuencia = c(
    paste(names(table(Base_limpia$`Education Level`)), collapse = ", "),
    paste(names(table(Base_limpia$Gender)), collapse = ", "),
    paste(names(table(Base_limpia$`Marital Status`)), collapse = ", ")
  )
)

# Luego resumimos la información en un cuadro, para presentarlo mejor.
tabla_resumen <- data.frame(
  Variable = c("Edad", "Ingresos", "Monto del Préstamo", resumen_categoricas$Variable),
  Media = c(resumen_numericas$Edad_Media, resumen_numericas$Ingreso_Medio, resumen_numericas$
  Varianza = c(NA, resumen_numericas$Ingreso_Varianza, resumen_numericas$Prestamo_Varianza, N
  Mínimo = c(resumen_numericas$Edad_Minima, resumen_numericas$Ingreso_Minimo, resumen_numeric
  Máximo = c(resumen_numericas$Edad_Maxima, resumen_numericas$Ingreso_Maximo, resumen_numeric
  Frecuencia = c(rep(NA, 3), resumen_categoricas$Frecuencia)
)

# Mostrar la tabla con la descripción
tabla_resumen_kable <- kable(tabla_resumen, caption = "Resumen de Variables Numéricas y Categ
  kable_styling() %>%
  add_header_above(c(" " = 1, "Resumen" = 5))
tabla_resumen_kable <- tabla_resumen_kable %>%
  kableExtra::footnote(general = "Fuente: Elaboración propia utilizando la base de datos de K
tabla_resumen_kable
```

Tabla 2.2: Resumen de Variables Numéricas y Categóricas

	Resumen				
Variable	Media	Varianza	Mínimo	Máximo	Frecuencia
Edad	43.5817	NA	18	69	NA
Ingresos	70190.3585	849685113	20014	119978	NA
Monto del Préstamo	27577.0679	168240584	5001	49978	NA
Nivel de Educación	NA	NA	NA	NA	Bachelor's, High School, Master's, PhD
Género	NA	NA	NA	NA	Female, Male, Non-binary
Estado Civil	NA	NA	NA	NA	Divorced, Married, Single, Widowed

Note:
Fuente: Elaboración propia utilizando la base de datos de Kaggle

Presentamos las variables de más importancia en nuestro estudio

Tabla 2.3: Resumen de Variables Numéricas y Categóricas

	Resumen				
Variable	Media	Varianza	Mínimo	Máximo	Frecuencia
Edad	43.5817	NA	18	69	NA
Ingresos	70190.3585	849685113	20014	119978	NA
Monto del Préstamo	27577.0679	168240584	5001	49978	NA
Nivel de Educación	NA	NA	NA	NA	Bachelor's, High School, Master's, PhD
Género	NA	NA	NA	NA	Female, Male, Non-binary
Estado Civil	NA	NA	NA	NA	Divorced, Married, Single, Widowed

Note:
Fuente: Elaboración propia utilizando la base de datos de Kaggle

Con nuestra base de datos limpia, vanmos a proceder a calcular algunos estadísticos importantes, por separado. Estos son los datos resumidos de la variable edad.

```
library(dplyr)

# Calculamos las estadísticas de interés para las variables y las almacenamos en un solo data
resultado <- Base_limpia %>%
  summarise(
    # Estadísticos para la variable edad/Age
    media_edad = mean(Age, na.rm = TRUE),
    varianza_edad = var(Age, na.rm = TRUE),
    min_edad = min(Age, na.rm = TRUE),
    max_edad = max(Age, na.rm = TRUE),

    # Estadísticos para la variable ingreso/Income
    media_ingresos = mean(Income, na.rm = TRUE),
    varianza_ingresos = var(Income, na.rm = TRUE),
    min_ingresos = min(Income, na.rm = TRUE),
    max_ingresos = max(Income, na.rm = TRUE),

    # Estadísticos para la variable record_crediticio/Credit Score
    media_record_crediticio = mean(`Credit Score`, na.rm = TRUE),
    varianza_record_crediticio = var(`Credit Score`, na.rm = TRUE),
    min_record_crediticio = min(`Credit Score`, na.rm = TRUE),
    max_record_crediticio = max(`Credit Score`, na.rm = TRUE),
```

```
    # Estadísticos para la variable monto del préstamo/Loan Amount
    media_monto_prestamo = mean(`Loan Amount`, na.rm = TRUE),
    varianza_monto_prestamo = var(`Loan Amount`, na.rm = TRUE),
    min_monto_prestamo = min(`Loan Amount`, na.rm = TRUE),
    max_monto_prestamo = max(`Loan Amount`, na.rm = TRUE),

    # Estadísticos para la variable años de trabajo/Years at Current Job
    media_anyos_trabajo = mean(`Years at Current Job`, na.rm = TRUE),
    varianza_anyos_trabajo = var(`Years at Current Job`, na.rm = TRUE),
    min_anyos_trabajo = min(`Years at Current Job`, na.rm = TRUE),
    max_anyos_trabajo = max(`Years at Current Job`, na.rm = TRUE)
  )

# Imprimimos el resultado
print(resultado)
```

```
# A tibble: 1 x 20
  media_edad varianza_edad min_edad max_edad media_ingresos varianza_ingresos
       <dbl>         <dbl>    <dbl>    <dbl>          <dbl>             <dbl>
1       43.6          218.       18       69         70190.        849685113.
# i 14 more variables: min_ingresos <dbl>, max_ingresos <dbl>,
#   media_record_crediticio <dbl>, varianza_record_crediticio <dbl>,
#   min_record_crediticio <dbl>, max_record_crediticio <dbl>,
#   media_monto_prestamo <dbl>, varianza_monto_prestamo <dbl>,
#   min_monto_prestamo <dbl>, max_monto_prestamo <dbl>,
#   media_anyos_trabajo <dbl>, varianza_anyos_trabajo <dbl>,
#   min_anyos_trabajo <dbl>, max_anyos_trabajo <dbl>
```

```
resultado$media_edad
```

```
[1] 43.5817
```

```
resultado$varianza_edad
```

```
[1] 217.7303
```

```
resultado$min_edad
```

```
[1] 18
```

```
resultado$max_edad
```

```
[1] 69
```

Estos son los datos resumidos de la variable Ingresos

```
resultado$media_ingresos
```

```
[1] 70190.36
```

```
resultado$varianza_ingresos
```

```
[1] 849685113
```

```
resultado$min_ingresos
```

```
[1] 20014
```

```
resultado$max_ingresos
```

```
[1] 119978
```

Estos son los datos resumidos de la variable Monto del Préstamo

```
resultado$media_monto_prestamo
```

```
[1] 27577.07
```

```
resultado$varianza_monto_prestamo
```

```
[1] 168240584
```

```
resultado$min_monto_prestamo
```

```
[1] 5001
```

```r
resultado$max_monto_prestamo
```

```
[1] 49978
```

Resumimos la información obtenida en un cuadro, para una mejor visualización de ellos.

```r
library(knitr)
library(kableExtra)

# Creamos una tabla para resumir la información obtenida
tabla_resumen <- data.frame(
  Variable = c("Edad", "Ingresos", "Monto del Préstamo"),
  Media = c(resultado$media_edad, resultado$media_ingresos, resultado$media_monto_prestamo),
  Varianza = c(NA, resultado$varianza_ingresos, resultado$varianza_monto_prestamo),
  Mínimo = c(resultado$min_edad, resultado$min_ingresos, resultado$min_monto_prestamo),
  Máximo = c(resultado$max_edad, resultado$max_ingresos, resultado$max_monto_prestamo)
)

# Mostramos la tabla con la descripción
tabla_resumen_kable <- kable(tabla_resumen, caption = "Resumen de Variables: Edad, Ingresos y
  kable_styling()

tabla_resumen_kable <- tabla_resumen_kable %>%
  kableExtra::footnote(general = "Fuente: Elaboración propia utilizando la base de datos de K

tabla_resumen_kable
```

Tabla 2.4: Resumen de Variables: Edad, Ingresos y Monto del Préstamo

Variable	Media	Varianza	Mínimo	Máximo
Edad	43.5817	NA	18	69
Ingresos	70190.3585	849685113	20014	119978
Monto del Préstamo	27577.0679	168240584	5001	49978

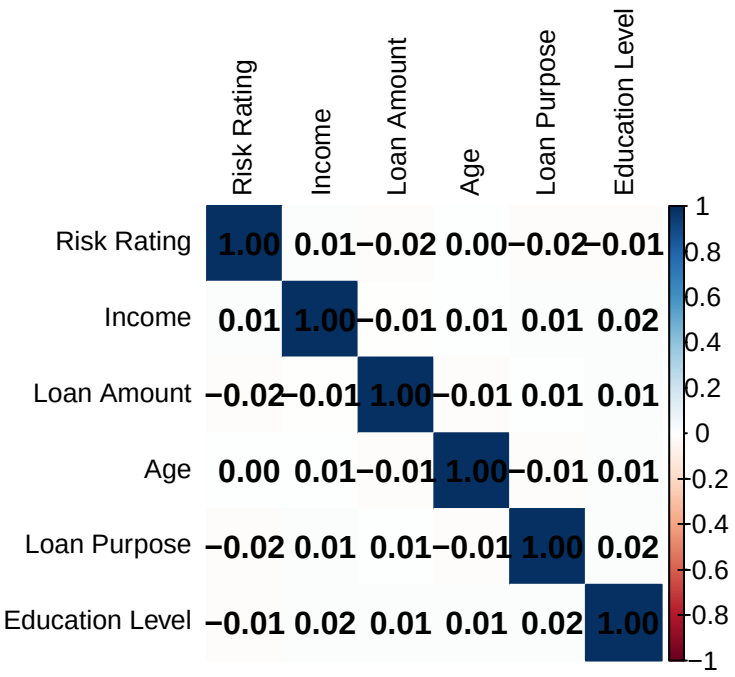
Note:
Fuente: Elaboración propia utilizando la base de datos de Kaggle

Por último, vamos a realizar una matriz de correlación de los datos, esto porque queremos observar la relación que tienen las variables, solo tomaremos las variables de interés, la justificación de dicha escogencia viene del lado teórico, pues son las variables que históricamente más se toman en los estudios de calificación de riesgo.

```r
library(dplyr)
library(corrplot)

# Escogemos las variables que nos interesan para la matriz de correlación.
Base_correlacion <- Base_limpia %>%
  select(`Risk Rating`, Income, `Loan Amount`, Age, `Loan Purpose`, `Education Level`)

# Como hay variables categóricas, entonces vamos a convertir las variables a numérico, para p
Base_correlacion$`Risk Rating` <- as.numeric(as.factor(Base_correlacion$`Risk Rating`))
Base_correlacion$`Loan Purpose` <- as.numeric(as.factor(Base_correlacion$`Loan Purpose`))
Base_correlacion$`Education Level` <- as.numeric(as.factor(Base_correlacion$`Education Level

# Calculamos la matriz de correlación
matriz_correlacion <- cor(Base_correlacion, use = "complete.obs", method = "pearson")
print(matriz_correlacion)
```

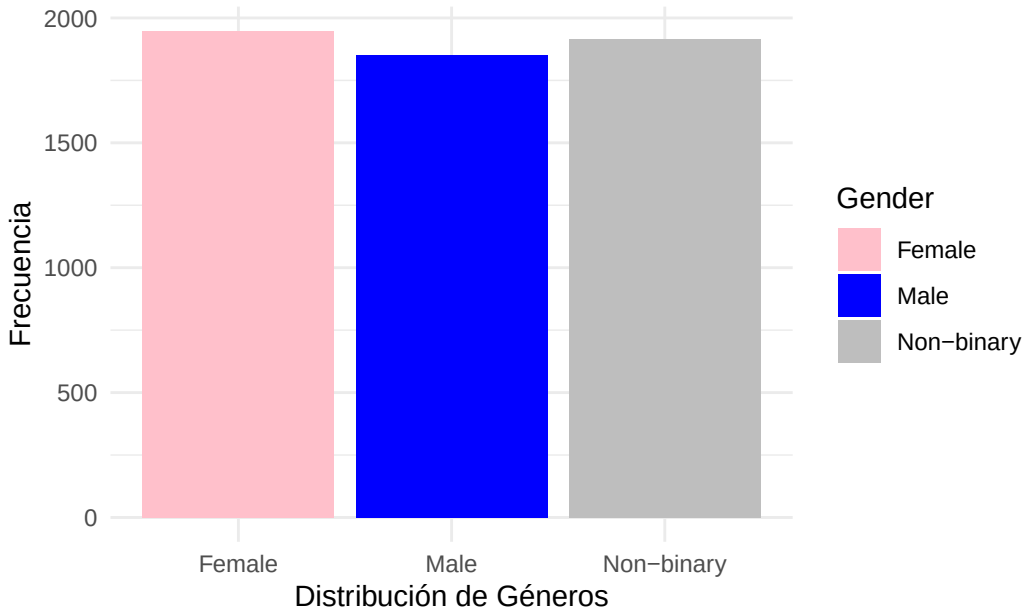
```
                Risk Rating       Income  Loan Amount          Age
Risk Rating      1.000000000  0.013528536 -0.015100412  0.003258428
Income           0.013528536  1.000000000 -0.008137282  0.005019572
Loan Amount     -0.015100412 -0.008137282  1.000000000 -0.011121494
Age              0.003258428  0.005019572 -0.011121494  1.000000000
Loan Purpose    -0.015622201  0.014753633  0.006311870 -0.013760821
Education Level -0.013449909  0.019406630  0.010511349  0.011114696
                Loan Purpose Education Level
Risk Rating      -0.01562220     -0.01344991
Income            0.01475363      0.01940663
Loan Amount       0.00631187      0.01051135
Age              -0.01376082      0.01111470
Loan Purpose      1.00000000      0.01934448
Education Level   0.01934448      1.00000000
```

```r
# Creamos el gráfico de la matriz de correlación
corrplot(matriz_correlacion, method = "color", addCoef.col = "black", tl.col = "black", tl.c

mtext("Fuente: Elaboración propia utilizando la base de datos de Kaggle", side = 1, line = 4
```



Ahora vamos a comenzar a ver los gráficos que se forman de nuestra base de datos. Para ello vamos analizar cómo es la distribución que siguen los géneros de nuestra base de datos, esto es solamente por sondear cómo es nuestra población. Al ser una variable categórica, lo recomendado es realizar un gráfico de barras.

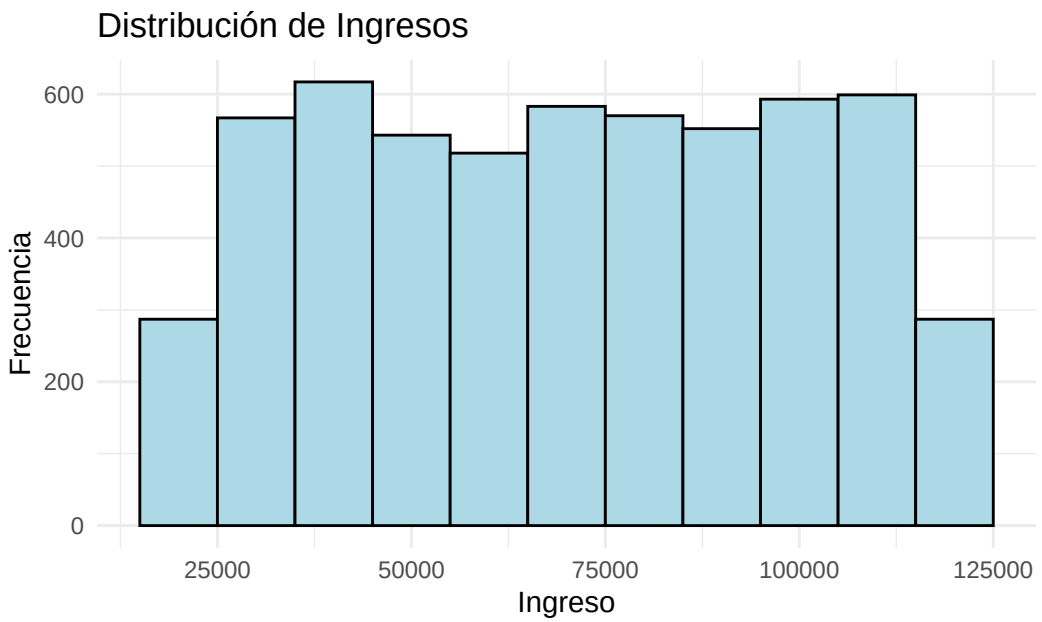
```
library(ggplot2)
# Gráfico con las distribuciones del género
ggplot(Base_limpia, aes(x = Gender, fill = Gender)) +
  geom_bar() +
  scale_fill_manual(values = c("Male" = "blue", "Female" = "pink", "Non-binary" = "gray")) +
  labs(x = "Distribución de Géneros", y = "Frecuencia") + theme_minimal() +
  labs(caption = "Fuente: Elaboración propia utilizando la base de datos de Kaggle") +
theme(plot.caption = element_text(hjust = 0.5))
```



Fuente: Elaboración propia utilizando la base de datos de Kaggle

Por otro lado, la variable "Income", es una variable continua, por lo que lo recomendado para visualizar la distribución de estos datos de una manera rápida es a través de los histogramas, por lo que adjuntamos el gráfico correspondiente:

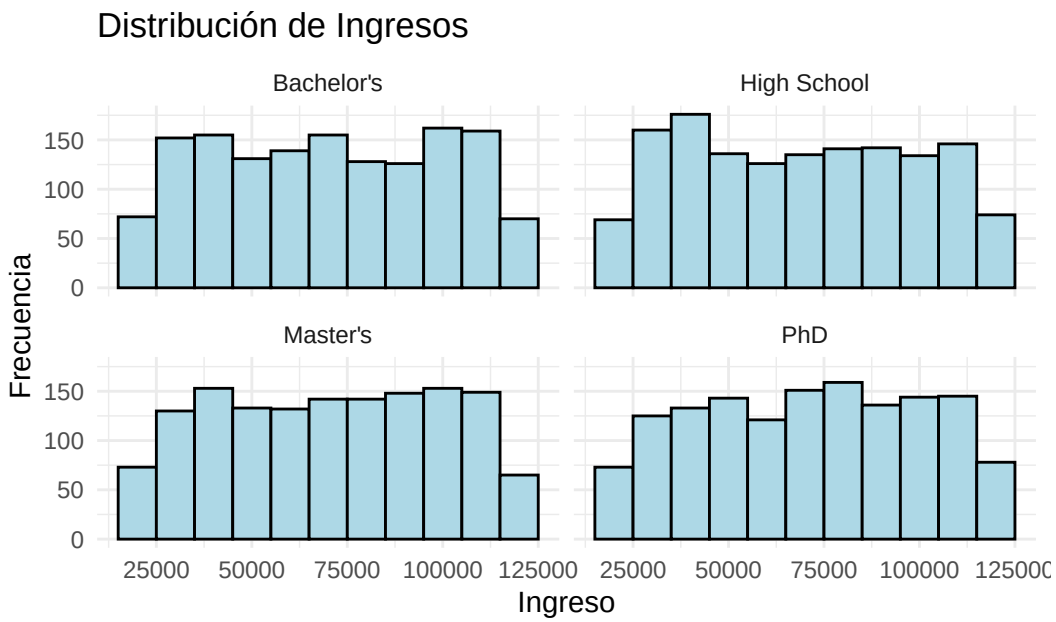
```
library(ggplot2)
ggplot(Base_limpia, aes(x = Income)) +
  geom_histogram(binwidth = 10000, fill = "lightblue", color = "black") +
  labs(x = "Ingreso", y = "Frecuencia", title = "Distribución de Ingresos") +
 theme_minimal() +
  labs(caption = "Fuente: Elaboración propia utilizando la base de datos de Kaggle") +
theme(plot.caption = element_text(hjust = 0.5))
```



Fuente: Elaboración propia utilizando la base de datos de Kaggle

Además decidimos realizar un facet en esta misma variable con respecto a la variable "Education Level", con el fin de visualizar la distribución en cada categoría, esto porque queremos descartar o validar que de alguna forma el nivel educativo tiene relación con nuestra variable objetivo, la cual es la calificación de riesgo.

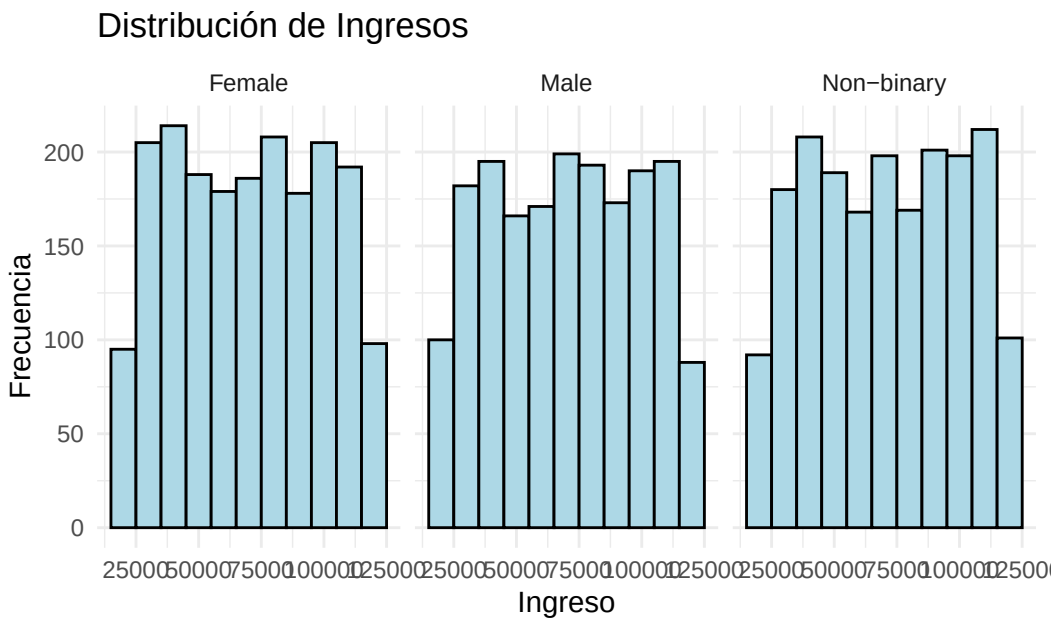
```
library(ggplot2)
ggplot(Base_limpia, aes(x = Income)) +
  geom_histogram(binwidth = 10000, fill = "lightblue", color = "black") +
  labs(x = "Ingreso", y = "Frecuencia", title = "Distribución de Ingresos") + theme_minimal()
  facet_wrap(~ `Education Level`) + # Facet por nivel educativo
  labs(caption = "Fuente: Elaboración propia utilizando la base de datos de Kaggle") +
theme(plot.caption = element_text(hjust = 0.5))
```



Fuente: Elaboración propia utilizando la base de datos de Kaggle

Adjuntamos un gráfico más de los ingresos, pero esta vez por género, con la misma curiosidad de observar cómo es la distribución de estos dada la mencionada variable.

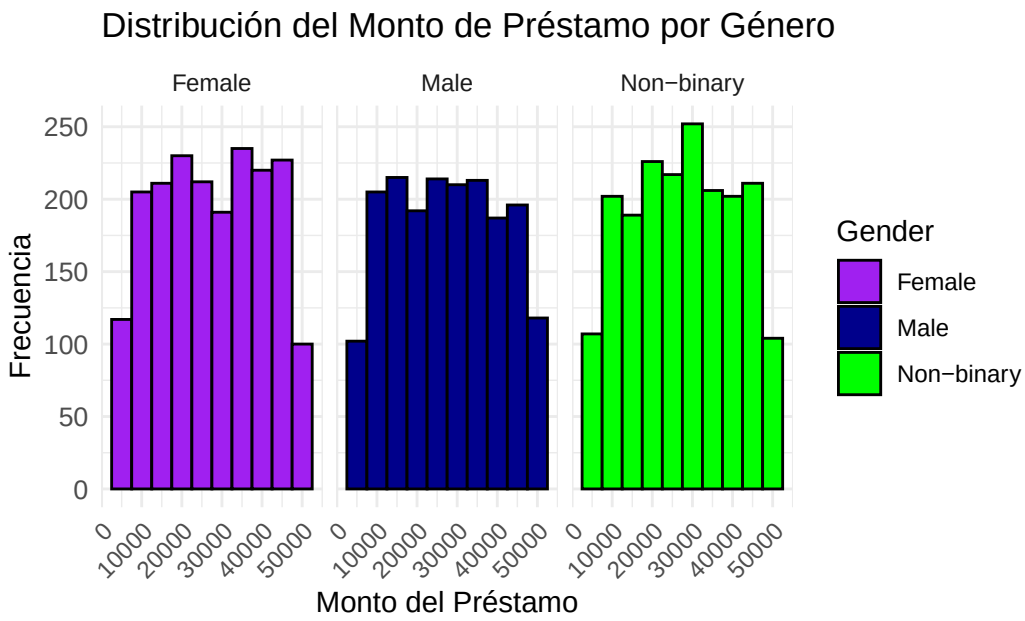
```
library(ggplot2)
ggplot(Base_limpia, aes(x = Income)) +
  geom_histogram(binwidth = 10000, fill = "lightblue", color = "black") +
  labs(x = "Ingreso", y = "Frecuencia", title = "Distribución de Ingresos") + theme_minimal()
  facet_wrap(~ Gender) + # Facet Género
  labs(caption = "Fuente: Elaboración propia utilizando la base de datos de Kaggle") +
theme(plot.caption = element_text(hjust = 0.5))
```



Fuente: Elaboración propia utilizando la base de datos de Kaggle

Otras variables a tener en cuenta aparte de las anteriores mencionadas con respecto al riesgo de crédito, son el monto del préstamos y el propósito del préstamo. Vamos analizar primero el Monto del préstamo, nos interesa ver qué distribución en general tiene.

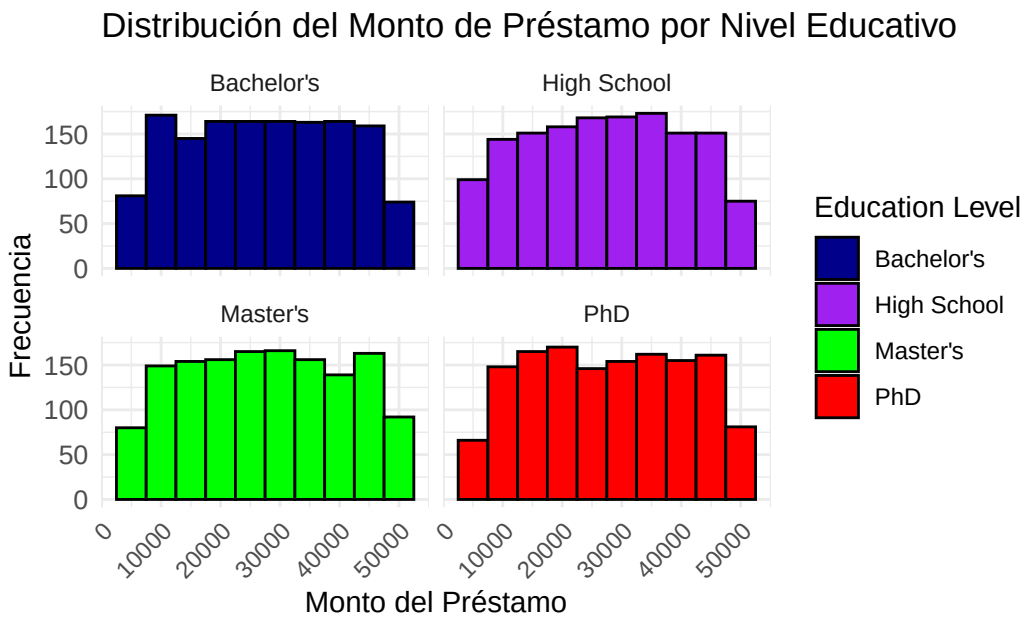
```
library(ggplot2)

ggplot(Base_limpia, aes(x = `Loan Amount`, fill = Gender)) +
  geom_histogram(binwidth = 5000, color = "black", position = "identity") +
  scale_fill_manual(values = c("Male" = "darkblue", "Female" = "purple", "Non-binary" = "gree
  facet_wrap(~ Gender) +
  labs(x = "Monto del Préstamo", y = "Frecuencia", title = "Distribución del Monto de Préstar
  theme_minimal() +
  theme(axis.text.y = element_text(size = 10), # Ajustamos el tamaño del texto
        axis.text.x = element_text(angle = 45, hjust = 1)) + #Rotamos el texto
  labs(caption = "Fuente: Elaboración propia utilizando la base de datos de Kaggle") +
theme(plot.caption = element_text(hjust = 0.5))
```



Fuente: Elaboración propia utilizando la base de datos de Kaggle

Haremos lo mismo, pero esta vez vamos a ver cómo se comportan los salarios cuando cambiamos la variable y utilizamos por ejemplo el grado académico.

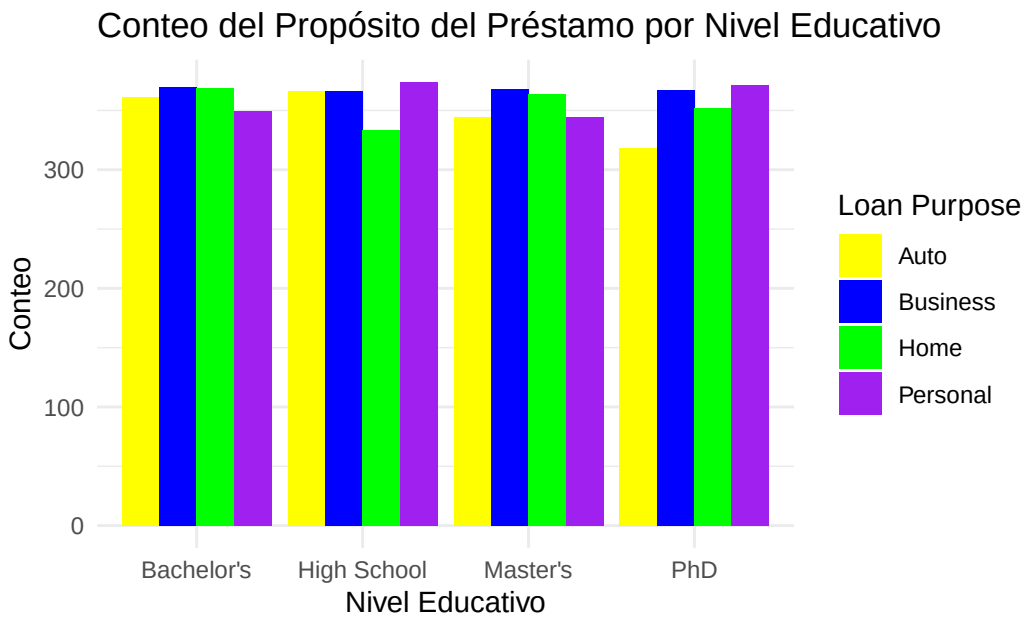
```
library(ggplot2)

ggplot(Base_limpia, aes(x = `Loan Amount`, fill = `Education Level`)) +
  geom_histogram(binwidth = 5000, color = "black", position = "identity") +
  scale_fill_manual(values = c("Bachelor's" = "darkblue", "High School" = "purple", "Master's
  facet_wrap(~ `Education Level`) +
  labs(x = "Monto del Préstamo", y = "Frecuencia", title = "Distribución del Monto de Préstar
  theme_minimal() +
  theme(axis.text.y = element_text(size = 10), # Ajustamos el tamaño del texto
        axis.text.x = element_text(angle = 45, hjust = 1)) + #Rotamos el texto
  labs(caption = "Fuente: Elaboración propia utilizando la base de datos de Kaggle") +
theme(plot.caption = element_text(hjust = 0.5))
```



Fuente: Elaboración propia utilizando la base de datos de Kaggle

Por otro lado, ahora queremos ver el gráfico de la variable propósito del préstamo en relación con el grado académico, al ser ambas variables categóricas, lo recomendado es utilizar una geometría que se adapte a esto, Sin embargo, al ser medidas que están muy cercanas, casi no se aprecia la diferencia, por lo que se decide adaptarlo a un gráfico de barras y apreciar mejor la diferencias de manera visual, también recurrimos al uso de colores, para identificar las variables.

```
library(ggplot2)
library(dplyr)

# Contar las combinaciones de Education Level y Loan Purpose
Base_count <- Base_limpia %>%
  group_by(`Education Level`, `Loan Purpose`) %>%
  summarise(Count = n(), .groups = 'drop')

# Gráfico de barras para visualizar el conteo
ggplot(Base_count, aes(x = `Education Level`, y = Count, fill = `Loan Purpose`)) +
  geom_bar(stat = "identity", position = "dodge") +  # Usa barras para mostrar el conteo
  labs(x = "Nivel Educativo", y = "Conteo", title = "Conteo del Propósito del Préstamo por Ni
  scale_fill_manual(values = c("Business" = "blue", "Personal" = "purple", "Home" = "green",
  theme_minimal() +
  labs(caption = "Fuente: Elaboración propia utilizando la base de datos de Kaggle") +
theme(plot.caption = element_text(hjust = 0.5))
```

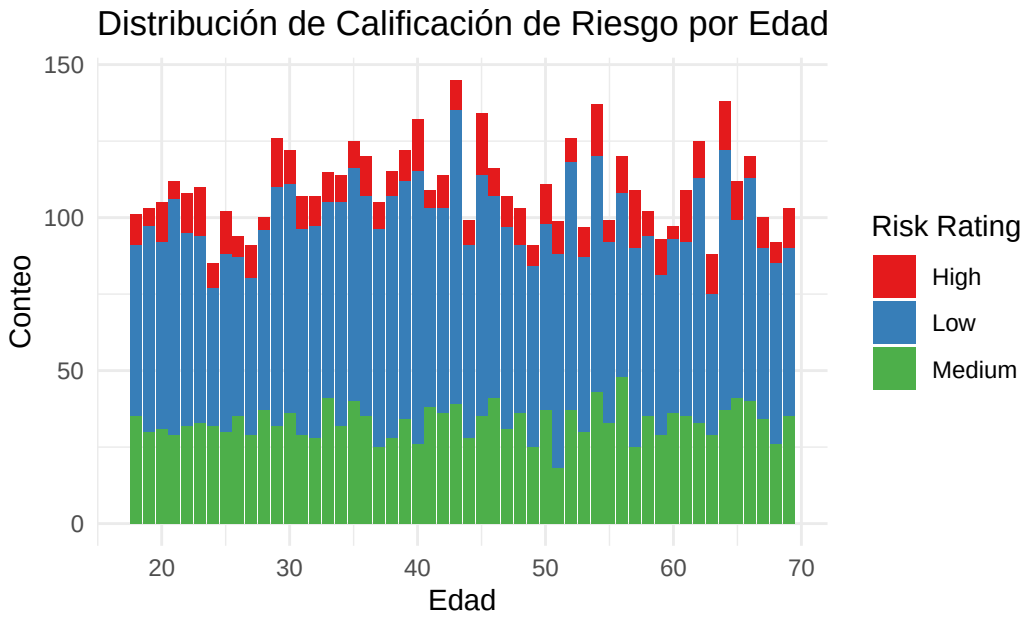


Fuente: Elaboración propia utilizando la base de datos de Kaggle

Para terminar esta sección, vamos a graficar con la variable objetivo de interés, la cual es la Calificación de riesgo, para ello, lo vamos a comparar a través de diversos gráficos con las variables de nivel educativo, ingresos, monto del préstamo, edad y género. Es importante mencionar, que las variables en conjunto afectan a esta calificación, al menos así es de manera teórica. En este apartado nos vamos a centrar en estos gráficos, en secciones posteriores nos ecargaremos de hacer la conexión entre la teoría, nuestras hipótesis y los datos obtenidos.

Para comenzar, nos interesa observar cómo se comporta la variable de Calificación de Riesgo, con respecto a la edad:

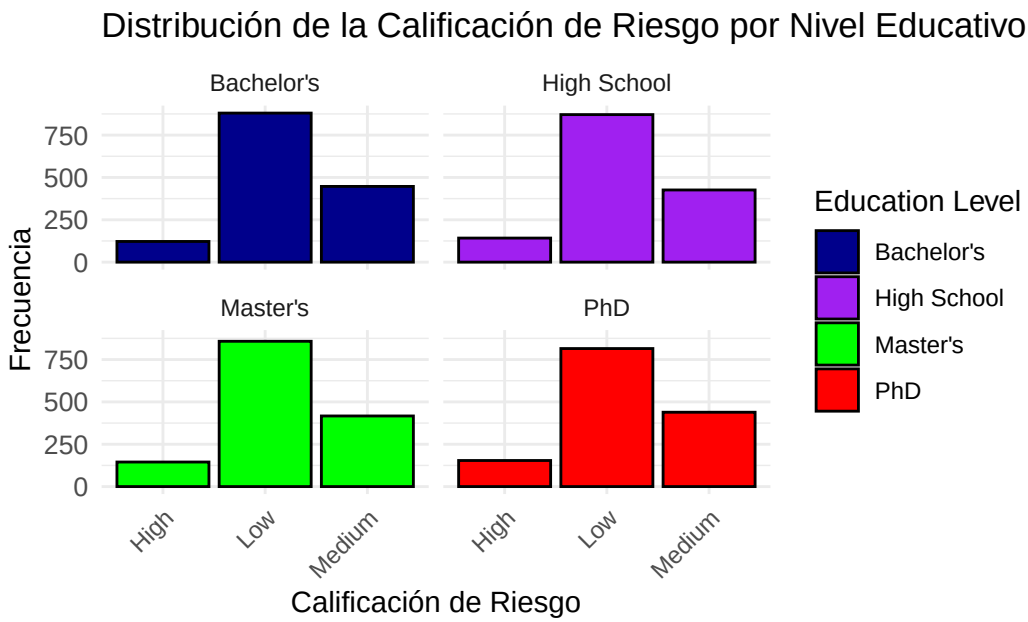
```
library(ggplot2)
ggplot(Base_limpia, aes(x = Age, fill = `Risk Rating`)) +
  geom_bar() +
  labs(x = "Edad", y = "Conteo", title = "Distribución de Calificación de Riesgo por Edad") +
  theme_minimal() +
  scale_fill_brewer(palette = "Set1")  +
  labs(caption = "Fuente: Elaboración propia utilizando la base de datos de Kaggle") +
theme(plot.caption = element_text(hjust = 0.5))
```



Fuente: Elaboración propia utilizando la base de datos de Kaggle

Por otro lado, vamos a ver cómo se comporta la distribución de la calificación de riesgo, haciendo un facet con la categoría Nivel Educativo. Al ser dos variables categóricas las que estamos comparando, lo recomendado es utilizar un gráfico de barras.

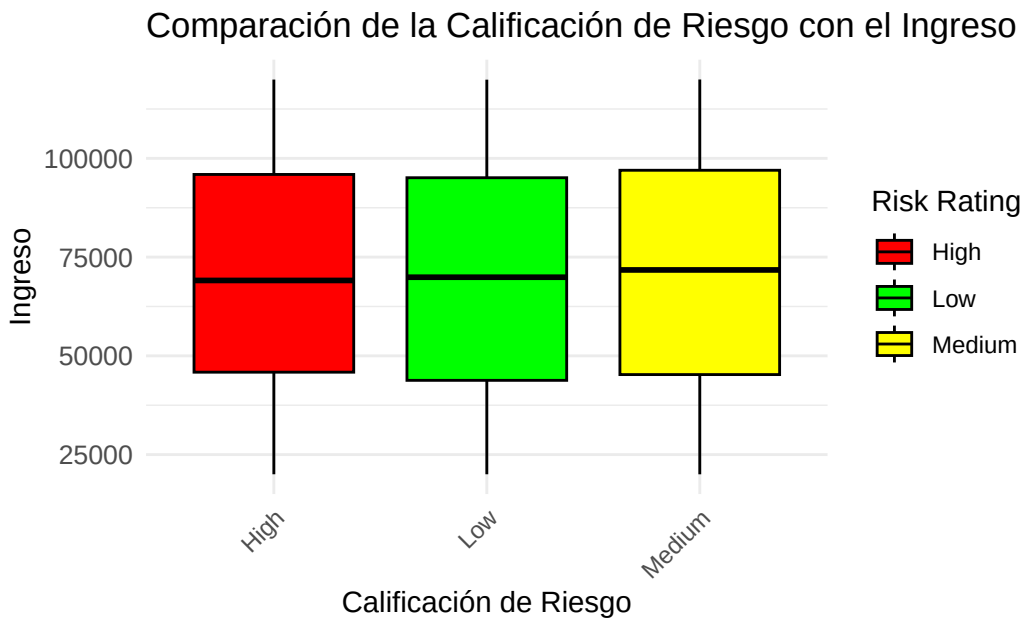
```
library(ggplot2)

ggplot(Base_limpia, aes(x = `Risk Rating`, fill = `Education Level`)) +
  geom_bar(color = "black", position = "identity") +
  scale_fill_manual(values = c("Bachelor's" = "darkblue", "High School" = "purple", "Master's
  facet_wrap(~ `Education Level`) +
  labs(x = "Calificación de Riesgo", y = "Frecuencia", title = "Distribución de la Calificaci
  theme_minimal() +
  theme(axis.text.y = element_text(size = 10), # Ajustamos el tamaño del texto
        axis.text.x = element_text(angle = 45, hjust = 1)) + # Rotamos el texto
  labs(caption = "Fuente: Elaboración propia utilizando la base de datos de Kaggle") +
theme(plot.caption = element_text(hjust = 0.5))
```



Fuente: Elaboración propia utilizando la base de datos de Kaggle

A su vez, nos interesa observar cómo se ve el gráifco de la variable calificación de riesgo con la variable contra la variable ingreso. Para este gráfico, vamos a utilizar el recomendado en las notas del profesor Maikol Solís, el cual indica que usar diagramas de cajas es útil para comparar las distribuciones.

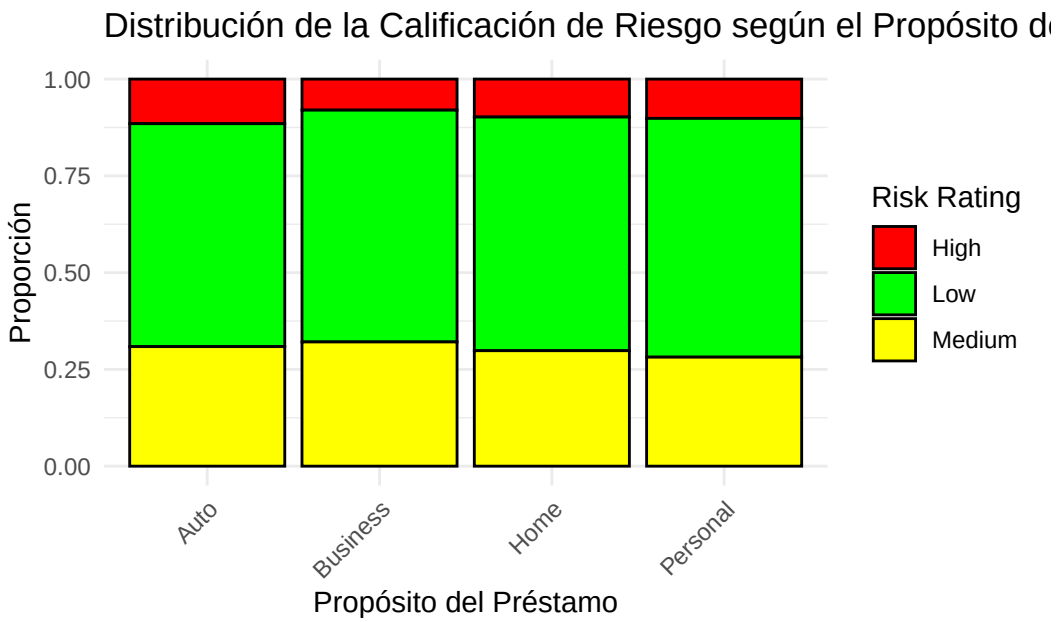
```
library(ggplot2)

ggplot(Base_limpia, aes(x = `Risk Rating`, y = Income, fill = `Risk Rating`)) +
  geom_boxplot(color = "black") +
  scale_fill_manual(values = c("Low" = "green", "Medium" = "yellow", "High" = "red", "Very Hi
  labs(x = "Calificación de Riesgo", y = "Ingreso", title = "Comparación de la Calificación d
  theme_minimal() +
  theme(axis.text.y = element_text(size = 10), # Ajustamos el tamaño del texto
        axis.text.x = element_text(angle = 45, hjust = 1))  +# Rotamos el texto
  labs(caption = "Fuente: Elaboración propia utilizando la base de datos de Kaggle") +
theme(plot.caption = element_text(hjust = 0.5))
```



Fuente: Elaboración propia utilizando la base de datos de Kaggle

Otra de las variables de interés, es la calificación de riesgo contra el propósito del préstamo.

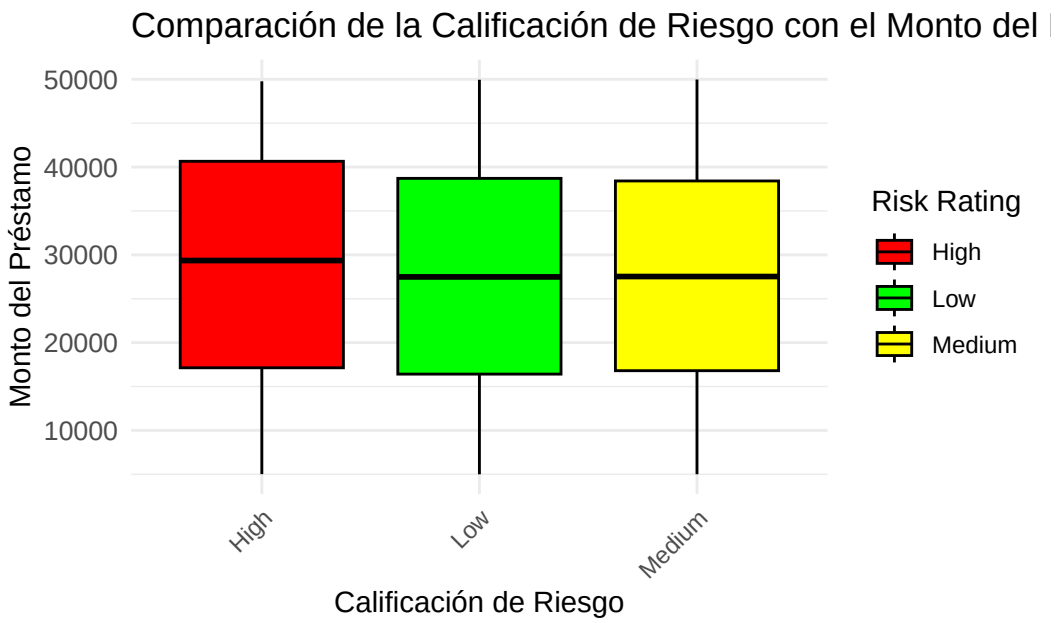
```
library(ggplot2)

ggplot(Base_limpia, aes(x = `Loan Purpose`, fill = `Risk Rating`)) +
  geom_bar(position = "fill", color = "black") +
  scale_fill_manual(values = c("Low" = "green", "Medium" = "yellow", "High" = "red", "Very Hi
  labs(x = "Propósito del Préstamo", y = "Proporción", title = "Distribución de la Calificaci
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) + # Rotamos el texto para mejorar
  labs(caption = "Fuente: Elaboración propia utilizando la base de datos de Kaggle") +
theme(plot.caption = element_text(hjust = 0.5))
```



Fuente: Elaboración propia utilizando la base de datos de Kaggle

Por último, vamos a ver la distribución que existe de la variable calificación de riesgo cuando la comparamos contra el monto del préstamo.

```
library(ggplot2)

ggplot(Base_limpia, aes(x = `Risk Rating`, y = `Loan Amount`, fill = `Risk Rating`)) +
  geom_boxplot(color = "black") +
  scale_fill_manual(values = c("Low" = "green", "Medium" = "yellow", "High" = "red", "Very Hi
  labs(x = "Calificación de Riesgo", y = "Monto del Préstamo", title = "Comparación de la Cal
  theme_minimal() +
  theme(axis.text.y = element_text(size = 10), # Ajustamos el tamaño del texto
        axis.text.x = element_text(angle = 45, hjust = 1))  + # Rotamos el texto
labs(caption = "Fuente: Elaboración propia utilizando la base de datos de Kaggle") +
theme(plot.caption = element_text(hjust = 0.5))
```



Fuente: Elaboración propia utilizando la base de datos de Kaggle

2.4 Propuesta Metodológica

Se utilizan los gráficos, como ayuda en las distribuciones de las variables, con el fin de observar su comportamiento. Sin embargo, hay que hacer un análisis más a profundidad, para ello utilizamos el coeficiente de correlación de Pearson, que fue justo el que utilizamos para crear nuestra matriz de correlación, a continuación enunciamos el proceso teórico de dicho procedimiento.

Como menciona el autor (Edgar Apaza 2022) "Los análisis de correlación son métodos estadísticos descriptivos utilizados en investigación de nivel relacional, con los que estima la magnitud y define la tendencia de la relación entre variables.". Como queremos encontrar alguna relación en nuestra variables de interés, entonces queremos utilizar esta metodología para encontrar dicha relación. Según este mismo autor "el método de correlación de Pearson es una técnica bivariada que emplea en circunstancia multivariada para la explicación de diversos fenómenos relacionados en el campo animal y vegetal. En la correlación de Pearson, los procedimientos guardan relación con la naturaleza de las variables utilizadas.". Como podemos observar el método nos sirve para este estudio, pues tenemos muchas variables, pero el índice sale de comparar dos a dos las variables, para obtener la correlación entre ellas, lo que luego acomodamos en una matriz para tener una mejor observación de ellas.

R ya posee una librería que calcula automáticamente este coeficiente, sin embargo, la manera teórica de hacerlo es mediante la fórmula:

$\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$

En donde ρ recibe el nombre de Coeficiente de correlación de Pearson y además tiene que ocurrir que $-1 \leq \rho \leq 1$. Según el autor (Edgar Apaza 2022) "... se deduce la magnitud de la relación lineal entre variables, los que pueden ser calificados como: Correlación Nula (0), Muy baja (0.01 a 0.1), Débil (0.11 a 0.5), Media (0.51 a 0.75), considerable (0.76 a 0.9), Muy fuerte (0.91 a 0.99) y Perfecta(1)". Otro aspecto importante a decir acerca de nuestro estudio y que tiene relación con lo que este mismo autor afirma y es que " = 0.000, no necesariamente implica que no exista relación entre las varables, sino que la relación podría ser no lineal". Esto de hecho es un buen punto en vista de los resultados que arrojó nuestra matriz de correlación. Además, algo que deberíamos de tomar en cuenta, es que estamos trabajando con variables categóricas, a la hora de la conversión puede haber fuga de información, por ello hay que tener cuidado con solo ver una cifra y animarse a dar una conclusión, cuando en realidad hay que analizar a detalle qué es lo que está pasando.

2.5 Construcción de Fichas de Resultados

```
if (!requireNamespace("kableExtra", quietly = TRUE)) {
  install.packages("kableExtra")
}
library(kableExtra)


data <- data.frame(
  Encabezado = c("Nombre de Su hallazgo",
                 "Resumen en una Oración",
                 "Problemas o Posibles Desafíos",
                 "Resumen en un párrafo"),
  Contenido = c("Poca Correlación entre las variables.",
                "Encontramos que las variables a utilizar no presentan una correlación fuerte
                "La conversión de variable categórica a variable numérica puede estar afectar
                "Al utilizar el índice de correlación de Pearson, hay que utilizar variables
)


if (knitr::is_html_output()) {
  # Si es HTML
  kable(data, col.names = c("Encabezado", "Contenido"),
        format = "html",
        escape = FALSE) %>%
    kable_styling(full_width = FALSE) %>%
```

```
    add_header_above(c("Hallazgo de Resultado 1" = 2), bold = TRUE)
} else {
  # Si es PDF
  kable(data, col.names = c("Encabezado", "Contenido"),
        format = "latex",
        booktabs = TRUE) %>%
    kable_styling(latex_options = c("striped", "hold_position"))
}
```

Encabezado	Contenido
Nombre de Su hallazgo	Poca Correlación entre las variables.
Resumen en una Oración	Encontramos que las variables a utilizar no presentan una correlación fuerte,
Problemas o Posibles Desafíos	La conversión de variable categórica a variable numérica puede estar afectand
Resumen en un párrafo	Al utilizar el índice de correlación de Pearson, hay que utilizar variables num

```
if (!requireNamespace("kableExtra", quietly = TRUE)) {
  install.packages("kableExtra")
}
library(kableExtra)


data <- data.frame(
  Encabezado = c("Nombre de Su hallazgo",
                 "Resumen en una Oración",
                 "Problemas o Posibles Desafíos",
                 "Resumen en un párrafo"),
  Contenido = c("Datos muy Iguales",
                "A la hora de graficar las variables se puede observar que los datos tienen c
                "Este problema se puede estar ocasionando debido a que las variables de la ba
                "Al realizar las diferentes gráficas, con el objetivo de observar la distribu
)


if (knitr::is_html_output()) {
  # Si es HTML
  kable(data, col.names = c("Encabezado", "Contenido"),
        format = "html",
        escape = FALSE) %>%
    kable_styling(full_width = FALSE) %>%
    add_header_above(c("Hallazgo de Resultado 2" = 2), bold = TRUE)
} else {
```

```
  # Si es PDF
  kable(data, col.names = c("Encabezado", "Contenido"),
        format = "latex",
        booktabs = TRUE) %>%
    kable_styling(latex_options = c("striped", "hold_position")) %>%
    add_header_above(c("Hallazgo de Resultado 2" = 2), bold = TRUE)
}
```

	Hallazgo de Resultado 2
Encabezado	Contenido
Nombre de Su hallazgo	Datos muy Iguales
Resumen en una Oración	A la hora de graficar las variables se puede observar que los datos tienen dist
Problemas o Posibles Desafíos	Este problema se puede estar ocasionando debido a que las variables de la ba
Resumen en un párrafo	Al realizar las diferentes gráficas, con el objetivo de observar la distribución

Para la siguiente ficha, es de importancia hacer la acotación, que este curso evalúa el tratamiento de datos, más que el tema de la investigación, pues es un curso de herramientas de datos, por ello, se nos hace pertinente mencionar que el tratamiento de las variables categóricas con variables numéricas puede llevar a problemas, sino se hace un buen tratamiento, además como mencionamos anteriormente en la metodología, que las variables presenten poca correlación, se puede deber a que la relación entre ellas no es lineal o al menos no es cuantifificable, pero a la hora de trabajar o de interpretarlas, si tiene sentido hacerlo o existe un cuerpo teórico que lo apoya. A continuación el cuadro del hallazgo:

```
if (!requireNamespace("kableExtra", quietly = TRUE)) {
  install.packages("kableExtra")
}
library(kableExtra)


data <- data.frame(
  Encabezado = c("Nombre de Su hallazgo",
                 "Resumen en una Oración",
                 "Problemas o Posibles Desafíos",
                 "Resumen en un párrafo"),
  Contenido = c("Dificultad a la hora de trabajar con variables categóricas y numéricas.",
                "Buscar correlaciones en variables que no son del mismo tipo puede ocasionar
                "A la hora de tratar variables categóricas con variables numéricas, se puede
                "Durante el proceso de tratamiento de datos, hemos observado como los datos
)
```

```r
if (knitr::is_html_output()) {
  # Si es HTML
  kable(data, col.names = c("Encabezado", "Contenido"),
        format = "html",
        escape = FALSE) %>%
    kable_styling(full_width = FALSE) %>%
    add_header_above(c("Hallazgo de Resultado 3" = 2), bold = TRUE)
} else {
  # Si es PDF
  kable(data, col.names = c("Encabezado", "Contenido"),
        format = "latex",
        booktabs = TRUE) %>%
    kable_styling(latex_options = c("striped", "hold_position")) %>%
    add_header_above(c("Hallazgo de Resultado 3" = 2), bold = TRUE)
}
```

	Hallaz
Encabezado	Contenido
Nombre de Su hallazgo	Dificultad a la hora de trabajar con variables categóricas y numéricas.
Resumen en una Oración	Buscar correlaciones en variables que no son del mismo tipo puede ocasionar
Problemas o Posibles Desafíos	A la hora de tratar variables categóricas con variables numéricas, se puede es
Resumen en un párrafo	Durante el proceso de tratamiento de datos, hemos observado como los datos

3 Bitácora 4

3.1 Introducción

3.1.1 Visualización y Limpieza de la Nueva Base de Datos

Queremos señalar que fue necesario cambiar la base de datos utilizada, ya que la anterior parecía haber sido generada artificialmente, sin provenir de datos reales. Para la presente bitácora, hemos optado por una base de datos auténtica que incluye información sobre pagos por defecto, factores demográficos, datos de crédito, historial de pagos y estados de cuenta de clientes de tarjetas de crédito en Taiwán, correspondiente al periodo de abril a septiembre de 2005.

Empezamos el estudio de la base de datos, para ello lo primodiarla es cargarla.

```
# Cargamos nuestra nueva base de datos.
library(readr)
data_credit <- read_csv("UCI_Credit_Card.csv")
```

```
Rows: 30000 Columns: 25
-- Column specification --------------------------------------------------------
Delimiter: ","
dbl (25): ID, LIMIT_BAL, SEX, EDUCATION, MARRIAGE, AGE, PAY_0, PAY_2, PAY_3,...

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# Visualizamos nuestro dataset.
data_credit
```

```
# A tibble: 30,000 x 25
      ID LIMIT_BAL   SEX EDUCATION MARRIAGE   AGE PAY_0 PAY_2 PAY_3 PAY_4 PAY_5
   <dbl>     <dbl> <dbl>     <dbl>    <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
 1     1     20000     2         2        1    24     2     2    -1    -1    -2
 2     2    120000     2         2        2    26    -1     2     0     0     0
```

```
 3     3     90000     2         2         2    34     0     0     0     0     0
 4     4     50000     2         2         1    37     0     0     0     0     0
 5     5     50000     1         2         1    57    -1     0    -1     0     0
 6     6     50000     1         1         2    37     0     0     0     0     0
 7     7    500000     1         1         2    29     0     0     0     0     0
 8     8    100000     2         2         2    23     0    -1    -1     0     0
 9     9    140000     2         3         1    28     0     0     2     0     0
10    10     20000     1         3         2    35    -2    -2    -2    -2    -1
# i 29,990 more rows
# i 14 more variables: PAY_6 <dbl>, BILL_AMT1 <dbl>, BILL_AMT2 <dbl>,
#   BILL_AMT3 <dbl>, BILL_AMT4 <dbl>, BILL_AMT5 <dbl>, BILL_AMT6 <dbl>,
#   PAY_AMT1 <dbl>, PAY_AMT2 <dbl>, PAY_AMT3 <dbl>, PAY_AMT4 <dbl>,
#   PAY_AMT5 <dbl>, PAY_AMT6 <dbl>, default.payment.next.month <dbl>
```

Una vez cargada la base de datos, es importante realizar algunas observaciones iniciales antes de comenzar a trabajar con ella. En primer lugar, queremos revisar los nombres de las variables y el tipo de datos que representan, ya que esto nos permite empezar a considerar qué técnicas estadísticas podríamos aplicar. Además, nos interesa verificar que la base de datos esté limpia; para ello, realizaremos un conteo de los valores faltantes en el dataset.

```r
# Verificamos los nombres de nuestro dataset
names(data_credit)
```

```
 [1] "ID"                         "LIMIT_BAL"
 [3] "SEX"                        "EDUCATION"
 [5] "MARRIAGE"                   "AGE"
 [7] "PAY_0"                      "PAY_2"
 [9] "PAY_3"                      "PAY_4"
[11] "PAY_5"                      "PAY_6"
[13] "BILL_AMT1"                  "BILL_AMT2"
[15] "BILL_AMT3"                  "BILL_AMT4"
[17] "BILL_AMT5"                  "BILL_AMT6"
[19] "PAY_AMT1"                   "PAY_AMT2"
[21] "PAY_AMT3"                   "PAY_AMT4"
[23] "PAY_AMT5"                   "PAY_AMT6"
[25] "default.payment.next.month"
```

Con los nombres hacemos un pequeño resumen de qué significa cada uno:

- ID: ID de cada cliente.
- LIMIT_BAL: Monto de crédito otorgado en dólares taiwaneses (NT) (incluye crédito individual y familiar/suplementario).

- SEX: Género (1=hombre, 2=mujer).
- EDUCATION: Nivel educativo (1=posgrado, 2=universidad, 3=preparatoria, 4=otros, 5=desconocido, 6=desconocido).
- MARRIAGE: Estado civil (1=casado, 2=soltero, 3=otros).
- AGE: Edad en años.
- PAY_0: Estado de reembolso en septiembre de 2005 (-1=pago puntual, 1=atraso de un mes, 2=atraso de dos meses, …, 8=atraso de ocho meses, 9=atraso de nueve meses o más).
- PAY_2: Estado de reembolso en agosto de 2005 (escala igual a la anterior).
- PAY_3: Estado de reembolso en julio de 2005 (escala igual a la anterior).
- PAY_4: Estado de reembolso en junio de 2005 (escala igual a la anterior).
- PAY_5: Estado de reembolso en mayo de 2005 (escala igual a la anterior).
- PAY_6: Estado de reembolso en abril de 2005 (escala igual a la anterior).
- BILL_AMT1: Monto del estado de cuenta en septiembre de 2005 (dólares taiwaneses, NT).
- BILL_AMT2: Monto del estado de cuenta en agosto de 2005 (dólares taiwaneses, NT).
- BILL_AMT3: Monto del estado de cuenta en julio de 2005 (dólares taiwaneses, NT).
- BILL_AMT4: Monto del estado de cuenta en junio de 2005 (dólares taiwaneses, NT).
- BILL_AMT5: Monto del estado de cuenta en mayo de 2005 (dólares taiwaneses, NT).
- BILL_AMT6: Monto del estado de cuenta en abril de 2005 (dólares taiwaneses, NT).
- PAY_AMT1: Monto del pago anterior en septiembre de 2005 (dólares taiwaneses, NT).
- PAY_AMT2: Monto del pago anterior en agosto de 2005 (dólares taiwaneses, NT).
- PAY_AMT3: Monto del pago anterior en julio de 2005 (dólares taiwaneses, NT).
- PAY_AMT4: Monto del pago anterior en junio de 2005 (dólares taiwaneses, NT).
- PAY_AMT5: Monto del pago anterior en mayo de 2005 (dólares taiwaneses, NT).
- PAY_AMT6: Monto del pago anterior en abril de 2005 (dólares taiwaneses, NT).
- default.payment.next.month: Pago en mora (1=sí, 0=no).

Por otro lado, veamos con qué tipo de datos contamos.

```
# Verificar el tipo de datos del dataset.
str(data_credit)
```

```
spc_tbl_ [30,000 x 25] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ ID                        : num [1:30000] 1 2 3 4 5 6 7 8 9 10 ...
 $ LIMIT_BAL                 : num [1:30000] 20000 120000 90000 50000 50000 50000 500000 1000
 $ SEX                       : num [1:30000] 2 2 2 2 1 1 1 2 2 1 ...
 $ EDUCATION                 : num [1:30000] 2 2 2 2 2 1 1 2 3 3 ...
 $ MARRIAGE                  : num [1:30000] 1 2 2 1 1 2 2 2 1 2 ...
 $ AGE                       : num [1:30000] 24 26 34 37 57 37 29 23 28 35 ...
 $ PAY_0                     : num [1:30000] 2 -1 0 0 -1 0 0 0 0 -2 ...
 $ PAY_2                     : num [1:30000] 2 2 0 0 0 0 0 -1 0 -2 ...
```

```
$ PAY_3                     : num [1:30000] -1 0 0 0 -1 0 0 -1 2 -2 ...
$ PAY_4                     : num [1:30000] -1 0 0 0 0 0 0 0 0 -2 ...
$ PAY_5                     : num [1:30000] -2 0 0 0 0 0 0 0 0 -1 ...
$ PAY_6                     : num [1:30000] -2 2 0 0 0 0 0 -1 0 -1 ...
$ BILL_AMT1                 : num [1:30000] 3913 2682 29239 46990 8617 ...
$ BILL_AMT2                 : num [1:30000] 3102 1725 14027 48233 5670 ...
$ BILL_AMT3                 : num [1:30000] 689 2682 13559 49291 35835 ...
$ BILL_AMT4                 : num [1:30000] 0 3272 14331 28314 20940 ...
$ BILL_AMT5                 : num [1:30000] 0 3455 14948 28959 19146 ...
$ BILL_AMT6                 : num [1:30000] 0 3261 15549 29547 19131 ...
$ PAY_AMT1                  : num [1:30000] 0 0 1518 2000 2000 ...
$ PAY_AMT2                  : num [1:30000] 689 1000 1500 2019 36681 ...
$ PAY_AMT3                  : num [1:30000] 0 1000 1000 1200 10000 657 38000 0 432 0 ...
$ PAY_AMT4                  : num [1:30000] 0 1000 1000 1100 9000 ...
$ PAY_AMT5                  : num [1:30000] 0 0 1000 1069 689 ...
$ PAY_AMT6                  : num [1:30000] 0 2000 5000 1000 679 ...
$ default.payment.next.month: num [1:30000] 1 1 0 0 0 0 0 0 0 0 ...
- attr(*, "spec")=
 .. cols(
 ..   ID = col_double(),
 ..   LIMIT_BAL = col_double(),
 ..   SEX = col_double(),
 ..   EDUCATION = col_double(),
 ..   MARRIAGE = col_double(),
 ..   AGE = col_double(),
 ..   PAY_0 = col_double(),
 ..   PAY_2 = col_double(),
 ..   PAY_3 = col_double(),
 ..   PAY_4 = col_double(),
 ..   PAY_5 = col_double(),
 ..   PAY_6 = col_double(),
 ..   BILL_AMT1 = col_double(),
 ..   BILL_AMT2 = col_double(),
 ..   BILL_AMT3 = col_double(),
 ..   BILL_AMT4 = col_double(),
 ..   BILL_AMT5 = col_double(),
 ..   BILL_AMT6 = col_double(),
 ..   PAY_AMT1 = col_double(),
 ..   PAY_AMT2 = col_double(),
 ..   PAY_AMT3 = col_double(),
 ..   PAY_AMT4 = col_double(),
 ..   PAY_AMT5 = col_double(),
 ..   PAY_AMT6 = col_double(),
```

```
  ..   default.payment.next.month = col_double()
  .. )
 - attr(*, "problems")=<externalptr>
```

Antes de hacer un pequeño conteo de la cantidad de missing value, vamos a realizar un pequeño resumen de la base de datos.

```
library(skimr)

# Resumen del dataset.
summary(data_credit)
```

```
       ID           LIMIT_BAL            SEX          EDUCATION
 Min.   :    1   Min.   :  10000   Min.   :1.000   Min.   :0.000
 1st Qu.: 7501   1st Qu.:  50000   1st Qu.:1.000   1st Qu.:1.000
 Median :15000   Median : 140000   Median :2.000   Median :2.000
 Mean   :15000   Mean   : 167484   Mean   :1.604   Mean   :1.853
 3rd Qu.:22500   3rd Qu.: 240000   3rd Qu.:2.000   3rd Qu.:2.000
 Max.   :30000   Max.   :1000000   Max.   :2.000   Max.   :6.000
    MARRIAGE          AGE            PAY_0             PAY_2
 Min.   :0.000   Min.   :21.00   Min.   :-2.0000   Min.   :-2.0000
 1st Qu.:1.000   1st Qu.:28.00   1st Qu.:-1.0000   1st Qu.:-1.0000
 Median :2.000   Median :34.00   Median : 0.0000   Median : 0.0000
 Mean   :1.552   Mean   :35.49   Mean   :-0.0167   Mean   :-0.1338
 3rd Qu.:2.000   3rd Qu.:41.00   3rd Qu.: 0.0000   3rd Qu.: 0.0000
 Max.   :3.000   Max.   :79.00   Max.   : 8.0000   Max.   : 8.0000
     PAY_3             PAY_4             PAY_5             PAY_6
 Min.   :-2.0000   Min.   :-2.0000   Min.   :-2.0000   Min.   :-2.0000
 1st Qu.:-1.0000   1st Qu.:-1.0000   1st Qu.:-1.0000   1st Qu.:-1.0000
 Median : 0.0000   Median : 0.0000   Median : 0.0000   Median : 0.0000
 Mean   :-0.1662   Mean   :-0.2207   Mean   :-0.2662   Mean   :-0.2911
 3rd Qu.: 0.0000   3rd Qu.: 0.0000   3rd Qu.: 0.0000   3rd Qu.: 0.0000
 Max.   : 8.0000   Max.   : 8.0000   Max.   : 8.0000   Max.   : 8.0000
   BILL_AMT1         BILL_AMT2         BILL_AMT3         BILL_AMT4
 Min.   :-165580   Min.   :-69777   Min.   :-157264   Min.   :-170000
 1st Qu.:   3559   1st Qu.:  2985   1st Qu.:   2666   1st Qu.:   2327
 Median :  22382   Median : 21200   Median :  20089   Median :  19052
 Mean   :  51223   Mean   : 49179   Mean   :  47013   Mean   :  43263
 3rd Qu.:  67091   3rd Qu.: 64006   3rd Qu.:  60165   3rd Qu.:  54506
 Max.   : 964511   Max.   :983931   Max.   :1664089   Max.   : 891586
   BILL_AMT5         BILL_AMT6          PAY_AMT1         PAY_AMT2
 Min.   :-81334   Min.   :-339603   Min.   :     0   Min.   :      0
```

```
 1st Qu.:  1763   1st Qu.:   1256   1st Qu.:  1000   1st Qu.:    833
 Median : 18105   Median :  17071   Median :  2100   Median :   2009
 Mean   : 40311   Mean   :  38872   Mean   :  5664   Mean   :   5921
 3rd Qu.: 50191   3rd Qu.:  49198   3rd Qu.:  5006   3rd Qu.:   5000
 Max.   :927171   Max.   : 961664   Max.   :873552   Max.   :1684259
    PAY_AMT3         PAY_AMT4         PAY_AMT5          PAY_AMT6
 Min.   :     0   Min.   :     0   Min.   :     0.0   Min.   :     0.0
 1st Qu.:   390   1st Qu.:   296   1st Qu.:   252.5   1st Qu.:   117.8
 Median :  1800   Median :  1500   Median :  1500.0   Median :  1500.0
 Mean   :  5226   Mean   :  4826   Mean   :  4799.4   Mean   :  5215.5
 3rd Qu.:  4505   3rd Qu.:  4013   3rd Qu.:  4031.5   3rd Qu.:  4000.0
 Max.   :896040   Max.   :621000   Max.   :426529.0   Max.   :528666.0
 default.payment.next.month
 Min.   :0.0000
 1st Qu.:0.0000
 Median :0.0000
 Mean   :0.2212
 3rd Qu.:0.0000
 Max.   :1.0000
```

```
# Generamos un cuadro resumen, con la información anterior.
skim(data_credit)
```

Tabla 3.1: Data summary

Name	data_credit
Number of rows	30000
Number of columns	25
_____	
Column type frequency:	
numeric	25
_____	
Group variables	None

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
ID	0	1	15000.50	8660.40	1	7500.75	15000.5	22500.25	30000	
LIMIT_BAL	0	1	167484.32	129747.66	10000	50000.00	140000.0	240000.00	1000000	
SEX	0	1	1.60	0.49	1	1.00	2.0	2.00	2	

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
EDUCATION	0	1	1.85	0.79	0	1.00	2.0	2.00	6	
MARRIAGE	0	1	1.55	0.52	0	1.00	2.0	2.00	3	
AGE	0	1	35.49	9.22	21	28.00	34.0	41.00	79	
PAY_0	0	1	-0.02	1.12	-2	-1.00	0.0	0.00	8	
PAY_2	0	1	-0.13	1.20	-2	-1.00	0.0	0.00	8	
PAY_3	0	1	-0.17	1.20	-2	-1.00	0.0	0.00	8	
PAY_4	0	1	-0.22	1.17	-2	-1.00	0.0	0.00	8	
PAY_5	0	1	-0.27	1.13	-2	-1.00	0.0	0.00	8	
PAY_6	0	1	-0.29	1.15	-2	-1.00	0.0	0.00	8	
BILL_AMT1	0	1	51223.33	73635.86	-165580	3558.75	22381.5	67091.00	964511	
BILL_AMT2	0	1	49179.08	71173.77	-69777	2984.75	21200.0	64006.25	983931	
BILL_AMT3	0	1	47013.15	69349.39	-157264	2666.25	20088.5	60164.75	1664089	
BILL_AMT4	0	1	43262.95	64332.86	-170000	2326.75	19052.0	54506.00	891586	
BILL_AMT5	0	1	40311.40	60797.16	-81334	1763.00	18104.5	50190.50	927171	
BILL_AMT6	0	1	38871.76	59554.11	-339603	1256.00	17071.0	49198.25	961664	
PAY_AMT1	0	1	5663.58	16563.28	0	1000.00	2100.0	5006.00	873552	
PAY_AMT2	0	1	5921.16	23040.87	0	833.00	2009.0	5000.00	1684259	
PAY_AMT3	0	1	5225.68	17606.96	0	390.00	1800.0	4505.00	896040	
PAY_AMT4	0	1	4826.08	15666.16	0	296.00	1500.0	4013.25	621000	
PAY_AMT5	0	1	4799.39	15278.31	0	252.50	1500.0	4031.50	426529	
PAY_AMT6	0	1	5215.50	17777.47	0	117.75	1500.0	4000.00	528666	
default.payment.next.month	0	1	0.22	0.42	0	0.00	0.0	0.00	1	

Apesar de qué el cuadro resumen anterior ya nos indica que las variables no tienen missing value, nos parece pertinente verificarlo de manera aislada, para ello.

```
#Verificamos la cantidad de datos nulos que hay nuestro dataset
sum(is.na(data_credit))
```

```
[1] 0
```

```
#Verificamos la cantidad de datos nulos que hay en cada columna
sapply(data_credit, function(x) sum(is.na(x)))
```

```
                        ID                        LIMIT_BAL
                         0                                0
                       SEX                        EDUCATION
                         0                                0
                  MARRIAGE                              AGE
                         0                                0
                     PAY_0                            PAY_2
                         0                                0
                     PAY_3                            PAY_4
                         0                                0
                     PAY_5                            PAY_6
                         0                                0
                 BILL_AMT1                        BILL_AMT2
                         0                                0
                 BILL_AMT3                        BILL_AMT4
                         0                                0
                 BILL_AMT5                        BILL_AMT6
                         0                                0
                  PAY_AMT1                         PAY_AMT2
                         0                                0
                  PAY_AMT3                         PAY_AMT4
                         0                                0
                  PAY_AMT5                         PAY_AMT6
                         0                                0
default.payment.next.month
                         0
```

Sin embargo, hay columnas que aunque no están vacías, contiene datos que no nos sirven, por eso hay que filtrar estos datos. Estos datos no sirven por el hecho de que no están definidos como parámetros significativos, es decir, si tenemos definidos la variable sexo como 1=hombre y 2=mujer, entonces aparecen números como el 3 y 0, por ello hay que filtrarlos, ya que afectan los análisis.

Nos damos cuenta de ello, gracias a ver el cuadro resumen, que aparecen valores que no deberían aparecer.

```r
library(tidyverse)

# Filtramos los datos que están definidos.
data_credit <- data_credit %>%
  filter(MARRIAGE %in% c(1, 2, 3))

# Filtramos los datos que están definidos.
```

```r
data_credit <- data_credit %>%
  filter(EDUCATION %in% c(1, 2, 3, 4, 5, 6))

data_credit <- data_credit %>%
  filter(PAY_0 %in% c(-1, 1, 2, 3, 4, 5, 6, 7, 8, 9))

data_credit <- data_credit %>%
  filter(PAY_2 %in% c(-1, 1, 2, 3, 4, 5, 6, 7, 8, 9))

data_credit <- data_credit %>%
  filter(PAY_3 %in% c(-1, 1, 2, 3, 4, 5, 6, 7, 8, 9))

data_credit <- data_credit %>%
  filter(PAY_4 %in% c(-1, 1, 2, 3, 4, 5, 6, 7, 8, 9))

data_credit <- data_credit %>%
  filter(PAY_5 %in% c(-1, 1, 2, 3, 4, 5, 6, 7, 8, 9))

data_credit <- data_credit %>%
  filter(PAY_6 %in% c(-1, 1, 2, 3, 4, 5, 6, 7, 8, 9))
```

Realizamos de nuevo nuestro cuadro resumen después de esta filtración con el objetivo de observar las nuevas tendencias de la base.

```r
library(skimr)

# Resumen del dataset.
summary(data_credit)
```

```
       ID          LIMIT_BAL          SEX          EDUCATION
 Min.   :   12   Min.   : 10000   Min.   :1.000   Min.   :1.000
 1st Qu.: 6941   1st Qu.: 60000   1st Qu.:1.000   1st Qu.:1.000
 Median :13671   Median :150000   Median :2.000   Median :2.000
 Mean   :14278   Mean   :171695   Mean   :1.592   Mean   :1.757
 3rd Qu.:21713   3rd Qu.:240000   3rd Qu.:2.000   3rd Qu.:2.000
 Max.   :29995   Max.   :740000   Max.   :2.000   Max.   :6.000
    MARRIAGE          AGE            PAY_0             PAY_2
 Min.   :1.000   Min.   :21.00   Min.   :-1.0000   Min.   :-1.0000
 1st Qu.:1.000   1st Qu.:29.00   1st Qu.:-1.0000   1st Qu.:-1.0000
 Median :1.000   Median :35.00   Median :-1.0000   Median :-1.0000
 Mean   :1.493   Mean   :36.53   Mean   : 0.1819   Mean   : 0.2869
```

```
 3rd Qu.:2.000   3rd Qu.:43.00   3rd Qu.: 2.0000   3rd Qu.: 2.0000
 Max.   :3.000   Max.   :72.00   Max.   : 8.0000   Max.   : 8.0000
     PAY_3             PAY_4             PAY_5             PAY_6
 Min.   :-1.0000   Min.   :-1.0000   Min.   :-1.0000   Min.   :-1.0000
 1st Qu.:-1.0000   1st Qu.:-1.0000   1st Qu.:-1.0000   1st Qu.:-1.0000
 Median :-1.0000   Median :-1.0000   Median :-1.0000   Median :-1.0000
 Mean   : 0.3188   Mean   : 0.2837   Mean   : 0.2424   Mean   : 0.2488
 3rd Qu.: 2.0000   3rd Qu.: 2.0000   3rd Qu.: 2.0000   3rd Qu.: 2.0000
 Max.   : 8.0000   Max.   : 8.0000   Max.   : 8.0000   Max.   : 8.0000
   BILL_AMT1         BILL_AMT2         BILL_AMT3         BILL_AMT4
 Min.   : -4316   Min.   :-24704   Min.   :-61506   Min.   : -3903
 1st Qu.:   931   1st Qu.:   856   1st Qu.:   835   1st Qu.:   828
 Median :  4394   Median :  4398   Median :  4192   Median :  4176
 Mean   : 22124   Mean   : 22296   Mean   : 22304   Mean   : 22641
 3rd Qu.: 22231   3rd Qu.: 22678   3rd Qu.: 22974   3rd Qu.: 22819
 Max.   :581775   Max.   :572677   Max.   :471175   Max.   :486776
   BILL_AMT5         BILL_AMT6          PAY_AMT1          PAY_AMT2
 Min.   : -3876   Min.   :-339603   Min.   :     0   Min.   :     0
 1st Qu.:   838   1st Qu.:    776   1st Qu.:   316   1st Qu.:   316
 Median :  4069   Median :   4120   Median :  1600   Median :  1595
 Mean   : 22589   Mean   :  22676   Mean   :  4669   Mean   :  4608
 3rd Qu.: 23341   3rd Qu.:  23710   3rd Qu.:  4427   3rd Qu.:  4398
 Max.   :503914   Max.   : 527711   Max.   :187206   Max.   :302961
    PAY_AMT3          PAY_AMT4          PAY_AMT5          PAY_AMT6
 Min.   :     0   Min.   :     0   Min.   :     0   Min.   :     0
 1st Qu.:   316   1st Qu.:   331   1st Qu.:   100   1st Qu.:     0
 Median :  1443   Median :  1443   Median :  1225   Median :  1044
 Mean   :  4756   Mean   :  4558   Mean   :  4594   Mean   :  4621
 3rd Qu.:  4200   3rd Qu.:  4100   3rd Qu.:  4000   3rd Qu.:  3710
 Max.   :417588   Max.   :193712   Max.   :303512   Max.   :345293
 default.payment.next.month
 Min.   :0.0000
 1st Qu.:0.0000
 Median :0.0000
 Mean   :0.3548
 3rd Qu.:1.0000
 Max.   :1.0000
```

```
# Generamos un cuadro resumen, con la información anterior.
skim(data_credit)
```

Tabla 3.3: Data summary

Name	data_credit
Number of rows	4047
Number of columns	25
_____	
Column type frequency:	
numeric	25
_____	
Group variables	None

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
ID	0	1	14277.57	8616.90	12	6941	13671	21712.5	29995	
LIMIT_BAL	0	1	171695.08	125912.17	10000	60000	150000	240000.0	740000	
SEX	0	1	1.59	0.49	1	1	2	2.0	2	
EDUCATION	0	1	1.76	0.75	1	1	2	2.0	6	
MARRIAGE	0	1	1.49	0.52	1	1	1	2.0	3	
AGE	0	1	36.53	9.19	21	29	35	43.0	72	
PAY_0	0	1	0.18	1.58	-1	-1	-1	2.0	8	
PAY_2	0	1	0.29	1.68	-1	-1	-1	2.0	8	
PAY_3	0	1	0.32	1.74	-1	-1	-1	2.0	8	
PAY_4	0	1	0.28	1.80	-1	-1	-1	2.0	8	
PAY_5	0	1	0.24	1.78	-1	-1	-1	2.0	8	
PAY_6	0	1	0.25	1.75	-1	-1	-1	2.0	8	
BILL_AMT1	0	1	22124.22	44383.68	-4316	931	4394	22231.0	581775	
BILL_AMT2	0	1	22296.39	44437.54	-24704	856	4398	22678.0	572677	
BILL_AMT3	0	1	22304.37	44520.48	-61506	835	4192	22974.0	471175	
BILL_AMT4	0	1	22640.65	45041.16	-3903	828	4176	22818.5	486776	
BILL_AMT5	0	1	22588.65	44576.24	-3876	838	4069	23341.0	503914	
BILL_AMT6	0	1	22675.85	45579.27	-339603	776	4120	23710.0	527711	
PAY_AMT1	0	1	4669.12	10921.61	0	316	1600	4427.0	187206	
PAY_AMT2	0	1	4608.06	11960.55	0	316	1595	4398.5	302961	
PAY_AMT3	0	1	4756.23	13590.04	0	316	1443	4200.0	417588	

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
PAY_AMT4	0	1	4557.61	11100.86	0	331	1443	4100.0	193712	
PAY_AMT5	0	1	4594.45	13511.57	0	100	1225	4000.0	303512	
PAY_AMT6	0	1	4620.82	15153.76	0	0	1044	3710.0	345293	
default.payment.next.month	0	1	0.35	0.48	0	0	0	1.0	1	

Gracias a todo lo anterior ya tenemos una base de datos limpia, lista para el análisis estadístico.

3.1.2 Análisis Estadístico de la Base de Datos

Para esta sección, nuestro objetivo es explorar el comportamiento de nuestro conjunto de datos de manera más profunda. Para ello, aplicaremos algunas técnicas estadísticas que nos permitirán obtener información relevante.

En primer lugar, vamos a construir una matriz de correlación. Esta herramienta nos ayudará a identificar las relaciones más fuertes entre las variables de nuestro dataset. Es importante recordar que la matriz de correlación solo tiene sentido cuando se aplica a variables numéricas, por lo que es esencial seleccionar adecuadamente las variables antes de proceder con este análisis.

3.1.2.1 Matriz de Correlación

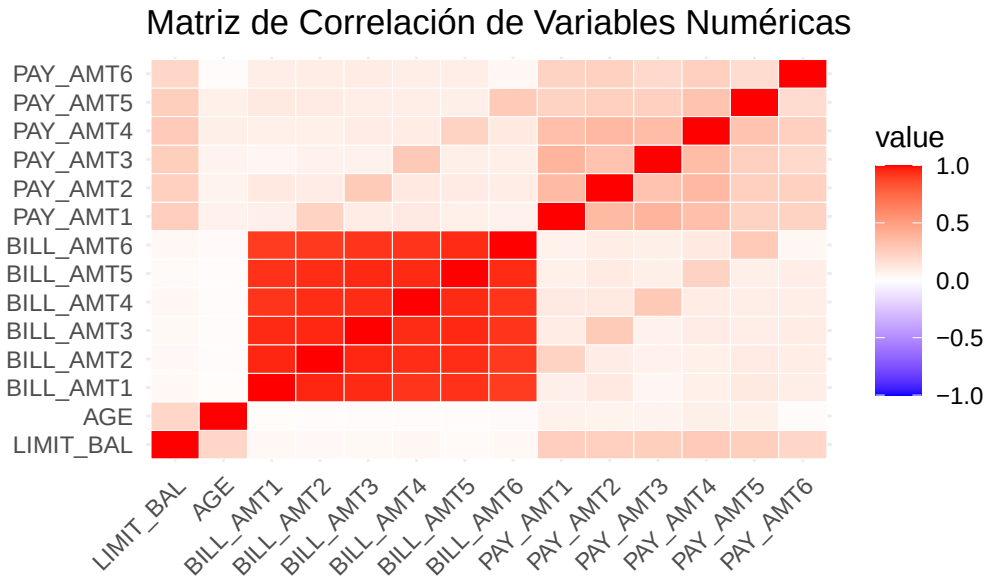
Vamos a seleccionar solo las variables que sean de interés, es decir, vamos a quitar variables como el id, y el sexo, ya que el id es único para cada persona en el data set, y la variable sexo es una variable de tipo categórico, lo mismo pasa con la variable de defaul.payment, la cual a pesar de representarse con números, estos significan que son categóricos.

```
library(tidyverse)
# Seleccionamos las variables.
data_credit_numerico <- data_credit %>% select(-ID, -SEX, -EDUCATION, -MARRIAGE, -PAY_0, -PA
```

Con los datos filtrados, podemos entonces realizar nuestra matriz de correlación.

```
library(dplyr)
library(ggplot2)
library(reshape2)


# Crear la matriz de correlación
```

```r
matriz_correlacion <- cor(data_credit_numerico, use = "complete.obs")

# Creamos la variable para ver la matriz de correlación como un gráfico de calor
base_matriz <- melt(matriz_correlacion)

# Hacemos el plot de la base
ggplot(base_matriz, aes(Var1, Var2, fill = value)) +
  geom_tile(color = "white") +
  scale_fill_gradient2(low = "blue", high = "red", mid = "white", midpoint = 0, limit = c(-1
  theme_minimal() +
  labs(title = "Matriz de Correlación de Variables Numéricas",
       x = "", y = "") + labs(caption = "Fuente: Elaboración propia utilizando la base de dat
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

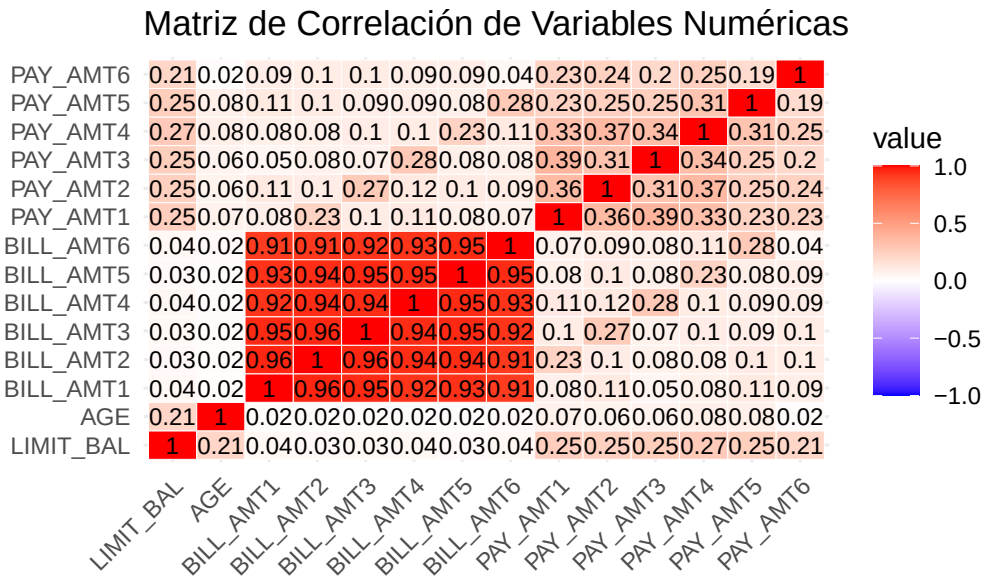


Fuente: Elaboración propia utilizando la base de datos de Kaggle

Agregamos la misma matriz, pero esta vez indicando los índices de correlación, para visualizar de manera gráfica y matemática la correlación existente.

```r
library(dplyr)
library(ggplot2)
library(reshape2)


# Crear la matriz de correlación
```

```r
matriz_correlacion <- cor(data_credit_numerico, use = "complete.obs")

# Creamos la variable para ver la matriz de correlación como un gráfico de calor
base_matriz <- melt(matriz_correlacion)

# Hacemos el plot de la base
ggplot(base_matriz, aes(Var1, Var2, fill = value)) +
  geom_tile(color = "white") +
  scale_fill_gradient2(low = "blue", high = "red", mid = "white", midpoint = 0, limit = c(-1
  theme_minimal() + geom_text(aes(label = round(value, 2)), color = "black", size = 3) +
  labs(title = "Matriz de Correlación de Variables Numéricas",
       x = "", y = "") + labs(caption = "Fuente: Elaboración propia utilizando la base de dat
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Fuente: Elaboración propia utilizando la base de datos de Kaggle

Figura 3.1: Figura 1: Matriz de Correlación

3.1.2.2 Gráficos relacionados a la Base de Datos

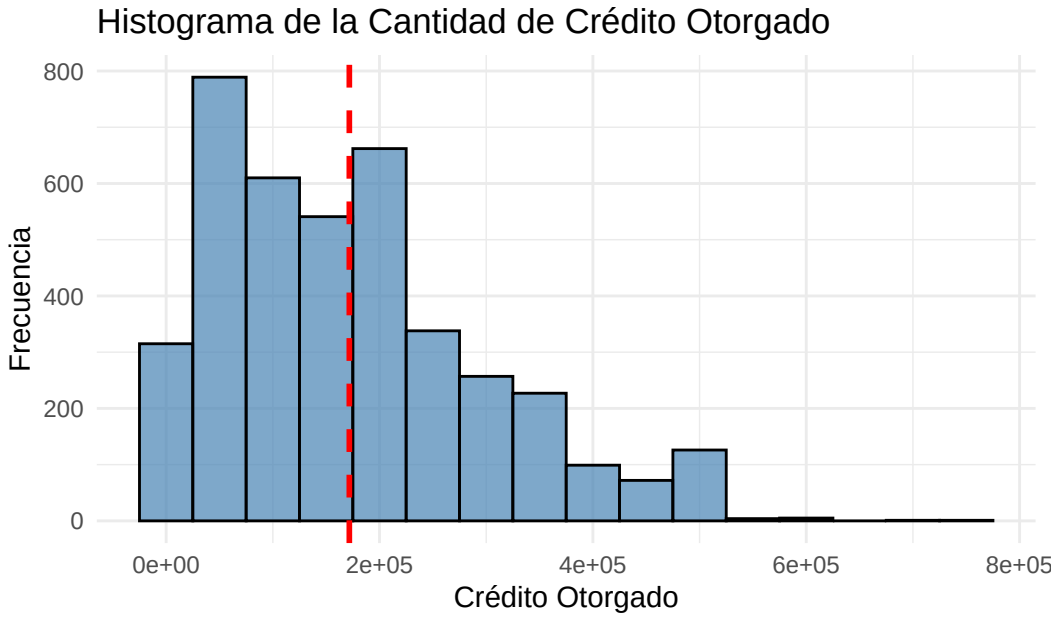
3.1.2.2.1 Gráficos de Variables Numéricas

Vamos a realizar algunos gráficos de la base, con el fin de observar cómo se comportan los datos, para ello, primero veamos algunos histogramas, recordemos que los histogramas están hechos para variables numéricas, por lo que trataremos de ir en orden a la hora de gráficarlas.

Damos inicio con la variable de “LIMIT_BAL”, la cual se refiere al crédito otorgado.

```
library(ggplot2)
library(dplyr)

data_credit %>%
  ggplot(aes(x = LIMIT_BAL)) +
  geom_histogram(binwidth = 50000, fill = "steelblue", color = "black", alpha = 0.7) +
  geom_vline(aes(xintercept = mean(LIMIT_BAL)), linetype = "dashed", color = "red", size = 1
  labs(title = "Histograma de la Cantidad de Crédito Otorgado",
       x = "Crédito Otorgado",
       y = "Frecuencia") + labs(caption = "Fuente: Elaboración propia utilizando la base de d
  theme_minimal()
```

```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
i Please use `linewidth` instead.
```

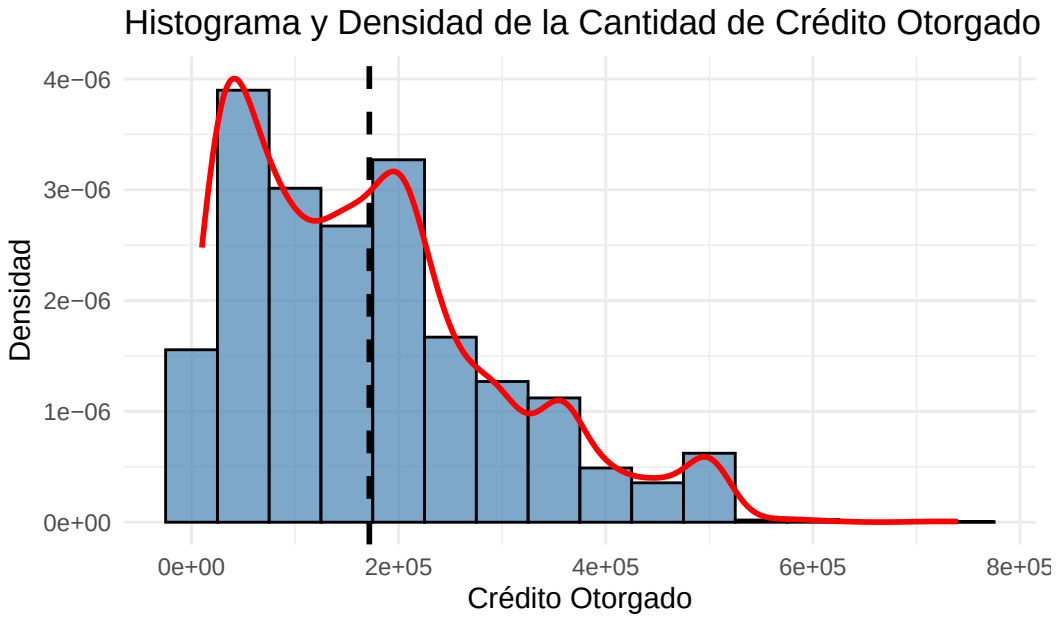


Fuente: Elaboración propia utilizando la base de datos de Kaggle

Agregamos un gráfico adicional, donde podemos observar como se comporta la densidad de esta variable.

```
library(ggplot2)
library(dplyr)
```

```r
data_credit %>%
  ggplot(aes(x = LIMIT_BAL)) +
  geom_histogram(aes(y = after_stat(density)), binwidth = 50000, fill = "steelblue", color =
  geom_density(color = "red", size = 1) + # Densidad en línea roja
  geom_vline(aes(xintercept = mean(LIMIT_BAL)), linetype = "dashed", color = "black", size =
  labs(title = "Histograma y Densidad de la Cantidad de Crédito Otorgado",
       x = "Crédito Otorgado",
       y = "Densidad") + labs(caption = "Fuente: Elaboración propia utilizando la base de dat
  theme_minimal()
```

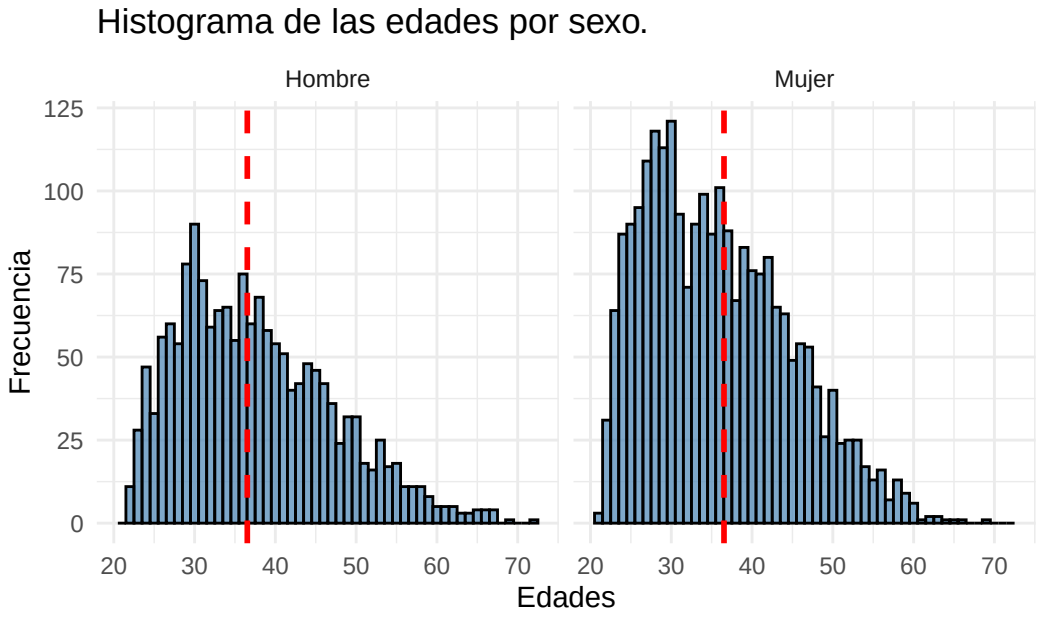


Fuente: Elaboración propia utilizando la base de datos de Kaggle

Procedemos ahora con la variable de la edad, esto porque queremos visualizar la distribución de las edades de nuestro data set.

```r
library(ggplot2)
library(dplyr)

data_credit %>%
  ggplot(aes(x = AGE)) +
  geom_histogram(binwidth = 1, fill = "steelblue", color = "black", alpha = 0.7) +
  geom_vline(aes(xintercept = mean(AGE, na.rm = TRUE)),
             color = "red",
             linetype = "dashed",
             size = 1) +
```

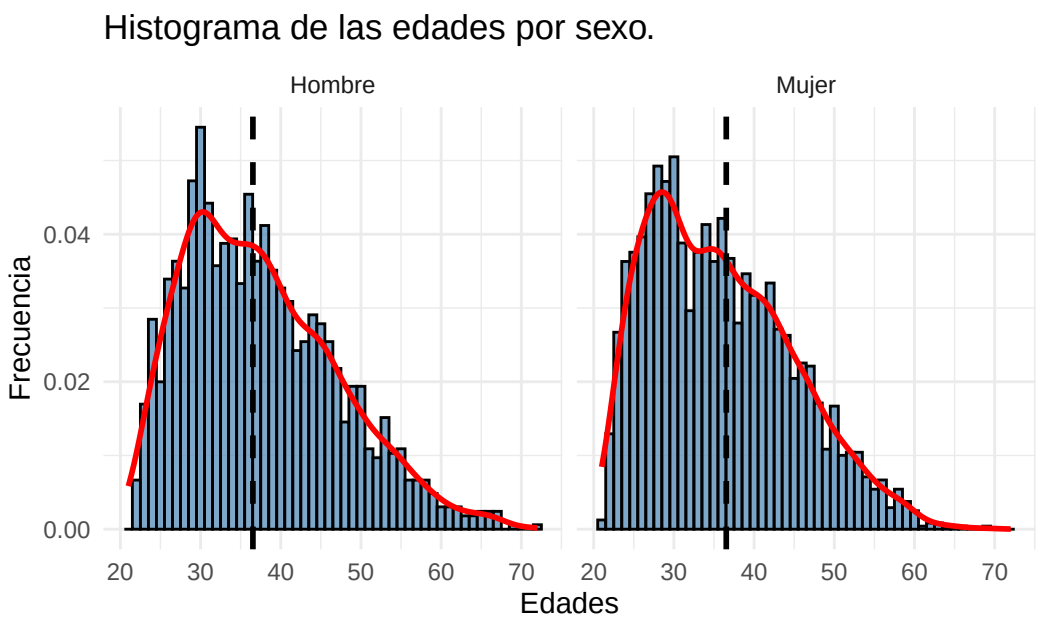
```
  labs(title = "Histograma de las edades por sexo.",
       x = "Edades",
       y = "Frecuencia") + labs(caption = "Fuente: Elaboración propia utilizando la base de d
  facet_wrap(~SEX, labeller = as_labeller(c("1" = "Hombre", "2" = "Mujer"))) +
  theme_minimal()
```



Fuente: Elaboración propia utilizando la base de datos de Kaggle

Agregamos un gráfico adicional donde podemos observar la densidad de la variable.

```
library(ggplot2)
library(dplyr)

data_credit %>%
  ggplot(aes(x = AGE)) +
  geom_histogram(aes(y = after_stat(density)), binwidth = 1, fill = "steelblue", color = "bla
  geom_density(aes(y = after_stat(density)), color = "red", size = 1) +  # Curva de densidad
  geom_vline(aes(xintercept = mean(AGE, na.rm = TRUE)),
             color = "black",
             linetype = "dashed",
             size = 1) +  # Línea roja punteada
  labs(title = "Histograma de las edades por sexo.",
       x = "Edades",
       y = "Frecuencia") +
  facet_wrap(~SEX, labeller = as_labeller(c("1" = "Hombre", "2" = "Mujer"))) + labs(caption =
  theme_minimal()
```

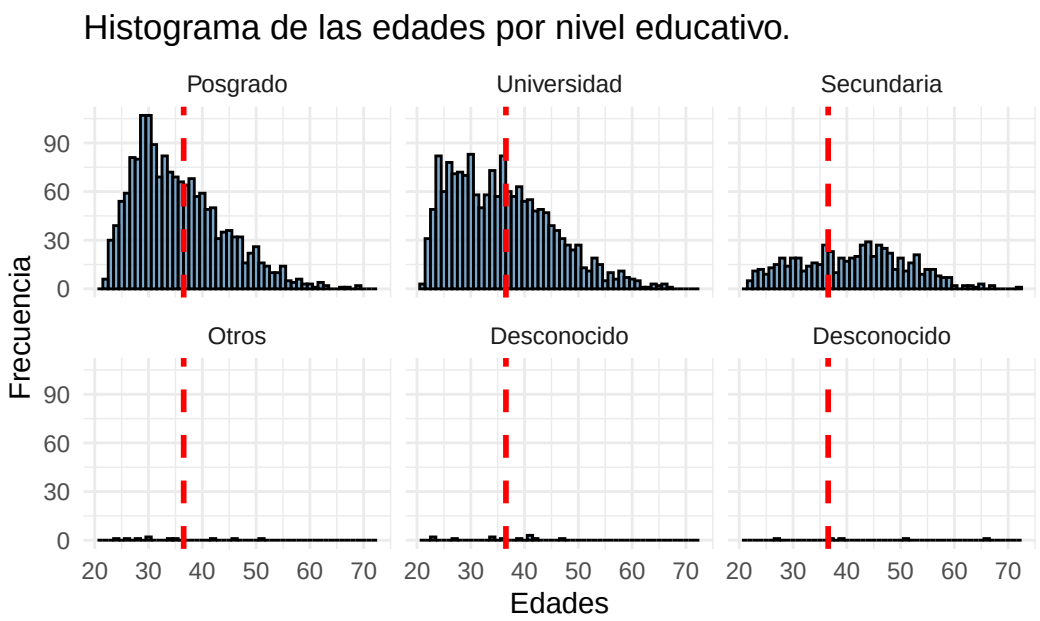


Fuente: Elaboración propia utilizando la base de datos de Kaggle

Con estas dos variables es sufiente de gráficos aislados, debido a la naturaleza de las demás variables numéricas que tenemos en el dataset, en vez de eso, vamos a ver cómo se comportan los gráficos, cuando se plotean las relaciones entre ellos.

```
library(ggplot2)
library(dplyr)

data_credit %>%
  ggplot(aes(x = AGE)) +
  geom_histogram(binwidth = 1, fill = "steelblue", color = "black", alpha = 0.7) +
  geom_vline(aes(xintercept = mean(AGE, na.rm = TRUE)),
             color = "red",
             linetype = "dashed",
             size = 1) +
  labs(title = "Histograma de las edades por nivel educativo.",
       x = "Edades",
       y = "Frecuencia") +
  facet_wrap(~EDUCATION, labeller = labeller(EDUCATION = c(
    "1" = "Posgrado",
    "2" = "Universidad",
    "3" = "Secundaria",
    "4" = "Otros",
```

```
    "5" = "Desconocido",
    "6" = "Desconocido"
  ))) + labs(caption = "Fuente: Elaboración propia utilizando la base de datos de Kaggle") +
```

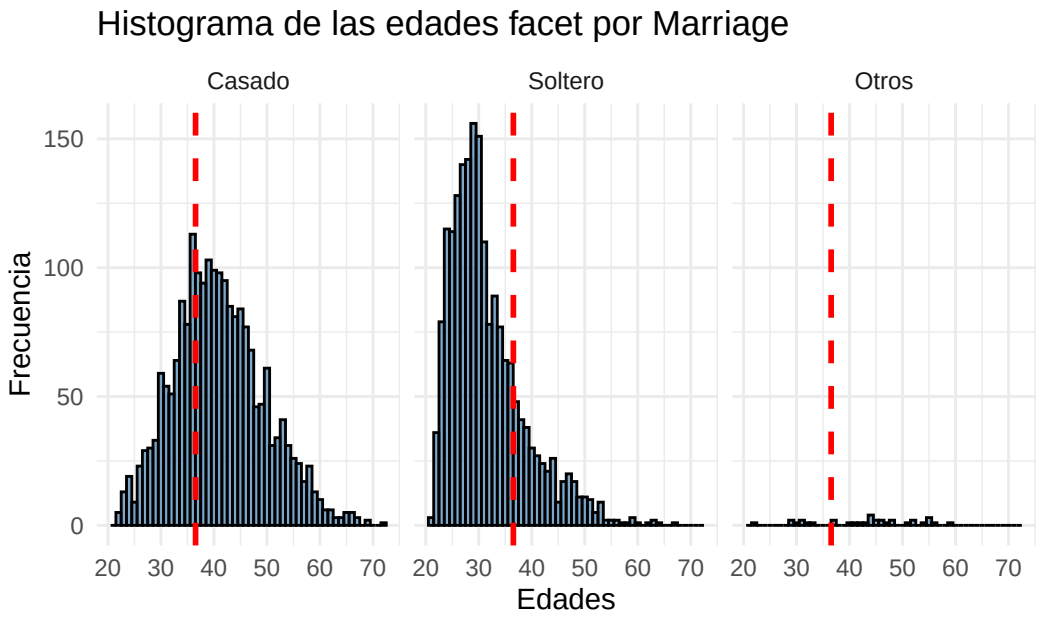


Fuente: Elaboración propia utilizando la base de datos de Kaggle

Gracias a este gráfico, podemos observar que la gran mayoría de los datos se encuentran en los subvariables de Universidad. Secundaria y Posgrado.

Ahora veamos como se distribuye la edad, con respecto a la variable de Estado Civil.

```
library(ggplot2)
library(dplyr)

data_credit %>%
  ggplot(aes(x = AGE)) +
  geom_histogram(binwidth = 1, fill = "steelblue", color = "black", alpha = 0.7) +
  geom_vline(aes(xintercept = mean(AGE, na.rm = TRUE)),
             color = "red",
             linetype = "dashed",
             size = 1) +  # Línea roja punteada para la media
  labs(title = "Histograma de las edades facet por Marriage",
       x = "Edades",
       y = "Frecuencia") +
  facet_wrap(~MARRIAGE, labeller = as_labeller(c("1" = "Casado", "2" = "Soltero", "3" = "Otro
  theme_minimal()
```

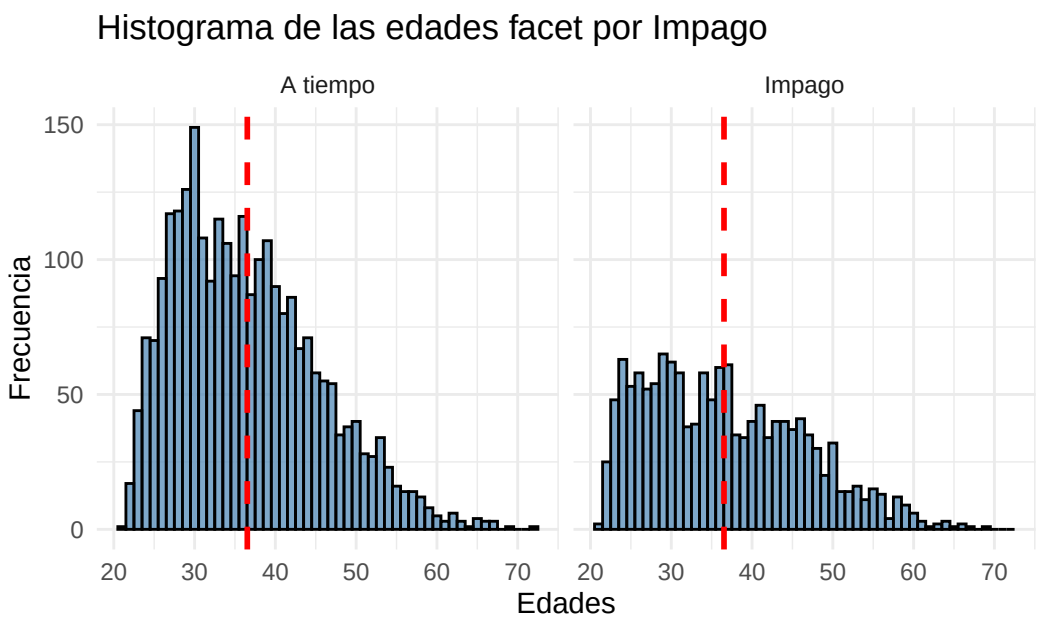


Fuente: Elaboración propia utilizando la base de datos de Kaggle

Conjeturando, podemos ver que la distribución de las edades de las personas que están casadas, sigue más o menos una distribución normal, mientras que las personas que están solteras lo hace como una distribución exponencial.

Por último nos interesa saber cómo se distribuyen las edades en relación a la variable de impago, que es la que tiene el principal peso de dicho estudio.

```
library(ggplot2)
library(dplyr)

data_credit %>%
  ggplot(aes(x = AGE)) +
  geom_histogram(binwidth = 1, fill = "steelblue", color = "black", alpha = 0.7) +
  geom_vline(aes(xintercept = mean(AGE, na.rm = TRUE)),
             color = "red",
             linetype = "dashed",
             size = 1) +
  labs(title = "Histograma de las edades facet por Impago",
       x = "Edades",
       y = "Frecuencia") +
  facet_wrap(~default.payment.next.month, labeller = as_labeller(c("0" = "A tiempo", "1" = "
  theme_minimal()
```

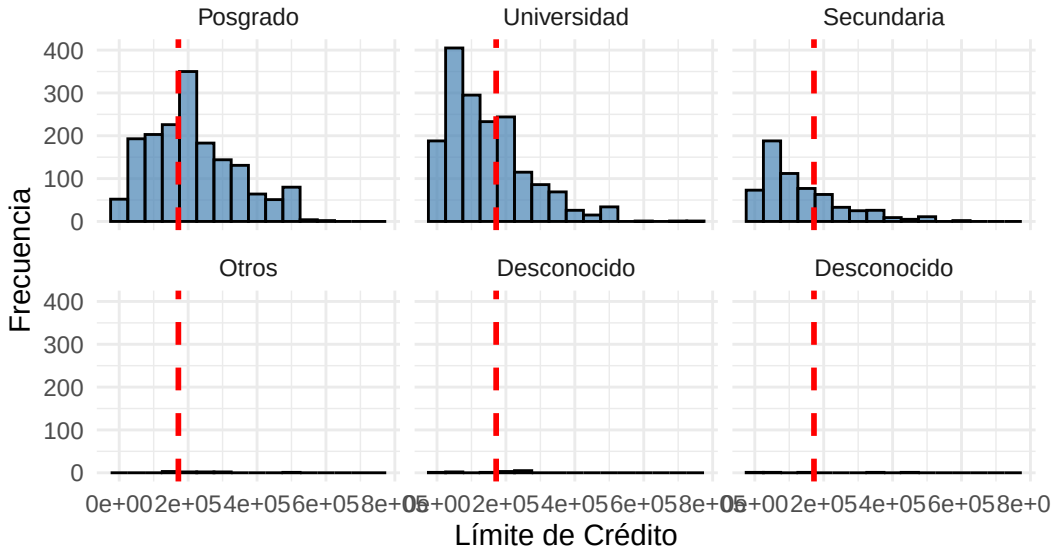


Fuente: Elaboración propia utilizando la base de datos de Kaggle

Vamos a realizar las mismas gráficas, pero esta vez con la la variable de LIMIT_BAL, para observar su comportamiento.

```
library(ggplot2)
library(dplyr)

data_credit %>%
  ggplot(aes(x = LIMIT_BAL)) +
  geom_histogram(binwidth = 50000, fill = "steelblue", color = "black", alpha = 0.7) +
  geom_vline(aes(xintercept = mean(LIMIT_BAL, na.rm = TRUE)),
             color = "red",
             linetype = "dashed",
             size = 1) +
  labs(title = "Histograma de los límites de crédito por nivel educativo.",
       x = "Límite de Crédito",
       y = "Frecuencia") +
  facet_wrap(~EDUCATION, labeller = labeller(EDUCATION = c(
    "1" = "Posgrado",
    "2" = "Universidad",
    "3" = "Secundaria",
    "4" = "Otros",
    "5" = "Desconocido",
    "6" = "Desconocido"
  ))) + labs(caption = "Fuente: Elaboración propia utilizando la base de datos de Kaggle") +
```

Histograma de los límites de crédito por nivel educativo.



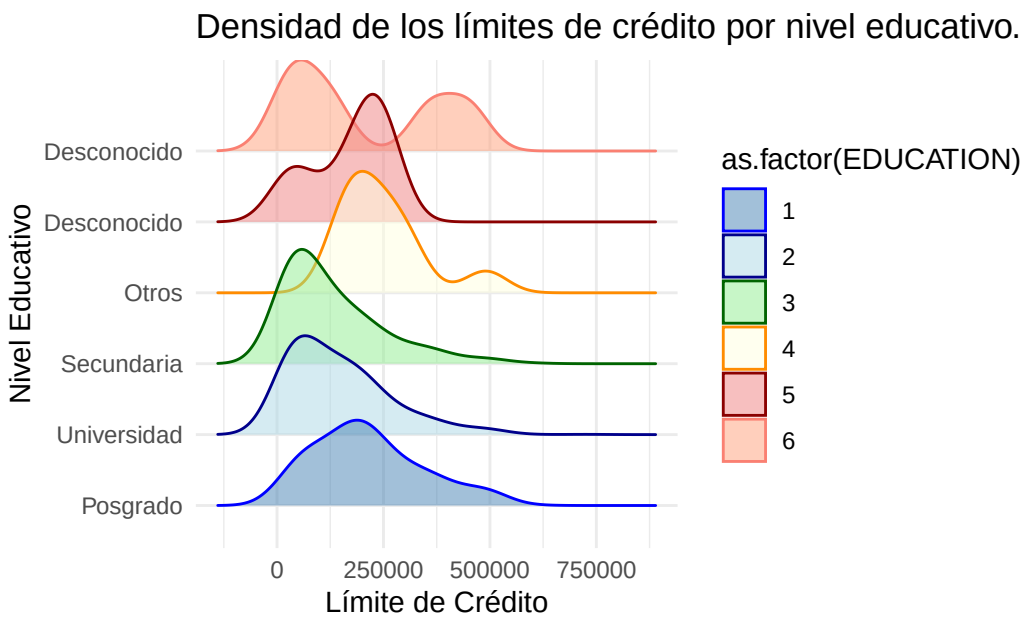
Fuente: Elaboración propia utilizando la base de datos de Kaggle

Agregamos también un gráfico de densidades, esto con el objetivo de ver la distribución.

```
library(ggplot2)
library(dplyr)
library(ggridges)

data_credit %>%
  ggplot(aes(x = LIMIT_BAL, y = as.factor(EDUCATION), fill = as.factor(EDUCATION), color = as
  geom_density_ridges(alpha = 0.5) +
  labs(title = "Densidad de los límites de crédito por nivel educativo.",
       x = "Límite de Crédito",
       y = "Nivel Educativo") +
  scale_fill_manual(values = c("1" = "steelblue", "2" = "lightblue", "3" = "lightgreen",
                               "4" = "lightyellow", "5" = "lightcoral", "6" = "lightsalmon"))
  scale_color_manual(values = c("1" = "blue", "2" = "darkblue", "3" = "darkgreen",
                                "4" = "darkorange", "5" = "darkred", "6" = "salmon")) +
  scale_y_discrete(labels = c(
    "1" = "Posgrado",
    "2" = "Universidad",
    "3" = "Secundaria",
    "4" = "Otros",
    "5" = "Desconocido",
    "6" = "Desconocido"
  )) +  labs(caption = "Fuente: Elaboración propia utilizando la base de datos de Kaggle") +
```

```
  theme_minimal()
```

```
Picking joint bandwidth of 49800
```

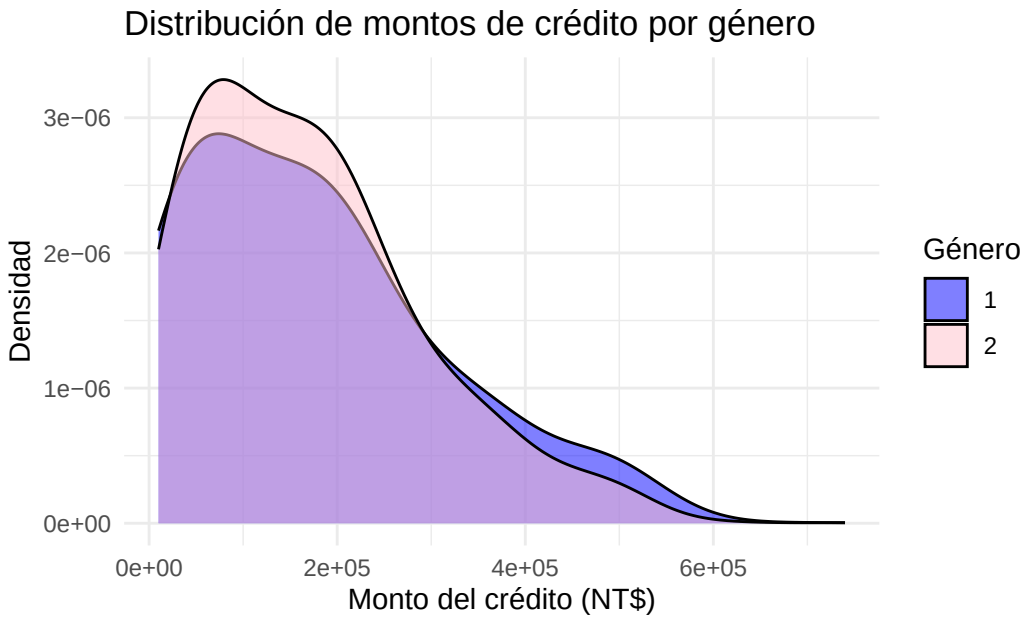


Fuente: Elaboración propia utilizando la base de datos de Kaggle

Las densidades que deben llamar nuestra atención son la de Secundaria, Universidad y Posgrado, esto por el hecho de que ellas son las que tienen la mayor concentración de datos.Además, veamos que las densidades tienen más o menos una distribución exponencial.

Por otro lado, observemos las densidades del límite de crédito con respecto al género.

```
library(ggplot2)


ggplot(data_credit, aes(x = LIMIT_BAL, fill = as.factor(SEX))) +
  geom_density(adjust = 2, alpha = 0.5) +
  labs(
    x = "Monto del crédito (NT$)",
    y = "Densidad",
    title = "Distribución de montos de crédito por género",
    fill = "Género"
  ) +
  scale_fill_manual(values = c("1" = "blue", "2" = "pink")) + labs(caption = "Fuente: Elabora
  theme_minimal()
```

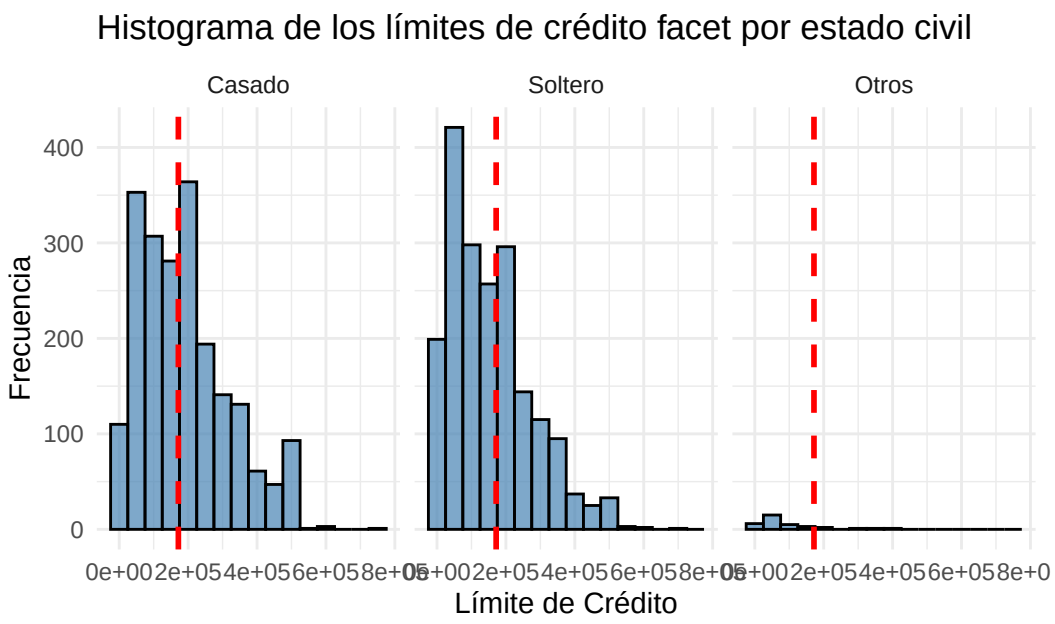


Fuente: Elaboración propia utilizando la base de datos de Kaggle

Del gráfico anterior, podemos determinar que a menores montos las mujeres tienen más crédito.

El siguiente gráfico es un histograma del límite de crédito con respecto al estado civil de las personas.

```
library(ggplot2)
library(dplyr)

data_credit %>%
  ggplot(aes(x = LIMIT_BAL)) +
  geom_histogram(binwidth = 50000, fill = "steelblue", color = "black", alpha = 0.7) +
  geom_vline(aes(xintercept = mean(LIMIT_BAL, na.rm = TRUE)),
             color = "red",
             linetype = "dashed",
             size = 1) +
  labs(title = "Histograma de los límites de crédito facet por estado civil",
       x = "Límite de Crédito",
       y = "Frecuencia") +
  facet_wrap(~MARRIAGE, labeller = as_labeller(c("1" = "Casado", "2" = "Soltero", "3" = "Otr
  theme_minimal()
```

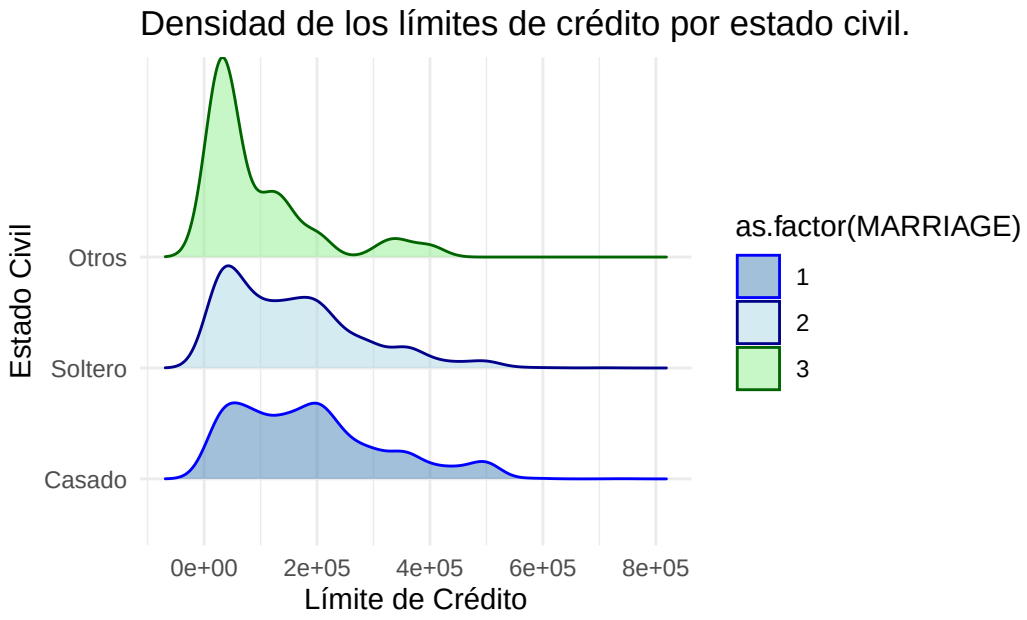


Fuente: Elaboración propia utilizando la base de datos de Kaggle

Agregamos densidades de las respectivas variables.

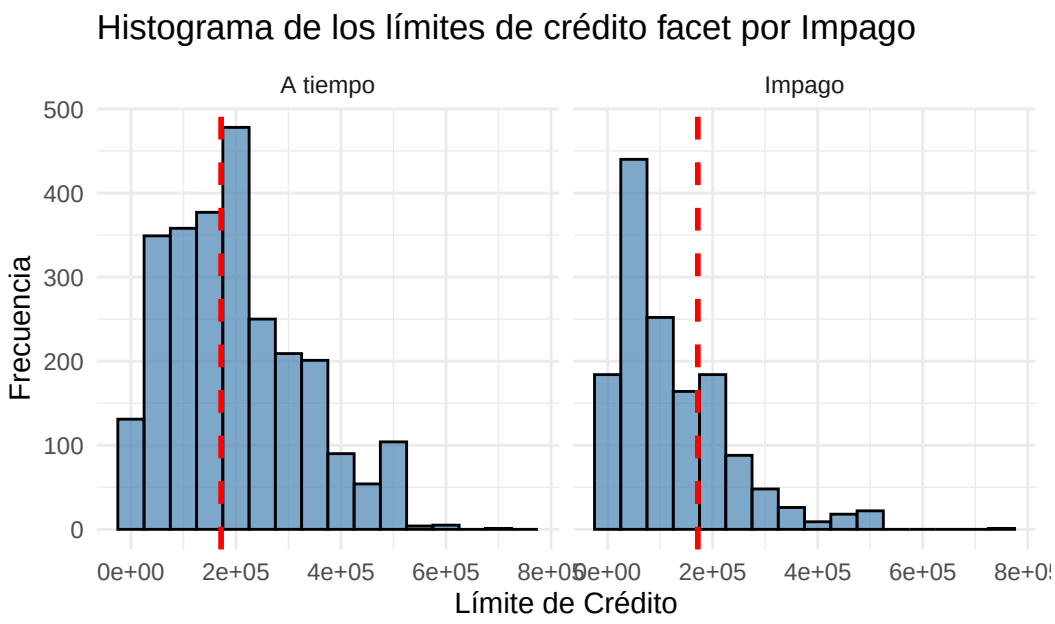
```
library(ggplot2)
library(dplyr)
library(ggridges)

data_credit %>%
  ggplot(aes(x = LIMIT_BAL, y = as.factor(MARRIAGE), fill = as.factor(MARRIAGE), color = as.
  geom_density_ridges(alpha = 0.5) +
  labs(title = "Densidad de los límites de crédito por estado civil.",
       x = "Límite de Crédito",
       y = "Estado Civil") +
  scale_fill_manual(values = c("1" = "steelblue", "2" = "lightblue", "3" = "lightgreen")) +
  scale_color_manual(values = c("1" = "blue", "2" = "darkblue", "3" = "darkgreen")) +
  scale_y_discrete(labels = c(
    "1" = "Casado",
    "2" = "Soltero",
    "3" = "Otros"
  )) +  labs(caption = "Fuente: Elaboración propia utilizando la base de datos de Kaggle") +
  theme_minimal()
```

```
Picking joint bandwidth of 26300
```



Fuente: Elaboración propia utilizando la base de datos de Kaggle

```
library(ggplot2)
library(dplyr)

data_credit %>%
  ggplot(aes(x = LIMIT_BAL)) +
  geom_histogram(binwidth = 50000, fill = "steelblue", color = "black", alpha = 0.7) +
  geom_vline(aes(xintercept = mean(LIMIT_BAL, na.rm = TRUE)),
             color = "red",
             linetype = "dashed",
             size = 1) +
  labs(title = "Histograma de los límites de crédito facet por Impago",
       x = "Límite de Crédito",
       y = "Frecuencia") +
  facet_wrap(~default.payment.next.month, labeller = as_labeller(c("0" = "A tiempo", "1" = "
  theme_minimal()
```

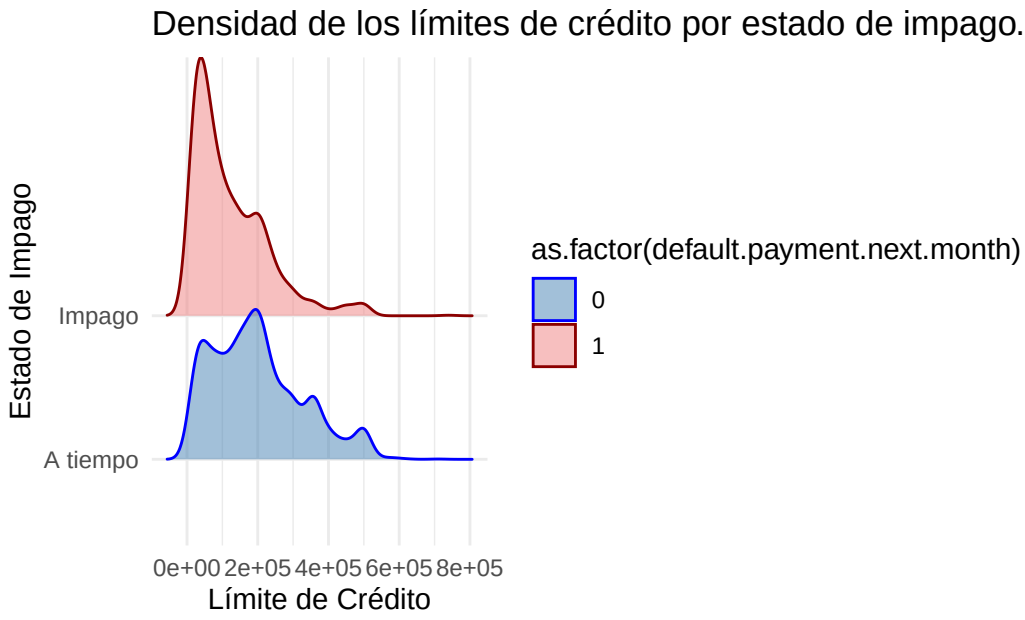


Fuente: Elaboración propia utilizando la base de datos de Kaggle

Agregamos dicho gráfico de densidades para las anteriores variables.

```
library(ggplot2)
library(dplyr)
library(ggridges)

data_credit %>%
  ggplot(aes(x = LIMIT_BAL, y = as.factor(default.payment.next.month), fill = as.factor(defau
  geom_density_ridges(alpha = 0.5) +
  labs(title = "Densidad de los límites de crédito por estado de impago.",
       x = "Límite de Crédito",
       y = "Estado de Impago") +
  scale_fill_manual(values = c("0" = "steelblue", "1" = "lightcoral")) +
  scale_color_manual(values = c("0" = "blue", "1" = "darkred")) +
  scale_y_discrete(labels = c(
    "0" = "A tiempo",
    "1" = "Impago"
  )) +  labs(caption = "Fuente: Elaboración propia utilizando la base de datos de Kaggle") +
  theme_minimal()
```

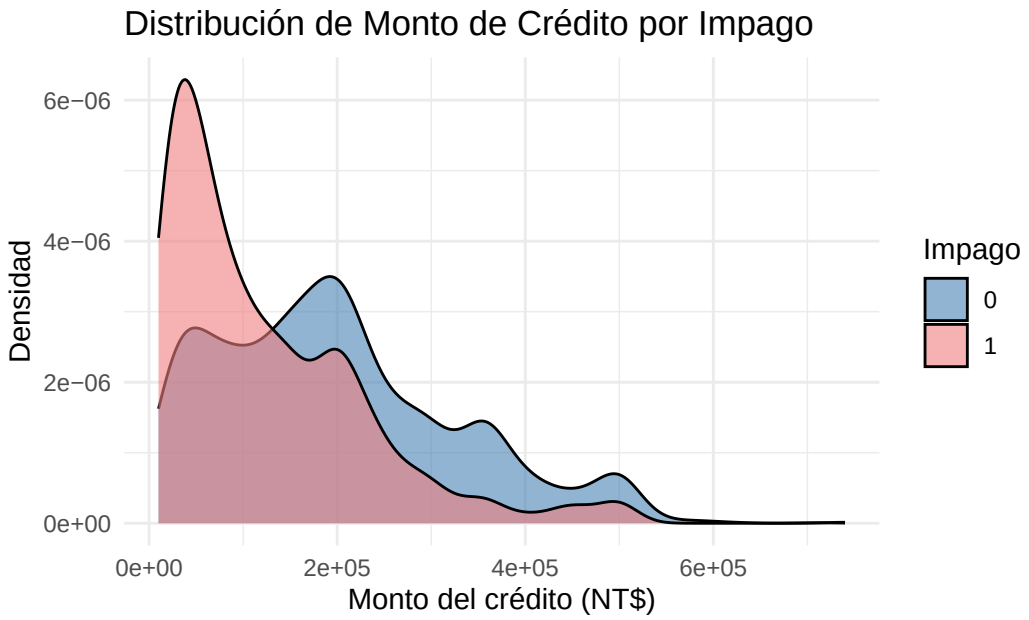
```
Picking joint bandwidth of 22100
```



n propia utilizando la base de datos de Kaggle

Ploteando los gráficos juntos, para una mejor visualización tenemos que:

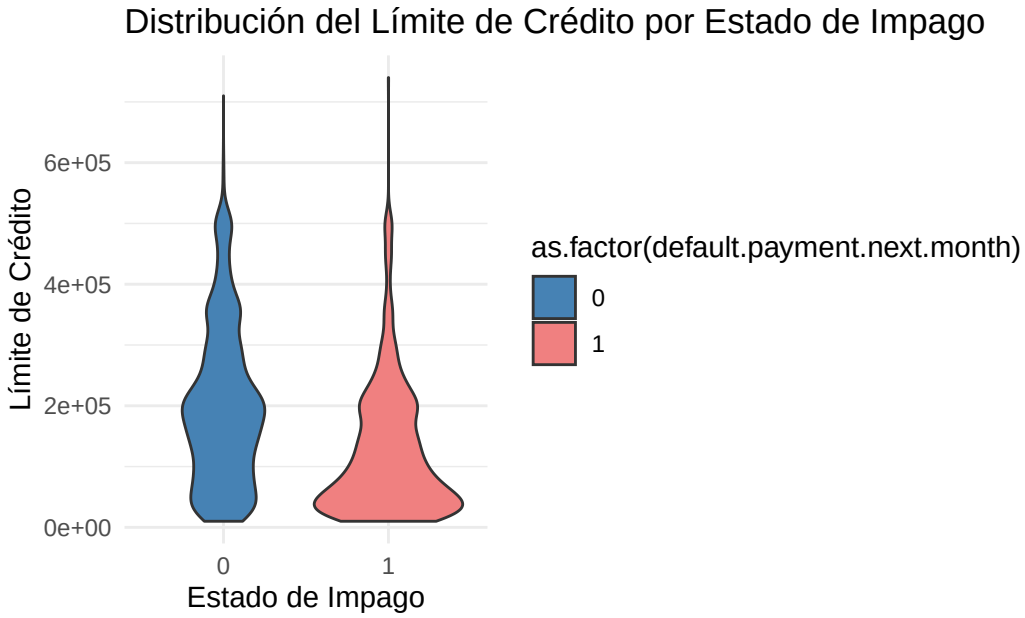
```
ggplot(data_credit, aes(x = LIMIT_BAL, fill = as.factor(default.payment.next.month))) +
  geom_density(alpha = 0.6) +
  labs(
    x = "Monto del crédito (NT$)",
    y = "Densidad",
    title = "Distribución de Monto de Crédito por Impago",
    fill = "Impago"
  ) +
  scale_fill_manual(values = c("0" = "steelblue", "1" = "lightcoral")) +
 labs(caption = "Fuente: Elaboración propia utilizando la base de datos de Kaggle") + theme_
```



Fuente: Elaboración propia utilizando la base de datos de Kaggle

Utilizando una gráfica de violín, veamos la distribución para compararlas visualmente.

```
ggplot(data_credit, aes(x = as.factor(default.payment.next.month), y = LIMIT_BAL, fill = as.
  geom_violin() +
  labs(title = "Distribución del Límite de Crédito por Estado de Impago", x = "Estado de Impa
  scale_fill_manual(values = c("0" = "steelblue", "1" = "lightcoral")) + labs(caption = "Fuen
  theme_minimal()
```

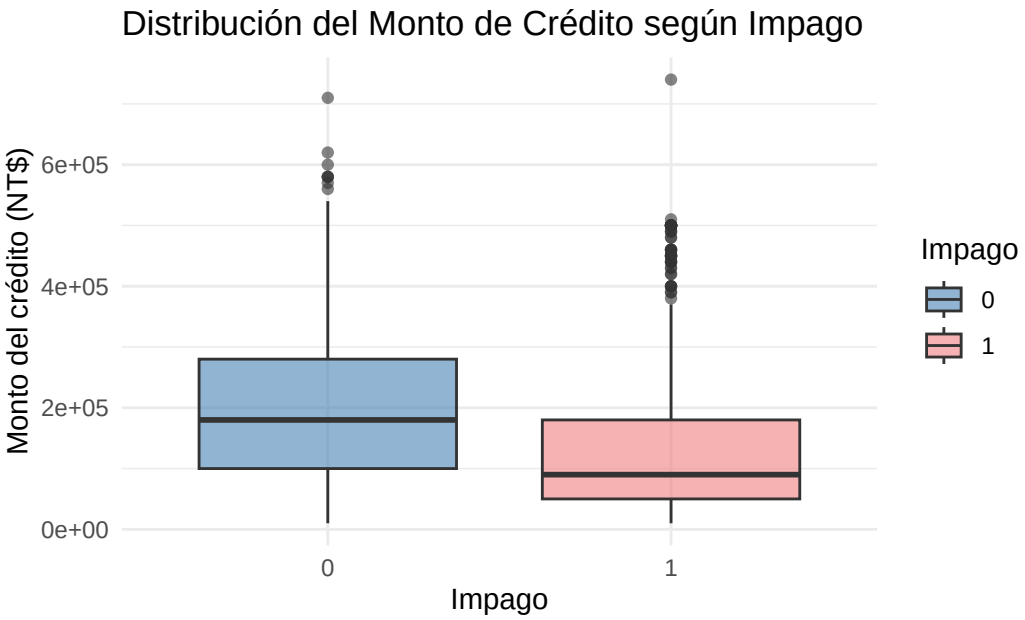


n propia utilizando la base de datos de Kaggle

Por el gráfico de la izquierda, podemos observar como a menores límite de créditos hay una concentración de las personas que caen en impago, reduciendose conforme el límite de crédito es mayor. Note además que un gráfico de violín al final son las densidades reflejadas, podemos obtener la misma información de un gráfico de violín que de un gráfico de densidades.

luego utilizando un diagrama de cajas para hacer otra comparación.

```
library(ggplot2)

ggplot(data_credit, aes(x = as.factor(default.payment.next.month), y = LIMIT_BAL, fill = as.
  geom_boxplot(alpha = 0.6) +
  labs(
    x = "Impago",
    y = "Monto del crédito (NT$)",
    title = "Distribución del Monto de Crédito según Impago",
    fill = "Impago"
  ) +
  scale_fill_manual(values = c("0" = "steelblue", "1" = "lightcoral")) + labs(caption = "Fuer
  theme_minimal()
```



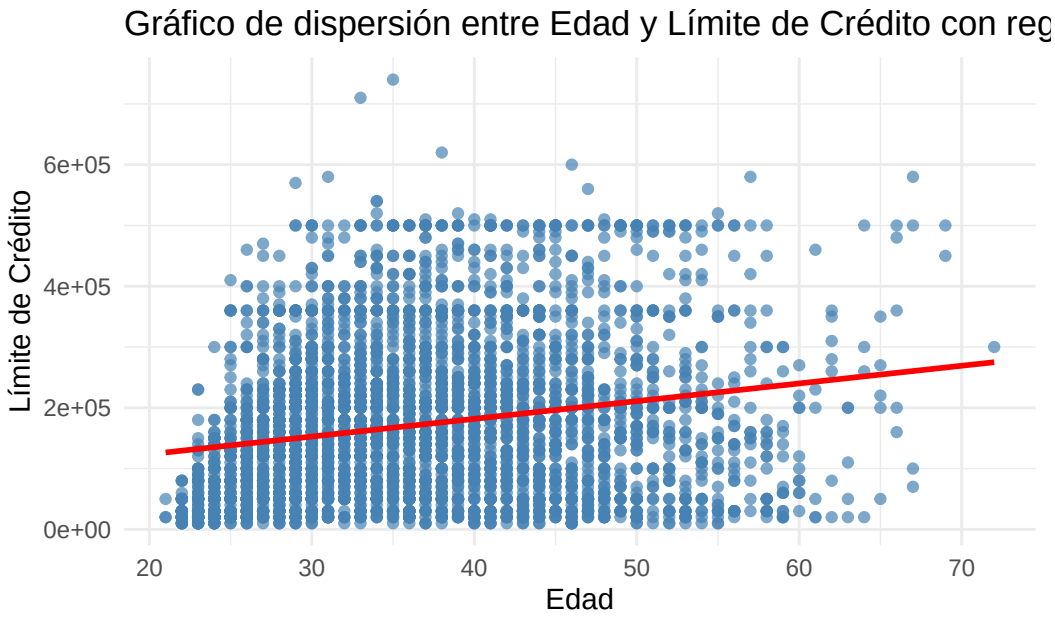
Fuente: Elaboración propia utilizando la base de datos de Kaggle

Del gráfico anterior entonces podemos observar que en promedio las personas que pagan a tiempo según el monto de crédito es mayor que el de las personas que cae en impago, lo cual nos ayudará a determinar más adelante si este factor es de importancia a la hora del riesgo de pago.

Por último veamos cómo de distribuyen las variables AGE y LIMIT_BAL. Usaremos un gráfico de dispersión

```
library(ggplot2)

ggplot(data_credit, aes(x = AGE, y = LIMIT_BAL)) +
  geom_point(color = "steelblue", alpha = 0.7) +
  geom_smooth(method = "lm", color = "red", se = FALSE) +
  labs(title = "Gráfico de dispersión entre Edad y Límite de Crédito con regresión lineal",
       x = "Edad",
       y = "Límite de Crédito") +labs(caption = "Fuente: Elaboración propia utilizando la bas
  theme_minimal()
```

```
`geom_smooth()` using formula = 'y ~ x'
```



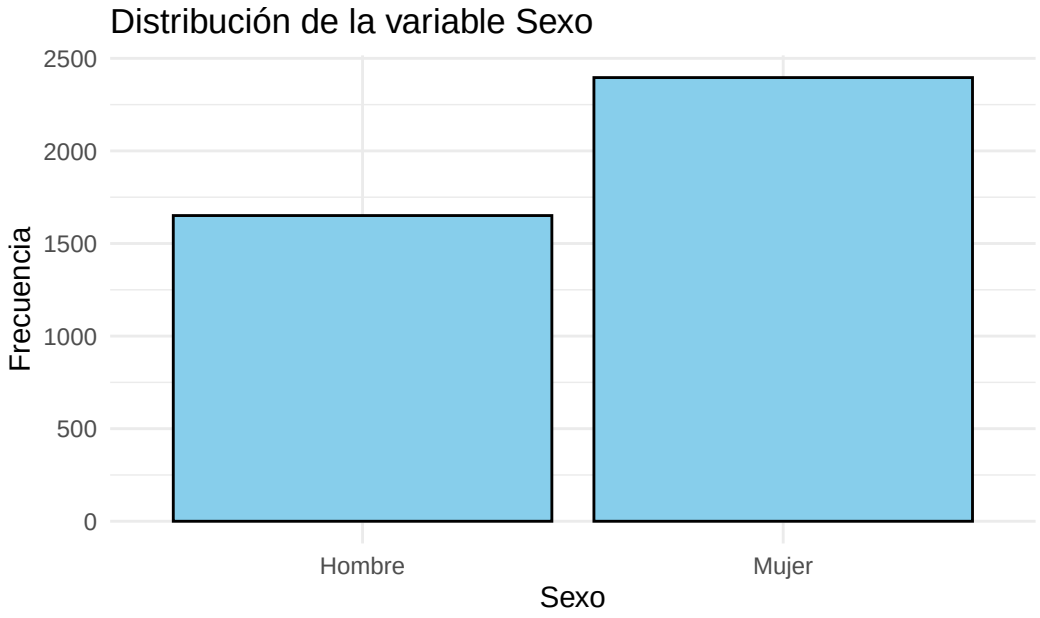
Fuente: Elaboración propia utilizando la base de datos de Kaggle

Cuando realizamos la matriz de correlación, vimos que la correlación entre estas dos variables es de 0.21, lo cual es algo bajo, podemos entonces ver esa tendencia en este gráfico, pues tenemos muchos puntos dispersos, al realizar una regresión lineal, podemos ver la linea que mejor se ajusta a estos puntos. Podemos inferir que hay una relación positiva débil entre las variables. Aunque un valor de 0.21 no ayuda a predecir qué pasaría cuando las variables aumenten.

3.1.2.2.2 Gráficos de Variables Categóricas.

Comenzaremos esta sección realizando gráficos de barras, con la intención de ver las frecuencias de las variables.

```
library(ggplot2)


ggplot(data = data_credit, aes(x = as.factor(SEX))) +
  geom_bar(fill = "skyblue", color = "black") +
  labs(title = "Distribución de la variable Sexo", x = "Sexo", y = "Frecuencia") +
  scale_x_discrete(labels = c("1" = "Hombre", "2" = "Mujer")) + labs(caption = "Fuente: Elabo
  theme_minimal()
```

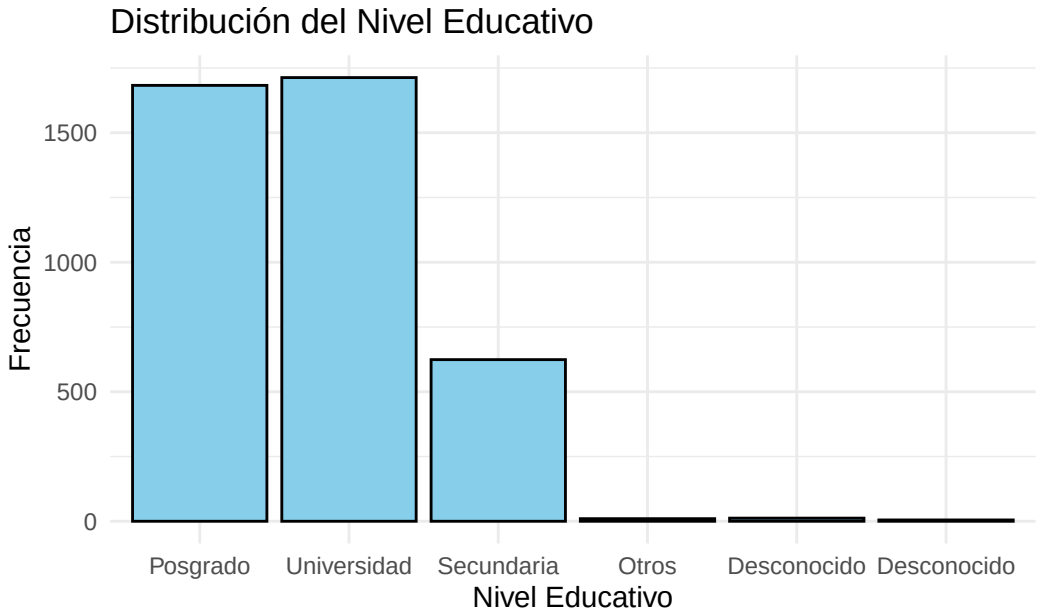


Fuente: Elaboración propia utilizando la base de datos de Kaggle

Con esto podemos ver que tenemos más datos de mujeres que de hombres.

Veamos como se comporta la variable de educación.

```
library(ggplot2)

ggplot(data = data_credit, aes(x = as.factor(EDUCATION))) +
  geom_bar(fill = "skyblue", color = "black") +
  labs(title = "Distribución del Nivel Educativo", x = "Nivel Educativo", y = "Frecuencia") +
  scale_x_discrete(labels = c("1" = "Posgrado",
                              "2" = "Universidad",
                              "3" = "Secundaria",
                              "4" = "Otros",
                              "5" = "Desconocido",
                              "6" = "Desconocido")) + labs(caption = "Fuente: Elaboración pro
  theme_minimal()
```

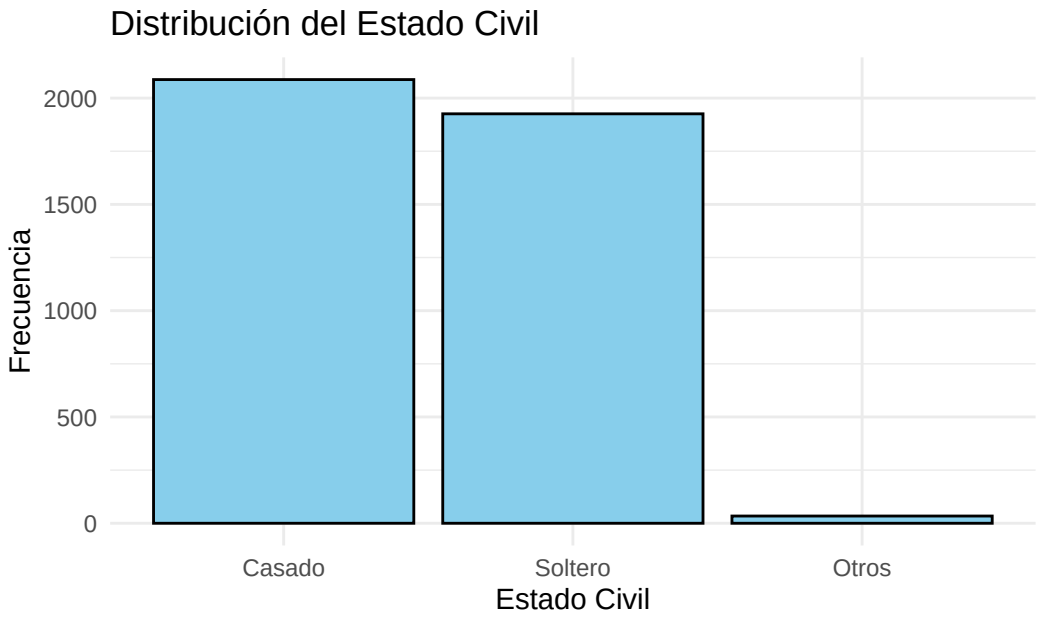


Fuente: Elaboración propia utilizando la base de datos de Kaggle

Observamos entonces que nuestra base de datos contiene más información de personas que están posgrados o que terminaron la universidad.

Luego para la variable de estado civil.

```
library(ggplot2)

ggplot(data = data_credit, aes(x = as.factor(MARRIAGE))) +
  geom_bar(fill = "skyblue", color = "black") +
  labs(title = "Distribución del Estado Civil", x = "Estado Civil", y = "Frecuencia") +
  scale_x_discrete(labels = c("1" = "Casado",
                              "2" = "Soltero",
                              "3" = "Otros")) + labs(caption = "Fuente: Elaboración propia u
  theme_minimal()
```

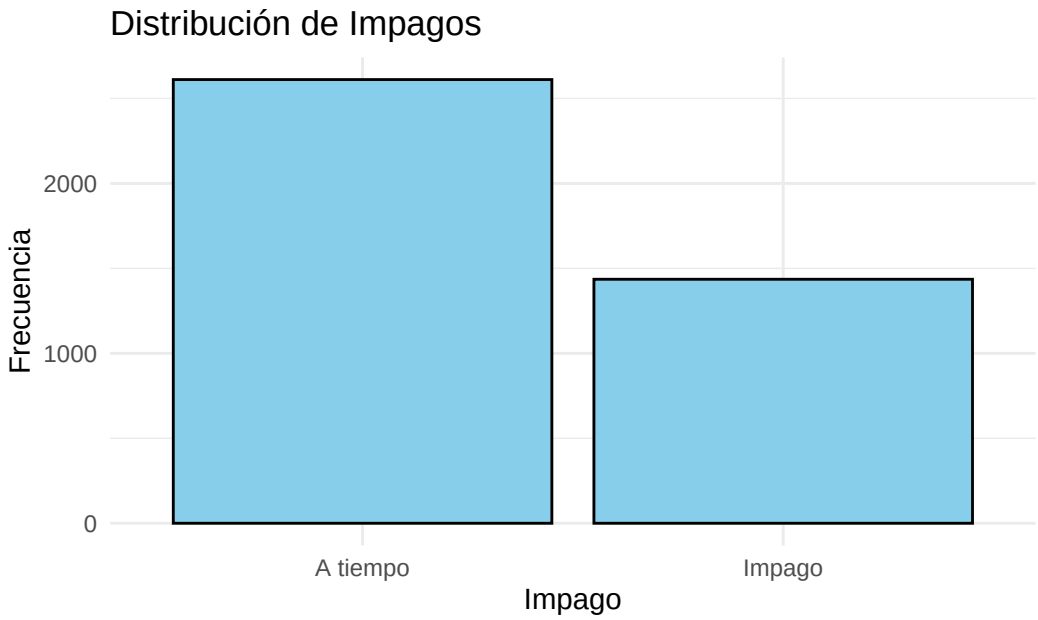


Fuente: Elaboración propia utilizando la base de datos de Kaggle

Al igual que antes, tenemos más información de las personas que están casadas y de las que están solteras.

Por último vamos a ver la gráfica de barras de la variable de interés, la cual es si cayó en impago o no lo hizo.

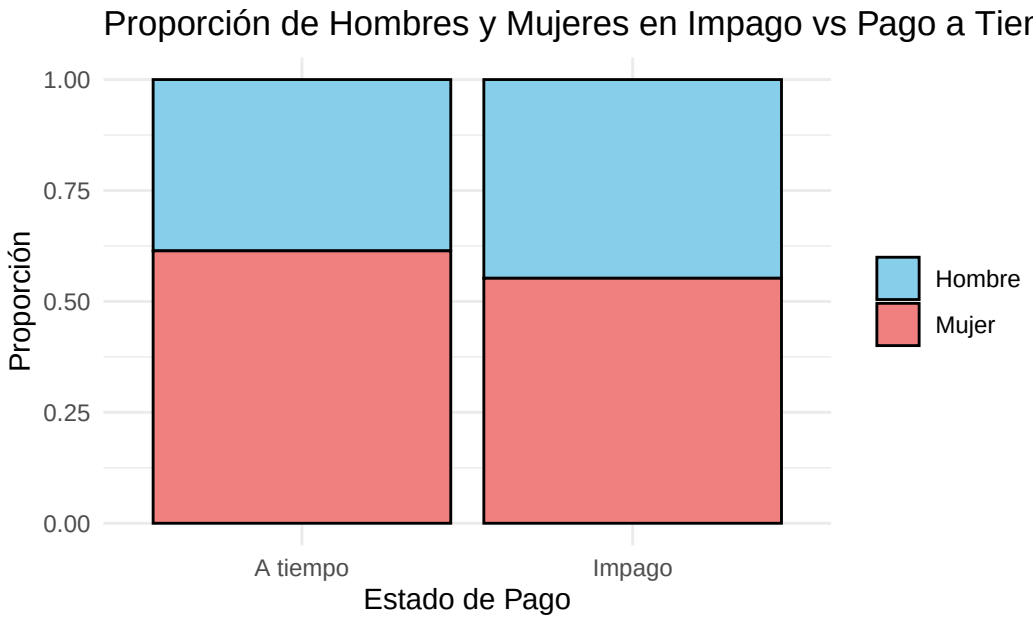
```
library(ggplot2)

ggplot(data = data_credit, aes(x = as.factor(default.payment.next.month))) +
  geom_bar(fill = "skyblue", color = "black") +
  labs(title = "Distribución de Impagos", x = "Impago", y = "Frecuencia") +
  scale_x_discrete(labels = c("0" = "A tiempo", "1" = "Impago")) + labs(caption = "Fuente: E
  theme_minimal()
```



Fuente: Elaboración propia utilizando la base de datos de Kaggle

Con esto terminamos los gráficos aislados de las variables categóricas y damos inicio a ver cómo se distribuyen cuando las relacionamos.

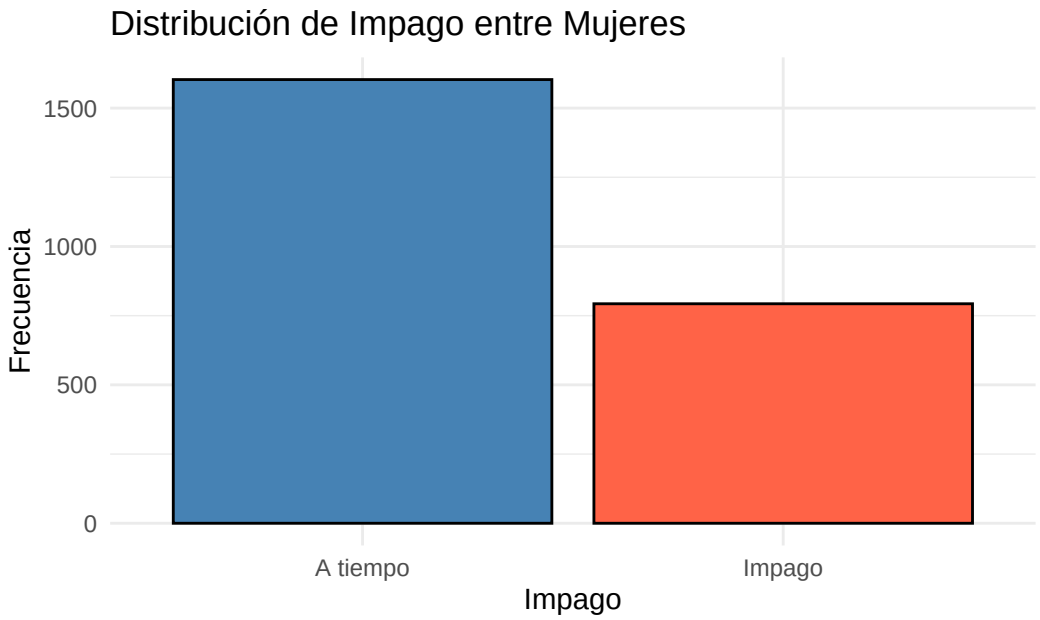
```
library(ggplot2)
library(dplyr)

ggplot(data = data_credit, aes(x = as.factor(default.payment.next.month), fill = as.factor(SE
  geom_bar(position = "fill", color = "black") +
  labs(title = "Proporción de Hombres y Mujeres en Impago vs Pago a Tiempo",
       x = "Estado de Pago", y = "Proporción") +
  scale_x_discrete(labels = c("0" = "A tiempo", "1" = "Impago")) +
  scale_fill_manual(labels = c("1" = "Hombre", "2" = "Mujer"), values = c("skyblue", "lightco
  theme_minimal() + labs(caption = "Fuente: Elaboración propia utilizando la base de datos de
  theme(legend.title = element_blank())
```



Fuente: Elaboración propia utilizando la base de datos de Kaggle

Del gráfico anterior, podemos observar entonces que de las personas que cayeron en impago, la mayoría son mujeres, al menos más del 50%, sin embargo, veamos de manera aislada esto.

```
library(ggplot2)
library(dplyr)

data_credit %>%
  filter(SEX == 2) %>%
  ggplot(aes(x = as.factor(default.payment.next.month), fill = as.factor(default.payment.next
  geom_bar(stat = "count", color = "black") +
  labs(title = "Distribución de Impago entre Mujeres",
       x = "Impago",
       y = "Frecuencia",
       fill = "Impago") +
  scale_x_discrete(labels = c("0" = "A tiempo", "1" = "Impago")) +
  scale_fill_manual(values = c("0" = "steelblue", "1" = "tomato")) +
  theme_minimal() + labs(caption = "Fuente: Elaboración propia utilizando la base de datos de
  theme(legend.position = "none")
```



Fuente: Elaboración propia utilizando la base de datos de Kaggle

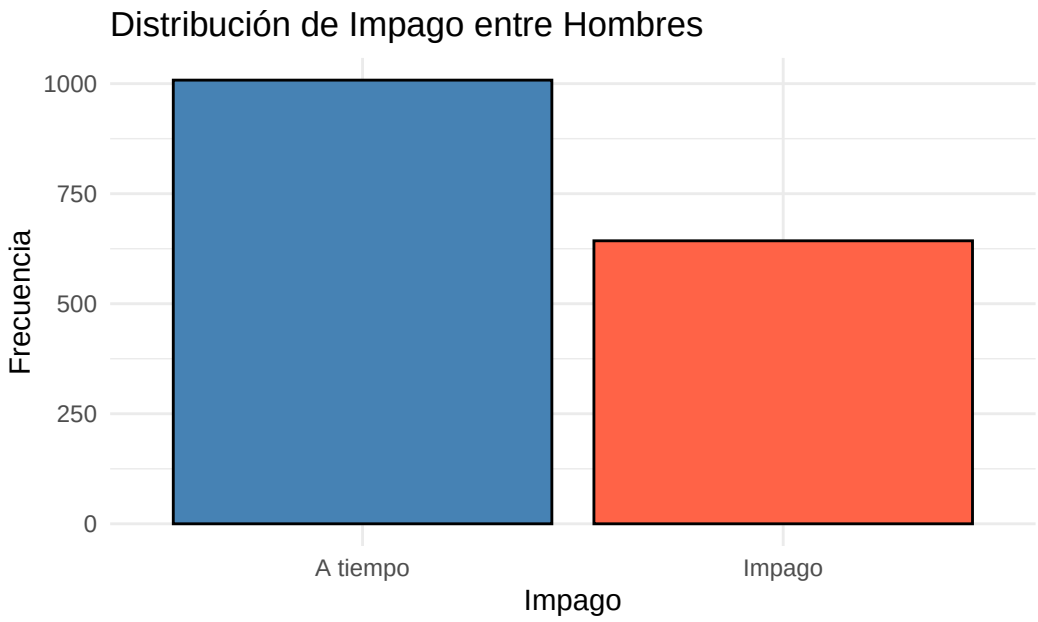
Numéricamente esto es:

```
library(dplyr)

mujeres_impago <- data_credit %>%
  filter(SEX == 2) %>%  # Filtramos los datos, porque nos interesan solo las mujeres
  summarise(
    total_mujeres = n(),
    mujeres_impago = sum(default.payment.next.month == 1)  # Número de mujeres en impago
  ) %>%
  mutate(porcentaje_impago = mujeres_impago / total_mujeres * 100)  # Calculamos el porcentaj

# Mostramos el resultado
mujeres_impago
```

```
# A tibble: 1 x 3
  total_mujeres mujeres_impago porcentaje_impago
          <int>          <int>             <dbl>
1          2396            793              33.1
```

Con esto podemos observar que de las mujeres totales, solo el 33% cayó en impago. Haremos un análisis similar con respecto a los hombre.

```r
library(ggplot2)
library(dplyr)

data_credit %>%
  filter(SEX == 1) %>%
  ggplot(aes(x = as.factor(default.payment.next.month), fill = as.factor(default.payment.next
  geom_bar(stat = "count", color = "black") +
  labs(title = "Distribución de Impago entre Hombres",
       x = "Impago",
       y = "Frecuencia",
       fill = "Impago") +
  scale_x_discrete(labels = c("0" = "A tiempo", "1" = "Impago")) +
  scale_fill_manual(values = c("0" = "steelblue", "1" = "tomato")) +
  theme_minimal() + labs(caption = "Fuente: Elaboración propia utilizando la base de datos de
  theme(legend.position = "none")
```



Fuente: Elaboración propia utilizando la base de datos de Kaggle

Numéricamente podemos observar que:

```r
library(dplyr)

hombres_impago <- data_credit %>%
  filter(SEX == 1) %>%
  summarise(
```

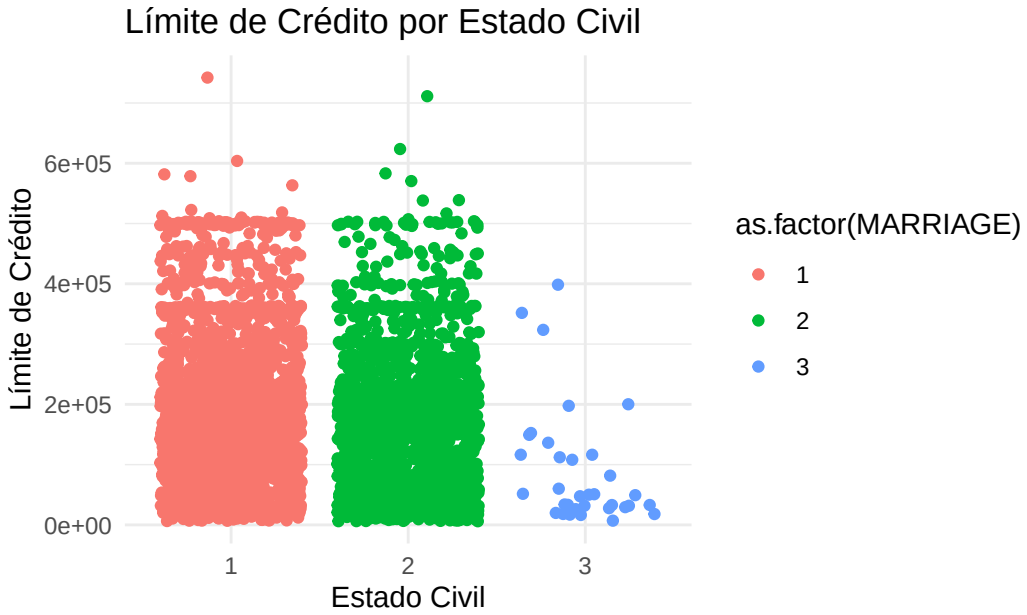
```
    total_hombres = n(),
    hombres_impago = sum(default.payment.next.month == 1)
  ) %>%
  mutate(porcentaje_impago = hombres_impago / total_hombres * 100)

hombres_impago
```

```
# A tibble: 1 x 3
  total_hombres hombres_impago porcentaje_impago
          <int>          <int>             <dbl>
1          1651            643              38.9
```

Con esto observamos que el porcentaje de los hombres que cayeron en impago, aunque es por poco, es mayor que el de las mujeres que cayeron en impago. Esto lo hicimos porque anteriormente se estaban comparando magnitudes que no se podían comparar, con los porcentajes podemos determinar que relativamente, los hombres tienden a caer más en impago que las mujeres, al menos eso podemos inferrir gracias a la evidencia de los datos.

Ahora vamos a visualizar el crédito con respecto a la variable de MARRIAGE.

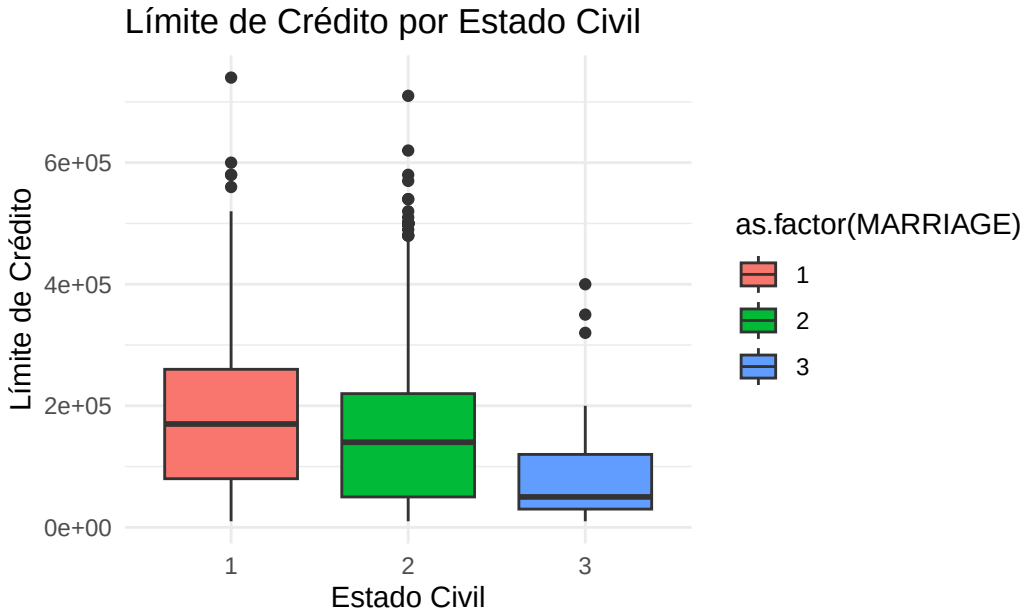
```
library(ggplot2)

ggplot(data_credit, aes(x = as.factor(MARRIAGE), y = LIMIT_BAL, color = as.factor(MARRIAGE))
  geom_jitter() +
  labs(title = "Límite de Crédito por Estado Civil", x = "Estado Civil", y = "Límite de Crédi
  theme_minimal()
```



Fuente: Elaboración propia utilizando la base de datos de Kaggle

Después analizaremos la relación entre las variables, ya que son variables de diferente naturaleza, es decir una categórica y una numérica, por lo que utilizaremos un análisis ANOVA para clarificar si las diferencias se deben al azar o si la evidencia estadística indican que están relacionados. Por el momento, haremos una comparación con gráficos de cajas, para observar de manera gráfica, como se siguen comportando.

```
library(ggplot2)

ggplot(data_credit, aes(x = as.factor(MARRIAGE), y = LIMIT_BAL, fill = as.factor(MARRIAGE)))
  geom_boxplot() +
  labs(title = "Límite de Crédito por Estado Civil", x = "Estado Civil", y = "Límite de Crédi
  theme_minimal()
```

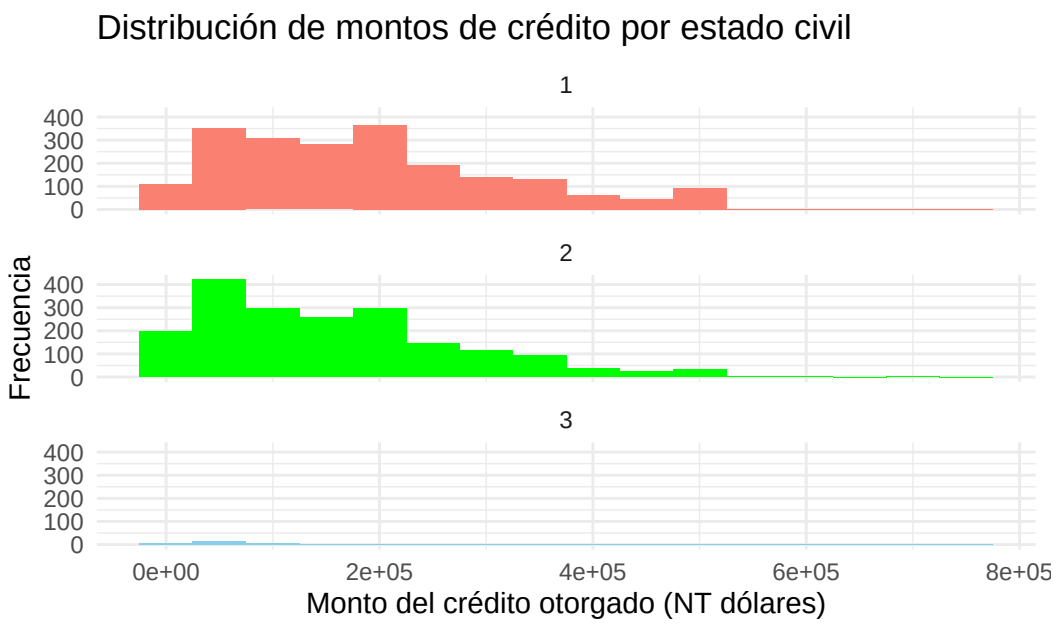


Fuente: Elaboración propia utilizando la base de datos de Kaggle

Podemos observar gráficamente que las diferencias no son tan significativas, sin embargo esto es una conjetura, nos ajustaremos a un análisis estadístico más riguroso en posteriores secciones. Por otro lado, podemos observar que el nivel 3 difiere de los otros niveles, esto se puede deber a las bajas observaciones que tenemos en este nivel.

Por último para tener la relación entre estas variables, observemos el siguiente gráfico.

```
library(ggplot2)

ggplot(data_credit, aes(x = LIMIT_BAL, fill = as.factor(MARRIAGE))) +
  geom_histogram(binwidth = 50000) +
  facet_wrap(~MARRIAGE, nrow = 3) +
  labs(
    x = "Monto del crédito otorgado (NT dólares)",
    y = "Frecuencia",
    title = "Distribución de montos de crédito por estado civil",
    fill = "Estado civil"
  ) +
  scale_fill_manual(values = c("1" = "salmon", "2" = "green", "3" = "skyblue")) +
  theme_minimal() + labs(caption = "Fuente: Elaboración propia utilizando la base de datos d
  theme(legend.position = "none")
```

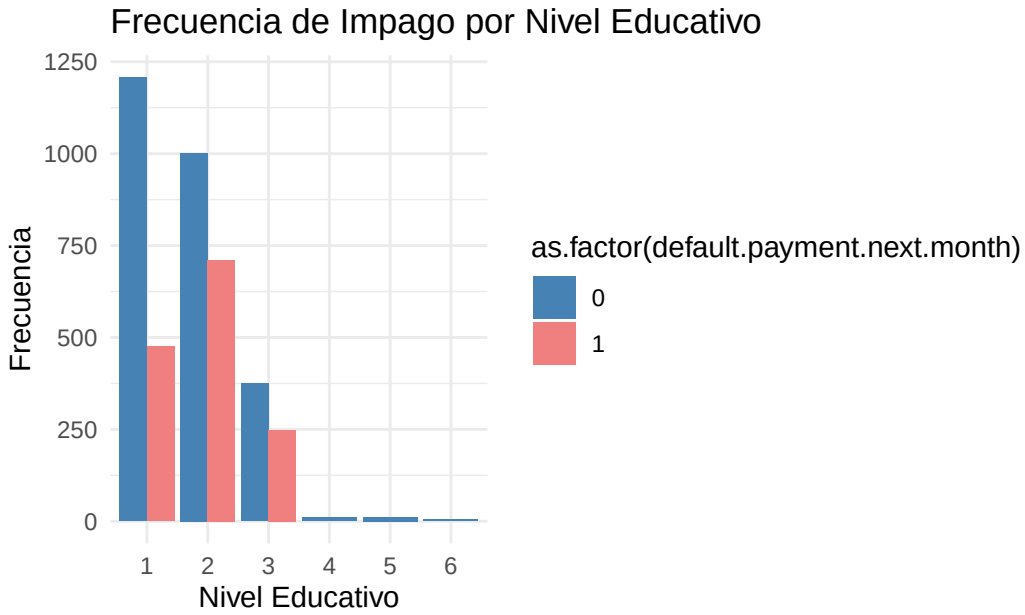


Fuente: Elaboración propia utilizando la base de datos de Kaggle

Como hemos mencionado, la distribución de las variables casado y soltero se parecen mucho visualmente, no podemos decir más del estado "otro".

Analizamos ahora la frecuencia de impago en relación con el nivel educativo.

```
ggplot(data_credit, aes(x = as.factor(EDUCATION), fill = as.factor(default.payment.next.montl
  geom_bar(position = "dodge") +
  labs(title = "Frecuencia de Impago por Nivel Educativo", x = "Nivel Educativo", y = "Frecue
  scale_fill_manual(values = c("0" = "steelblue", "1" = "lightcoral")) + labs(caption = "Fuer
  theme_minimal()
```

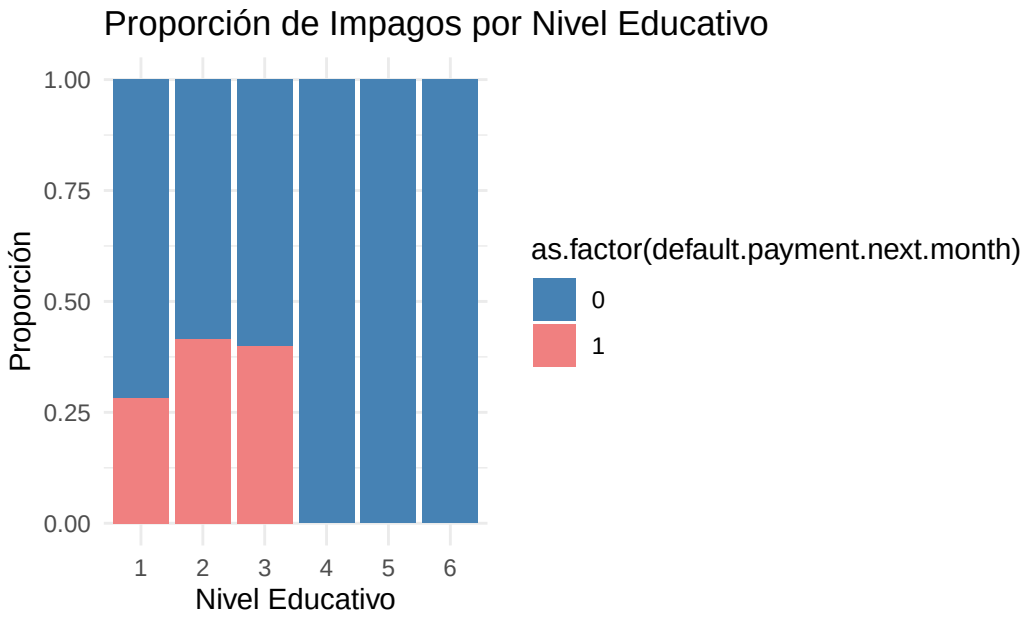


n propia utilizando la base de datos de Kaggle

Recordemos que la etiqueta 2 equivale a las personas que están en grado de haber terminado o concluido la universidad. Con este gráfico podemos interpretar que las personas que terminaron la universidad tienen una alta proporción de haber caido en impago, esto se puede deber a créditos estudiantiles y la dificultad de conseguir empleo, sin embargo esto es una conjetura y no vamos analizar esta consecuencia, ya que solo nos importa ver qué están diciendo nuestros datos.

Con el siguiente gráfico queremos observar la proporción de las personas que caen en impago, según el nievl educativo.

```
ggplot(data_credit, aes(x = as.factor(EDUCATION), fill = as.factor(default.payment.next.month
  geom_bar(position = "fill") +
  labs(title = "Proporción de Impagos por Nivel Educativo", x = "Nivel Educativo", y = "Propo
  scale_fill_manual(values = c("0" = "steelblue", "1" = "lightcoral")) + labs(caption = "Fuen
  theme_minimal()
```

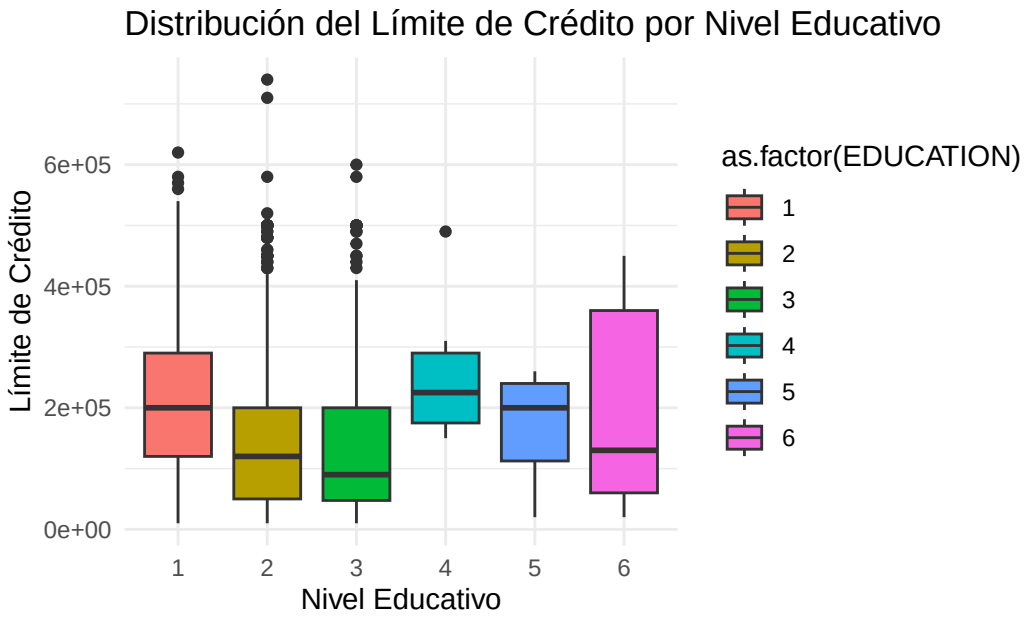


n propia utilizando la base de datos de Kaggle

Justamente, este gráfico refleja que las personas que están en el nivel de universidad presentan una mayor proporción de impago.

Por otro lado, vamos analizar gráficamente la relación de la variable educación con el límite de crédito.

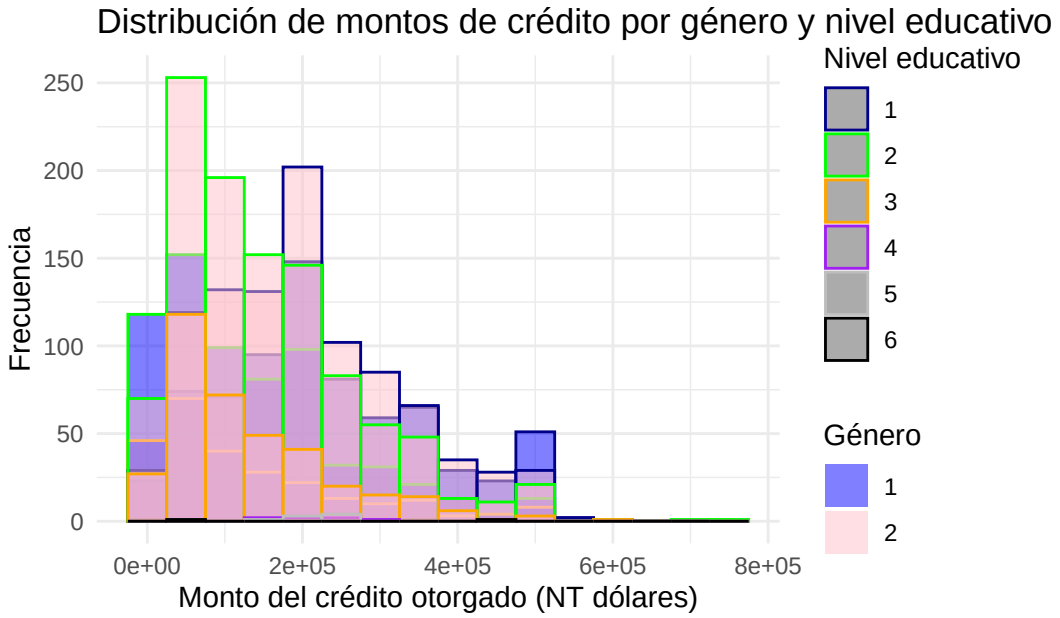
```
ggplot(data_credit, aes(x = as.factor(EDUCATION), y = LIMIT_BAL, fill = as.factor(EDUCATION)
  geom_boxplot() +
  labs(title = "Distribución del Límite de Crédito por Nivel Educativo", x = "Nivel Educativ
  theme_minimal()
```



Fuente: Elaboración propia utilizando la base de datos de Kaggle

Observando las medias, inferimos que en promedio el límite de crédito otorgado a las personas que están en un nivel de posgrado es mayor que las personas que tiene solo universidad o secundaria. A su vez, la media de Universidad es mayor que la media de secundaria. Por la calidad de los datos de la base de datos, no podemos analizar con lujo de detalle los niveles 3,4 y 5.

```
library(ggplot2)

ggplot(data_credit, aes(x = LIMIT_BAL, fill = as.factor(SEX), color = as.factor(EDUCATION)))
  geom_histogram(binwidth = 50000, alpha = 0.5, position = "identity") +
  labs(
    x = "Monto del crédito otorgado (NT dólares)",
    y = "Frecuencia",
    title = "Distribución de montos de crédito por género y nivel educativo",
    fill = "Género",
    color = "Nivel educativo"
  ) +
  scale_fill_manual(values = c("1" = "blue", "2" = "pink")) +  # Colores para sexo
  scale_color_manual(values = c("1" = "darkblue", "2" = "green", "3" = "orange", "4" = "purpl
  theme_minimal()
```



Fuente: Elaboración propia utilizando la base de datos de Kaggle

Con el gráfico anterior podemos ver la relación existente entre el nivel educativo, el género y el monto del crédito otorgado, es decir, 3 variables graficadas.

3.1.3 Análisis Matemático de las correlaciones

Para esta sección, primero vamos analizar tablas de contingencias, esto con el fin de eencontrar relaciones entre las variables, utilizaremos las pruebas chi-cuadrado y la prueba ” — “, para determinar esta relaciones entre variables categóricas.

Posteriormente, utilizaremos los índices de correlación obtenidos en la matriz de correlación para el análisis de la base de datos y con ello apoyarnos en la evidencia teórica que existe.

3.1.3.1 Tablas de Contigencias y p-values

Para esta sección vamos hacer tablas de contigencias, esto con el objetivo de buscar las relaciones que tienen las variables categóricas.

```
library(gmodels)

# Realizamos la tabla de contingencia
CrossTable(data_credit$EDUCATION, data_credit$default.payment.next.month, prop.chisq = FALSE)
```

```
   Cell Contents
|-------------------------|
|                       N |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  4047


                     | data_credit$default.payment.next.month
data_credit$EDUCATION |         0 |         1 | Row Total |
----------------------|-----------|-----------|-----------|
                    1 |      1207 |       476 |      1683 |
                      |     0.717 |     0.283 |     0.416 |
                      |     0.462 |     0.331 |           |
                      |     0.298 |     0.118 |           |
----------------------|-----------|-----------|-----------|
                    2 |      1002 |       711 |      1713 |
                      |     0.585 |     0.415 |     0.423 |
                      |     0.384 |     0.495 |           |
                      |     0.248 |     0.176 |           |
----------------------|-----------|-----------|-----------|
                    3 |       375 |       249 |       624 |
                      |     0.601 |     0.399 |     0.154 |
                      |     0.144 |     0.173 |           |
                      |     0.093 |     0.062 |           |
----------------------|-----------|-----------|-----------|
                    4 |        10 |         0 |        10 |
                      |     1.000 |     0.000 |     0.002 |
                      |     0.004 |     0.000 |           |
                      |     0.002 |     0.000 |           |
----------------------|-----------|-----------|-----------|
                    5 |        12 |         0 |        12 |
                      |     1.000 |     0.000 |     0.003 |
                      |     0.005 |     0.000 |           |
                      |     0.003 |     0.000 |           |
----------------------|-----------|-----------|-----------|
                    6 |         5 |         0 |         5 |
```

```
                     |     1.000 |     0.000 |     0.001 |
                     |     0.002 |     0.000 |           |
                     |     0.001 |     0.000 |           |
---------------------|-----------|-----------|-----------|
         Column Total |      2611 |      1436 |      4047 |
                     |     0.645 |     0.355 |           |
---------------------|-----------|-----------|-----------|
```

De la tabla anterior podemos ver entonces las relaciones que hay entre las variables, por ejemplo, podemos ver que de las personas de educación, que pertencen al nivel de posgrado, un 71% de esas personas no imcuplieron su pago, es decir pagaron a tiempo. Y un 29% de esas personas si cayeron en impago.

Haremos la prueba Exacta de Fisher para realizar una Prueba de Hipótesis donde nuestra hipótesis nula es. H_0, no existe correlación entre las variables, y donde nuestra hipótesis alternativa, H_1, es que hay relación entre las variables.

```r
tabla_education <- table(data_credit$EDUCATION, data_credit$default.payment.next.month)

# Realizar la prueba exacta de Fisher con simulación de Monte Carlo
set.seed(2024)  # Fijamos la semilla
fisher_test_education <- fisher.test(tabla_education, simulate.p.value = TRUE, B = 10000)

print(fisher_test_education)
```

```
    Fisher's Exact Test for Count Data with simulated p-value (based on
    10000 replicates)

data:  tabla_education
p-value = 9.999e-05
alternative hypothesis: two.sided
```

Gracias al test anterior podemos inferir entonces que existe una dependencia de las variables. Rechazamos la hipótesis nula, hay evidencia estadística suficiente para decir que las variables tienen una relación.

Más explicado aún, el p-value que obtuvimos fue de 0.0001, lo que es mucho más pequeño que el 5% del nivel de significancia, por lo que entonces podemos rechazar la hipótesis nula.

Por otro lado, ahora vamos a comparar las variables de sexo y de impago, esto con el objetivo de observar si el sexo influye o tiene relación con la probabilidad de caer en impago.

```r
library(gmodels)

# Realizamos la tabla de contingencia
CrossTable(data_credit$SEX, data_credit$default.payment.next.month, prop.chisq = FALSE)
```

```
   Cell Contents
|-------------------------|
|                       N |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  4047


                | data_credit$default.payment.next.month
data_credit$SEX |         0 |         1 | Row Total |
----------------|-----------|-----------|-----------|
              1 |      1008 |       643 |      1651 |
                |     0.611 |     0.389 |     0.408 |
                |     0.386 |     0.448 |           |
                |     0.249 |     0.159 |           |
----------------|-----------|-----------|-----------|
              2 |      1603 |       793 |      2396 |
                |     0.669 |     0.331 |     0.592 |
                |     0.614 |     0.552 |           |
                |     0.396 |     0.196 |           |
----------------|-----------|-----------|-----------|
   Column Total |      2611 |      1436 |      4047 |
                |     0.645 |     0.355 |           |
----------------|-----------|-----------|-----------|
```

Con la tabla vemos la distribución de las variables entre ellas. Aplicamos la prueba de Fisher con simulación al igual que en el caso anterior.

```r
tabla_sex <- table(data_credit$SEX, data_credit$default.payment.next.month)

# Realizar la prueba exacta de Fisher con simulación de Monte Carlo
set.seed(2024)
fisher_test_sex <- fisher.test(tabla_sex, simulate.p.value = TRUE, B = 10000)

print(fisher_test_sex)
```

```
    Fisher's Exact Test for Count Data

data:  tabla_sex
p-value = 0.0001387
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.6792443 0.8856355
sample estimates:
odds ratio
 0.7755741
```

Gracias a la prueba anterior, obtenemos que el p-value es de 0.0001387, siendo bastante menor que el valor de significancia, el cual es de 5%, por lo que hay suficiente evidencia estadística para rechazar la hipótesis nula, así decimos entonces que existe una cierta dependencia entre estas variables.

Vamos a ver ahora como se comporta la variable MARRIAGE con el impago. Veamos primero su tabla de contingencia.

```r
library(gmodels)

# Realizamos la tabla de contingencia
CrossTable(data_credit$MARRIAGE, data_credit$default.payment.next.month, prop.chisq = FALSE)
```

```
   Cell Contents
|-------------------------|
|                       N |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
```

```
|-------------------------|


Total Observations in Table:  4047


                     | data_credit$default.payment.next.month
data_credit$MARRIAGE |         0 |         1 | Row Total |
---------------------|-----------|-----------|-----------|
                   1 |      1359 |       728 |      2087 |
                     |     0.651 |     0.349 |     0.516 |
                     |     0.520 |     0.507 |           |
                     |     0.336 |     0.180 |           |
---------------------|-----------|-----------|-----------|
                   2 |      1235 |       691 |      1926 |
                     |     0.641 |     0.359 |     0.476 |
                     |     0.473 |     0.481 |           |
                     |     0.305 |     0.171 |           |
---------------------|-----------|-----------|-----------|
                   3 |        17 |        17 |        34 |
                     |     0.500 |     0.500 |     0.008 |
                     |     0.007 |     0.012 |           |
                     |     0.004 |     0.004 |           |
---------------------|-----------|-----------|-----------|
        Column Total |      2611 |      1436 |      4047 |
                     |     0.645 |     0.355 |           |
---------------------|-----------|-----------|-----------|
```

Hagamos la misma prueba de hipótesis para determinar la relación de las variables.

```r
tabla_marriage <- table(data_credit$MARRIAGE, data_credit$default.payment.next.month)

# Realizar la prueba exacta de Fisher con simulación de Monte Carlo
set.seed(2024)
fisher_test_marriage <- fisher.test(tabla_marriage, simulate.p.value = TRUE, B = 10000)

print(fisher_test_marriage)
```

```
    Fisher's Exact Test for Count Data with simulated p-value (based on
```

```
    10000 replicates)

data:  tabla_marriage
p-value = 0.1634
alternative hypothesis: two.sided
```

En este caso, el valor de significancia que hemos estado utilizando es del 5%, es decir, un 0,05, como el p-value nos dió un valor de 0.1634, el p-value es mayor que el nivel de significancia, por lo que no hay evidencia estadística suficientes para rechazar la hipótesis nula, es decir, las diferencias que hemos encontrado en las categorías se pueden deber al azar. En conclusión, no podemos afirmar que el estado civil afecte al riesdo de impago.

Con esto terminamos la sección de las variables categóricas, procedemos entonces con las variables numéricas y la variable de impago.

3.1.3.2 Variables Numéricas con la Variable de Riesgo de Pago

Vamos a realizar una prueba de hipótesis entonces para las variables de LIMIT_BAL e Impago, para ello vamos a utilizar una prueba t no pareada, ya que estamos comparando poblaciones diferentes.

```r
t_test_result <- t.test(LIMIT_BAL ~ default.payment.next.month, data = data_credit)

print(t_test_result)
```

```
    Welch Two Sample t-test

data:  LIMIT_BAL by default.payment.next.month
t = 19.866, df = 3418.3, p-value < 2.2e-16
alternative hypothesis: true difference in means between group 0 and group 1 is not equal to
95 percent confidence interval:
 67451.30 82223.61
sample estimates:
mean in group 0 mean in group 1
       198249.7        123412.3
```

Este resultado no debería ser una sorpresa, pues gráficamente habíamos visto que a mayores límites de crédito, las personas tendían a incumplir menos que las personas que si lo hacían. Analicemos, el p-value tomó un valor demasiado pequeño en comparación con 0.05, por lo que hay evidencia estadística suficiente para rechazar la hipótesis nula. En conclusión, el límite de

crédito tiene relación con la probabilidad de impago, en general, las personas con mayor límite de crédito tienen una menor porbabilidad de caer en impago.

Haremos un análisis similar, pero esta vez cambiando el límite de crédito por la variable de edad.

```
t_test_result <- t.test(AGE ~ default.payment.next.month, data = data_credit)

print(t_test_result)
```

```
    Welch Two Sample t-test

data:  AGE by default.payment.next.month
t = -0.10677, df = 2734.6, p-value = 0.915
alternative hypothesis: true difference in means between group 0 and group 1 is not equal to
95 percent confidence interval:
 -0.6408875  0.5746956
sample estimates:
mean in group 0 mean in group 1
       36.51704        36.55014
```

De la prueba anterior podemos inferir que la edad no influye en el incumplimiento de los pagos, obtuvimos un p-value de 0.915 lo que es mucho mayor que nuestro nivel de significancia del 5%, así no hay evidencia estadística suficiente para rechazar la hipótesis nula, entonces no rechazamos la hipótesis nula. Concluímos que la edad no influye dado los datos en la probabilidad de caer en impago.

Con esto concluímos el análisis estadístico de la base de datos.

3.2 Parte de Planificación

3.2.1 Fichas literarias nuevas

Título: Metodología para un scoring de clientes sin referencias crediticias.

- Autor: Osvaldo Espin García, Carlos Rodríguez
- Año: 2013.
- Nombre del tema: Perfil de clientes crediticios.
- Cronología: 2011-2013.

- Metodología: Recolección y comparación de datos.
- Temática: Estudios económicos.
- Teórica: Valoración de riesgos.
- Resumen en una oración: El enfoque principal consiste en asignar un puntaje a cada solicitante, indicando la probabilidad de impago.
- Argumento central: Desarrollar un modelo predictivo que permita anticipar el comportamiento de impago de los clientes.
- Problema con el argumento o el tema: La mayor problemática que existe es que al basarse en la información personal de las personas, la existencia de personas con poca información en la base de datos o que acaban de iniciar en el sistema de manera que no tengan un historial de crédito, puede generar un sesgo en la muestra, lo cual llevar a que existan resultados menos precisos, lo cual termina por afectar el modelo.
- Resumen en un párrafo: Esta investigación gira en torno a desarrollar un modelo predictivo que permita de cierta manera lograr anticipar el comportamiento que pueda existir de impago en función de la información personal del cliente. De manera que las instituciones financieras como lo son los bancos o cooperativas puedan asignar cierto puntaje a cada solicitante donde se indique la probabilidad de impago.

3.2.2 Construcción de las Fichas de Resultados

```
if (!requireNamespace("kableExtra", quietly = TRUE)) {
  install.packages("kableExtra")
}
library(kableExtra)


data <- data.frame(
  Encabezado = c("Nombre de Su hallazgo",
                 "Resumen en una Oración",
                 "Problemas o Posibles Desafíos",
                 "Resumen en un párrafo"),
  Contenido = c("Correlación de las Variables",
                "Encontramos que algunas variables presentan una correlación alta entre sí po
                "Podría ser que la relación entre las variables en realidad no es lineal.",
                "Al utilizar el índice de correlación de Pearson, podemos observar que existe
)
```

```r
if (knitr::is_html_output()) {
  # Si es HTML
  kable(data, col.names = c("Encabezado", "Contenido"),
        format = "html",
        escape = FALSE) %>%
    kable_styling(full_width = FALSE) %>%
    add_header_above(c("Hallazgo de Resultado 1" = 2), bold = TRUE)
} else {
  # Si es PDF
  kable(data, col.names = c("Encabezado", "Contenido"),
        format = "latex",
        booktabs = TRUE) %>%
    kable_styling(latex_options = c("striped", "hold_position"))
}
```

Encabezado	Contenido
Nombre de Su hallazgo	Correlación de las Variables
Resumen en una Oración	Encontramos que algunas variables presentan una correlación alta entre sí pe
Problemas o Posibles Desafíos	Podría ser que la relación entre las variables en realidad no es lineal.
Resumen en un párrafo	Al utilizar el índice de correlación de Pearson, podemos observar que existe u

```r
if (!requireNamespace("kableExtra", quietly = TRUE)) {
  install.packages("kableExtra")
}
library(kableExtra)


data <- data.frame(
  Encabezado = c("Nombre de Su hallazgo",
                 "Resumen en una Oración",
                 "Problemas o Posibles Desafíos",
                 "Resumen en un párrafo"),
  Contenido = c("Relación entre la Densidad y la Cantidad de Crédito Otorgado",
                "Mientras más alta sea la cantidad de crédito otorgada, menor es la densidad
                "Esto no siempre sería cierto para cantidades pequeñas ó no podría seguir la
                "Al estimar la cantidad de personas que adquieren un crédito, la mayor canti
)


if (knitr::is_html_output()) {
  # Si es HTML
```

```
  kable(data, col.names = c("Encabezado", "Contenido"),
        format = "html",
        escape = FALSE) %>%
    kable_styling(full_width = FALSE) %>%
    add_header_above(c("Hallazgo de Resultado 2" = 2), bold = TRUE)
} else {
  # Si es PDF
  kable(data, col.names = c("Encabezado", "Contenido"),
        format = "latex",
        booktabs = TRUE) %>%
    kable_styling(latex_options = c("striped", "hold_position"))
}
```

Encabezado	Contenido
Nombre de Su hallazgo	Relación entre la Densidad y la Cantidad de Crédito Otorgado
Resumen en una Oración	Mientras más alta sea la cantidad de crédito otorgada, menor es la densidad
Problemas o Posibles Desafíos	Esto no siempre sería cierto para cantidades pequeñas ó no podría seguir la
Resumen en un párrafo	Al estimar la cantidad de personas que adquieren un crédito, la mayor canti

```
if (!requireNamespace("kableExtra", quietly = TRUE)) {
  install.packages("kableExtra")
}
library(kableExtra)


data <- data.frame(
  Encabezado = c("Nombre de Su hallazgo",
                 "Resumen en una Oración",
                 "Problemas o Posibles Desafíos",
                 "Resumen en un párrafo"),
  Contenido = c("Relación entre la Frecuencia de Crédito o con el Nivel educativo.",
                "Un mayor nivel educativo permite ó requiere una mayor cantidad de créditos.
                "No tener certeza de si el motivo del(los) crédito(s) es porque la persona pu
                "Al utilizar el histograma de edades por nivel educativo con la frecuencia de
)


if (knitr::is_html_output()) {
  # Si es HTML
  kable(data, col.names = c("Encabezado", "Contenido"),
        format = "html",
```

```
        escape = FALSE) %>%
    kable_styling(full_width = FALSE) %>%
    add_header_above(c("Hallazgo de Resultado 3" = 2), bold = TRUE)
} else {
  # Si es PDF
  kable(data, col.names = c("Encabezado", "Contenido"),
        format = "latex",
        booktabs = TRUE) %>%
    kable_styling(latex_options = c("striped", "hold_position"))
}
```

Encabezado	Contenido
Nombre de Su hallazgo	Relación entre la Frecuencia de Crédito o con el Nivel educativo.
Resumen en una Oración	Un mayor nivel educativo permite ó requiere una mayor cantidad de créditos
Problemas o Posibles Desafíos	No tener certeza de si el motivo del(los) crédito(s) es porque la persona pued
Resumen en un párrafo	Al utilizar el histograma de edades por nivel educativo con la frecuencia de c

```
if (!requireNamespace("kableExtra", quietly = TRUE)) {
  install.packages("kableExtra")
}
library(kableExtra)


data <- data.frame(
  Encabezado = c("Nombre de Su hallazgo",
                 "Resumen en una Oración",
                 "Problemas o Posibles Desafíos",
                 "Resumen en un párrafo"),
  Contenido = c("Relación entre el nivel educativo y los límites de crédito.",
                "Mientras más alto sea el nivel educativo, más alto tienden a ser los límites
                "No tener certeza de si el motivo del(los) crédito(s) es porque la persona pu
                "Al utilizar el gráfico de densidad por nivel educativo con el límite de cré
)


if (knitr::is_html_output()) {
  # Si es HTML
  kable(data, col.names = c("Encabezado", "Contenido"),
        format = "html",
        escape = FALSE) %>%
    kable_styling(full_width = FALSE) %>%
```

```
    add_header_above(c("Hallazgo de Resultado 4" = 2), bold = TRUE)
} else {
  # Si es PDF
  kable(data, col.names = c("Encabezado", "Contenido"),
        format = "latex",
        booktabs = TRUE) %>%
    kable_styling(latex_options = c("striped", "hold_position"))
}
```

Encabezado	Contenido
Nombre de Su hallazgo	Relación entre el nivel educativo y los límites de crédito.
Resumen en una Oración	Mientras más alto sea el nivel educativo, más alto tienden a ser los límites de
Problemas o Posibles Desafíos	No tener certeza de si el motivo del(los) crédito(s) es porque la persona pued
Resumen en un párrafo	Al utilizar el gráfico de densidad por nivel educativo con el límite de crédito,

```
if (!requireNamespace("kableExtra", quietly = TRUE)) {
  install.packages("kableExtra")
}
library(kableExtra)


data <- data.frame(
  Encabezado = c("Nombre de Su hallazgo",
                 "Resumen en una Oración",
                 "Problemas o Posibles Desafíos",
                 "Resumen en un párrafo"),
  Contenido = c("Relación de la variable de impago con los límites de crédito",
                "Un mejor historial crediticio, es decir, menor probabilidad de impago; permi
                "No necesariamente puede generalizarse a todas las bases de datos",
                "Al utilizar el histograma de límite de crédito por la variable de impago, pu
)


if (knitr::is_html_output()) {
  # Si es HTML
  kable(data, col.names = c("Encabezado", "Contenido"),
        format = "html",
        escape = FALSE) %>%
    kable_styling(full_width = FALSE) %>%
    add_header_above(c("Hallazgo de Resultado 5" = 2), bold = TRUE)
} else {
```

```
  # Si es PDF
  kable(data, col.names = c("Encabezado", "Contenido"),
        format = "latex",
        booktabs = TRUE) %>%
    kable_styling(latex_options = c("striped", "hold_position"))
}
```

Encabezado	Contenido
Nombre de Su hallazgo	Relación de la variable de impago con los límites de crédito
Resumen en una Oración	Un mejor historial crediticio, es decir, menor probabilidad de impago; permit
Problemas o Posibles Desafíos	No necesariamente puede generalizarse a todas las bases de datos
Resumen en un párrafo	Al utilizar el histograma de límite de crédito por la variable de impago, pudi

```
if (!requireNamespace("kableExtra", quietly = TRUE)) {
  install.packages("kableExtra")
}
library(kableExtra)


data <- data.frame(
  Encabezado = c("Nombre de Su hallazgo",
                 "Resumen en una Oración",
                 "Problemas o Posibles Desafíos",
                 "Resumen en un párrafo"),
  Contenido = c("Relación entre el sexo y la variable de impago.",
                "Las mujeres tienen más frecuencia de impago que los hombres.",
                "Esto no se puede generalizar a las demás bases de datos ni países del resto
                "Al analizar la frecuecia de impago por sexo, podemos ver que la cantidad de
)


if (knitr::is_html_output()) {
  # Si es HTML
  kable(data, col.names = c("Encabezado", "Contenido"),
        format = "html",
        escape = FALSE) %>%
    kable_styling(full_width = FALSE) %>%
    add_header_above(c("Hallazgo de Resultado 6" = 2), bold = TRUE)
} else {
  # Si es PDF
  kable(data, col.names = c("Encabezado", "Contenido"),
```

```
        format = "latex",
        booktabs = TRUE) %>%
    kable_styling(latex_options = c("striped", "hold_position"))
}
```

Encabezado	Contenido
Nombre de Su hallazgo	Relación entre el sexo y la variable de impago.
Resumen en una Oración	Las mujeres tienen más frecuencia de impago que los hombres.
Problemas o Posibles Desafíos	Esto no se puede generalizar a las demás bases de datos ni países del resto d
Resumen en un párrafo	Al analizar la frecuecia de impago por sexo, podemos ver que la cantidad de

3.2.3 Construcción de la UVE de Gowin Modificada

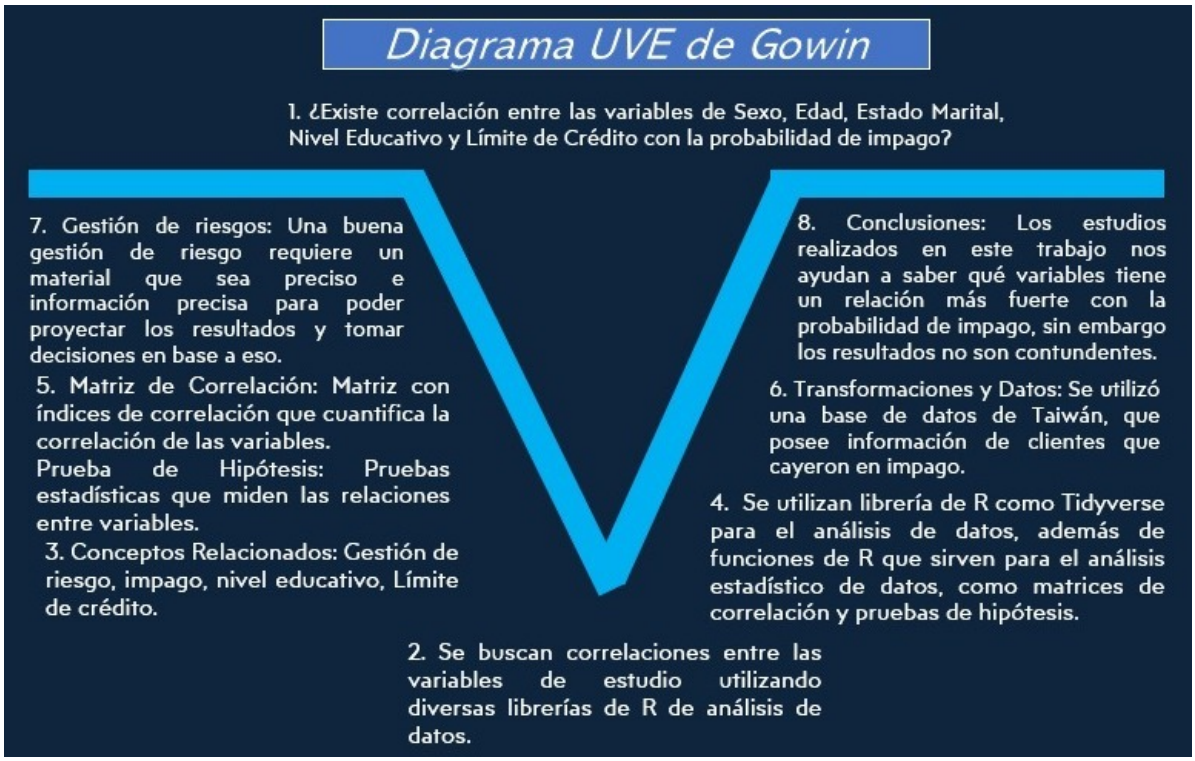


Figura 3.2: UVE de Gowin Nueva

3.3 Parte de Escritura

3.3.1 Escribir, escribir, escribir

3.3.1.1 La valoración del riesgo en el sector financiero

El problema que se va a tratar en el presente trabajo es de determinar la relación existente entre las variables de Monto de un crédito, nivel de educación y situaciones de pago con la calificación de riesgo. Desde el punto de vista teórico, el autor (Palacios 2012), menciona que "La principal función que radica en las calificaciones crediticias es la evaluación de la mayor o menor probabilidad de pago de la deuda y los intereses, proporcionando indicadores que sirvan de referencia a los inversores con el fin de que puedan tener conocimiento del riesgo crediticio de una forma simple y accesible". Desde este punto de vista, hay un apoyo en la investigación que tratamos de realizar, pues la calificación de riesgo es de suma importancia en el mundo financiero. Este mismo autor menciona que "Su importancia deriva de su implantación dentro de la regulación, lo que afecta a todo el entramado institucional, y sectores clave de la sociedad como son el bancario y las agencias de seguros y reaseguros".

Inicialmente mencionamos que la teoría respaldaba nuestra conjetura pues no había un análisis de datos que afirmara lo antes explicado. Después del cambio de base de datos y un nuevo análisis exploratorio y estadístico, podemos observar que realmente sí hay una relación como se planteaba en las Bitácoras 1 y 2. Si bien una variable no es determinante para las demás, si guardan una relación tal y como se pueden observar en las fichas de resultados.

El estudio del análisis financiero es de suma importancia en la actualidad, ya que las transacciones de los flujos de dinero cada vez son mayores, es decir, vender deuda para obtener financiamiento en el corto plazo es una de las estrategias más aplicadas. Por ello tanto inversores como prestatarios, según (Palacios 2012), "hacen uso de las calificaciones crediticias como un indicador de la probabilidad de recuperar su dinero. Adicionalmente, los prestatarios pueden beneficiarse de tener calificada su deuda, con el objetivo de colocarla con mayor facilidad y eliminar las dudas que haya relación a ellos." Asimismo, ambas parten obtienen beneficio de que exista este rating en el mundo de la información financiera. Y desde el punto de vista del inversor, como menciona la autora (Chicu 2020), "...a la hora de analizar una inversión, debemos valorar la rentabilidad esperada, así como la liquidez que perdemos y el riesgo que estamos dispuestos a asumir". Por lo tanto, poseer la información de rating es de suma utilidad, pues ayuda a los inversores a realizar mejores proyecciones.

Incluso podríamos pensar que, según los resultados del análisis, mientras mejor calificación tenga una persona en su historial de crédito, las entidades tienden a ofrecer mejores condiciones de préstamo e inversión. Por eso, según los diferentes histogramas y gráficos de densidad realizados, podemos ver que algunos créditos tienen un monto alto y tener la variable de impago inactiva.

En adición, haciendo referencia a esta misma autora "…la gestión de riesgos tiene un lugar cada vez que un inversor analiza e intenta cuantificar las pérdidas potenciales en una inversión y luego toma las medidas apropiadas, considerando sus objetivos de inversión y su tolerancia al riesgo." Esto último viene de la mano con lo que son las proyecciones, pues le permite al inversionista hacer un mejor análisis y una gestión de riesgos adecuada, que podemos definir según Chicu como "El proceso de identificación, análisis e incorporación de la incertidumbre en las decisiones de inversión" (Chicu 2020). Reforzando lo que menciona Chicu, la autora (Maria de los Ángeles Herrera 2024), menciona en adición a la gestión de riesgos "el contexto de incertidumbre genera inevitablemente un riesgo, y es ahí cuando la institución financiera debe preservar su valor económico y la integridad de los recursos confiados por los depositantes y socios." Y añadiendo la definición de esta misma autora tenemos que la gestión de riesgos es "la denominación que se utiliza para el conjunto de técnicas y herramientas que pretenden maximizar el valor económico de la institución financiera, en un contexto de incertidumbre". Concluyendo, la gestión de riesgos depende íntimamente de la calificación de riesgo, pues permite tener un parámetro ante la incertidumbre que representa invertir. Con esto, podemos proceder a explicar algunas de estas herramientas para el análisis de las variables en este estudio.

3.3.1.2 Prueba Exacta de Fisher para Datos de Conteo

Según la documentación de R y las t'ecnicas de pruebas de hip'otesis vistas en el libro de (Pértega Díaz y Pita Fernández 2004), esta es una prueba para variables categóricas donde hemos usado tablas de contingencias, de ahí nuestra escogencia de esta prueba. Si bien el método está pensando para que se realice a pruebas 2x2, donde la prueba chi-cuadrado falla, hemos decidio utilizar la de fisher para tablas más grandes, usando simulaciones.

En nuestro caso no hace falta utilizar simulaciones cuando hacemos la prueba para las variables de sexo y riesgo de pago, ya que hacen una tabla 2x2, entonces el cálculo da exacto, pues es para el tipo de tablas para las que fue creado el método. Sin embargo, cuando utilizamos la prueba en las tablas de Educación con Impago y de Estado Civil con Impago, hay que utilizar valores simulados, pues estas tablas no cumplen ser 2x2.

Para el caso de 2x2, el cálculo es exacto y para calcular el p-value se utiliza una distribución hipergeométrica, utilizando las observaciones de la tabla de contigencias.

$$p = \frac{(a+b)! \times (c+d)! \times (a+c)! \times (b+d)!}{n!a!b!c!d!}$$

La fórmula anterior es la manera en cómo se calcula para una tabla 2x2. Para tablas que no son de esta forma no tenemos la fórmula de manera explícita, pues es una función de R, donde indica que se simula varias veces la tabla, a saber un valor b, donde entre más grande sea este b mejor aproximación del p valor vamos a tener.

3.3.1.3 Prueba T

La siguiente prueba tomamos con referencia el método del libro de Probabilidad y Estadística de (Walpole et al. 1999). Esta prueba la utilizamos cuando queremos comparar la media de dos poblaciones, digamos que se utiliza cuando queremos ver la relación entre variables numéricas con variables categóricas, ya que dividimos entre las poblaciones, que viene siendo las poblaciones y las medias salen de la variable numérica.

De maneta teórica, el método para calcularlo es el siguiente:

Primero el estadístico a utilizar, el cual es:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Este estadístico sigue una distribución t, es decir una distribución sin parámetros, lo cual es de mucha utilidad.

Posteriormente, para calcular el p-value se utiliza la siguiente fórmula:

$$p\text{-valor} = 2 \times P(T > |t|)$$

Así entonces hemos explicado las dos pruebas de hipótesis utilizadas en nuestro trabajo. Cabe recalcar que las funciones de R aplican otros procedimientos para ajustar los valores.

3.3.2 Conclusion

Este trabajo tuvo como objetivo explorar la correlación entre diversas características de los perfiles de los clientes y su relación con la probabilidad de impago en el contexto de un banco de Taiwán. Para ello, se aplicaron técnicas estadísticas y funciones en R para abordar la pregunta de investigación: "¿Existe correlación entre las variables de Sexo, Edad, Estado Marital, Nivel Educativo y Límite de Crédito con la probabilidad de impago?". La motivación detrás de esta pregunta radica en la importancia de identificar características que permitan a los bancos y entidades financieras perfilar a clientes con mayor probabilidad de caer en impago, dado que muchas de estas instituciones basan su negocio en el otorgamiento de créditos.

El análisis buscó identificar patrones entre los clientes que caen en impago y determinar si tales patrones eran atribuibles al azar o si había una correlación estadísticamente significativa. Los resultados mostraron que, en el caso de las variables numéricas, solo la edad no presentaba relación con la probabilidad de impago, mientras que en las variables categóricas, el estado marital también resultó no estar relacionado. Esto indica que dichas variables no representan un factor significativo en el riesgo de crédito o probabilidad de impago.

Por otro lado, los hallazgos coinciden en gran medida con nuestras expectativas teóricas, dado que variables como el nivel educativo, el sexo y el límite de crédito sí influyen en la probabilidad de impago. Una limitación de este estudio es que no contamos con una metodología robusta de cálculo de probabilidades de impago o de evaluación del riesgo de crédito, ya que estas metodologías son conservadas de forma confidencial por las empresas y varían entre instituciones. No obstante, logramos responder a nuestra pregunta de investigación identificando correlaciones relevantes y descartando variables con menor impacto a través de pruebas de hipótesis.

Para futuros estudios, recomendamos incluir variables adicionales, tales como el motivo del préstamo, los años del crédito, el salario actual del solicitante y su historial crediticio previo, ya que estos factores podrían enriquecer los resultados y permitir una caracterización más precisa de los perfiles de riesgo.

Este trabajo resulta útil para instituciones financieras y casas de préstamo en sus etapas iniciales de operación, ya que pueden apoyarse en estos resultados para desarrollar un perfil de cliente con mayor probabilidad de pago. Al observar que variables como el sexo, el nivel educativo y el límite de crédito influyen en el riesgo de impago, estas entidades pueden empezar a construir criterios de préstamo más sólidos y orientados a la disminución del riesgo.

3.3.3 Introducción

Este estudio se enfoca en desarrollar un estudio en el cual se avale si variables como lo son la edad, el género, el estado marital y el nivel de educación, tienen algún peso o relevancia a la hora de medir el riesgo financiero de los prestatarios. De manera que lo que se busca es identificar patrones y con ello obtener una visión mas completa que no se limite únicamente a las características financieras de las personas, como lo es su nivel de ingresos, sino que también considera el contexto y características propias en las que se desenvuelve cada individuo.

Por ello, es que este estudio comparar variables cuantitativas con variables cualitativas y tratar de ver la relación o correlación que pueda existir entre ellas, ya que al final todo ello se relaciona intrínsecamente con aspectos propios del prestatario, lo cual a su vez ayuda a generar una perspectiva mas robusta del riesgo que el mismo puede generar para alguna entidad financiera llamase banco u otra.

Por lo cual, la presente investigación tiene como finalidad lograr determinar si variables cualitativas tienen una verdadera relevancia a la hora de estimar el riesgo de impago de las personas, lo cual ha su vez llega a ser pertinente en lo que es la carrera de ciencias actuariales, por lo que se pudieron usar herramientas que irán viendo a lo largo de los siguientes cursos, pero que llegan a ser muy útiles para el propósito de este trabajo.

De manera que el objetivo principal de esta investigación es determinar si existe una correlación positiva entre algunas variables cualitativas y factores cuantitativos que se utilizan para

determinar el riesgo de impago, esto a través del uso de una base de datos de un banco de Taiwán la cual brinda suficiente información para nuestro propósito.

Por lo cual, esta investigación se fundamenta en un marco teórico y empírico con el que se busca lograr determinar a partir de varios métodos y pruebas la existencia o no de correlación entre nuestras variables. A su vez, el marco teórico que se maneja para esta investigación parece muy pertinente, esto debido a que los autores desean ir más allá de los modelos tradicionales y explorar factores adicionales al contexto económico de los prestatarios, lo cual da una mayor perspectiva que permite deslumbrar elementos de riesgo que podrían llegar a pasar desapercibidos en estudios un poco más convencionales.

3.3.4 Resumen

Este estudio busca determinar si variables cualitativas como lo son la edad, el género, el estado marital y el nivel educativo influyen en el riesgo financiero de los prestatarios, explorando su relación con variables cuantitativas, como lo es el límite de crédito. Con ello, se pretende identificar patrones que contribuyan a una evaluación integral del riesgo de impago. Para este fin, se utilizó una base de datos de clientes de un banco en Taiwán del año 2005. De manera que los datos se analizaron en R, aplicando representaciones gráficas, matrices de correlación para variables numéricas y tablas de contingencia para variables categóricas. A su vez, se utilizaron pruebas de hipótesis, como lo es la prueba T, la cual permitió evaluar la significancia de las diferencias entre niveles de variables cualitativas. De esta forma, se pudo revelar que el nivel educativo, el sexo y el límite de crédito tienen una correlación significativa con el riesgo de impago. Estos hallazgos indican que ciertas características no financieras sí influyen en la probabilidad de incumplimiento. En conclusión, este estudio aporta ciertas percepciones para entidades financieras que buscan perfilar clientes que tengan mayor probabilidad de pago. Para ello, se tiene como resultado fundamental del que estudio que variables como el sexo, el nivel educativo y el límite de crédito son útiles en la predicción del riesgo de impago, sugiriendo criterios adicionales para lo que es el análisis de crédito.

3.3.5 Ordenamiento Final

3.3.5.1 Titulo

Análisis de variables cualitativas en relación al riesgo crediticio

3.3.5.2 Resumen

Este estudio busca determinar si variables cualitativas como lo son la edad, el género, el estado marital y el nivel educativo influyen en el riesgo financiero de los prestatarios, explorando su relación con variables cuantitativas, como lo es el límite de crédito. Con ello, se pretende

identificar patrones que contribuyan a una evaluación integral del riesgo de impago. Para este fin, se utilizó una base de datos de clientes de un banco en Taiwán del año 2005. De manera que los datos se analizaron en R, aplicando representaciones gráficas, matrices de correlación para variables numéricas y tablas de contingencia para variables categóricas. A su vez, se utilizaron pruebas de hipótesis, como lo es la prueba T, la cual permitió evaluar la significancia de las diferencias entre niveles de variables cualitativas. De esta forma, se pudo revelar que el nivel educativo, el sexo y el límite de crédito tienen una correlación significativa con el riesgo de impago. Estos hallazgos indican que ciertas características no financieras sí influyen en la probabilidad de incumplimiento. En conclusión, este estudio aporta ciertas percepciones para entidades financieras que buscan perfilar clientes que tengan mayor probabilidad de pago. Para ello, se tiene como resultado fundamental del que estudio que variables como el sexo, el nivel educativo y el límite de crédito son útiles en la predicción del riesgo de impago, sugiriendo criterios adicionales para lo que es el análisis de crédito.

3.3.5.3 Palabras Clave

- Nivel de educacion
- Edad
- Estado civil
- Monto de credito
- Limite de Credito
- Riesgo de impago

3.3.5.4 Introduccion

Este estudio se enfoca en desarrollar un estudio en el cual se avale si variables como lo son la edad, el género, el estado marital y el nivel de educación, tienen algún peso o relevancia a la hora de medir el riesgo financiero de los prestatarios. De manera que lo que se busca es identificar patrones y con ello obtener una visión mas completa que no se limite únicamente a las características financieras de las personas, como lo es su nivel de ingresos, sino que también considera el contexto y características propias en las que se desenvuelve cada individuo.

Por ello, es que este estudio comparar variables cuantitativas con variables cualitativas y tratar de ver la relación o correlación que pueda existir entre ellas, ya que al final todo ello se relaciona intrínsecamente con aspectos propios del prestatario, lo cual a su vez ayuda a generar una perspectiva mas robusta del riesgo que el mismo puede generar para alguna entidad financiera llamase banco u otra.

Por lo cual, la presente investigación tiene como finalidad lograr determinar si variables cualitativas tienen una verdadera relevancia a la hora de estimar el riesgo de impago de las personas, lo cual ha su vez llega a ser pertinente en lo que es la carrera de ciencias actuariales, por lo

que se pudieron usar herramientas que irán viendo a lo largo de los siguientes cursos, pero que llegan a ser muy útiles para el propósito de este trabajo.

De manera que el objetivo principal de esta investigación es determinar si existe una correlación positiva entre algunas variables cualitativas y factores cuantitativos que se utilizan para determinar el riesgo de impago, esto a través del uso de una base de datos de un banco de Taiwán la cual brinda suficiente información para nuestro propósito.

Por lo cual, esta investigación se fundamenta en un marco teórico y empírico con el que se busca lograr determinar a partir de varios métodos y pruebas la existencia o no de correlación entre nuestras variables. A su vez, el marco teórico que se maneja para esta investigación parece muy pertinente, esto debido a que los autores desean ir más allá de los modelos tradicionales y explorar factores adicionales al contexto económico de los prestatarios, lo cual da una mayor perspectiva que permite deslumbrar elementos de riesgo que podrían llegar a pasar desapercibidos en estudios un poco más convencionales.

3.3.5.5 Metodología

El presente trabajo es un estudio correlacional, cuyo objetivo es identificar las relaciones entre las variables seleccionadas en la base de datos de un banco de Taiwán correspondiente al año 2005, la cual contiene información sobre el riesgo crediticio de los clientes. En particular, el interés radica en determinar si existen relaciones significativas entre ciertas variables (edad, estado civil, nivel educativo y límite de crédito) y la probabilidad de caer en impago.

Mediante esta metodología, buscamos analizar la correlación entre estas variables y la probabilidad de incumplimiento en el pago. La muestra utilizada es la base de datos mencionada, proveniente de un banco en Taiwán, que proporciona datos relevantes para este análisis de riesgo.

Para el análisis de los datos, se empleará el lenguaje de programación R, dada su facilidad de uso y compatibilidad con herramientas de análisis estadístico avanzadas. El procedimiento comienza con una limpieza de la base de datos, un paso esencial ya que las bases de datos suelen contener información que no aporta al análisis estadístico; por lo tanto, es necesario asegurar que los datos estén en una forma óptima para el análisis. Una vez limpia, se procede a un análisis gráfico, generando representaciones visuales de las variables tanto de manera individual como conjunta, con el fin de observar cómo se relacionan gráficamente y formular algunas hipótesis preliminares.

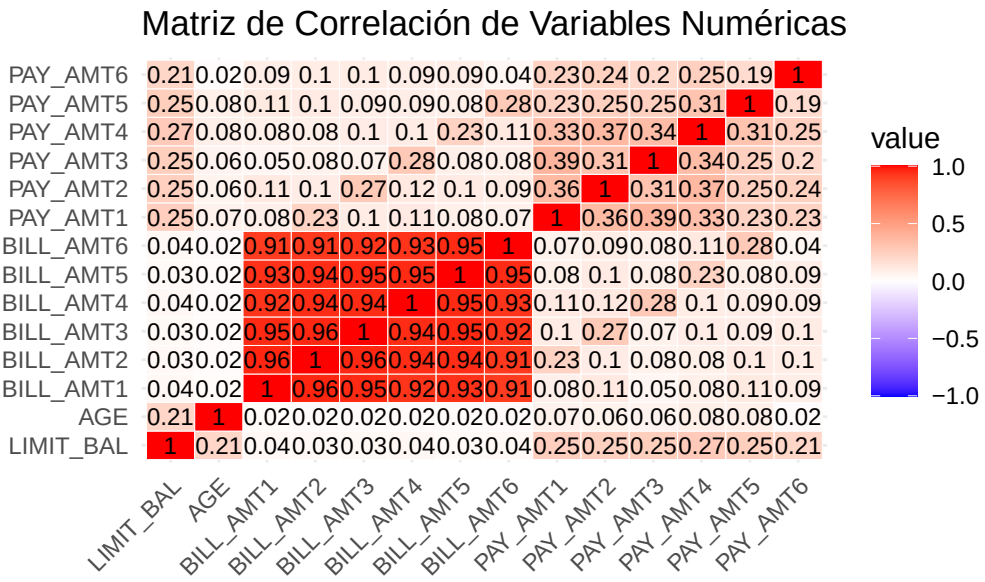
Posteriormente, se aplicará una matriz de correlación para identificar relaciones entre las variables numéricas. Para las variables categóricas, se utilizarán tablas de contingencia y, en caso de tablas mayores que 2x2, el método exacto de Fisher con simulación para obtener resultados precisos. Además, para explorar la dependencia entre variables numéricas y categóricas, se llevarán a cabo pruebas de hipótesis como la prueba t, evaluando si existen diferencias significativas en las medias de los distintos niveles de las variables categóricas.

Finalmente, se realizará un análisis bibliográfico para contrastar los hallazgos empíricos con la teoría existente sobre el tema. Este análisis nos permitirá evaluar si nuestros resultados se alinean con lo documentado previamente en la literatura y aportar un criterio fundamentado sobre las relaciones encontradas en la base de datos.

3.3.5.6 Resultados

Correlación de las variables

```
library(dplyr)
library(ggplot2)
library(reshape2)


# Crear la matriz de correlación
matriz_correlacion <- cor(data_credit_numerico, use = "complete.obs")

# Creamos la variable para ver la matriz de correlación como un gráfico de calor
base_matriz <- melt(matriz_correlacion)

# Hacemos el plot de la base
ggplot(base_matriz, aes(Var1, Var2, fill = value)) +
  geom_tile(color = "white") +
  scale_fill_gradient2(low = "blue", high = "red", mid = "white", midpoint = 0, limit = c(-1
  theme_minimal() + geom_text(aes(label = round(value, 2)), color = "black", size = 3) +
  labs(title = "Matriz de Correlación de Variables Numéricas",
       x = "", y = "") + labs(caption = "Fuente: Elaboración propia utilizando la base de dat
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Fuente: Elaboración propia utilizando la base de datos de Kaggle

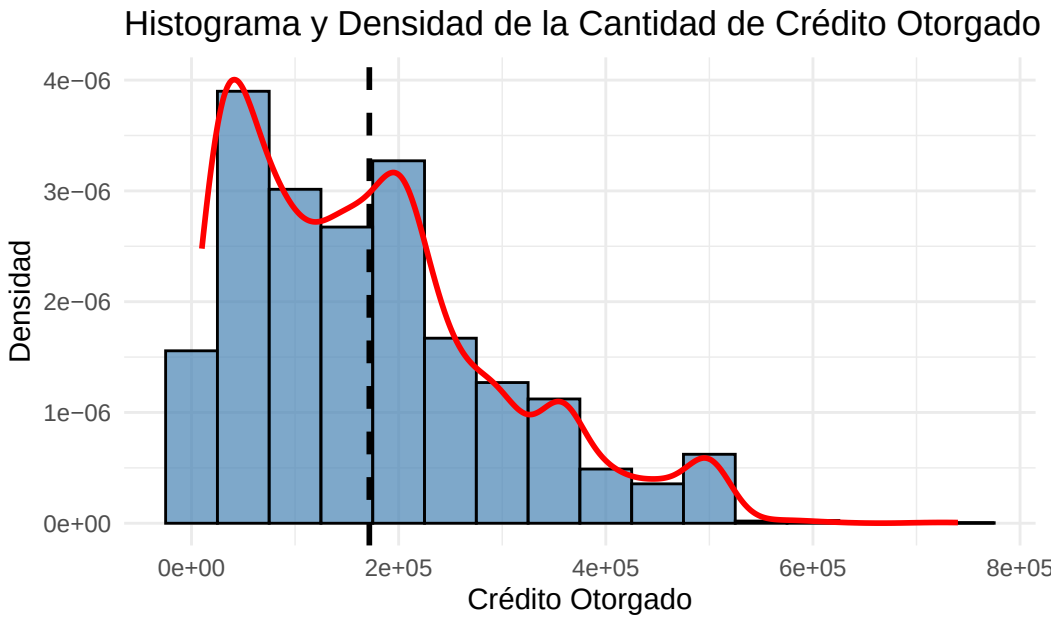
Figura 3.3: Figura 1: Matriz de Correlación

Al utilizar el índice de correlación de Pearson, se puede observar la correlacion que existe entre las distintas variables, más en especifico entre las situaciones de reembolso y entre los estados de cuenta. Además, se observa a simple vista como la edad podría llegar a tener algún tipo de relación con el monto del préstamo.

Relación entre la Densidad y la Cantidad de Crédito Otorgado

```
library(ggplot2)
library(dplyr)

data_credit %>%
  ggplot(aes(x = LIMIT_BAL)) +
  geom_histogram(aes(y = after_stat(density)), binwidth = 50000, fill = "steelblue", color =
  geom_density(color = "red", size = 1) + # Densidad en línea roja
  geom_vline(aes(xintercept = mean(LIMIT_BAL)), linetype = "dashed", color = "black", size =
  labs(title = "Histograma y Densidad de la Cantidad de Crédito Otorgado",
       x = "Crédito Otorgado",
       y = "Densidad") + labs(caption = "Fuente: Elaboración propia utilizando la base de dat
  theme_minimal()
```

```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
i Please use `linewidth` instead.
```



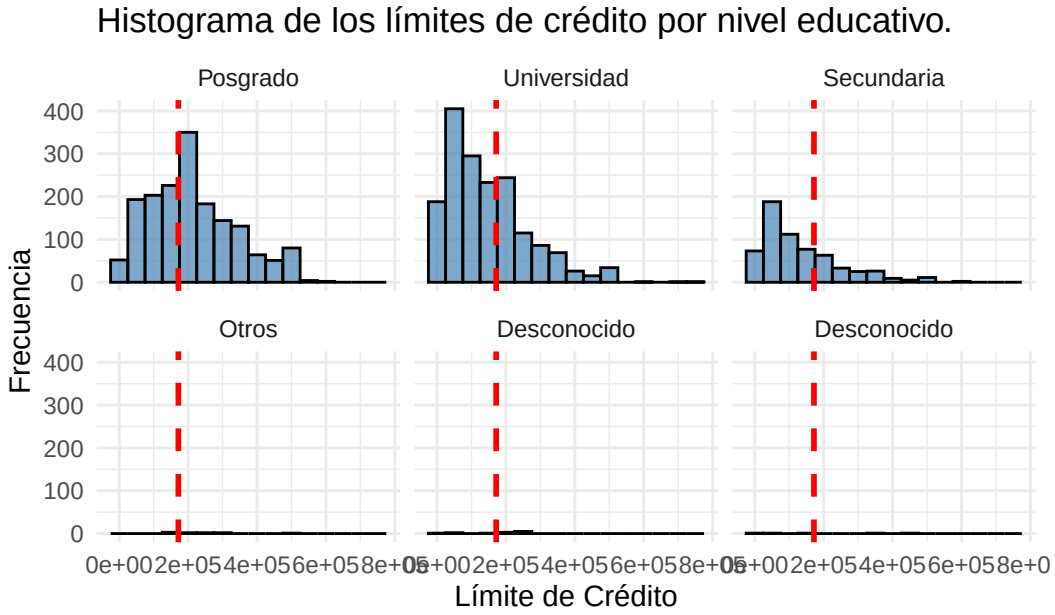
Fuente: Elaboración propia utilizando la base de datos de Kaggle

De lo anterior se entiende que al momento de estimar la cantidad de personas que llegan a adquirir un crédito, en su gran mayoria lo hacen por un monto relativamente pequeño, en comparacion con la cantidad de personas que adequieren un credito de mayor valor. Lo cual puede deberse a las necesidades de las personas ó al propio riesgo que viene intrinseco al adquirir un crédito de mayor valor.

Relación entre la Frecuencia de impago con el nivel educativo.

```
library(ggplot2)
library(dplyr)

data_credit %>%
  ggplot(aes(x = LIMIT_BAL)) +
  geom_histogram(binwidth = 50000, fill = "steelblue", color = "black", alpha = 0.7) +
  geom_vline(aes(xintercept = mean(LIMIT_BAL, na.rm = TRUE)),
             color = "red",
             linetype = "dashed",
             size = 1) +
  labs(title = "Histograma de los límites de crédito por nivel educativo.",
       x = "Límite de Crédito",
       y = "Frecuencia") +
  facet_wrap(~EDUCATION, labeller = labeller(EDUCATION = c(
    "1" = "Posgrado",
    "2" = "Universidad",
```

```
    "3" = "Secundaria",
    "4" = "Otros",
    "5" = "Desconocido",
    "6" = "Desconocido"
  ))) + labs(caption = "Fuente: Elaboración propia utilizando la base de datos de Kaggle") +
```



Fuente: Elaboración propia utilizando la base de datos de Kaggle

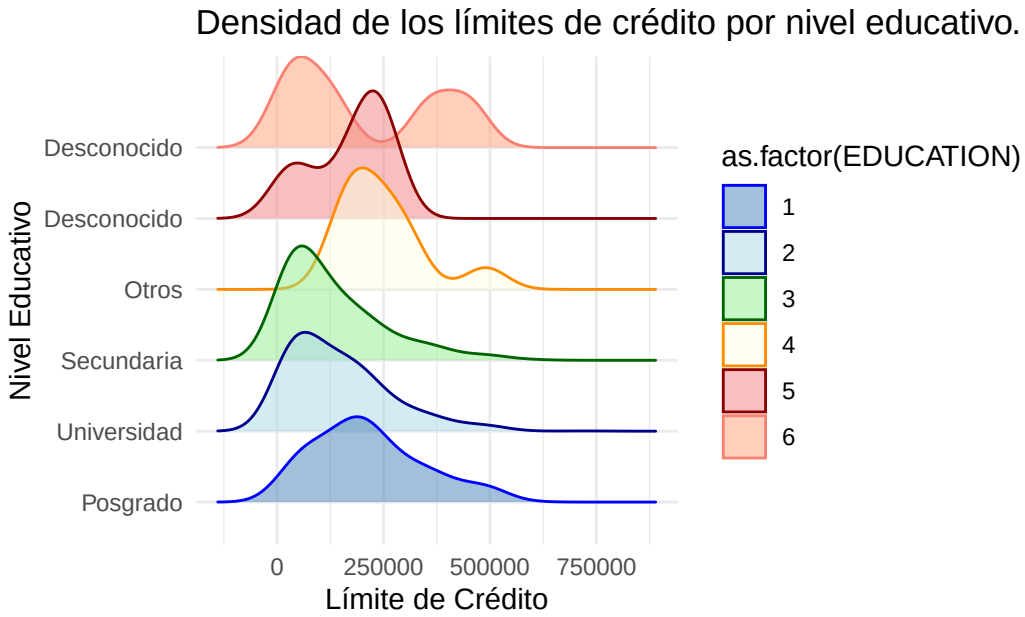
Al utilizar el histograma de edades por nivel educativo con la frecuencia de créditos, se puede observar que las personas con posgrados tienen una mayor frecuencia de caer en impago que las personas que no lo tienen. Esto puede ser dado que las personas con posgrado necesiten un crédito para pagar sus estudios de posgrado (los cuales generalmente son costosos) ó que puedan permitirse un crédito dados sus ingresos.

Relación entre el nivel educativo y los límites de crédito.

```
library(ggplot2)
library(dplyr)
library(ggridges)

data_credit %>%
  ggplot(aes(x = LIMIT_BAL, y = as.factor(EDUCATION), fill = as.factor(EDUCATION), color = as
  geom_density_ridges(alpha = 0.5) +
  labs(title = "Densidad de los límites de crédito por nivel educativo.",
       x = "Límite de Crédito",
       y = "Nivel Educativo") +
```

```
  scale_fill_manual(values = c("1" = "steelblue", "2" = "lightblue", "3" = "lightgreen",
                               "4" = "lightyellow", "5" = "lightcoral", "6" = "lightsalmon")) +
  scale_color_manual(values = c("1" = "blue", "2" = "darkblue", "3" = "darkgreen",
                                "4" = "darkorange", "5" = "darkred", "6" = "salmon")) +
  scale_y_discrete(labels = c(
    "1" = "Posgrado",
    "2" = "Universidad",
    "3" = "Secundaria",
    "4" = "Otros",
    "5" = "Desconocido",
    "6" = "Desconocido"
  )) +  labs(caption = "Fuente: Elaboración propia utilizando la base de datos de Kaggle") +
  theme_minimal()
```

```
Picking joint bandwidth of 49800
```



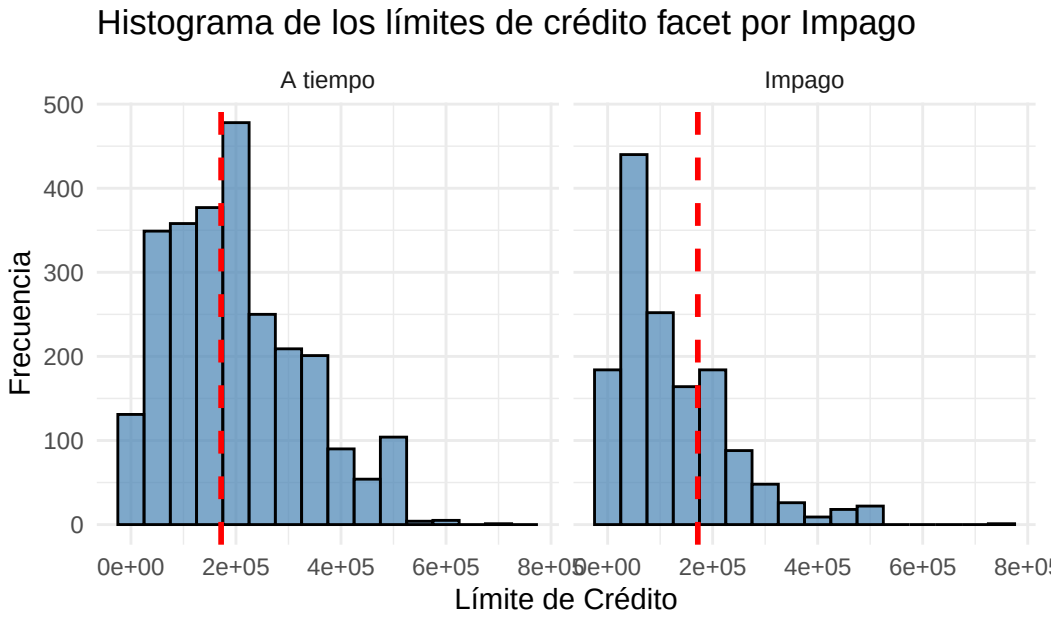
Fuente: Elaboración propia utilizando la base de datos de Kaggle

Al utilizar el gráfico de densidad por nivel educativo con el límite de crédito, podemos observar que las personas con un nivel de Posgrado tienen mayores niveles de crédito. Sin embargo, esto puede ser dado que las personas con mayores niveles educativos puedan permitirse los niveles de crédito dado su ingreso ó más bien que las personas necesiten ese crédito para pagar sus estudios de posgrado. Estudios que generalmente son muy costosos.

Ademas, las densidades que m'as llaman nuestra atención son las de secundaria, universidad y posgrado, esto por el hecho de que ellas son las que tienen la mayor concentración de datos.Además, veamos que las densidades tienen más o menos una distribución exponencial.

Relación de la variable de impago con los límites de crédito

```
library(ggplot2)
library(dplyr)

data_credit %>%
  ggplot(aes(x = LIMIT_BAL)) +
  geom_histogram(binwidth = 50000, fill = "steelblue", color = "black", alpha = 0.7) +
  geom_vline(aes(xintercept = mean(LIMIT_BAL, na.rm = TRUE)),
             color = "red",
             linetype = "dashed",
             size = 1) +
  labs(title = "Histograma de los límites de crédito facet por Impago",
       x = "Límite de Crédito",
       y = "Frecuencia") +
  facet_wrap(~default.payment.next.month, labeller = as_labeller(c("0" = "A tiempo", "1" = "I
  theme_minimal()
```

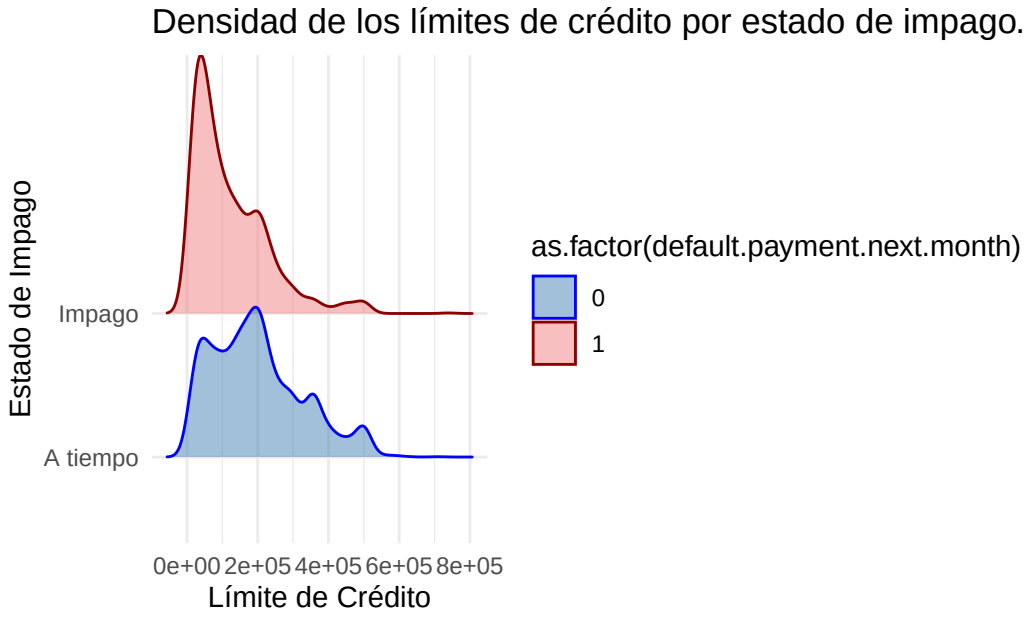


Fuente: Elaboración propia utilizando la base de datos de Kaggle

Agregamos dicho gráfico de densidades para las anteriores variables.

```r
library(ggplot2)
library(dplyr)
library(ggridges)

data_credit %>%
  ggplot(aes(x = LIMIT_BAL, y = as.factor(default.payment.next.month), fill = as.factor(defau
  geom_density_ridges(alpha = 0.5) +
  labs(title = "Densidad de los límites de crédito por estado de impago.",
       x = "Límite de Crédito",
       y = "Estado de Impago") +
  scale_fill_manual(values = c("0" = "steelblue", "1" = "lightcoral")) +
  scale_color_manual(values = c("0" = "blue", "1" = "darkred")) +
  scale_y_discrete(labels = c(
    "0" = "A tiempo",
    "1" = "Impago"
  )) +  labs(caption = "Fuente: Elaboración propia utilizando la base de datos de Kaggle") +
  theme_minimal()
```

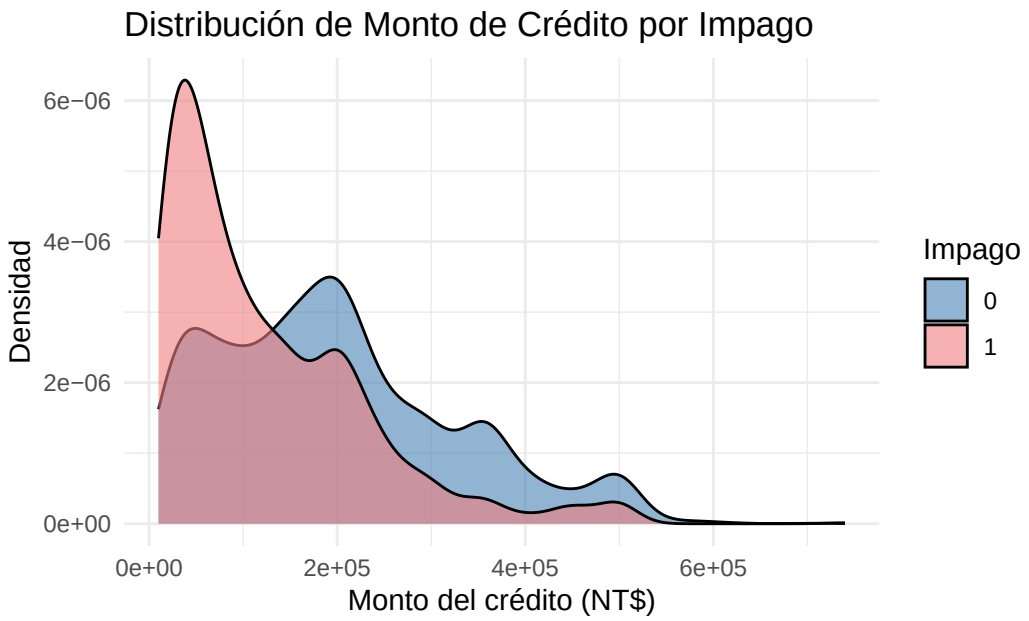
```
Picking joint bandwidth of 22100
```



n propia utilizando la base de datos de Kaggle

Ploteando los gráficos juntos, para una mejor visualización tenemos que:

```r
ggplot(data_credit, aes(x = LIMIT_BAL, fill = as.factor(default.payment.next.month))) +
  geom_density(alpha = 0.6) +
  labs(
    x = "Monto del crédito (NT$)",
    y = "Densidad",
    title = "Distribución de Monto de Crédito por Impago",
    fill = "Impago"
  ) +
  scale_fill_manual(values = c("0" = "steelblue", "1" = "lightcoral")) +
 labs(caption = "Fuente: Elaboración propia utilizando la base de datos de Kaggle") + theme_
```



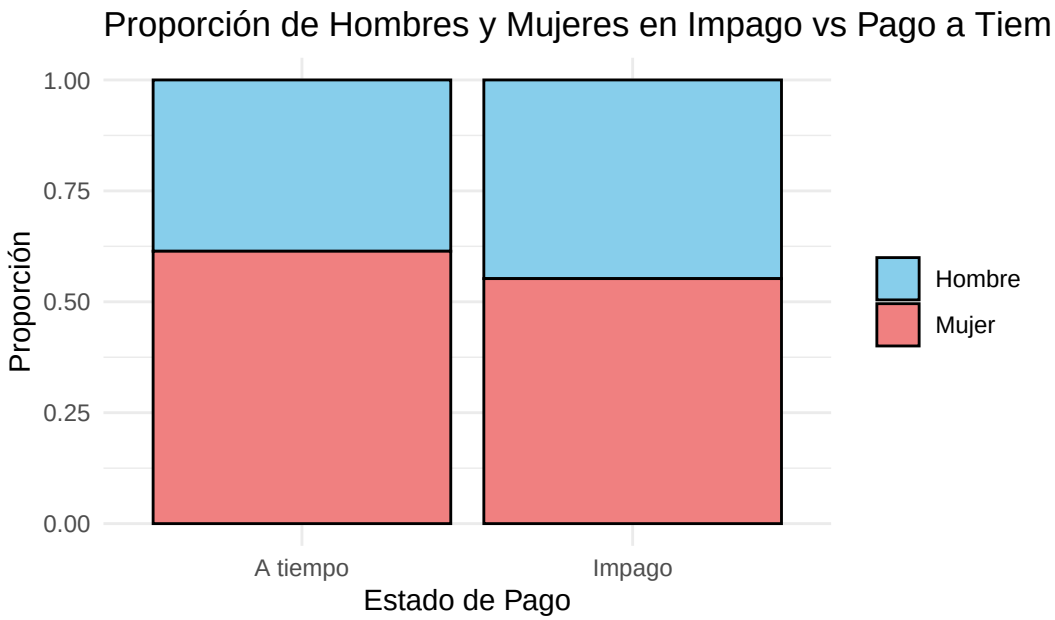
Fuente: Elaboración propia utilizando la base de datos de Kaggle

Al utilizar el histograma de límite de crédito por la variable de impago, pudimos observar que, mientras más alto sea el límite de crédito, entonces menos personas tienen el estado de cuenta "impago''. Esto puede deberse a que las personas que tienen un mayor acceso a créditos de un monto alto tienden a tener un historial de crédito exitoso y por ello pueden optar por mejores opciones de financiamiento.

Relación entre el sexo y la variable de impago.

```r
library(ggplot2)
library(dplyr)

ggplot(data = data_credit, aes(x = as.factor(default.payment.next.month), fill = as.factor(SE
  geom_bar(position = "fill", color = "black") +
```

```
  labs(title = "Proporción de Hombres y Mujeres en Impago vs Pago a Tiempo",
       x = "Estado de Pago", y = "Proporción") +
  scale_x_discrete(labels = c("0" = "A tiempo", "1" = "Impago")) +
  scale_fill_manual(labels = c("1" = "Hombre", "2" = "Mujer"), values = c("skyblue", "lightc
  theme_minimal() + labs(caption = "Fuente: Elaboración propia utilizando la base de datos d
  theme(legend.title = element_blank())
```



Fuente: Elaboración propia utilizando la base de datos de Kaggle

Del gráfico anterior, podemos observar entonces que de las personas que cayeron en impago, la mayoría son mujeres, al menos más del 50%, sin embargo, veamos de manera aislada esto.

De manera que al analizar la frecuecia de impago por sexo, podemos ver que la cantidad de mujeres en estado de impago supera a la cantidad de los hombres en ese mismo estado. Si bien no es una buena medida para todo el mundo, podemos ver que en el país de estudio, las mujeres pueden tener un peor historial crediticio que los hombres.

3.3.5.7 Conclusiones

Este trabajo tuvo como objetivo explorar la correlación entre diversas características de los perfiles de los clientes y su relación con la probabilidad de impago en el contexto de un banco de Taiwán. Para ello, se aplicaron técnicas estadísticas y funciones en R para abordar la pregunta de investigación: "¿Existe correlación entre las variables de sexo, edad, estado marital, nivel educativo y límite de crédito con la probabilidad de impago?". La motivación detrás de esta pregunta radica en la importancia de identificar características que permitan a los bancos y

entidades financieras perfilar a clientes con mayor probabilidad de caer en impago, dado que muchas de estas instituciones basan su negocio en el otorgamiento de créditos.

El análisis buscó identificar patrones entre los clientes que caen en impago y determinar si tales patrones eran atribuibles al azar o si había una correlación estadísticamente significativa. Los resultados mostraron que, en el caso de las variables numéricas, solo la edad no presentaba relación con la probabilidad de impago, mientras que en las variables categóricas, el estado marital también resultó no estar relacionado. Esto indica que dichas variables no representan un factor significativo en el riesgo de crédito o probabilidad de impago.

Por otro lado, los hallazgos coinciden en gran medida con nuestras expectativas teóricas, dado que variables como el nivel educativo, el sexo y el límite de crédito sí influyen en la probabilidad de impago. Una limitación de este estudio es que no contamos con una metodología robusta de cálculo de probabilidades de impago o de evaluación del riesgo de crédito, ya que estas metodologías son conservadas de forma confidencial por las empresas y varían entre instituciones. No obstante, logramos responder a nuestra pregunta de investigación identificando correlaciones relevantes y descartando variables con menor impacto a través de pruebas de hipótesis.

Para futuros estudios, recomendamos incluir variables adicionales, tales como el motivo del préstamo, los años del crédito, el salario actual del solicitante y su historial crediticio previo, ya que estos factores podrían enriquecer los resultados y permitir una caracterización más precisa de los perfiles de riesgo.

Este trabajo resulta útil para instituciones financieras y casas de préstamo en sus etapas iniciales de operación, ya que pueden apoyarse en estos resultados para desarrollar un perfil de cliente con mayor probabilidad de pago. Al observar que variables como el sexo, el nivel educativo y el límite de crédito influyen en el riesgo de impago, estas entidades pueden empezar a construir criterios de préstamo más sólidos y orientados a la disminución del riesgo.

3.3.5.8 Agradecimientos

Nos gustaría hacer un especial agradecimiento al profesor Maikol Solís por siempre encaminar y aconsejarnos a lo largo de este trabajo, además de siempre estar ahi para brindarnos su tutoria y recomendarnos la bibliografía más adecuada. También, agradecer a los compañeros que realizaron un trabajo similar donde la retroalimentación fue muy gratificante y de mucha ayuda para tener críticas constructivas y mejorar el estudio. De igual forma, nos parece oportuno agradecer a la asistente Ana Laura López, ya que se tomó el tiempo de darnos retroalimentación para mejorar considerablemente las bitácoras, a la vez que nos brindó material de referencia el cual nos ayudó mucho.

3.4 3. Revisiones Finales

4 Anexo

4.1 Anexo 1 (CHANGELOG Bitacora 1)

4.1.1 Chore

- Agrega archivo configuracion pre-commit
- Agrega configuracion repo-actions
- Agrega cambios en docs/
- Modifica la carpeta docs
- Modifica la carpeta docs
- Modifica la carpeta docs

4.1.2 Feat

- Agrega documentos para quarto
- Agrega carpeta docs
- Agrego la base de datos en formato csv
- Agrego la base de datos en formato txt a manera de respaldo del archivo en formato csv. Cualquier cambio realizado al archivo original sera realizado en este tambien.
- Agrega documentos Bitacora_1
- Agrega comentario en Bitacora_1
- Elimina la base de datos
- Agrega la base de datos
- Agrega la definición de la idea
- Agrega en bitácora 1 identificación de tensiones
- Agrega en bitácora 1, identificación de Tensiones
- Agrega en bitácora 1, Reformulación de la idea en modo pregunta
- Agrega en bitácora 1, argumentación de las preguntas
- Agrega en bitácora 1, argumentación a través de datos
- Agrega en Bitacora 1 avance de revisión bibliografica
- Agrega en bitácora 1, parte de escritura
- Agrega el archivo de references.bib
- Agrega las referencias bibliograficas
- Agrega en bitacora 1 principios y teorias

- Agrega en bitacora 1 busqueda bibliografica
- Agrega la introduccion
- Se agrega a bitácora 1, la UVE de Gowin
- Agrega imagen de la V de Gowin

4.1.3 Fix

- Arregla una funcionalidad de .gitignore
- Corrige la numeración de la bitacora 1
- Arreglo de index
- Realiza correciones diversas en los documentos
- Corrige error ortografico

4.2 Anexo 2 (Participacion Bitacora 1)

```
project     : Bitacora-Grupo-5-CA-0204-II-2024-
repo age    : 7 days
branch:     : main
last active : 40 minutes ago
active on   : 4 days
commits     : 36
files       : 41
uncommitted : 10
authors     :
   26  Jeikel Navarro  72.2%
    8  Erick Venegas   22.2%
    2  Gara            5.6%
```

Figura 4.1: Summary Bitacora 1

4.3 Anexo 3 (CHANGELOG Bitacora 2)

```
### Chore

- Agrega archivo configuracion pre-commit
- Agrega configuracion repo-actions
- Agrega cambios en docs/
- Modifica la carpeta docs
- Modifica la carpeta docs
- Modifica la carpeta docs
- Modifica quarto para agregar bitacora 2
```

Figura 4.2: Chore Bitacora 2

```
### Feat

- Agrega documentos para quarto
- Agrega carpeta docs
- Agrego la base de datos en formato csv
- Agrego la base de datos en formato txt a manera de respaldo del archivo en formato csv. Cualquier cambio realizado al archivo original sera realizado en este tambien.
- Agrega documentos Bitacora_1
- Agrega comentario en Bitacora_1
- Elimina la base de datos
- Agrega la base de datos
- Agrega la definición de la idea
- Agrega en bitácora 1 identificación de tensiones
- Agrega en bitácora 1, identificación de Tensiones
- Agrega en bitácora 1, Reformulación de la idea en modo pregunta
- Agrega en bitácora 1, argumentación de las preguntas
- Agrega en bitácora 1, argumentación a través de datos
- Agrega en Bitacora 1 avance de revisión bibliografica
- Agrega en bitácora 1, parte de escritura
- Agrega el archivo de references.bib
- Agrega las referencias bibliograficas
- Agrega en bitacora 1 principios y teorias
- Agrega en bitacora 1 busqueda bibliografica
- Agrega la introduccion
- Se agrega a bitácora 1, la UVE de Gowin
- Agrega imagen de la V de Gowin
- Agrega archivos necesarios para el changelog y summary
- Agrega títulos y contenido en el análisis estadistico
- Agrega base de datos
- Agrega datos resumidos de las variables
- Agrega nuevas referencias
- Agrega variables resumen de interés
- Agrega gráfico de distribución por género
- Agrega gráfica de la variable ingreso
- Agrega gráfico de la variable ingreso y facet por nivel académico
- Agrega gráfico de la variable ingreso y facet por género
- Agrega gráfico de la variable monto del préstamo facet por nivel educativo y género
- Agrega gráfico de la variable propósito del préstamos por nivel educativo
- Agrega gráfico de la variable calificación por riesgo contra edad
- Se agrega el archivo html de la bitacora 1
- Se actualiza el archivo .css de la bitacora 1
- Se agrega una nueva ficha de literatura
- Se agrega una ficha de literatura
- Se actualiza el archivo html de la bitacora 1
- Agrega gráfico de variable calificación de riesgo facet nivel educativo
- Agrega gráfico de cajas de variable calificación de riesgo contra salario
- Agrega gráfico de la variable calificación de riesgo contra propósito del préstamo
- Agrega la tabla de organizacion y literatura
- Agrega el archivo de Apendice
- Se agrega el archivo de anexo
- Se agregan subtitulo y nuevos capitulos
- Se agregan nuevas referencias
- Se agrega una nueva ficha de literatura
- Se actualiza el archivo html de la bitacora 1
- Se agrega el primer enlace de literatura
- Se agregan 2 enlaces de literatura
- Se agregan los ultimos enlaces de literatura
```

Figura 4.3: Feat Bitacora 2

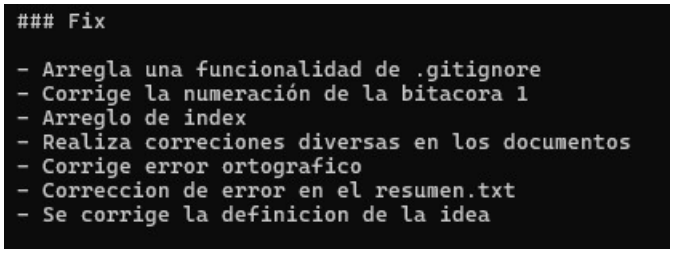


Figura 4.4: Fix Bitacora 2

4.4 Anexo 4 (Participacion Bitacora 2)

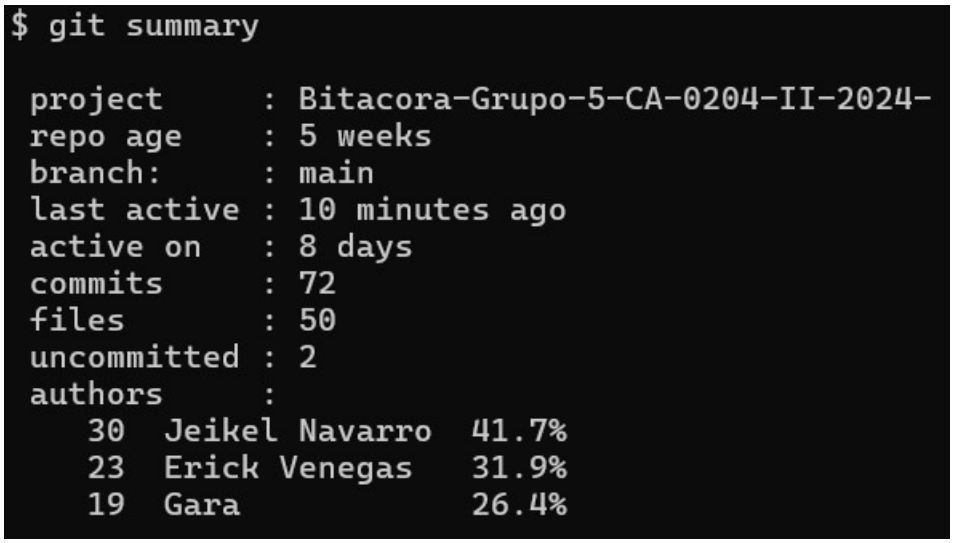


Figura 4.5: Summary Bitacora 2

4.5 Referencias bibliográfica

Chicu, Dorina. 2020. «La valoración del riesgo financiero». 2020. https://openaccess.uoc.edu/bitstream/10609/150126/1/LaValoracionDelRiesgoFinanciero.pdf.

Edgar Apaza, César Condori, Samuel Cazorla. 2022. «La Correlación de Pearson o de Spearman en caracteres físicos y textiles de la fibra de alpacas». 2022. http://www.scielo.org.pe/pdf/rivep/v33n3/1609-9117-rivep-33-03-e22908.pdf.

Hadley Wickham, Garrett Grolemund. 2019. «R for Data Science (2nd ed.)». 2019. https://digitallibrary.tsu.ge/book/2019/september/books/R-for-Data-Science.pdf.

Maria de los Ángeles Herrera, Juan Terán. 2024. «Conceptualización del riesgo de los mercados financieros». 2024. https://www.redalyc.org/pdf/900/90075920006.pdf.

Palacios, Alberto. 2012. «Calificación de riesgo: definición e influencia en la última década». 2012. https://digibuo.uniovi.es/dspace/bitstream/handle/10651/4017/ACC-.pdf;jsessionid=723581A47435AFB6D2FEC05A70379F77?sequence=1.

Pértega Díaz, S., y S. Pita Fernández. 2004. «Asociación de variables cualitativas: El test exacto de Fisher y el test de McNemar». *Cadernos de Atención Primaria* 11: 304-8. https://www.agamfec.com/wp/wp-content/uploads/2015/07/14_Invest_N11_5.pdf.

Solis, Maikol. 2024. «Guía del curso: Herramienta para Ciencia de Datos». 2024. https://maikolsolis.com/libros/hpcd/.

Walpole, Ronald E., Raymond H. Myers, Sharon L. Myers, y Keying E. Ye. 1999. *Probabilidad y estadística*. Prentice Hall. https://books.google.co.cr/books/about/Probabilidad_y_estad%C3%ADstica.html?hl=es&id=kz1VAAAACAAJ&redir_esc=y.