Actividad 6 (Regresión Lineal Múltiple y No Lineal)

Nombre del Equipo

CAPT

Miembros

Danna Paola Arciniega Zúñiga | A01731987

Victoria Eugenia Téllez Castillo | A01732258

Mitzi Castelán Chávez | A01731147

Elizabeth Pérez García | A01366971

Fecha

12 de Octubre de 2023

Analítica de datos y herramientas de inteligencia artificial II

PREPROCESAMIENTO

Durante esta actividad se realizaron diferentes acciones de preprocesamiento para poder tener un archivo sin datos nulos y outliers, para así trabajar sin datos externos, con esto se pudo realizar una extracción de características para las columnas categóricas,en relación se analizó una correlación y un modelo de regresión lineal múltiple entre todas las variables numéricas, esto se observará de manera visual mediante un mapa de calor donde se mostrarán los mejores modelos lineales simples en conjunto con sus coeficientes de correlación y una breve descripción para ellos, para todas las columnas se generará un análisis descriptivo de los hallazgos obtenidos empleando tablas y gráficos, con base en todos estos resultados se mostrará un modelo no lineal para predecir cada variable numérica del data frame, esto se mostrará mediante una gráfica comparando con los coeficientes obtenidos con respecto a los coeficientes obtenidos en los modelos lineales, a continuación se mostrará paso a paso el proceso de cada código para poder obtener óptimos resultados.

Como se mencionó anteriormente, se comenzó limpiando nuestra base de datos "BD_Socio formador (TrainingDataComplete)", primero se imprimió nuestra base de datos inicial, en ella obtuvimos las primeras líneas e información de la misma para saber de qué tipo de dato se trataba en cada columna, en ella observamos que existían datos tipo, objeto e int.

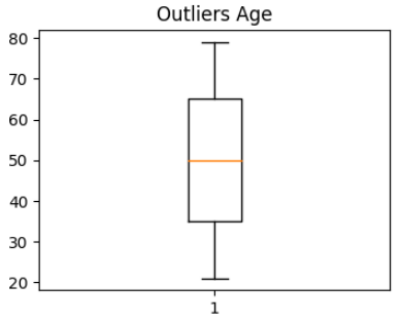
```
#OBTENER INFORMACION DEL DATA FRAME
df.info()
```

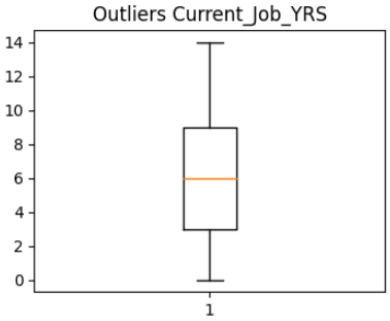
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 252000 entries, 0 to 251999
Data columns (total 13 columns):
 #   Column             Non-Null Count   Dtype 
---  ------             --------------   ----- 
 0   Id                 252000 non-null  int64 
 1   Income             252000 non-null  int64 
 2   Age                252000 non-null  int64 
 3   Experience         252000 non-null  int64 
 4   Married/Single     252000 non-null  object
 5   House_Ownership    252000 non-null  object
 6   Car_Ownership      252000 non-null  object
 7   Profession         252000 non-null  object
 8   CITY               252000 non-null  object
 9   STATE              252000 non-null  object
 10  CURRENT_JOB_YRS    252000 non-null  int64 
 11  CURRENT_HOUSE_YRS  252000 non-null  int64 
 12  Risk_Flag          252000 non-null  int64 
dtypes: int64(7), object(6)
memory usage: 25.0+ MB
```

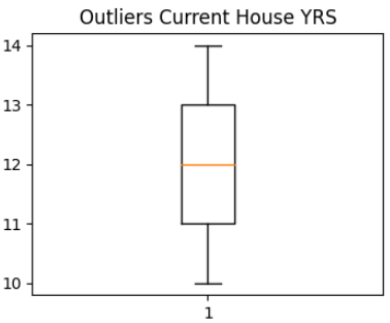
Para la continuación del preprocesamiento, identificamos los valores tanto nulos como atípicos para así analizar las columnas y reemplazarlos de la mejor manera para comenzar con la manipulación de datos, sin embargo, no fue necesario ningún tipo de limpieza de valores, ya que dentro del dataframe no se contaba con valores nulos ni con valores atípicos. Para tener una representación visual de la falta de valores atípicos de cada una de las columnas principales del conjunto de datos, se realizaron histogramas de cajas y de barras, las cuales se muestran a continuación.

VALORES ATÍPICOS

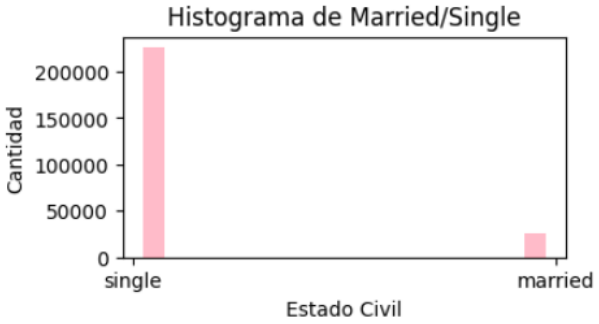
Outliers 'Caja' | Income, Age, Experience, current house, Current Job

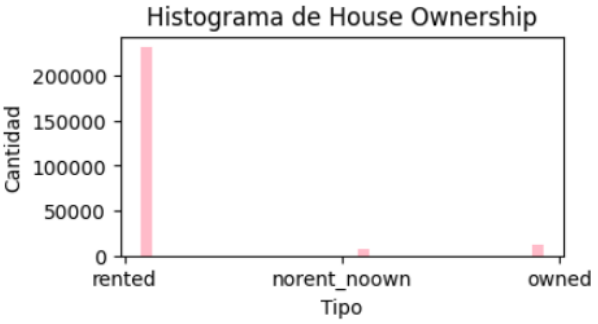


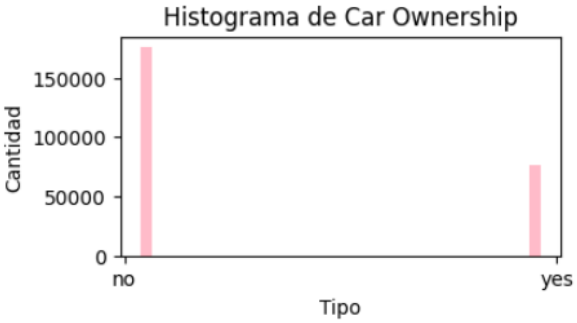


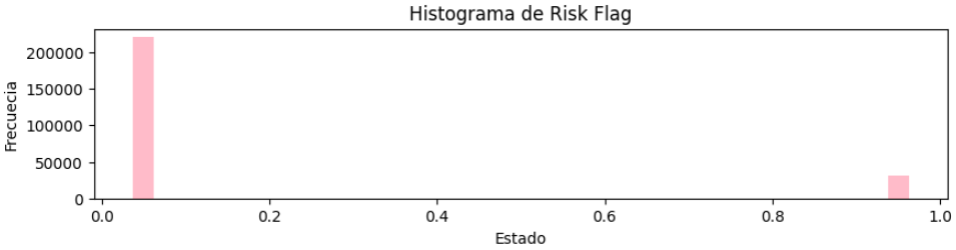


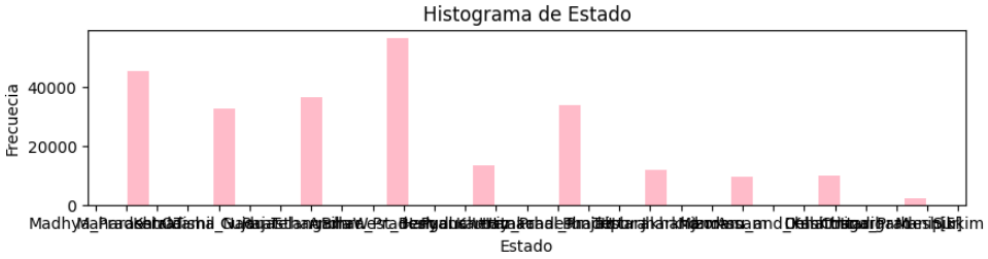
Histograma | Married/Single, Howse Ownership, Car Ownership, Profession,City, State, Risk Flag

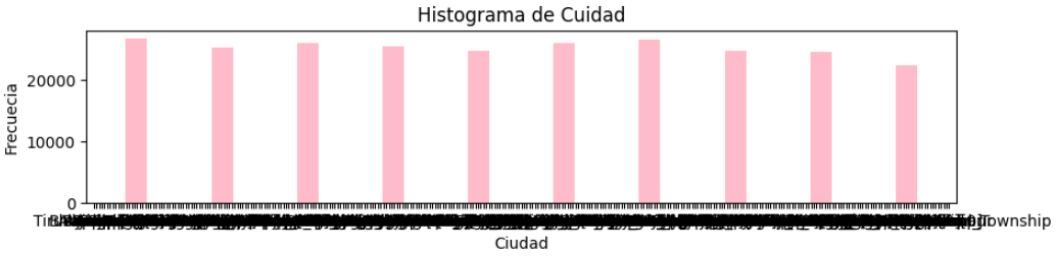










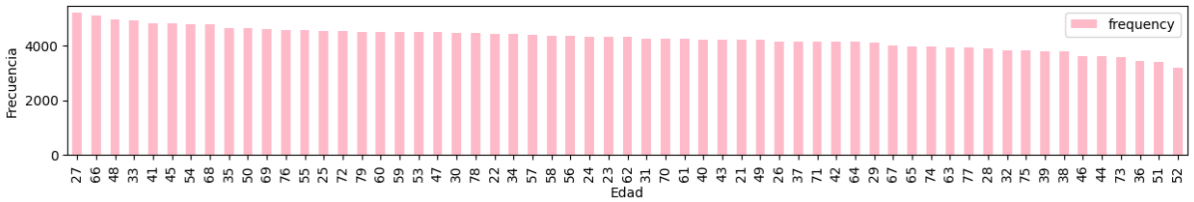


La creación de histogramas para las previas variables "Married/Single," "House Ownership," "Car Ownership," "Profession," "City," "State," y "Risk Flag" proporciona una representación visual de cómo se distribuyen las categorías en cada variable categórica, estos histogramas nos permite observar la frecuencia de cada categoría, identificar desequilibrios o tendencias en la distribución y destacar categorías dominantes. Además, la comparación de las distribuciones entre diferentes variables categóricas puede revelar patrones y relaciones potenciales, es importante crearlas y observarlas para la toma de decisiones en futuros análisis.

EXTRACCIÓN DE CARACTERÍSTICAS

Tras la creación de un nuevo data frame, el cual contiene un total de 9 columnas, consideradas como las más relevantes y de mayor importancia para el posterior análisis, comenzamos con los análisis univariados de cada una de las columnas junto con sus gráficos, como se muestra a continuación:

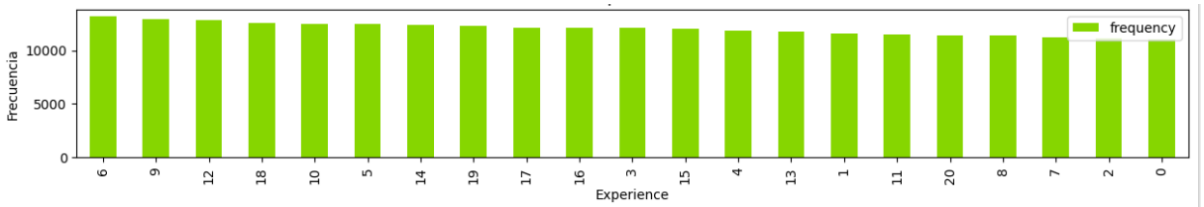
Edad | Age



Los resultados arrojan que la edad con mayor registro es la de 27 años, mientras que la edad con menores registros es de 52 años. Asimismo, se puede observar que no hay un patrón de comportamiento de edades, es decir, no depende de la edad para que su frecuencia sea en constante crecimiento o decrecimiento.

Otro de los hallazgos fue que las primeras cinco edades con mayor número de frecuencia o registros en la base de datos son dispersas a excepción de una. Las edades son 27, 66, 48, 33 y 41 años.

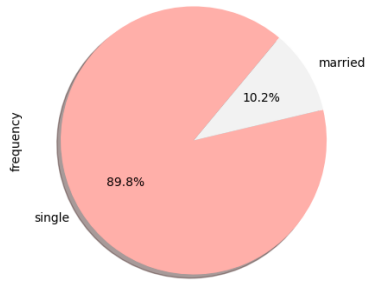
Experiencia | Experience



La experiencia tiene un comportamiento menos disperso, los resultados o la frecuencia entre los años de experiencia tienen una diferencia no tan pronunciada. La mayoría de las personas de la base de datos tienen seis años de experiencia, y la frecuencia con cero años de experiencia es la más baja.

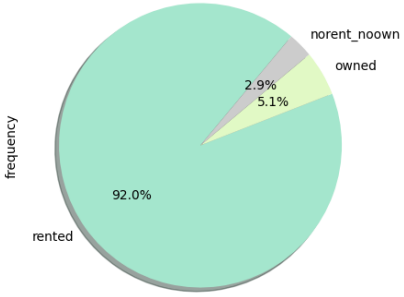
20 años de experiencia, es la experiencia mayor, es decir, la más alta en años, más no en frecuencia, con 11350 registros. 6 años de experiencia, con el mayor número de frecuencia tiene 13158 registros, y con cero años de experiencia, menor número de registro y menor número de años, cuenta con 11043 registros.

Casado / Soltero | Married/Single



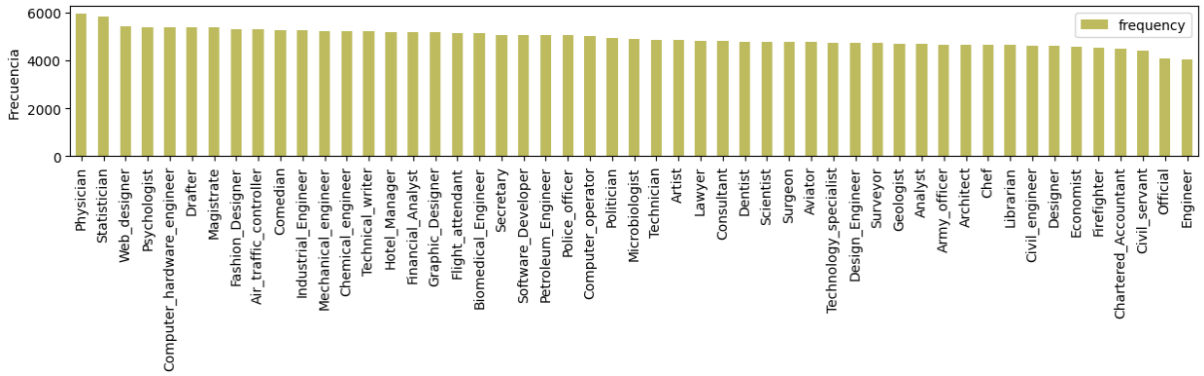
Se puede observar que la mayoría de las personas de la base de datos son personas solteras (89%) mientras que un porcentaje significativamente menor (30%) están casados. Estos datos indican un escenario demográfico de la población y puede ayudar a comprender la relación y los resultados de las otras columnas.

Propietario de Casa | House Ownership



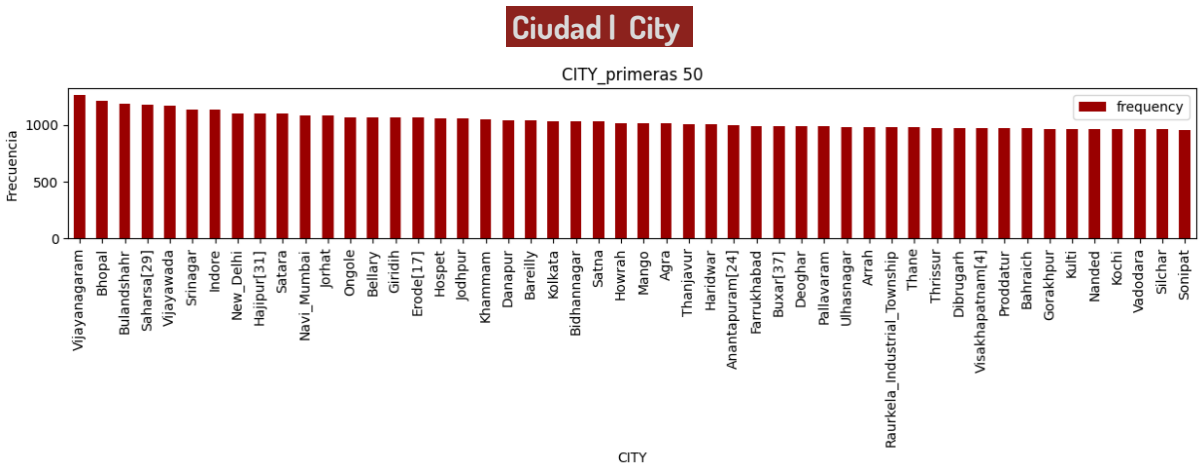
Más del 90% de las personas alquilan sus viviendas, un 5% son propietarios y un porcentaje muy bajo, dos terceras parte de las personas que son propietarios, no rentan, ni son dueños.

Profesión | Profession

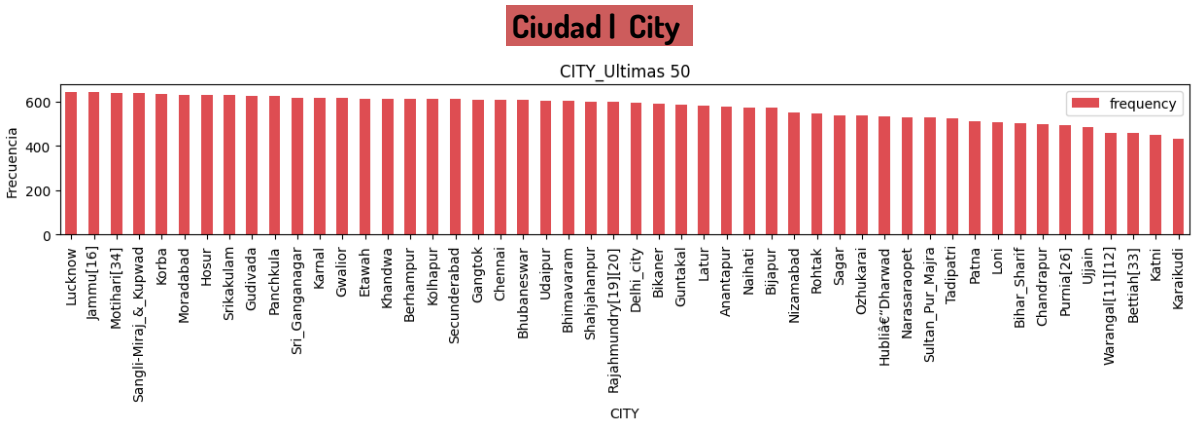


Se observa que hay 51 profesiones, todas con distintas frecuencias. Physician (Médico/ Doctor) es la profesión con mayor frecuencia, teniendo un registro total de 5957 personas con esa profesión; mientras que la profesión con menos registros es Engineer (Ingeniero).

Por otro lado, ninguna profesión tiene la misma frecuencia, es decir cada profesión tiene un registro único. Las diferencias entre profesiones encuentro a frecuencia de una a otra, de manera decreciente no es tan prolongada.



Para observar de mejor manera los datos, se filtraron los datos de tal manera que se pudieran visualizar los registros con las frecuencias más altas. En este caso la cuidad con el mayor número de registros es Vijayanagaram, ubicado en la India, con 1259 apariciones.

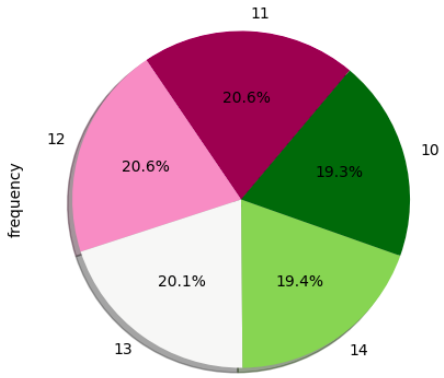


Mientras que la ciudad con el menor número de registros es Karaikudi, univaca en la India con 431 registros. La diferencia entre los límites es de casi de 828 registros.

La gráfica muestra los comportamientos en términos de frecuencia con relación a los años. La mayoría de los registros ha permanecido en su trabajo por tres años, seguido por cuatro, cinco y seis años, respectivamente. 14 años, a pesar que representa el número de año mayor, tiene un número de registros más bajo, es decir muy pocas personas han permanecido en el mismo trabajo por catorce años.

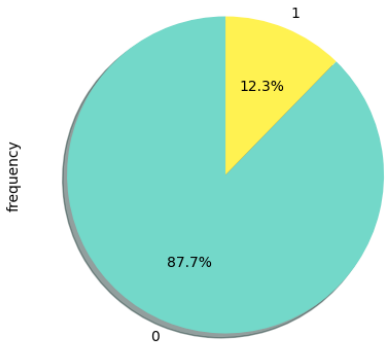
10 años es la edad mediana, es decir esta edad es la posición central del conjunto de datos. De la misma manera se puede observar que del rango de 3 a 10 años, se posicionan las frecuencias más altas; mientras que del rango de 11 a 14, añadiendo el rango de 0 a 2, se encuentran las frecuencias más bajas.

Años Actuales en su Hogar | Current House Yrs



Los datos muestran un comportamiento de frecuencia parcialmente equitativo, es decir sus frecuencias son una quinta parte del total. Los años 11 y 12 tienen la misma frecuencia en porcentaje, además de ser las más altas. Seguido del año 13, 14 y 10, en donde el año diez, es quien tiene la menor frecuencia. El año 11 tiene el mayor número de registros con 51873, y el año diez con 48674. Asimismo, se puede decir que la mayoría de las personas han vivido 11 años en su residencia actual.

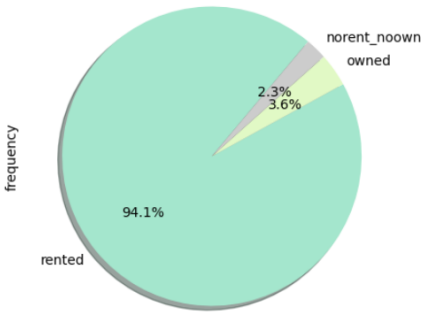
Indicador de Riesgo | Risk Flag



Gran parte de las personas no es considerada como un riesgo, ya que no ha incurrido en ningún incumplimiento significativo en sus obligaciones financieras. Mientras que un 12% si son consideradas como un riesgo crediticio, ya que han incumplido en obligaciones financieras.

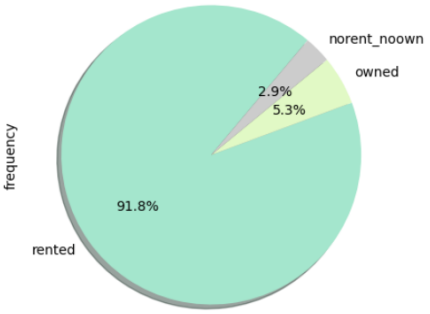
ANÁLISIS COLUMNA "Married/Single"

MARRIED -> Propietario de Casa | House Ownership



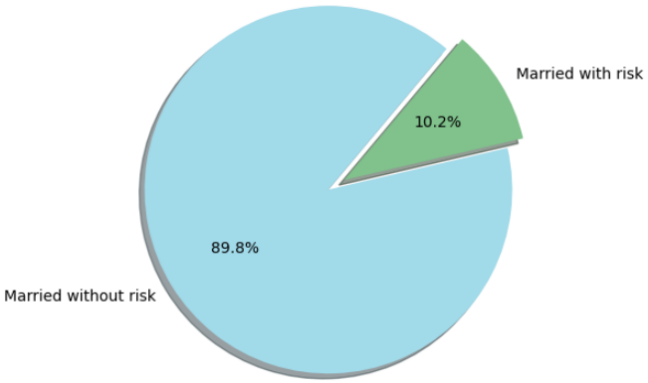
Esta gráfica muestra los resultados del comportamiento de un dataframe únicamente con registros de personas casadas "Married" en relación con la columna House Ownership, en donde los resultados muestran que gran porcentaje de las personas rentan sus hogares, seguido de un porcentaje muy bajo, que representan las personas que son dueñas de sus hogares pero relativamente más alto a las personas que no rentan, ni son dueños.

SINGLE -> Propietario de Casa | House Ownership



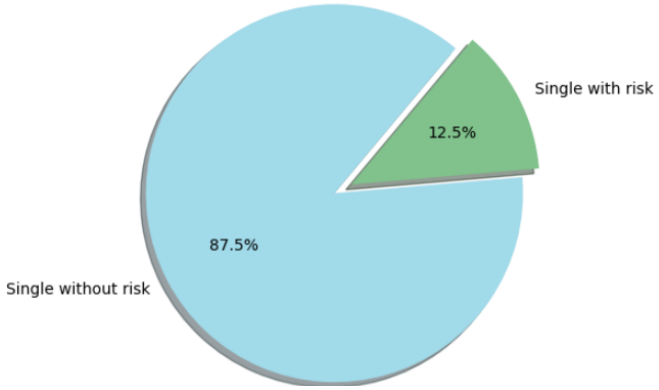
A comparación de la gráfica anterior, los resultados de esta, están tomando únicamente a las personas solteras. Nuevamente el porcentaje de personas que rentan es superior a las otras dos categorías, sin embargo reduce un 3% a comparación de las personas casadas en donde esa reducción de porcentaje, se le incrementa a las personas dueñas, mientras que el porcentaje de personas que no rentan, ni son dueños, se mantiene en un comportamiento similar

MARRIED -> Indicador de Riesgo | Risk Flag



La gráfica muestra el comportamiento de las personas casadas con la categoría de indicador de riesgo. Se puede observar que gran parte de las personas casadas no son consideradas como riesgosas, sin embargo un 10% de ellas si lo son.

SINGLE -> Indicador de Riesgo | Risk Flag



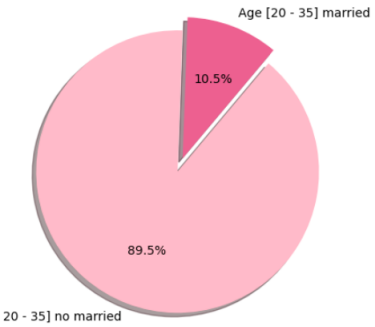
Esta nueva gráfica ahora muestra el comportamiento de las personas solteras. Al igual que el comportamiento anterior, la mayoría renta, sin embargo el porcentaje a comparación de las personas casadas es menor, lo que indica que las personas solteras son más propensas a no rentar el inmueble a donde habitan.

MARRIED -> Actuales en el Trabajo | Current Job Yrs

La gráfica muestra los comportamientos de las personas casadas en términos de frecuencia con relación a los años. La mayoría de los registros ha permanecido en su trabajo por tres años, seguido por cuatro y seis años, respectivamente. 14 años, a pesar que representa el número de año mayor, tiene un número de registros más bajo, es decir muy pocas personas han permanecido en el mismo trabajo por catorce años.

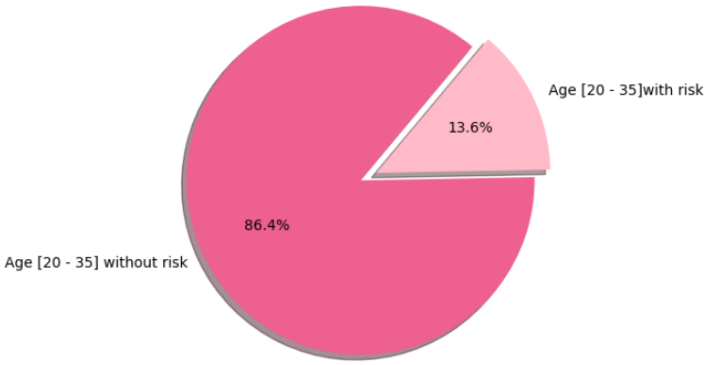
ANÁLISIS COLUMNA "Edad [20 -35]

Edad = [20-35] -> Married /Single



Analizando a las personas del rango de edad de entre los 20 a los 35 años, la mayoría de ellos son personas solteras, mientras que solo un 10% son personas casadas, para ambos se puede percibir mucha diferencia ya que viene siendo más de la mitad el cambio de porcentaje..

Edad = [20-35] -> Indicador de Riesgo | Risk Flag



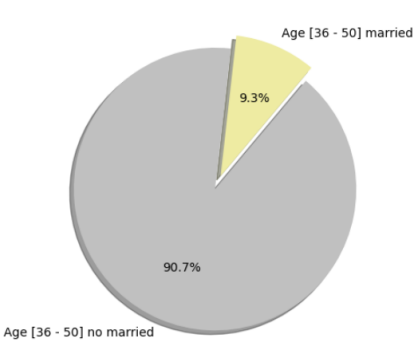
Analizando a las personas del rango de edad de entre los 20 a los 35 años, la mayoría de ellas no representan o son consideradas como personas con riesgo crediticio, únicamente un 13% de ellas lo son.

Edad = [20-35] -> Años Actuales en el Trabajo | Current Job Yrs

La gráfica muestra los comportamientos de las personas dentro del rango de edad entre los 20 a los 35 años, en términos de frecuencia con relación a los años. La mayoría de los registros ha permanecido en su trabajo por tres años, seguido por cuatro, cinco y seis años, respectivamente. 14 años, a pesar que representa el número de año mayor, tiene un número de registros más bajo, es decir muy pocas personas han permanecido en el mismo trabajo por catorce años. Además, de los años cero, uno y dos también cuentan con frecuencias bajas.

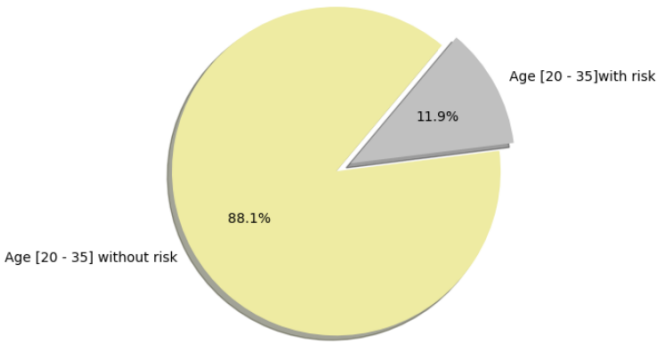
ANÁLISIS COLUMNA "Edad [36 -50]

Edad = [36 - 50] -> Married /Single



Analizando a las personas del rango de edad de entre los 36 a los 50 años, la mayoría de ellos son personas solteras, mientras que sólo un 9% son personas casadas, se observa que es bastante la diferencia y esta incliniado a la zona gris que significa que son personas solteras.

Edad = [36 - 50] -> Indicador de Riesgo | Risk Flag



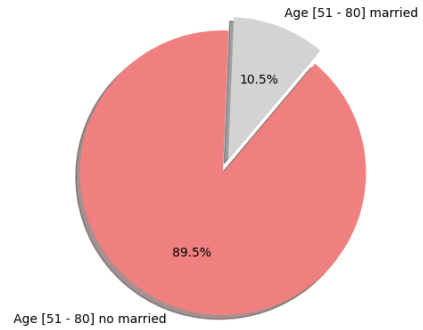
Analizando a las personas del mismo rango de edad, se puede observar que solo un 12% de ellas son personas con riesgo a un crédito, mientras que el 88% restante no lo es.

Edad = [36 - 50] -> Actuales en el Trabajo I Current Job Yrs..

La gráfica muestra los comportamientos de las personas del rango de edad seleccionado en términos de frecuencia con relación a los años. A comparación del rango anterior aquí las personas han permanecido más tiempo en su trabajo por cuatro años, seguido de tres años y después por cinco y seis años respectivamente. Asimismo el año 12, 13 y 14, son los años con menores frecuencias.

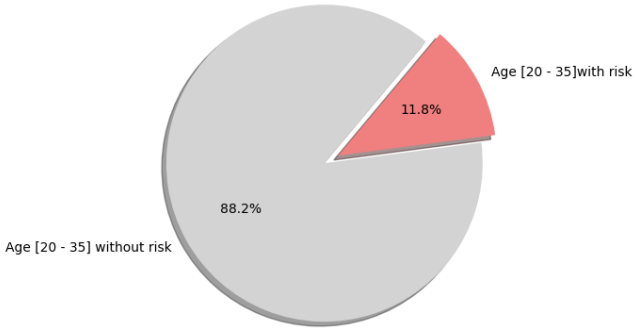
ANÁLISIS COLUMNA "Edad [51 - 80]

Edad = [51-80] / Married /Single = married



Analizando a las personas del rango de edad de entre los 51 a los 80 años, la mayoría de ellos (89.5%) son personas solteras, mientras que sólo un 10.5% son personas casadas, existe una gran diferencia pero no tanta comparada con la gráfica pasada.

Edad = [51 - 80] -> Indicador de Riesgo I Risk Flag..



Analizando a las personas del mismo rango de edad, se puede observar que solo un 11.8% de ellas son personas con riesgo a un crédito, mientras que el 88.2% restante no lo es, no es un riesgo a que se les proporcionen un crédito.

Edad = [51 - 80] -> Actuales en el Trabajo | Current Job Yrs..

La gráfica muestra los comportamientos de las personas del rango de edad seleccionado en términos de frecuencia con relación a los años. En esta gráfica se puede observar que la mayoría de personas han permanecido por más tiempo en su trabajo por tres años, seguido de cuatro años y después por cinco y seis años respectivamente. Asimismo se observa que en el año 1 y 2, fueron los años con una similitud en personas que se quedaron, en continuación los años 13 y 14 son los años con menores frecuencias ya que fueron las más bajas.

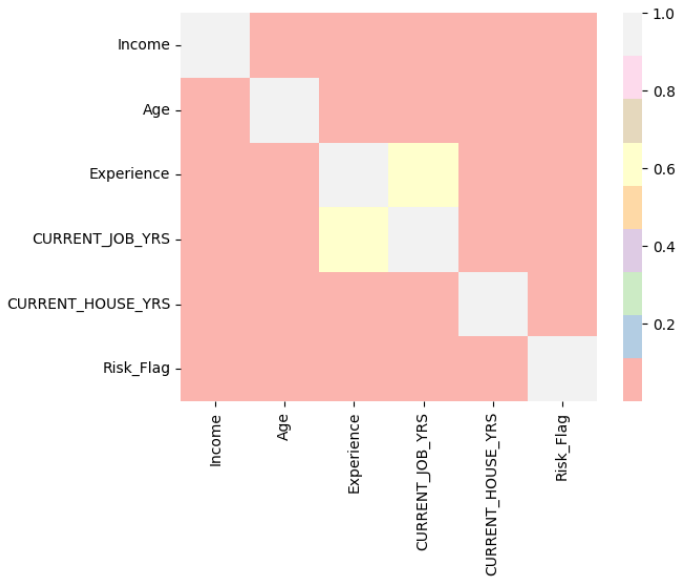
ANÁLISIS DE CORRELACIÓN ENTRE LAS VARIABLES NUMÉRICAS

A continuación, se realizará un análisis de correlación entre las variables numéricas Income, Age, Experience, CURRENT JOB_YRS, CURRENT_HOUSE_YRS y Risk Flag para nuestro data frame, en este se visualizará en un mapa de calor, esto para que se puedan identificar los tres mejores modelos de regresión lineal simple, junto con sus coeficientes de correlación y descripciones correspondientes.

Par comenzar este paso se creó un nuevo data frame utilizando la función .loc donde únicamente incluyen las columnas previamente mencionadas, con ellas se mostraron los primeros datos para identificar que efectivamente solo incluya variables numéricas, posteriormente se desea calcular la correlación entre variables, para ello se utilizó la función .corr, esta nos fue de ayuda, ya que nos imprime una matriz que contendrá el coeficiente de correlación que indica la relación entre dos variables, en consiguiente se realizó un nuevo código donde solicitamos encontrar el valor absoluto de todos los elementos en la matriz de correlación.

La función abs convierte todos los valores de correlación en números positivos, lo que facilita la identificación de las correlaciones, independientemente de si son positivas o negativas, a continuación se puede observar el resultado al imprimir la nueva variable Corr_Factores_1.

	Income	Age	Experience	CURRENT_JOB_YRS	CURRENT_HOUSE_YRS	Risk_Flag
Income	1.000000	0.000652	0.006422	0.007045	0.002397	0.003091
Age	0.000652	1.000000	0.001118	0.002154	0.020134	0.021809
Experience	0.006422	0.001118	1.000000	0.646098	0.019309	0.034523
CURRENT_JOB_YRS	0.007045	0.002154	0.646098	1.000000	0.005372	0.016942
CURRENT_HOUSE_YRS	0.002397	0.020134	0.019309	0.005372	1.000000	0.004375
Risk_Flag	0.003091	0.021809	0.034523	0.016942	0.004375	1.000000



Para continuar con las funciones pasadas graficamos un mapa de calor de los coeficientes de correlación en la última matriz "Corr_Factores_1".este mapa lo que nos indica es que mientras más obscuro se encuentre el color menor correlación hay, por el contrario, mientras menor es el color existe una correlación mayor entre las variables, en este caso se muestra en el medio de la gráfica, donde existe mayor correlación es en el punto 0.6 de color amarillo del cuadrante principal junto con el color gris que se encuentra en diagonal, los cuadrados en diagonal son todos 1 porque se están comparando las mismas variables y por lo tanto están perfectamente relacionadas.

Para tener mayor información de lo que significa es que en cada tipo de correlación existe un espectro representado por valores de -1 a +1 donde las características de correlación positiva leve o alta pueden ser como 0,5 o 0,7, por otro lado una correlación positiva muy fuerte y perfecta está representada por una puntuación de correlación de 0,9 o 1 y si hay una fuerte correlación negativa, se representará con un valor de -0,9 o -1, se sabe que los valores cercanos a cero indican que no hay correlación.

Para continuar podemos observar un gráfico con mayor información, en él podemos determinar en qué cuadrante se encuentra la v

La variable más alta con respecto a la correlación con otra variable, como antes mencionado, un valor mayor que 0 indica que existe una correlación positiva, en este caso es el valor de 0.65.

El gráfico muestra que existe una correlación positiva moderada entre la experiencia y los años en el trabajo actual (0.65). Esto significa que, en general, las personas que llevan más tiempo en el mismo trabajo también tienen experiencia más alta o tienen más años en el mismo empleo.

Las variables que tienen mayor peso en el gráfico son las que tienen una correlación más fuerte con la experiencia y los años actuales en su trabajo, los años en el trabajo actual también tienen una correlación positiva moderada con los ingresos, quiere decir que las personas que llevan más tiempo en el mismo trabajo también tienen ingresos más altos, toda la diagonal en gris tiene una correlación fuerte, puesto que están siendo estimadas entre la misma variable.

MODELO MATEMÁTICO

Para continuar se realizó el modelo matemático, con la finalidad de comprender, analizar y llevar a cabo una mejor toma de decisiones dentro de un campo que cuenta con diversas variables, proporcionando una representación numérica.

El primer paso para la creación de modelos fue ordenar de forma descendente las correlaciones obtenidas de las variables cuantitativas ('Experience','Age'), esto para saber cuál fue la mayor correlación jerárquicamente, sabemos que cuanto más cerca esté el valor absoluto de la correlación de 1 (positivo o negativo), más fuerte es la relación, una correlación cercana a 0 indica una relación débil. Como resultado se obtiene lo siguiente:

Experience		Age	
Experience	1.000000	Age	1.000000
CURRENT_JOB_YRS	0.646098	Risk_Flag	0.021809
Risk_Flag	0.034523	CURRENT_HOUSE_YRS	0.020134
CURRENT_HOUSE_YRS	0.019309	CURRENT_JOB_YRS	0.002154
Income	0.006422	Experience	0.001118
Age	0.001118	Income	0.000652
Name: Experience, dtype: float64		Name: Age, dtype: float64	

Para continuar, se selecciona la variable con la correlación más alta, sin tomar en cuenta age y experience, respectivamente, al observar los resultados se aprecia que existen dos correlaciones con Experience y una con age, cada uno se observa con su respectivo resultado.

Experience		Age	
CURRENT_JOB_YRS	0.646098		
Risk_Flag	0.034523	Risk_Flag	0.021809
Name: Experience, dtype: float64		Name: Age, dtype: float64	

Teniendo estos resultados se realizó el modelo de regresión lineal, la cual fue importante para poder analizar datos y comprender las relaciones entre variables, en él nos puede proporcionar información valiosa para la toma de decisiones, la predicción y la identificación de variables relevantes en función a nuestros objetivos.

Para su creación se utilizó un 'for' para crear 2 modelos de regresión lineal simples para diferentes variables independientes, en este caso, se deseaba comparar cada variable independiente y saber cómo se relaciona con las variables dependientes "Experience" y "Age.", nos fue de gran ayuda el 'for', ya que nos permite automatizar este proceso y calcular los modelos de regresión para todas las variables seleccionadas en un conjunto de datos, como resultado a este modelo nos dio lo siguiente:

```
Experience
Para CURRENT_JOB_YRS:
Coeficiente de correlación: 0.6461
Modelo de regresión lineal simple: 1.06x + 3.35

Para Risk_Flag:
Coeficiente de correlación: 0.0345
Modelo de regresión lineal simple: -0.63x + 10.16

Age
Para Risk_Flag:
Coeficiente de correlación: 0.0218
Modelo de regresión lineal simple: -1.13x + 50.09
```

Los resultados indican cómo las variables current job y risk flag se relacionan con las variables de experience y age. Mientras que la relación entre current job and experience es más fuerte y positiva, las relaciones con riskflag son más débiles y negativas, pero en ambos

casos, la relación es débil, a continuación se analiza la información detallada sobre los resultados.

Para "Experience" en función de "CURRENT_JOB_YRS", los resultados indican que existe una correlación positiva fuerte entre experiencia y trabajo actual con un coeficiente de correlación de 0.6461. Esto significa que a medida que la variable de trabajo actual aumenta, se espera que su experiencia también aumente, y viceversa, el modelo de regresión lineal sugiere que, en promedio, por cada año adicional en trabajo actual la variable experiencia aumenta en 1.06 unidades, y el valor de experiencia comienza en 3.35.

Para "Experience" en función de "Risk_Flag", En este caso, la correlación entre experiencia y riesgo es positiva, pero más débil, con un coeficiente de correlación de 0.0345, el modelo de regresión lineal sugiere un incremento en el riesgo y en cómo se relaciona con una disminución en experiencia, de igual forma se observa que el coeficiente de regresión es pequeño, lo que indica que la relación es débil, por otra parte, el valor de experiencia comienza en 10.16.

Para "Age" en función de "Risk_Flag", para este caso la correlación entre ambas también es positiva, pero muy débil, con un coeficiente de correlación de 0.0218, el modelo indica que pueda existir un incremento en la variable de riesgo, en función de cómo se relaciona con una disminución en la variable de edad. Sin embargo, al igual que en el caso anterior, la relación es muy débil, y el valor de edad comienza en 50.09.

Continuamos corroborando el coeficiente de determinación para nuestro modelo, para ello se tomó en cuenta la variable independiente 'Risk_flag' y la variable dependiente 'Experience',

Modelo Matematico | y = 1.0634x + 3.3490

Se tomó en cuenta el modelo matemático a la izquierda, consideramos que el coeficiente de determinación tiene un rango de 0 a 1 donde 0 indica que no existe variabilidad en el modelo mientras que 1 nos dice que explica toda la variabilidad, en este caso el resultado fue de 0.034522612890712595, este resultado indica que aproximadamente el 3.45% de la variabilidad se encuentra en la variable dependiente que es 'Experience'.

ANÁLISIS DE REGRESIÓN LINEAL MÚLTIPLE

Para este punto se generará un análisis comparativo entre modelos lineales y modelos no lineales propuestos para predecir variables numéricas clave en el data frame, incluyendo "Income," "Age," "Experience," "CURRENT_JOB_YRS," "CURRENT_HOUSE_YRS," y "Risk_Flag." En este análisis, se examinarán y comprobarán los coeficientes obtenidos de ambos tipos de modelos.
Los modelos no lineales nos dan una perspectiva alternativa y distinta para capturar relaciones entre las variables, en ellas se explorará cómo difieren de los modelos lineales tradicionales en términos de precisión y capacidad predictiva, para este análisis se respaldará visualmente a través de gráficas que ilustran las diferencias en las relaciones entre las variables.

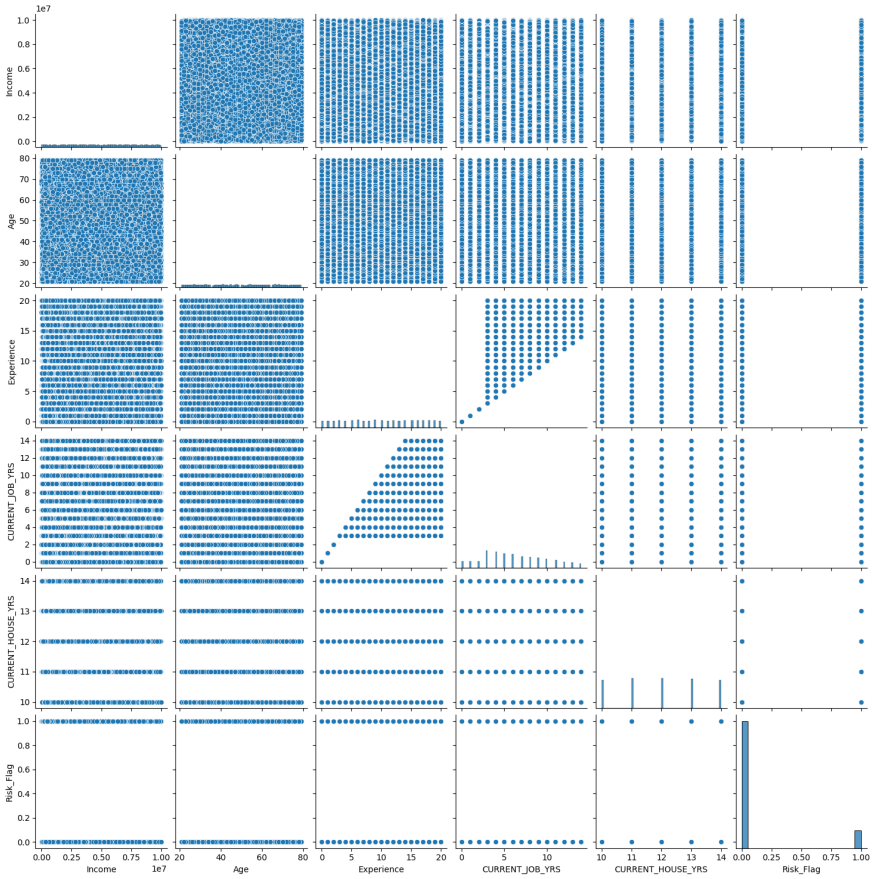
Se realizará ya que los modelos no lineales son capaces de capturar relaciones más complejas y no lineales entre variables, en diferentes situaciones, las relaciones entre variables no siguen una línea recta, y un modelo no lineal puede ajustarse mejor a los datos, es por ello que se va a corroborar con los datos con los que ya contamos.

Para este punto se comenzó creando un nuevo data frame con las columnas previamente mencionadas e importando todas las bibliotecas necesarias para este método, se imprimió el nuevo data frame para visualizar las variables categóricas que contiene cada columna.

	Income	Age	Experience	CURRENT_JOB_YRS	CURRENT_HOUSE_YRS	Risk_Flag
0	1303834	23	3	3	13	0
1	7574516	40	10	9	13	0
2	3991815	66	4	4	10	0
3	6256451	41	2	2	12	1
4	5768871	47	11	3	14	1

En continuación se graficaron todas las dispersiones entre todas las variables, la función de 'pair plot' proporciona una vista visual completa de cómo todas las variables se relacionan entre sí, nos es útil para tener una idea inicial de la estructura de los datos y posibles patrones.

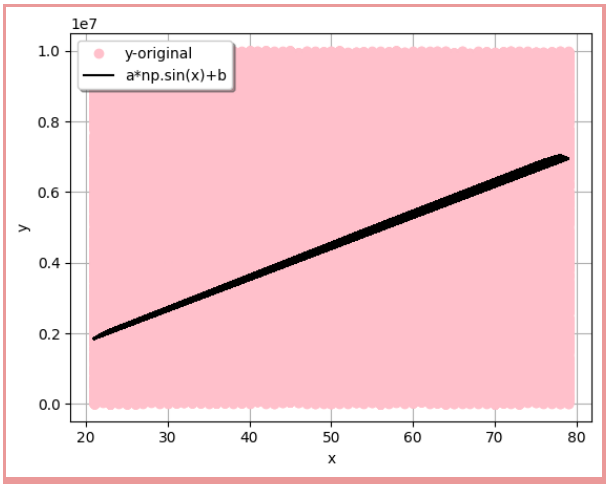
Consideramos que el hacer un pair plot es una etapa importante en la exploración de datos y en la comprensión de la estructura de los datos, ya que no puede ayudar a identificar patrones, relaciones y valores atípicos, con esto podemos tener decisiones y análisis certeros para el futuro, a continuación se puede observar los gráficos generados del nuevo data frame con todas las nuevas variables cuantitativas.



Se generaron modelos con distintas funciones, desde ser cuadrática, exponencial, inversa, senoidal, tangencial, valor absoluto, cociente entre polinomios, logarítmica, lineal con producto de coeficientes, cuadrática, inversa, polinomial inversa, al generar diferentes funciones y seleccionar la que tenga el mejor coeficiente de correlación es beneficioso, ya que permite encontrar el modelo que se ajusta de manera óptima a los datos, esto nos ayuda a tener como resultado predicciones más precisas y fiables, al tener un coeficiente de correlación más alto generalmente significa que tendremos un menor error residual y predicciones más confiables, lo que es sumamente importante para nuestro análisis y toma de decisiones. Además, el tener variedad de comparaciones con distintas funciones entre variables ayuda a tener certeza de la toma de decisiones, el porqué se obtuvo ese resultado y como es que fue la más adecuada para describir los datos.

Modelos

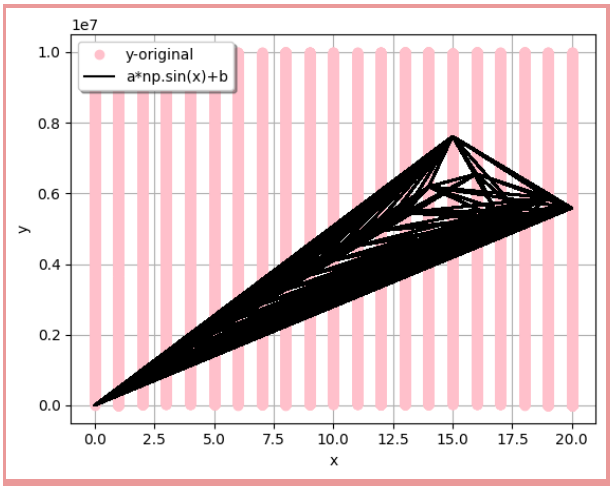
Income - Age | Función Cociente entre polinomios




```
Coeficiente de determinacion del modelo -0.3153735384874745
Coeficiente de correlacion del modelo 0.5615812839540457
```

El resultado del primer modelo es el que se muestra en la imagen superior, para su creación tomamos como variable objetivo la columna de 'Income', y como variable independiente la columna 'Age'. La función que mejor coeficiente de correlación arrojó fue cociente entre polinomios, con una ponencia de 16. Dentro de esta función contamos con un total de tres parámetros, siendo el modelo matemático el siguiente: y=(6.269 * x **16 + 3.131) / 3.497 * x

Income - Experience | Cociente entre polinomios

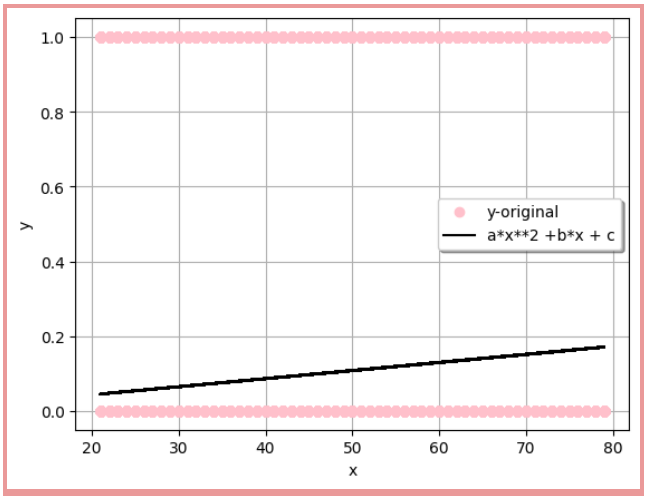



```
Coeficiente de determinacion del modelo -0.6946031150441663
Coeficiente de correlacion del modelo 0.833428530255694
```

El segundo modelo arroja el coeficiente de correlación más elevado de todo el conjunto y manipulación de datos, siendo este el más cercano a 1. La variable objetivo fue Income y la independiente Experience. Para este modelo de igual manera se utilizó la función de cociente entre

polinomios, con 16 de ponencia, donde obtuvimos el siguiente modelo matemático: y= (3.79 * x **16 + 1.039) / 2.54 * x

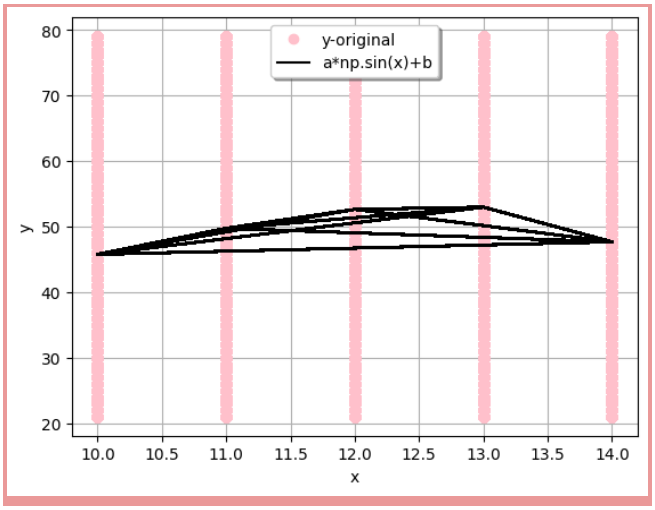
Risk Flag - Age I Cociente entre polinomios




```
Coeficiente de determinacion del modelo -0.01960186321930002
Coeficiente de correlacion del modelo 0.14000665419650604
```

Este tercer modelo arroja un resultado de coeficiente de correlación muy por debajo de los dos anteriores, se hizo uso de la misma función de cociente entre polinomios, y se jugó con tu ponencia hasta que se encontró el hallazgo que de igual manera, la ponencia de 16 era la más óptima para obtener una buena correlación. Su modelo matemático es: y= (3.017 * x ** 16 +8. 202) / 3.798 * x

Age - Current House Yrs I Cociente entre polinomios

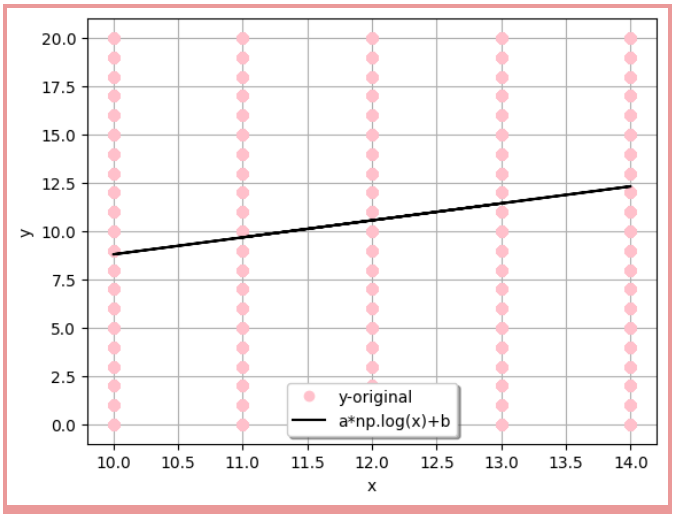



```
Coeficiente de determinacion del modelo -0.026781512751074432
Coeficiente de correlacion del modelo 0.16365058127325557
```

Continuamos con el cuarto modelo, con un coeficiente similar al modelo anterior. La función utilizada para su creación fue la misma que las anteriores, ya que demostró ser la que máximo

coeficiente de correlación obtenía, sin embargo, ahora su ponencia disminuye a 11, su modelo matemático fue el siguiente: y= (-2.411 * x ** 11 + 3.746) / 8.134 * x

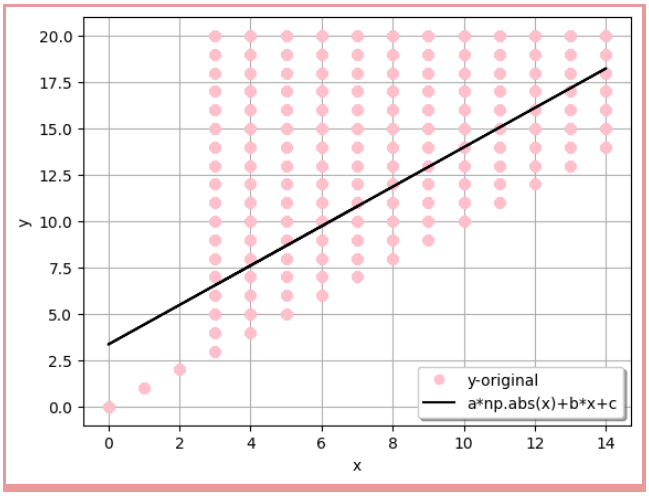
Experience - Current House Yrs I Función Inversa




```
Coeficiente de determinacion del modelo -0.040237495530901723
Coeficiente de correlacion del modelo 0.2005928601194512
```

El modelo que marca a la mitad del total, muestra un coeficiente de correlación mejor que los dos modelos anteriores, pero no tan eficiente como los dos primeros. La función inversa fue la que mejor desarrollo obtuvo en el aumento de los coeficientes, dejando su modelo matemático de : y= 1 / 1.137 * x

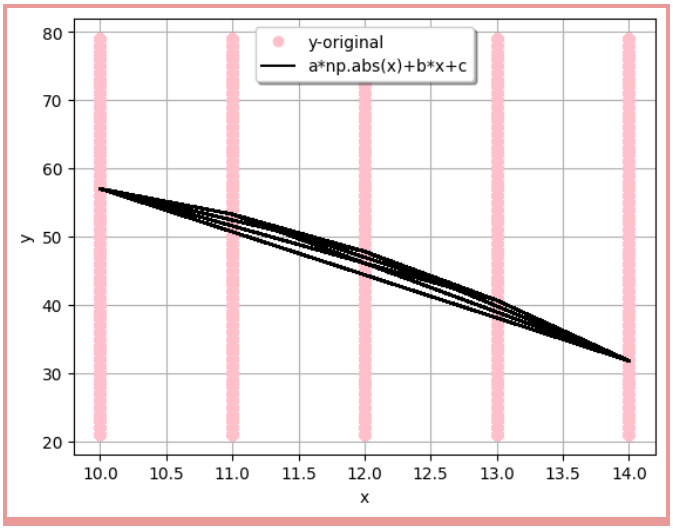
Experience - Current Job Yrs I Valor absoluto




```
Coeficiente de determinacion del modelo 0.4174420012200908
Coeficiente de correlacion del modelo 0.6460975168038419
```

La gráfica muestra que existe una correlación positiva entre la experiencia y las horas de trabajo actuales, significa que las personas con más experiencia tienden a trabajar más horas, su coeficiente de determinación indica que el 41,7% de la variación en las horas de trabajo actuales se explica por la experiencia, dejando su modelo matemático de : y= -99.935 *np.abs(x) + 100.999 * x + 3.349

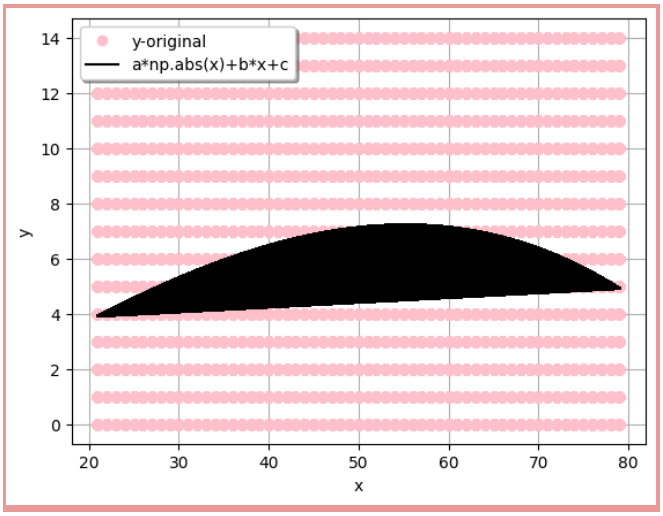
Age - Current House Yrs I Regresión no lineal




```
Coeficiente de determinacion del modelo -0.3023632432194343
Coeficiente de correlacion del modelo 0.5498756615994513
```

La gráfica muestra que existe una correlación positiva entre la edad y el número de casas, las personas mayores han tenido más casas, quedando su modelo matemático como: y = (-0.489 * x *1 + 11.473) / 2.653 * x * 2

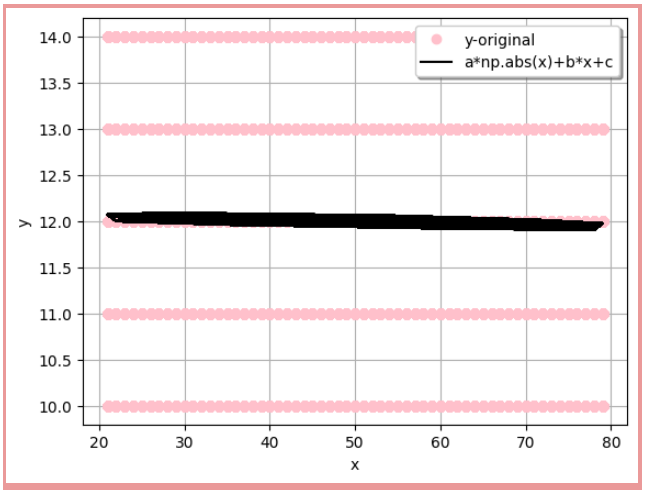
Current Job Yrs-Age I Cociente entre polinomios




```
Coeficiente de determinacion del modelo -0.06664890547446922
Coeficiente de correlacion del modelo 0.2581644930552403
```

La gráfica nos muestra la relación entre la variable dependiente e independiente de trabajo actual con respecto a la edad, podemos observar la figura que se forma en forma de curva invertida, el modelo mejoro hasta que la potencia se elevó dos veces, llegando a un resultado de coeficiente de correlación casi a la mitad comparado con el pasado, y=(- 2.985 * x ** 2 +2.731) / 1.392 * x

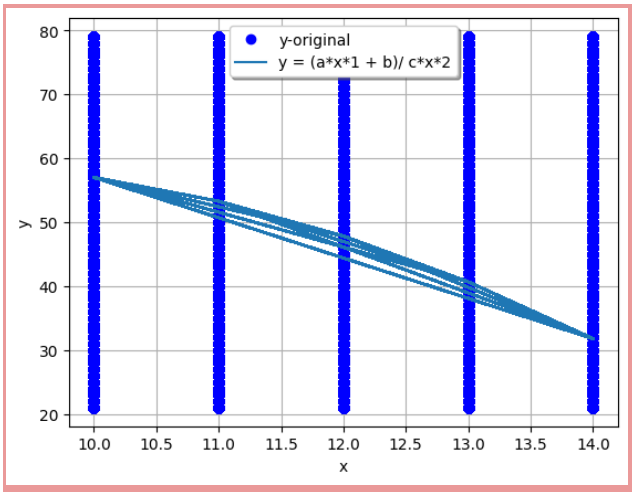
Current House Yrs-Age I Funcion cuadrática




```
Coeficiente de determinacion del modelo 0.0009529580469859189
Coeficiente de correlacion del modelo 0.030870018577673693
```

En este caso, la función que mejor se ajustaba a estos datos con respecto a si la persona tiene una casa actualmente con respecto a la edad nos indica que, entre mayor edad, más viviendas llegan a tener propias, el coeficiente de correlación resulto un poco bajo, pero no tanto como lo teníamos en un inicio en el mapa de calor (0.019), el modelo matemático resulto mejor hasta elevarlo a la 20, quedando de la siguiente manera: y= 6.97 * x **20 -1.190 *x + 1.206

Age- Current House Yrs- I Regresión no lineal



Para este modelo se tomó como variable objetivo Age y como variable independiente current house years, para este caso se utilizó el método no lineal para poder comprobar cuál era el modelo que mejor resultado tenía al momento de realizar el coeficiente de correlación, nuestro modelo matemático quedo de la siguiente manera: y= (-0.489 * x * 1 + 11.47) / 2.652 * x * 2

TABLA COMPARATIVA REGRESIÓN LINEAL SIMPLE

VARIABLES	COEFICIENTE DE CORRELACIÓN	MODELO MATEMÁTICO REGRESIÓN LINEAL SIMPLE
Experience / Current Job Years	0.6461	y= 1.06x + 3.35
Experience / Risk Flag	0.0345	y= -0.63x + 10.16
Age / Risk Flag	0.0218	y= -1.13x + 50.09

TABLA COMPARATIVA REGRESIÓN LINEAL MÚLTIPLE

VARIABLES	COEFICIENTE DE CORRELACIÓN	MODELO MATEMÁTICO
Income	0.00842	y= -130.44x + 1513.74x + 3922.85x + -5170.61x + -25640.99x + 5028710.48
Age	0.030	y= -0.00x + -0.01x + 0.02x + -0.25x + -1.14x + 53.06
Experience	0.647	y= 0.00x + -0.00x + 1.06x + 0.07x + -0.43x + 2.63
Current Job Years	0.646	y=0.00x + 0.00x + 0.39x + -0.02x + 0.06x + 2.54
Current House Years	0.0298	y=-0.00x + -0.00x + 0.01x + -0.00x + -0.02x + 12.05
Risk Flag	0.0418	y= -0.00x + -0.00x + -0.00x + 0.00x + -0.00x + 0.17

TABLA COMPARATIVA REGRESIÓN NO LINEAL

VARIABLES	FUNCIÓN UTILIZADA	FUNCIÓN UTILIZADA	COEFICIENTE CORRELACIÓN	MODELO MATEMÁTICO
x= Age y= Income	Cociente entre Polinomios	(a*x**16+ b)/c*x	0.562	y=(6.269 * x **16 + 3.131) / 3.497 * x
x= Experience y= Income	Cociente entre Polinomios	(a*x**16+ b)/c*x	0.833	y= (3.79 * x **16 + 1.039) / 2.54 * x
x= Age y= Risk Flag	Cociente entre Polinomios	(a*x**16+ b)/c*x	0.140	y= (3.017 * x ** 16 +8. 202) / 3.798 * x
x= Current House Years y= Age	Cociente entre Polinomios	(a*x**11+ b)/c*x	0.164	y= (-2.411 * x ** 11 + 3.746) / 8.134 * x
x= Current House Years y=Experience	Función Inversa	1 / a*x	0.201	y= 1 / 1.137 * x
x= Current Job Years y=Experience	Función de Valor Absoluto	a * np.abs (x)+b*x+c	0.646	y= -99.935 *np.abs(x) + 100.999 * x + 3.349
x= Current House Years y= Age	Cociente entre Polinomios	(a * x * 1 +b) / c * x * 2	0.550	y = (-0.489 * x *1 + 11.473) / 2.653 * x * 2
x= Age y= Current Job Years	Cociente entre Polinomios	(a*x** 2+ b)/c*x	0.258	y=(- 2.985 * x ** 2 +2.731) / 1.392 * x
x= Age y= Current House Years	Función Cuadrática	a * x **20 + b * x + c	0.031	y= 6.97 * x **20 -1.190 *x + 1.206
x= Current House Years	Cociente entre Polinomios	(a * x * 1 + b) / c * x * 2	0.550	y= (-0.489 * x * 1 + 11.47) / 2.652 * x * 2

y= Age				

TABLA COMPARATIVA CORRELACIONES

VARIABLES	TIPO DE REGRESIÓN	COEFICIENTE DE CORRELACIÓN.
Experience / Current Job Years	Regresión Lineal Simple	0.6461
Experience	Regresión Lineal Múltiple	0.647
x= Experience y= Income	Regresión No Lineal	0.833