ANÁLISIS DE DESERCIÓN DE CLIENTES

OBJETIVO

- Usar la Ciencia de Datos para identificar características que influyan en la deserción de clientes (Churning)
- También se buscarán Insights que puedan ayudar a dar pautas para optimizar el servicio existente o acciones para proceder

OBJETIVO PREGUNTA

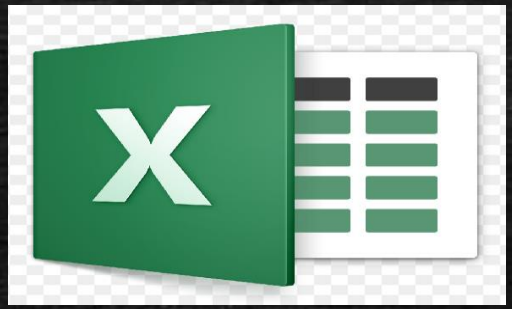
- ¿Se puede generar un modelo de clasificación que pueda detectar qué clientes desertarán y cuales son las variables que tienen más influencia en ello?

CONTEXTO

- Vivimos en un mundo muy demandante y con muchas opciones a la hora de elegir un servicio
- La deserción de clientes es un tema que está presente
- Hay una gran cantidad de variables y factores que pueden influir en esta determinación por parte de los clientes

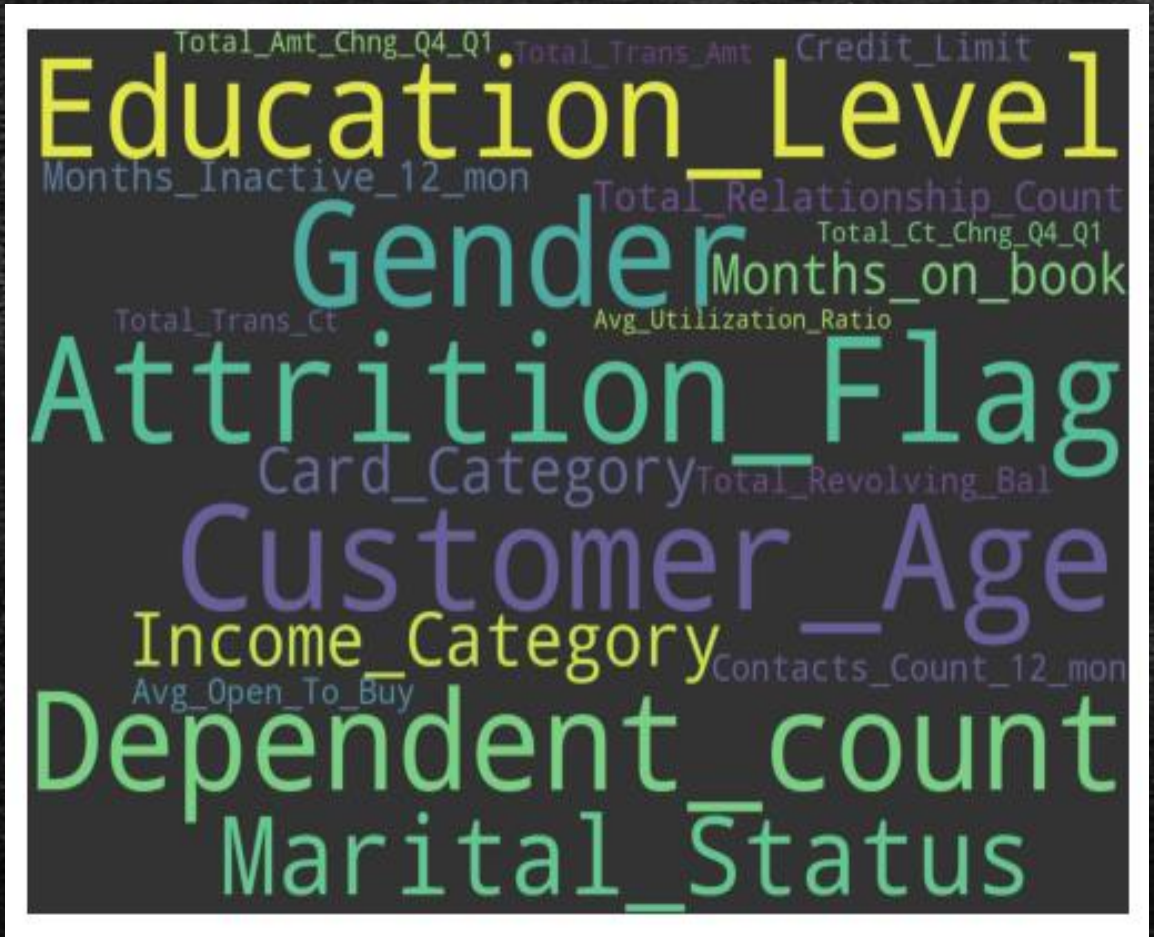
DATOS GENERALES

- El set de datos "Credit_card_churn.xlsx" obtenido de Kaggle cuenta con 10.127 filas y 23 columnas



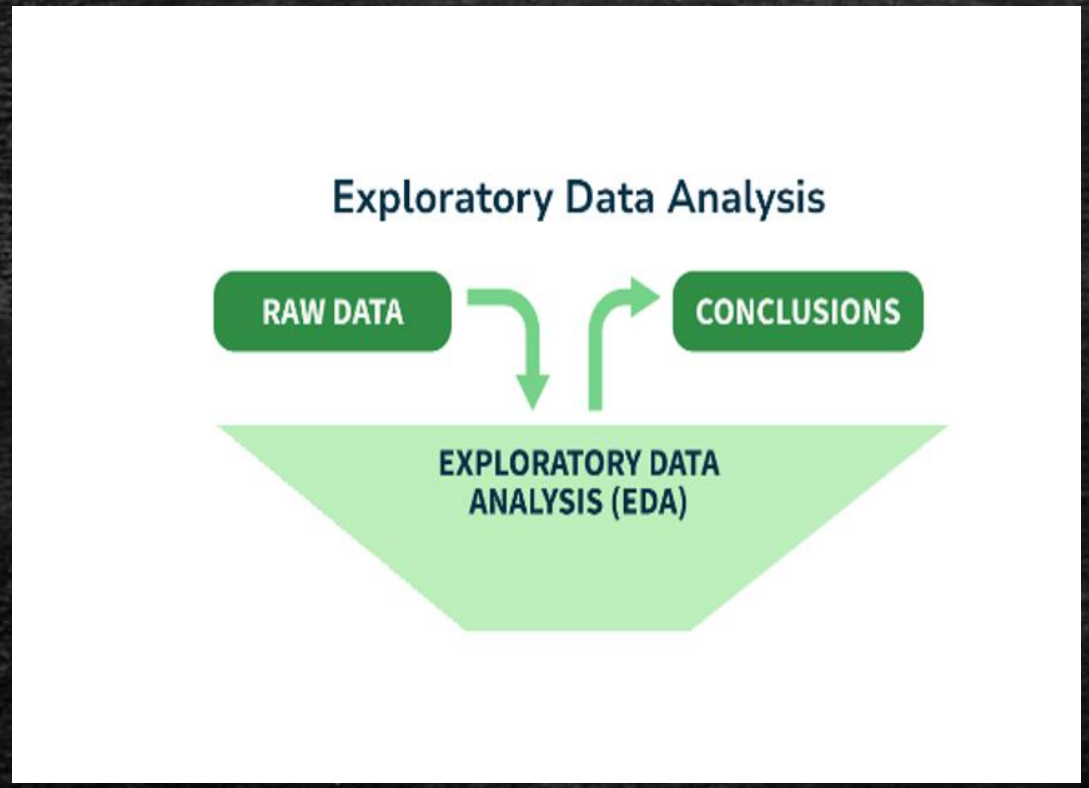
- El set de datos cuenta con la edad de los clientes, límite de crédito, si han sido contactados dentro de cierto tiempo, nivel de educación y si abandonaron o no

COMPOSICIÓN DEL SET DE DATOS



- Se visualizó que no habían datos Nulos
- Se buscó agrupar las categorías para ver cuántos valores tenía cada una
- Se hizo limpieza e imputación en las columnas en que fue necesario

ANÁLISIS EXPLORATORIO DE DATOS



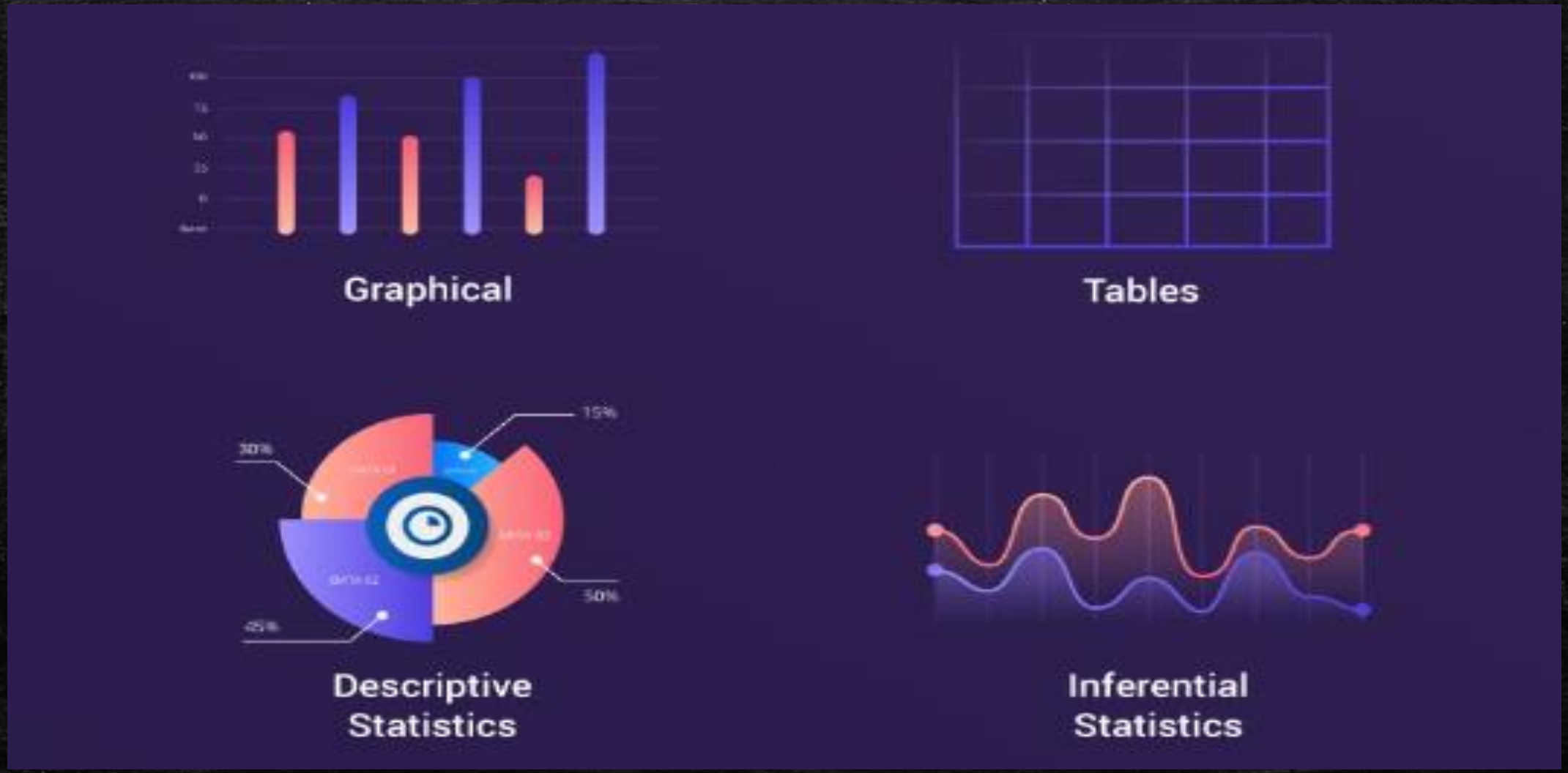
ANÁLISIS UNIVARIADO

ANÁLISIS BIVARIADO

ANÁLISIS MULTIVARIADO

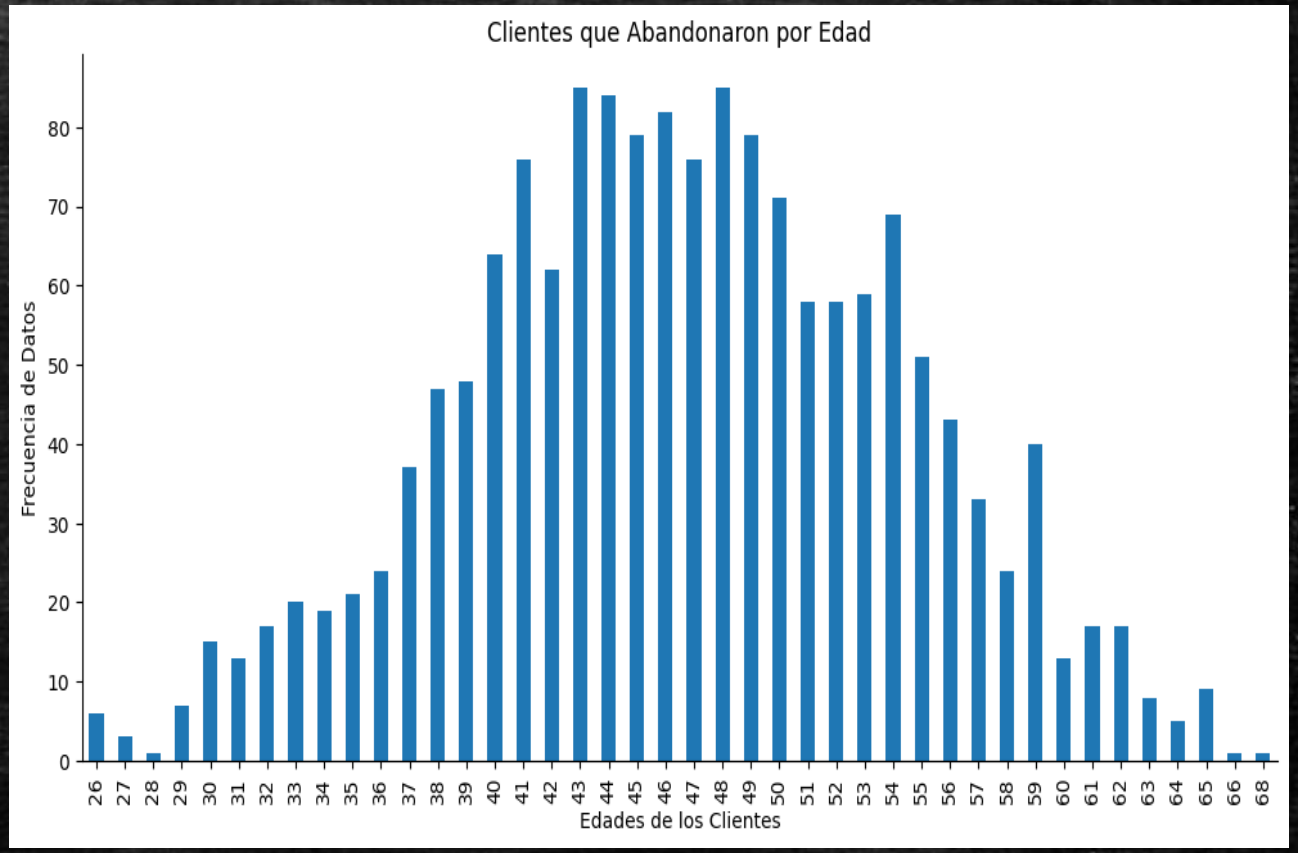
BÚSQUEDA DE CORRECIONES

ANÁLISIS UNIVARIADO



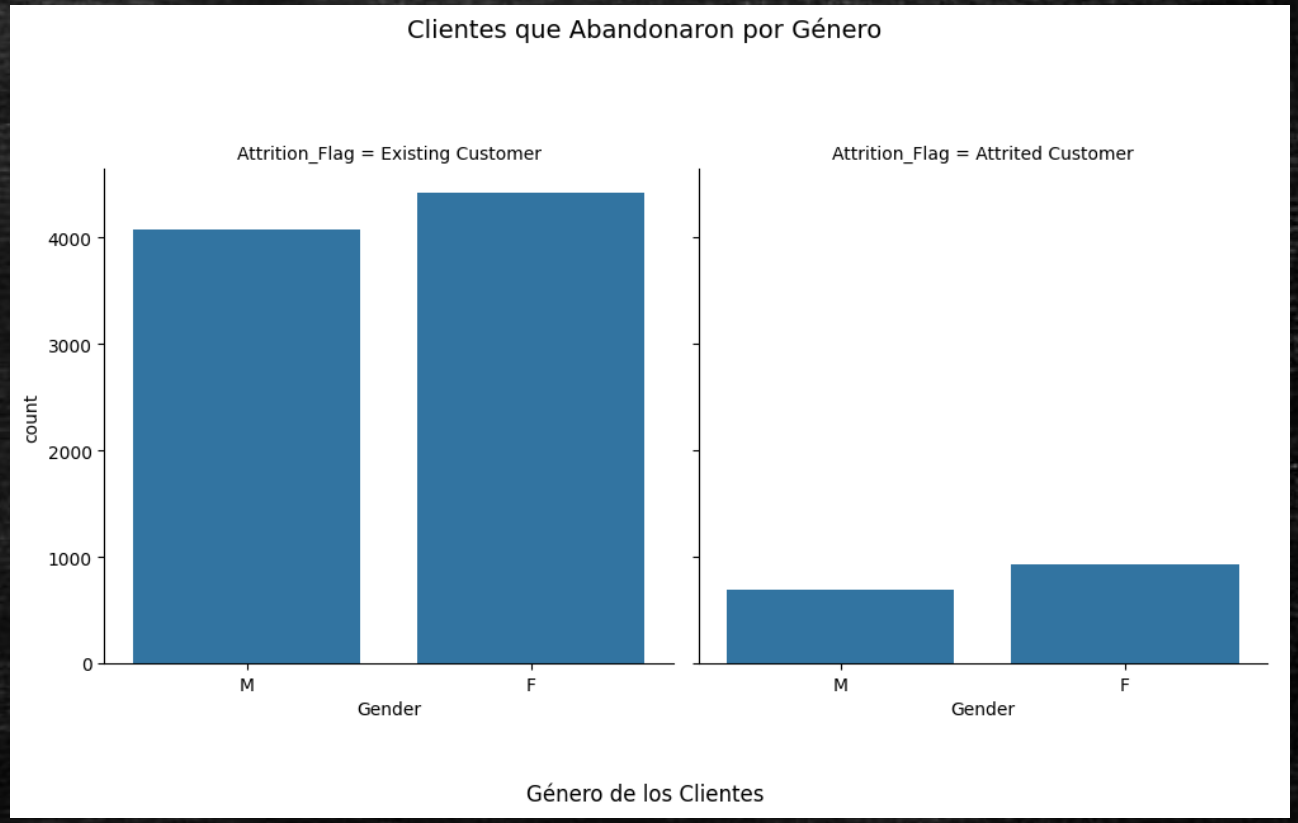
¿HAY ALGUNA DIFERENCIA SOBRSALIENTE EN EL GÉNERO QUE PUEDA ESTAR RELACIONADA A LA DESERCIÓN?

- Se puede observar una distribución normal
- Los clientes son de edades variadas entre 26 y 73 años
- Por lo que se puede observar la edad no pareciera ser de gran incidencia



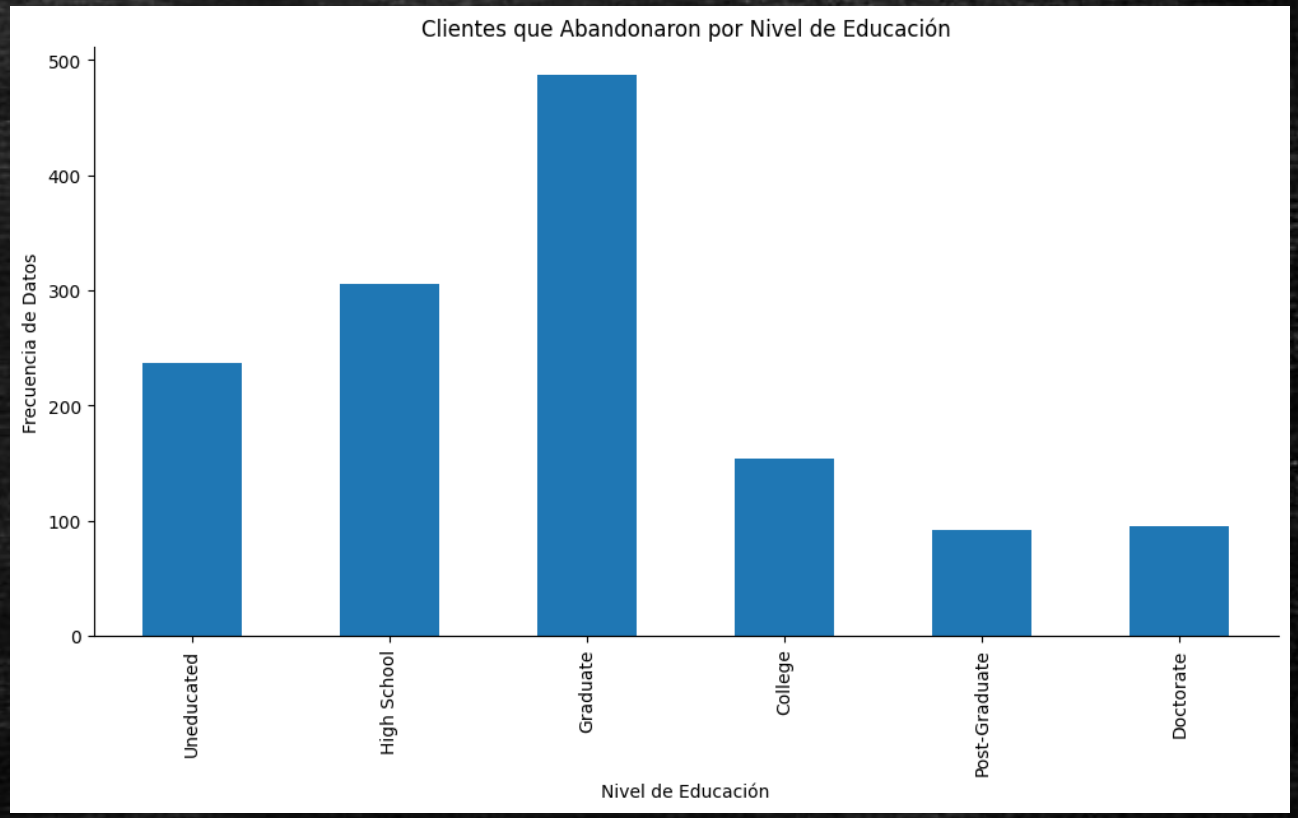
¿HAY ALGUNA DIFERENCIA SOBRESALIENTE EN EL GÉNERO QUE PUEDA ESTAR RELACIONADA A LA DESERCIÓN?

- A la izquierda están los clientes existentes, a la derecha los clientes que han abandonado
- Se observa que la cantidad de mujeres que han abandonado es un poco mayor
- 930 Mujeres vs 697 Hombres



¿CÓMO AFECTA EL NIVEL DE EDUCACIÓN DEL CLIENTE A LAS TASAS DE DESERCIÓN?

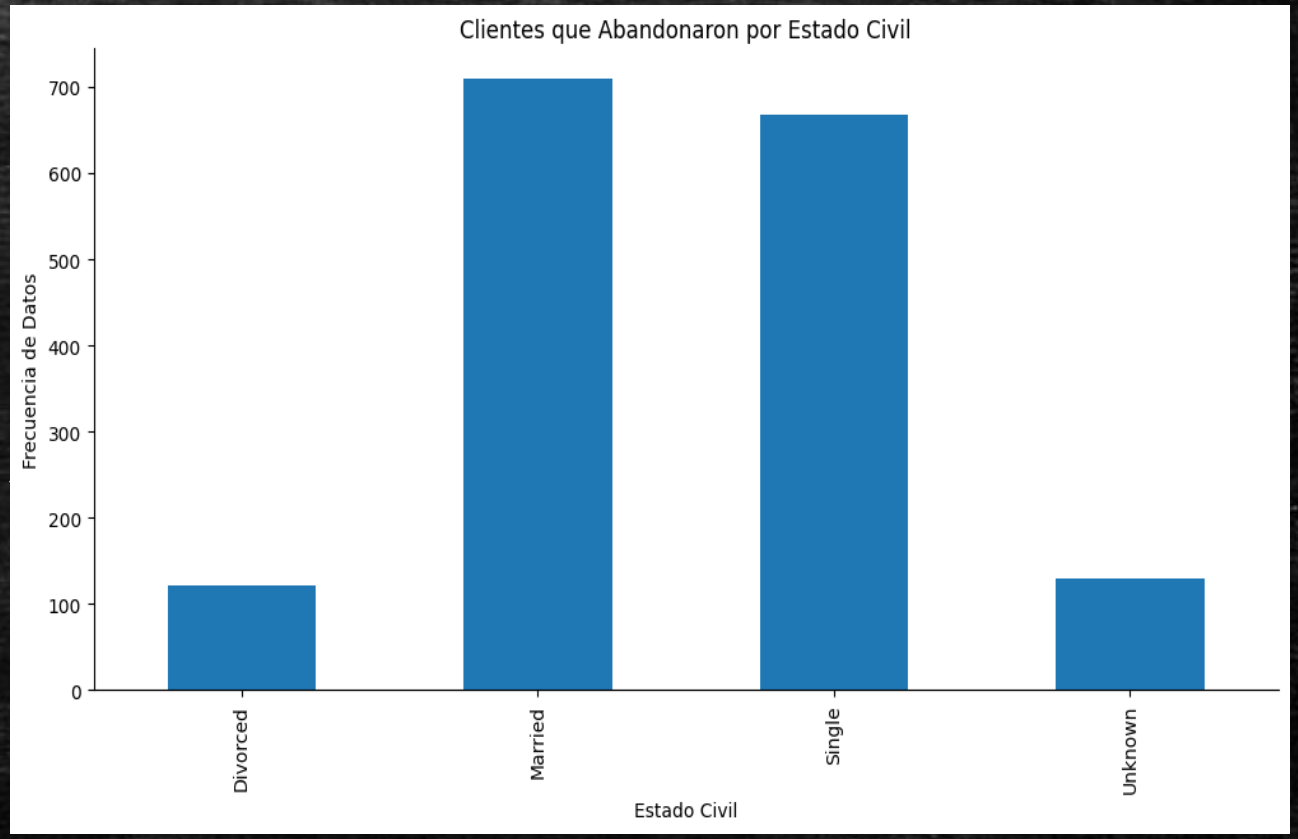
- Se observa que la mayor Cantidad de Clientes que abandonaron con Graduados: 487, seguidos por High School: 306.
- Puede ser que al tener mejor educación, se tenga mayores ingresos y puede ser mejores hábitos crediticios



¿AFECTA EL ESTADO CIVIL DEL CLIENTE A LAS TASAS DE DESERCIÓN?

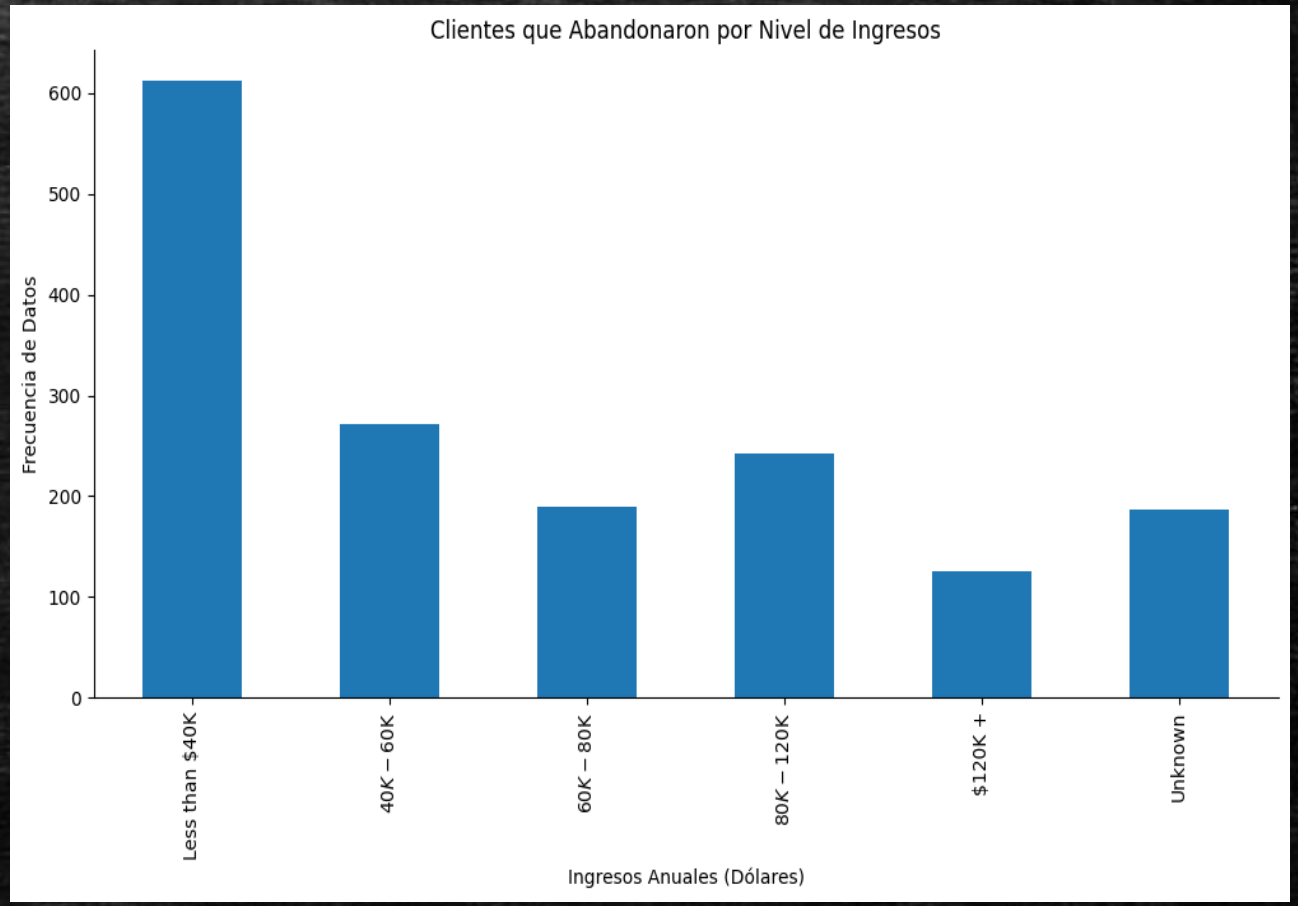
- Se puede observar que la mayor parte está repartida entre casados y solteros

- No se observa que haya una gran diferencia entre solteros casados que por el momento pueda servir de Insight

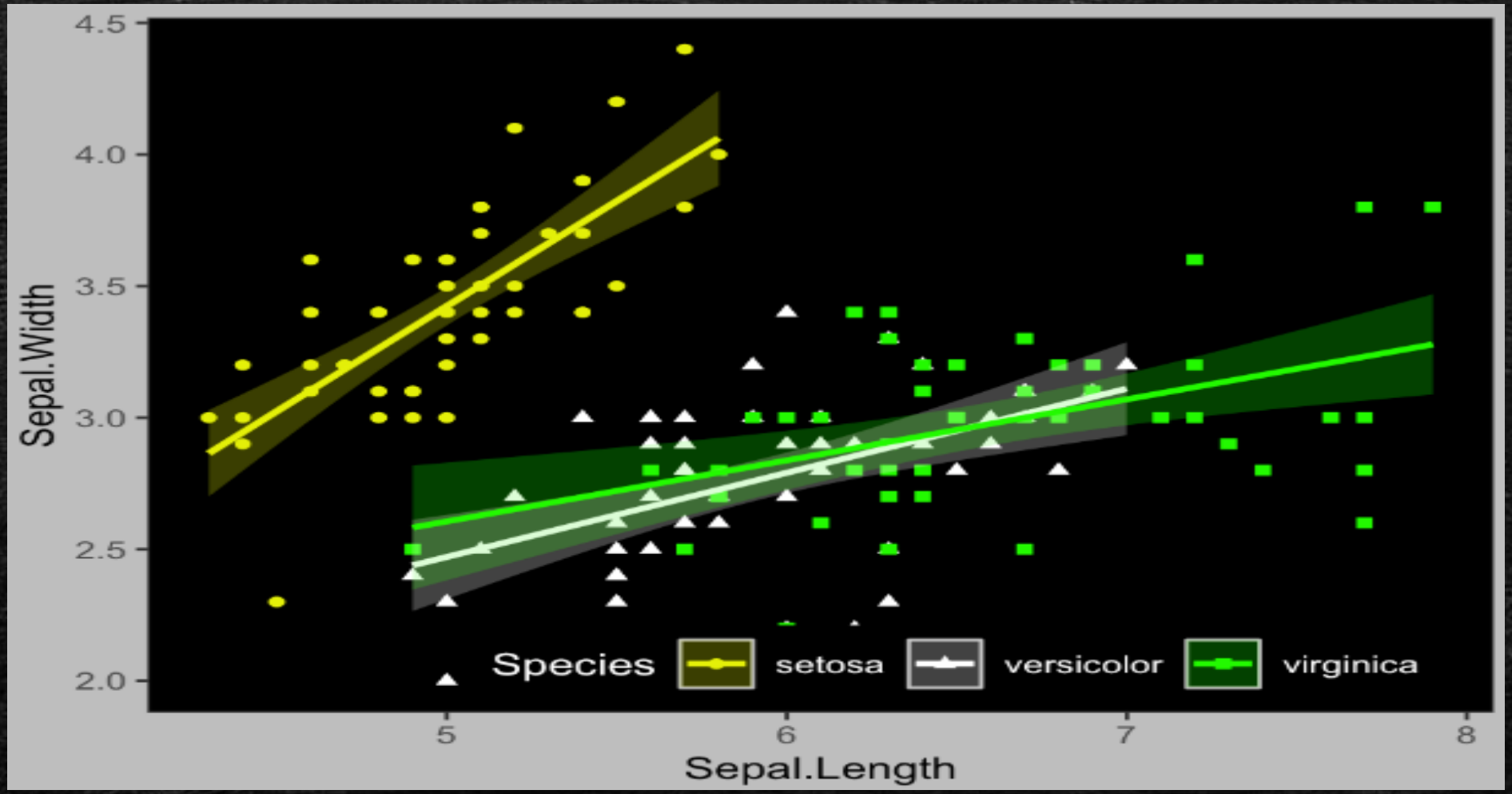


¿CÓMO AFECTA EL NIVEL DE INGRESOS DEL CLIENTE A LAS TASAS DE DESERCIÓN?

- Se observa que las personas que ganan menos de 40 mil dólares al año casi doblan a las otras categorías.
- Puede que sean más sensibles al cambio de precio en los productos, puede ser debido a problemas financieros, que no le encuentren el verdadero valor que ofrece el servicio



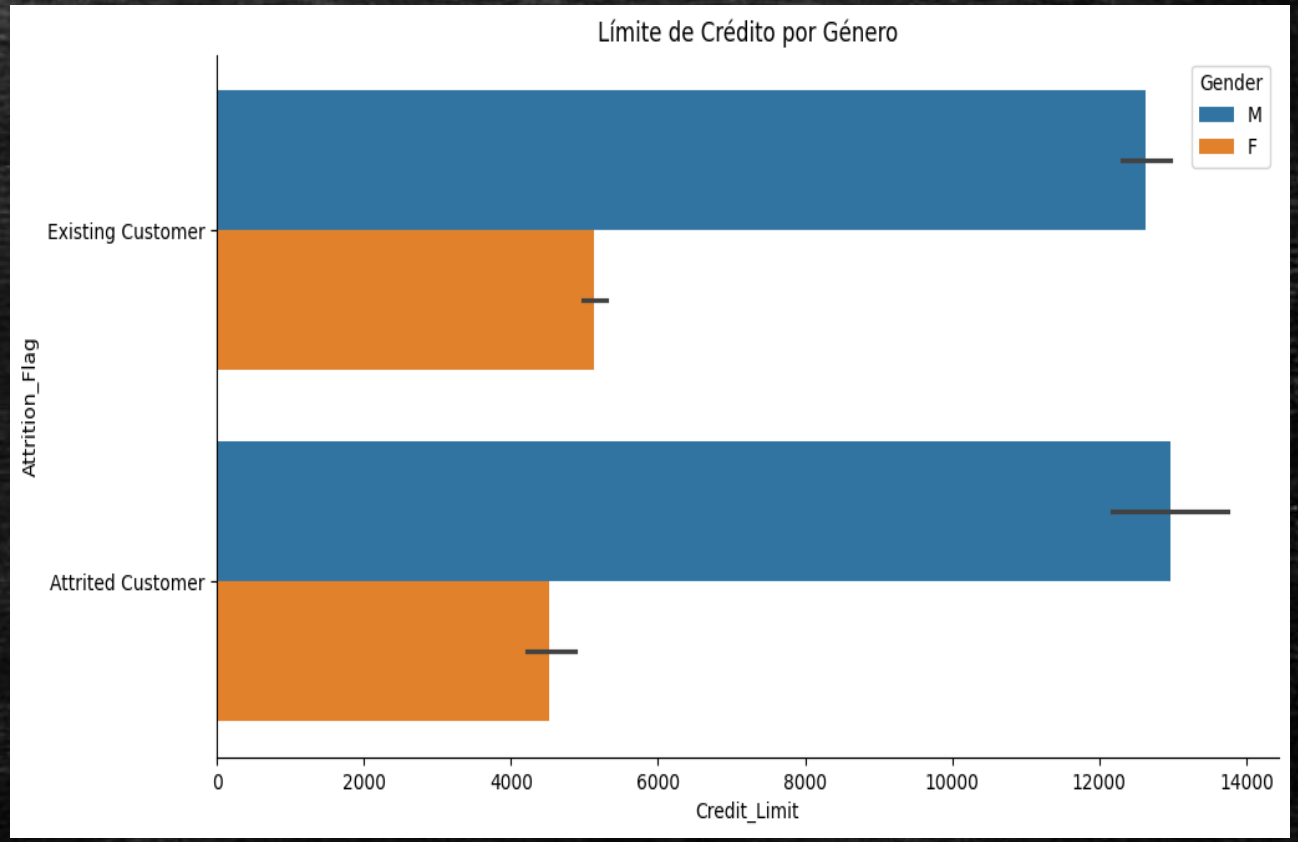
ANÁLISIS BIVARIADO



¿TIENE ALGUNA RELEVANCIA EL GÉNERO EN EL LÍMITE DE CRÉDITO?

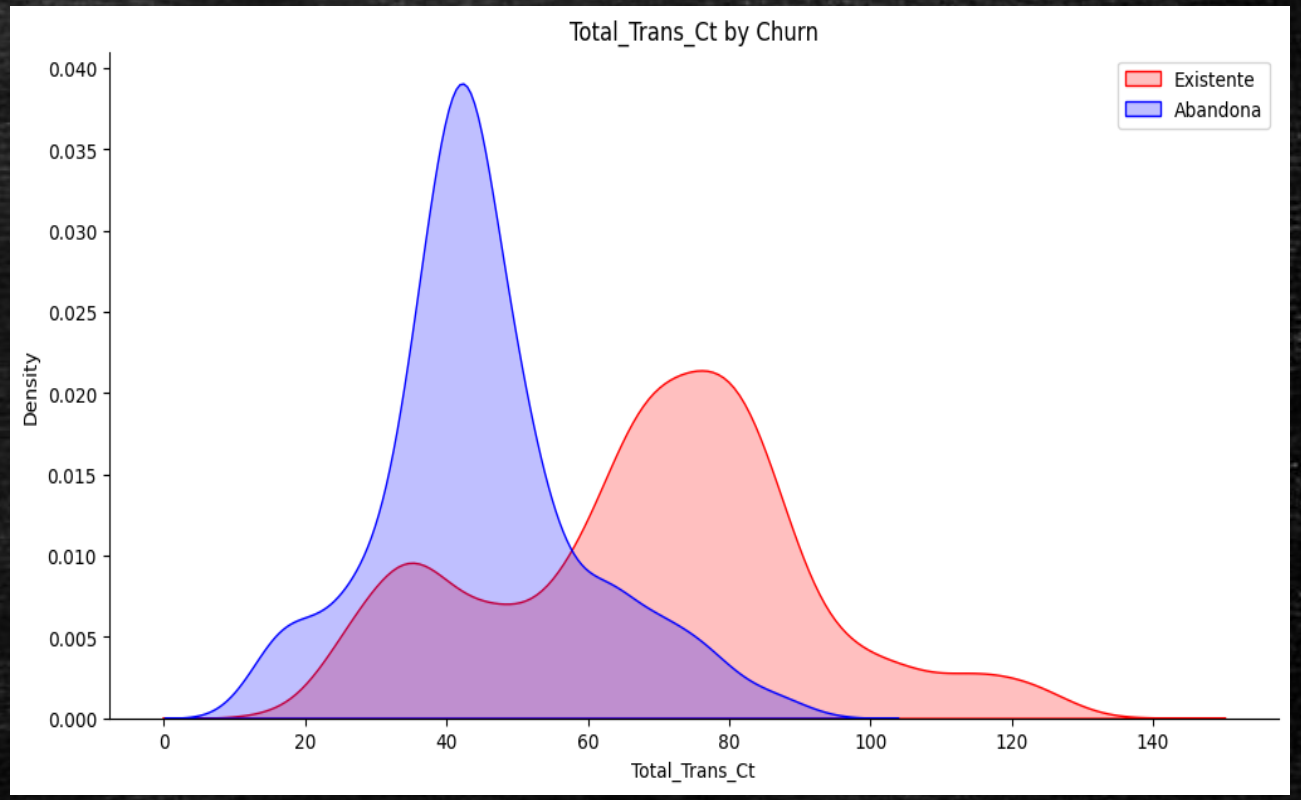
- Se graficó dividiendo en género Clientes Existentes y Clientes que Abandonaron

- En los Hombres el Límite de Crédito es mayor en ambos casos



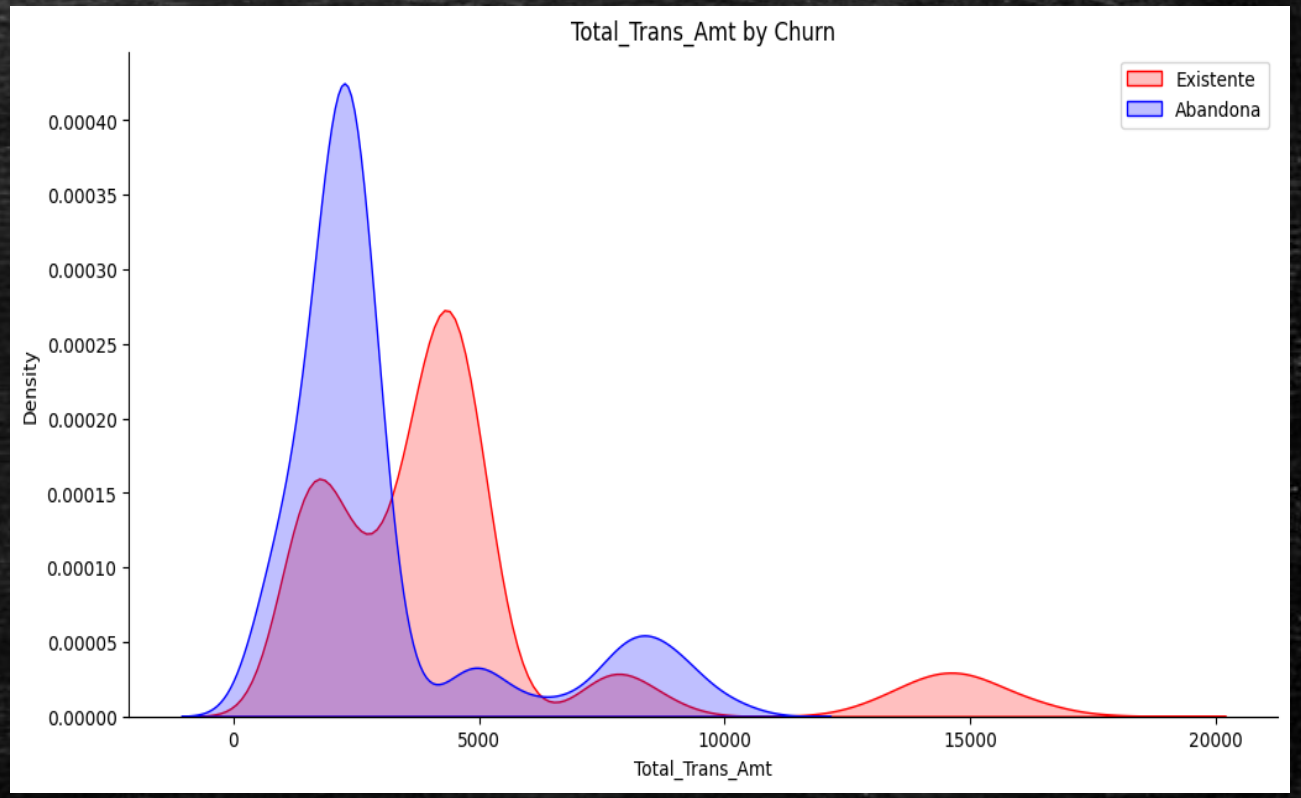
¿CÓMO AFECTA TOTAL_TRANS_CT A LAS TASAS DE DESERCIÓN?

- Se puede observar de forma gráfica que cuando las transacciones Totales son bajas entre 25 y 50 la probabilidad de que un cliente abandone es mucho más alta
- Se observa que entre más amenten las transacciones el riesgo de Abandonar disminuye



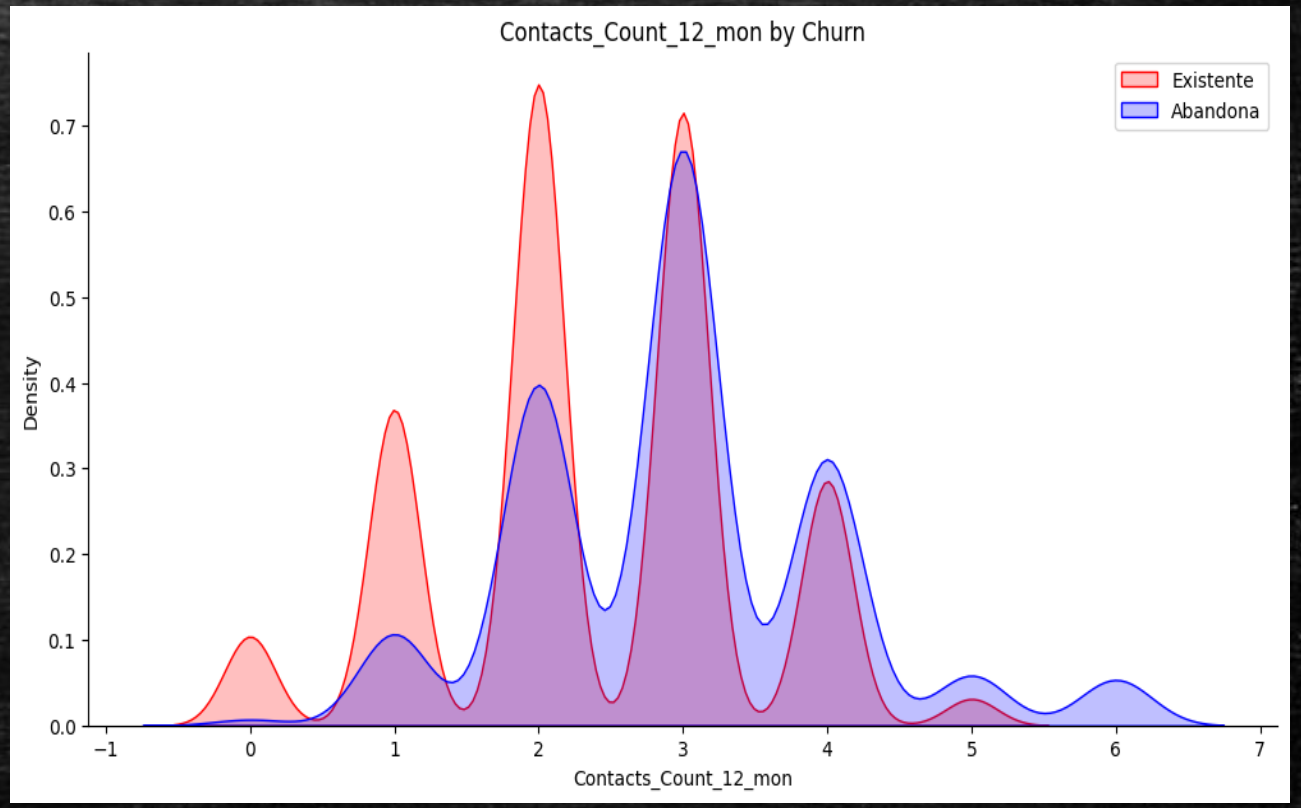
¿CÓMO AFECTA TOTAL_TRANS_AMT A LAS TASAS DE DESERCIÓN?

- Se puede observar que los clientes que han gastado menos en transacciones entre o a 4000 USD son más propensos a abandonar el servicio
- Podría ser que la cantidad de transacciones reflejada en monto luego de cierto límite cree una especie de lealtad



¿CÓMO AFECTA CONTACTS_COUNT_12_MON A LAS TASAS DE DESERCIÓN?

- El gráfico sugiere que los clientes que han tenido menos contactos en los últimos 12 meses entre 2 a 4 meses, son más propensos a abandonar dentro de ese periodo de tiempo
- La correlación positiva implica que los clientes que han tenido más contactos con la empresa en los últimos 12 meses tienen menos probabilidades de abandonar



DISTRIBUCIONES

DISTRIBUCIONES NUMÉRICAS

- Distribución Normal:

  Customer_Age

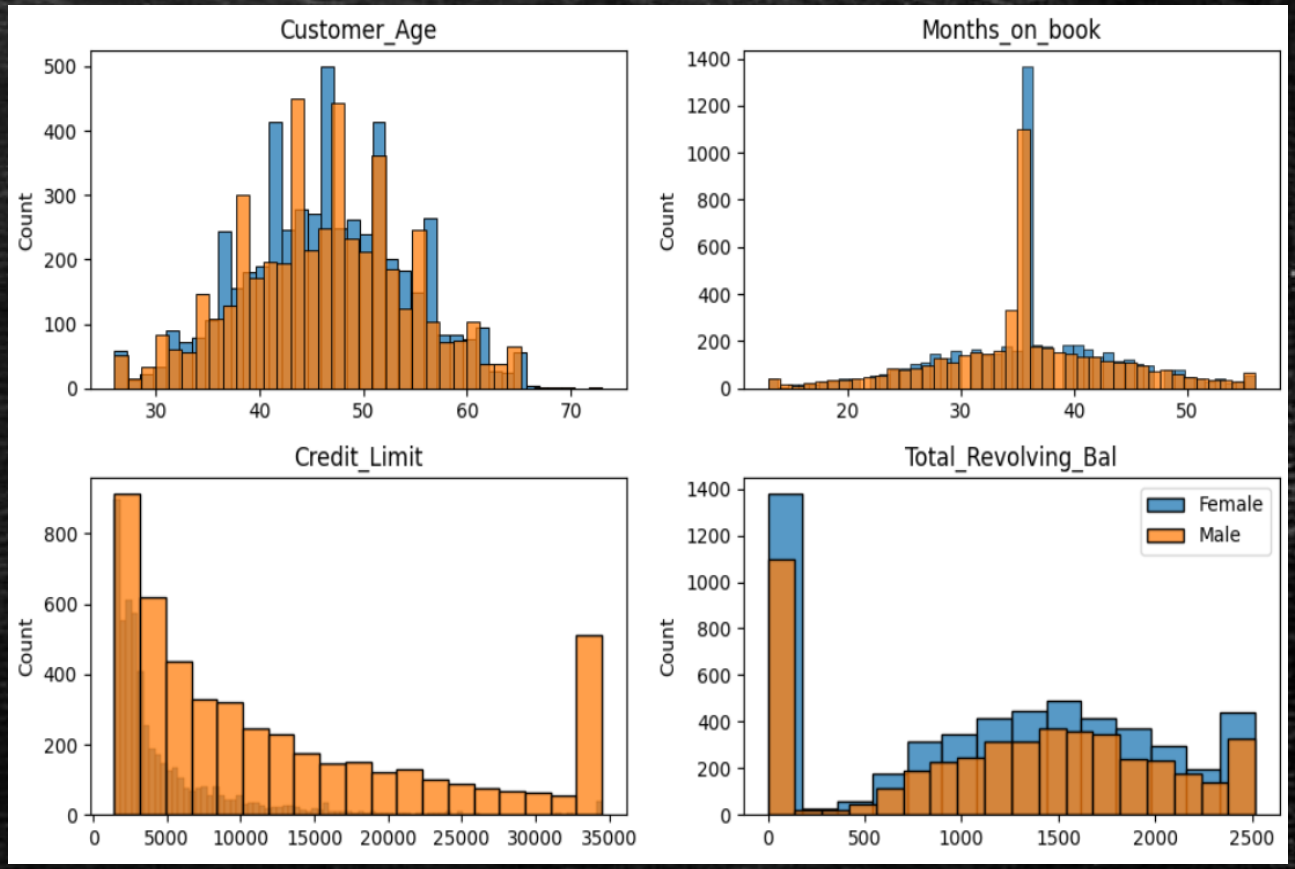
- Right Skewed:

  Credit_Limit,

  Total_Revolving_Bal

- Left Skewed:

  Months_on_book



DISTRIBUCIONES NUMÉRICAS

- Distribución Bimodal:

  Total_Trans_Ct,

  Total_Trans_Amt

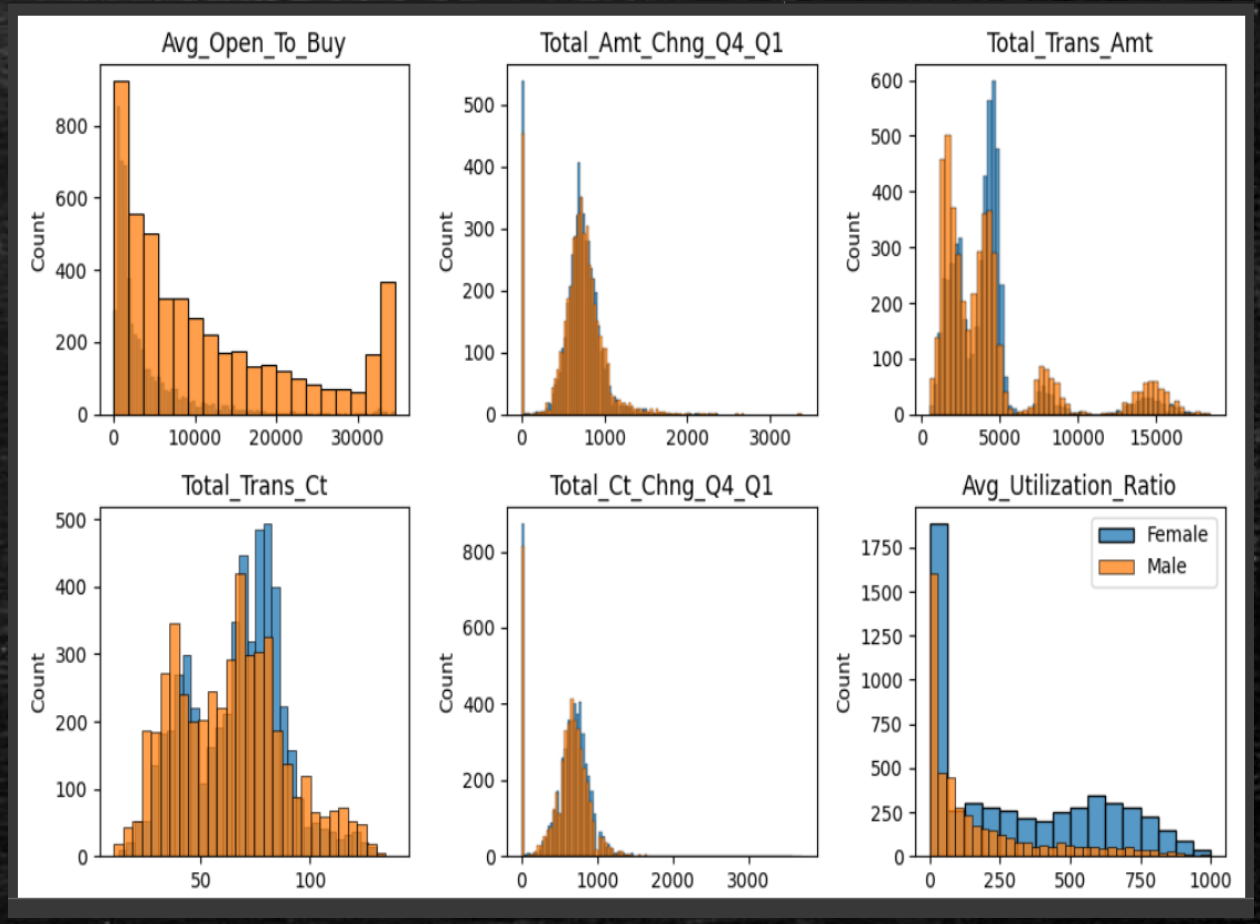
- Right Skewed:

  Avg_Open_To_Buy,

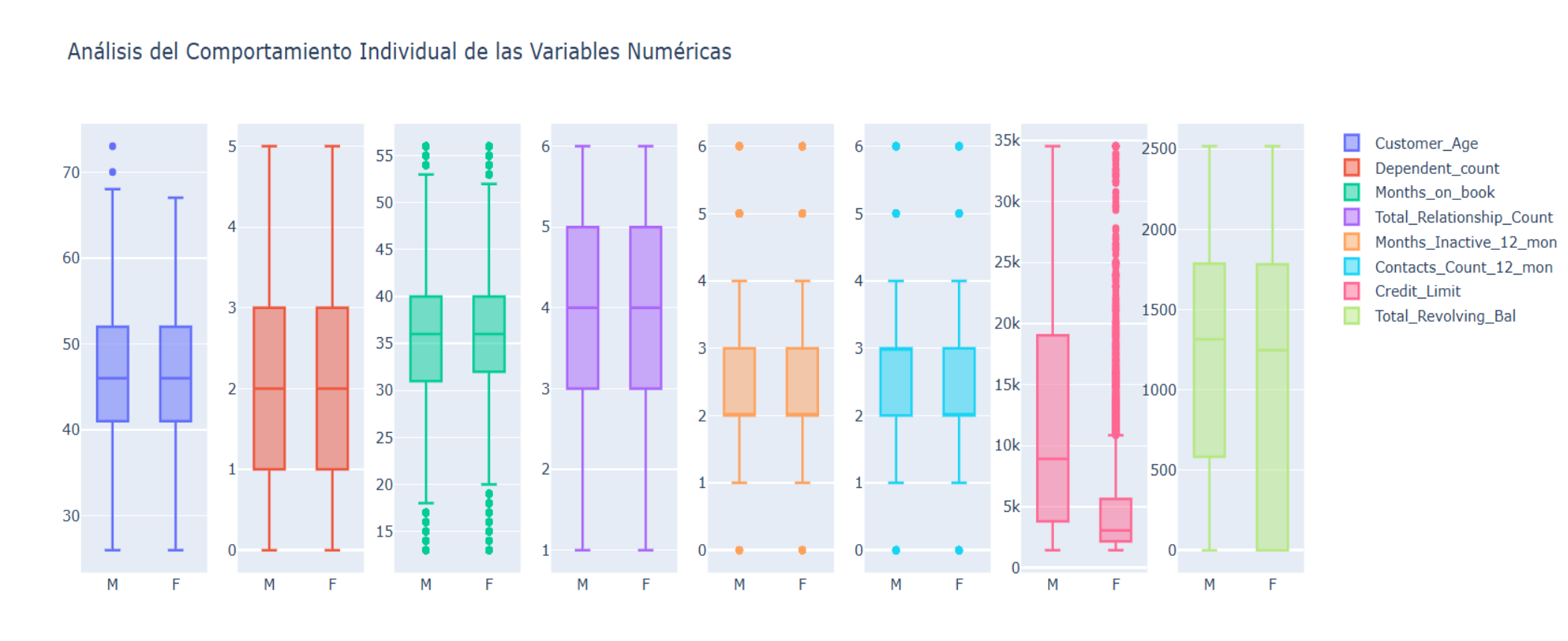
  Total_Amt_Chng_Q4_Q1,

  Total_Ct_Chng_Q4_Q1,

  Avg_Utilization_Ratio



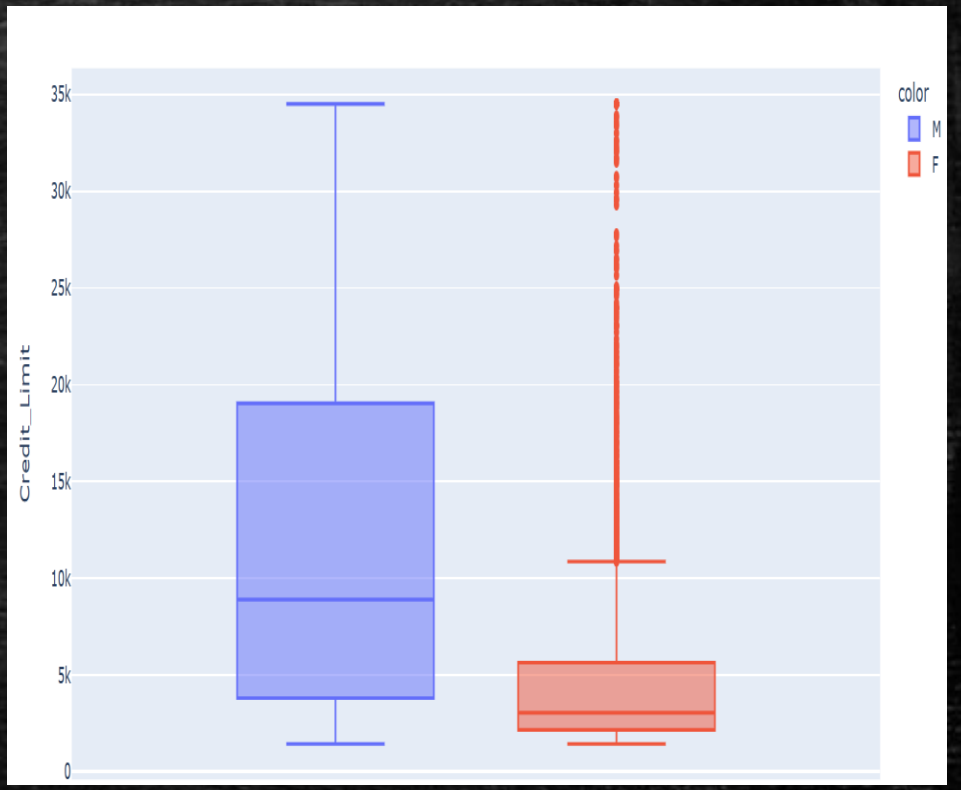
COMPORTAMIENTO INDIVIDUAL DE LAS VARIABLES NUMÉRICAS



COMPORTAMIENTO INDIVIDUAL DE LAS VARIABLES NUMÉRICAS

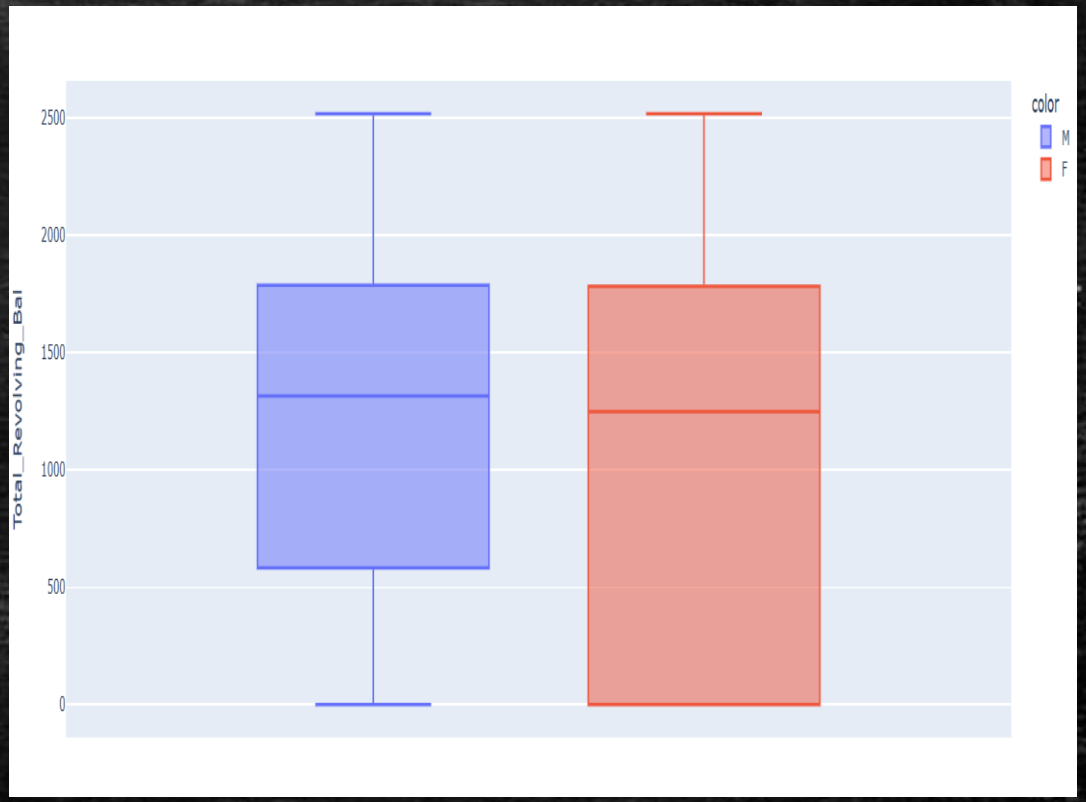
- El límite de crédito medio para las mujeres (F) es significativamente inferior al de los hombres (M). Esto indica que una clienta tiene un límite de crédito inferior en comparación con un cliente masculino.

- La distribución de los límites de crédito para las mujeres es más amplia que para los hombres, lo que sugiere una mayor variabilidad en los límites de crédito entre las clientas.

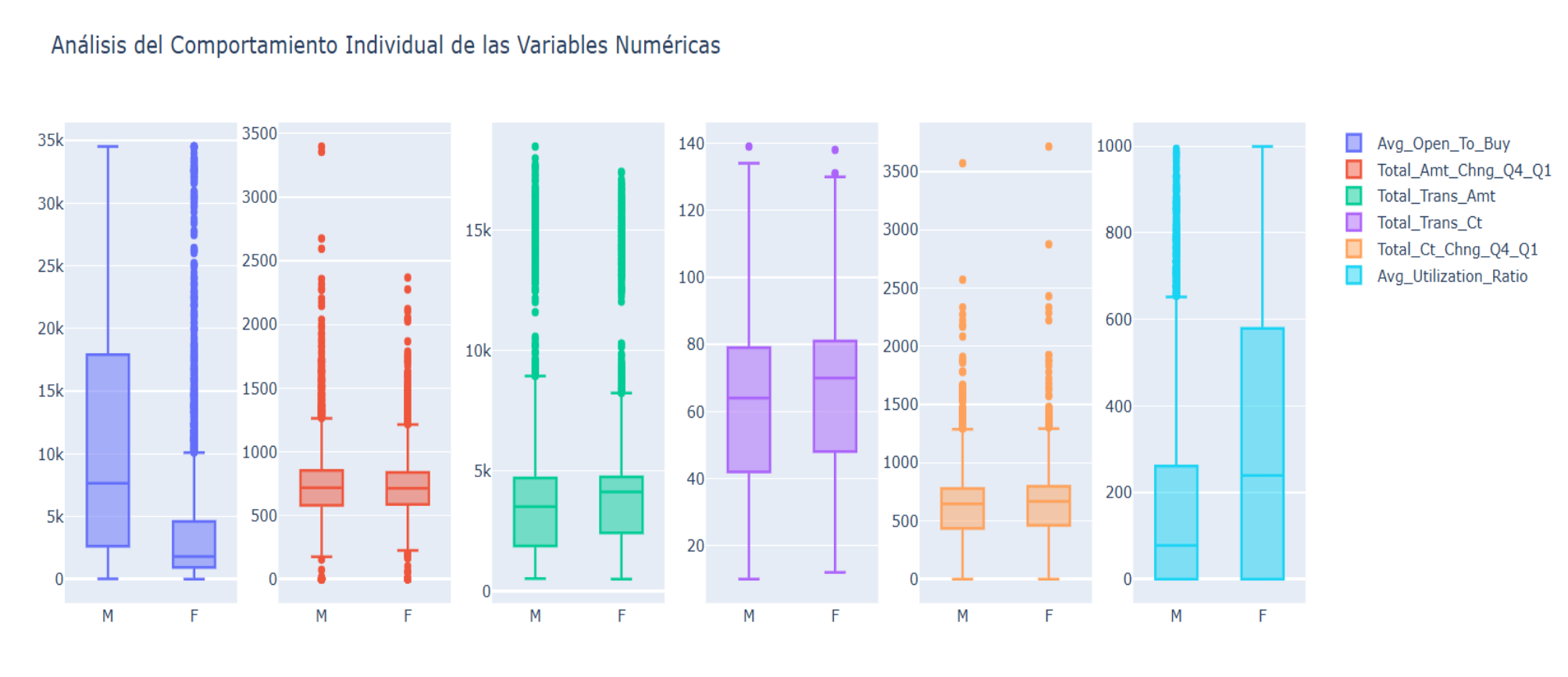


COMPORTAMIENTO INDIVIDUAL DE LAS VARIABLES NUMÉRICAS

- En promedio 'Total_Revolving_Bal' de los hombres es significativamente mayor que el de las mujeres. Esto indica que el cliente masculino típico tiene un saldo pendiente más alto en cuentas de crédito en comparación con el cliente femenino típico.
- Los saldos de los hombres y las mujeres se distribuyen de manera similar, y la mayoría de los clientes se encuentran dentro del rango intercuartil



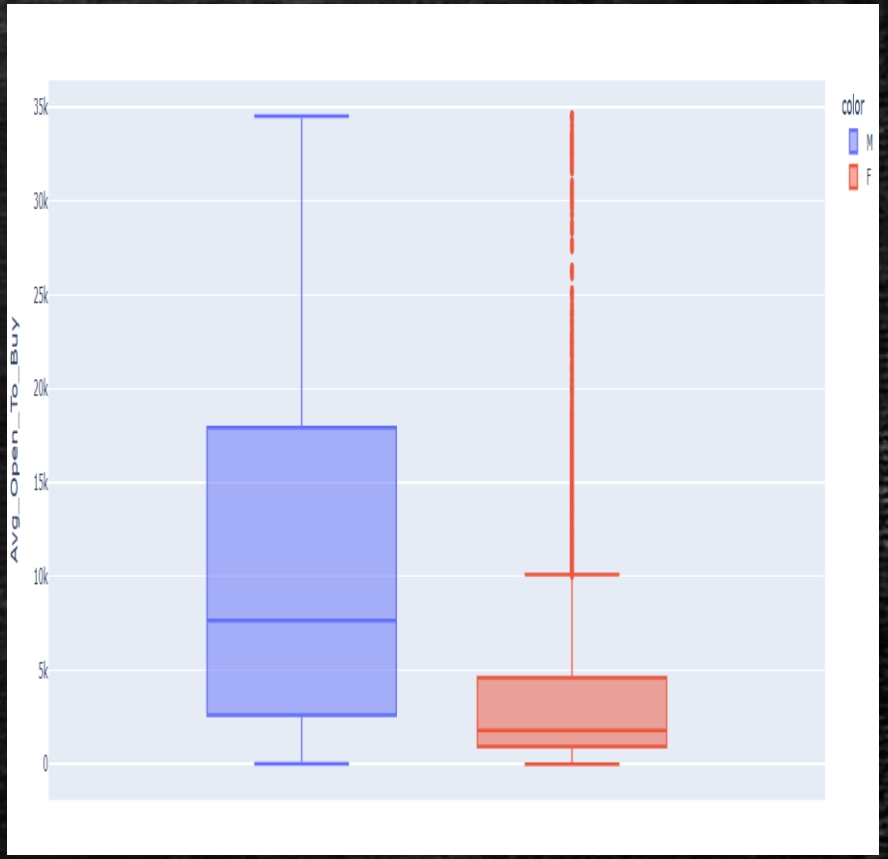
COMPORTAMIENTO INDIVIDUAL DE LAS VARIABLES NUMÉRICAS



COMPORTAMIENTO INDIVIDUAL DE LAS VARIABLES NUMÉRICAS

- La media de 'Avg_Open_To_Buy' para las mujeres es significativamente menor que la de los hombres. Esto indica que la clienta típica tiene un límite de crédito disponible más bajo en comparación con el cliente típico masculino.

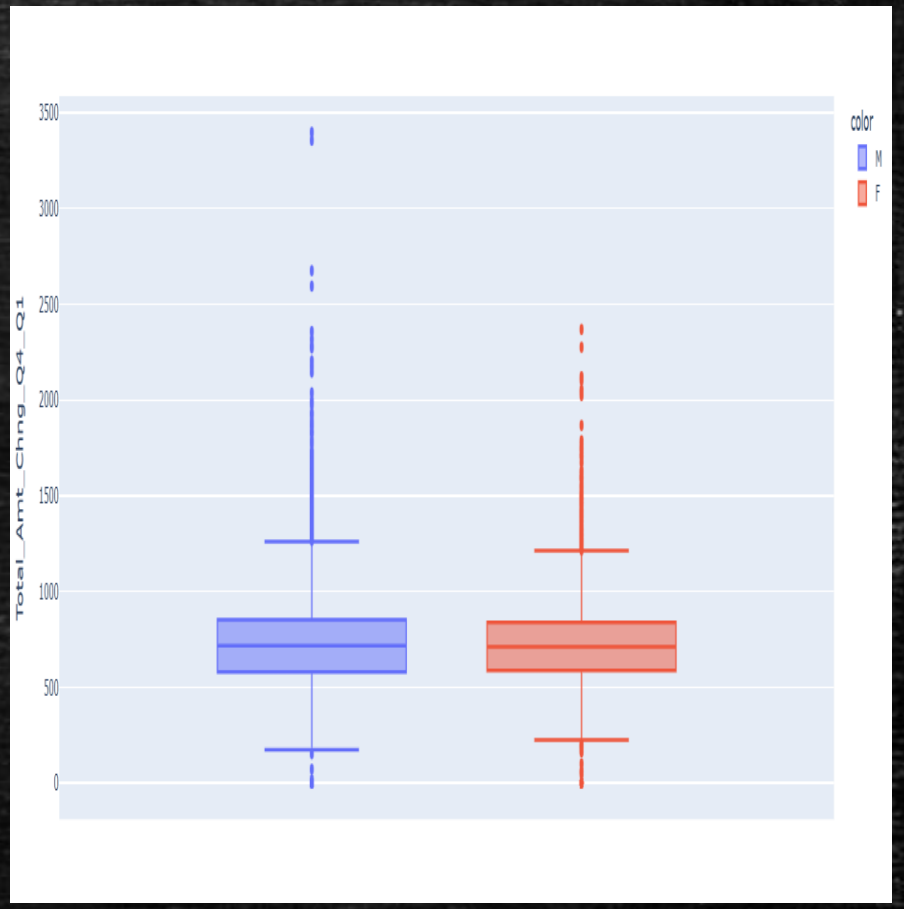
- La concentración de casos atípicos es significativamente mayor en el caso de las mujeres. Se trata de clientes con límites de crédito disponibles muy altos en comparación con la mayoría.



COMPORTAMIENTO INDIVIDUAL DE LAS VARIABLES NUMÉRICAS

- En 'Total_Amt_Chng_Q4_Q1' la presencia de más valores atípicos en el grupo masculino podría indicar cambios más extremos en los hábitos de gasto de algunos clientes del género.

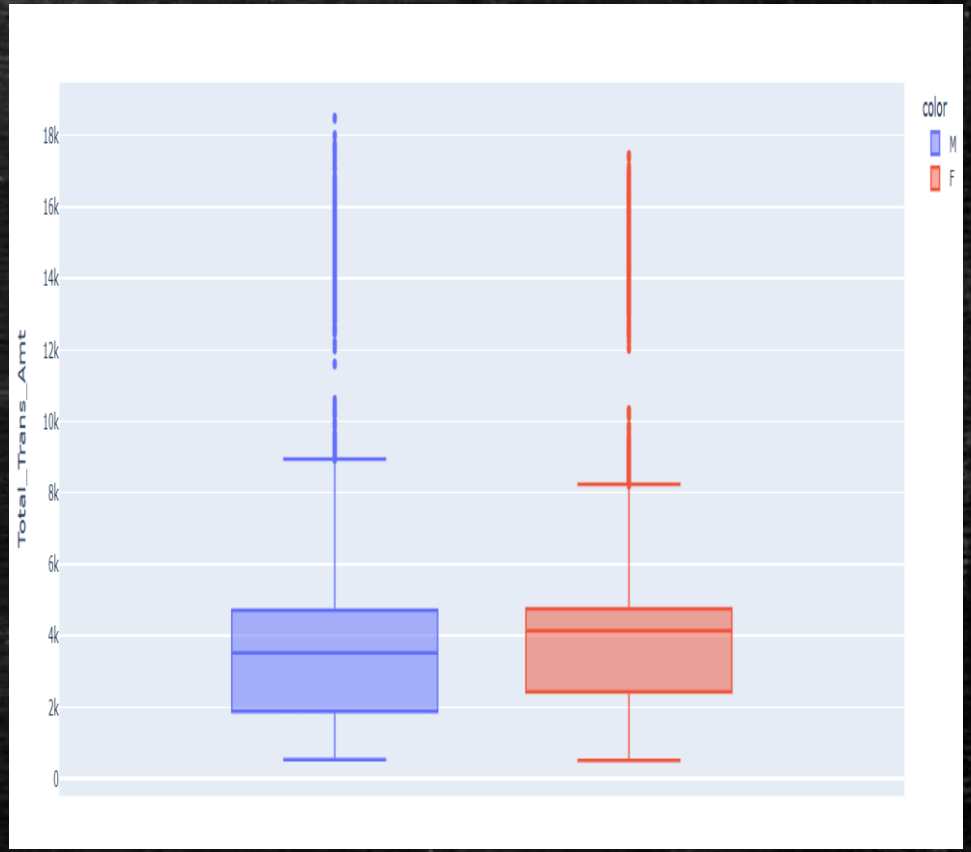
- A pesar de la diferencia en los valores atípicos, la distribución en general para ambos géneros es bastante similar, lo que sugiere que la tendencia en sí de 'Total_Amt_Chng_Q4_Q1' es comparable entre hombres y mujeres.



COMPORTAMIENTO INDIVIDUAL DE LAS VARIABLES NUMÉRICAS

- La distribución está sesgada hacia la derecha, lo que indica que hay más clientes con montos de transacción más bajos y menos clientes con montos más altos..

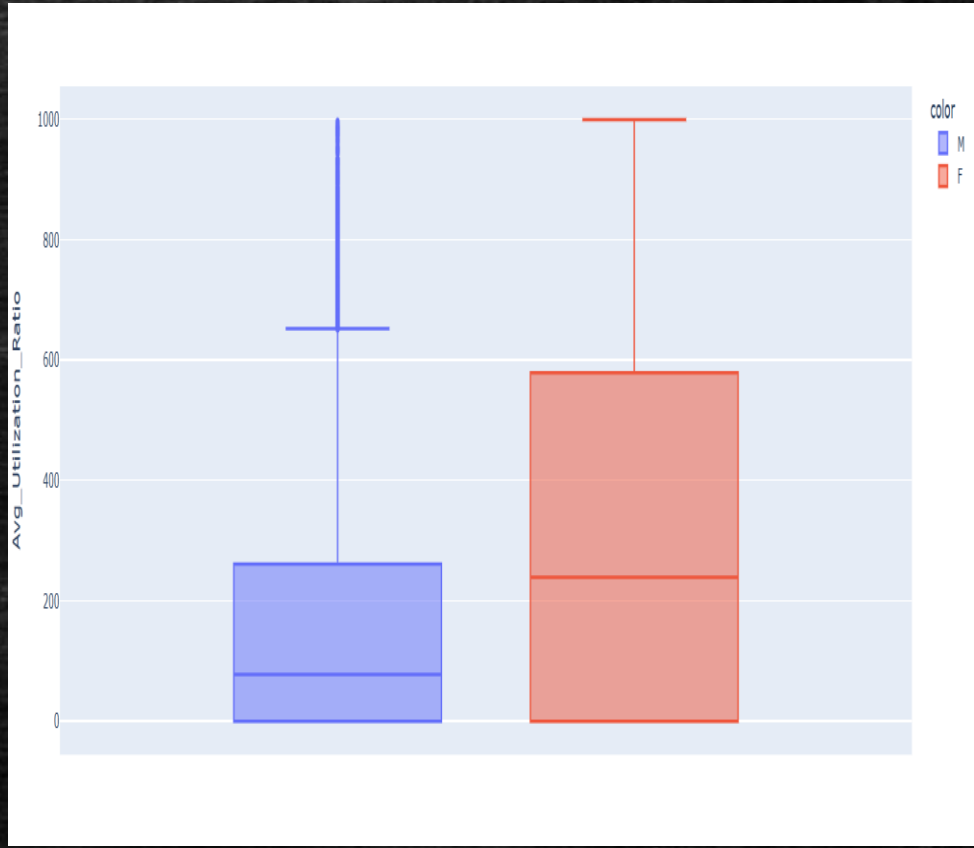
- En ambos géneros de 'Total_Trans_Amt' se observan outliers.



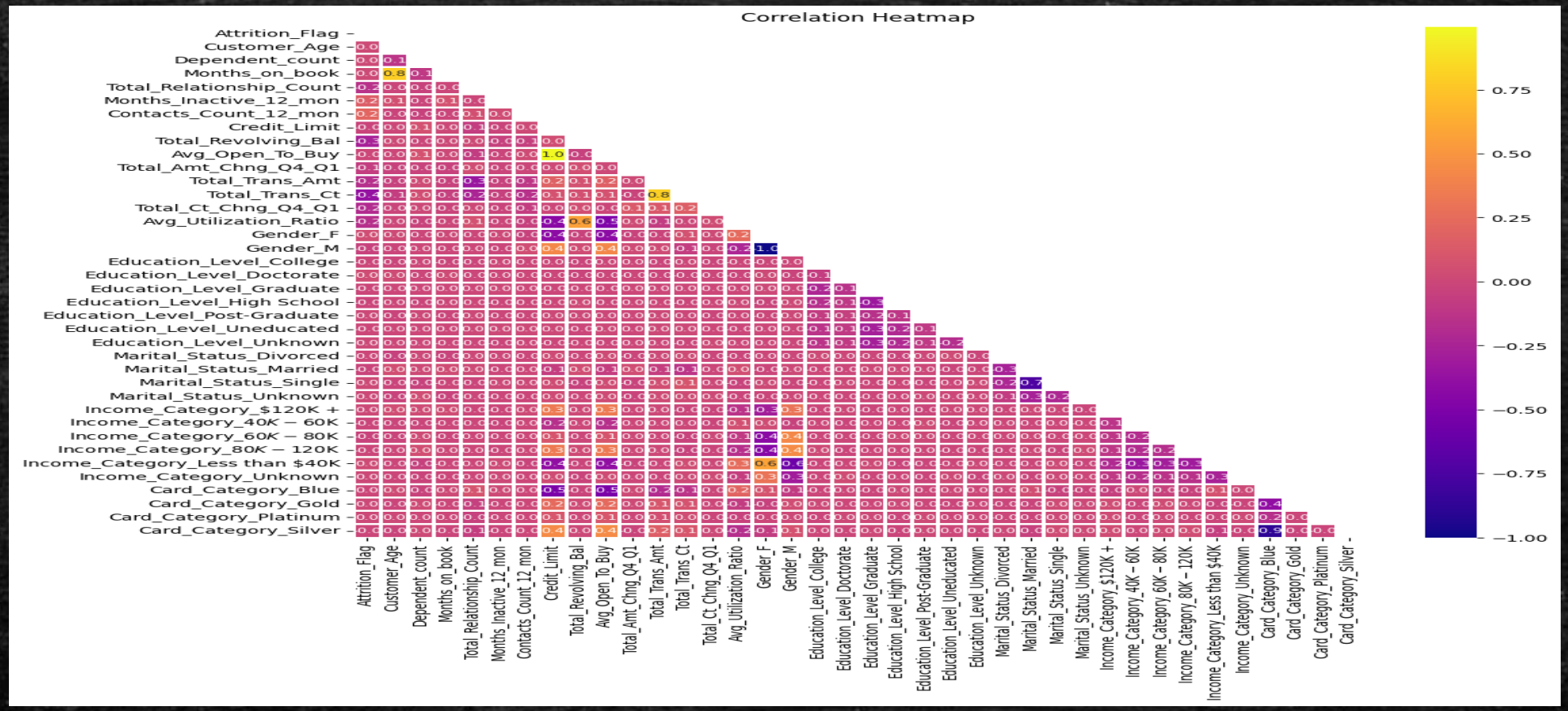
COMPORTAMIENTO INDIVIDUAL DE LAS VARIABLES NUMÉRICAS

- En promedio 'Avg_Utilization_Ratio' es menor en los hombres, que tienden a utilizar una proporción menor del crédito disponible en comparación con las mujeres.

- Hay algunos individuos masculinos (outliers) que utilizan una proporción significativamente mayor del crédito disponible en comparación con la mayoría de los otros hombres.

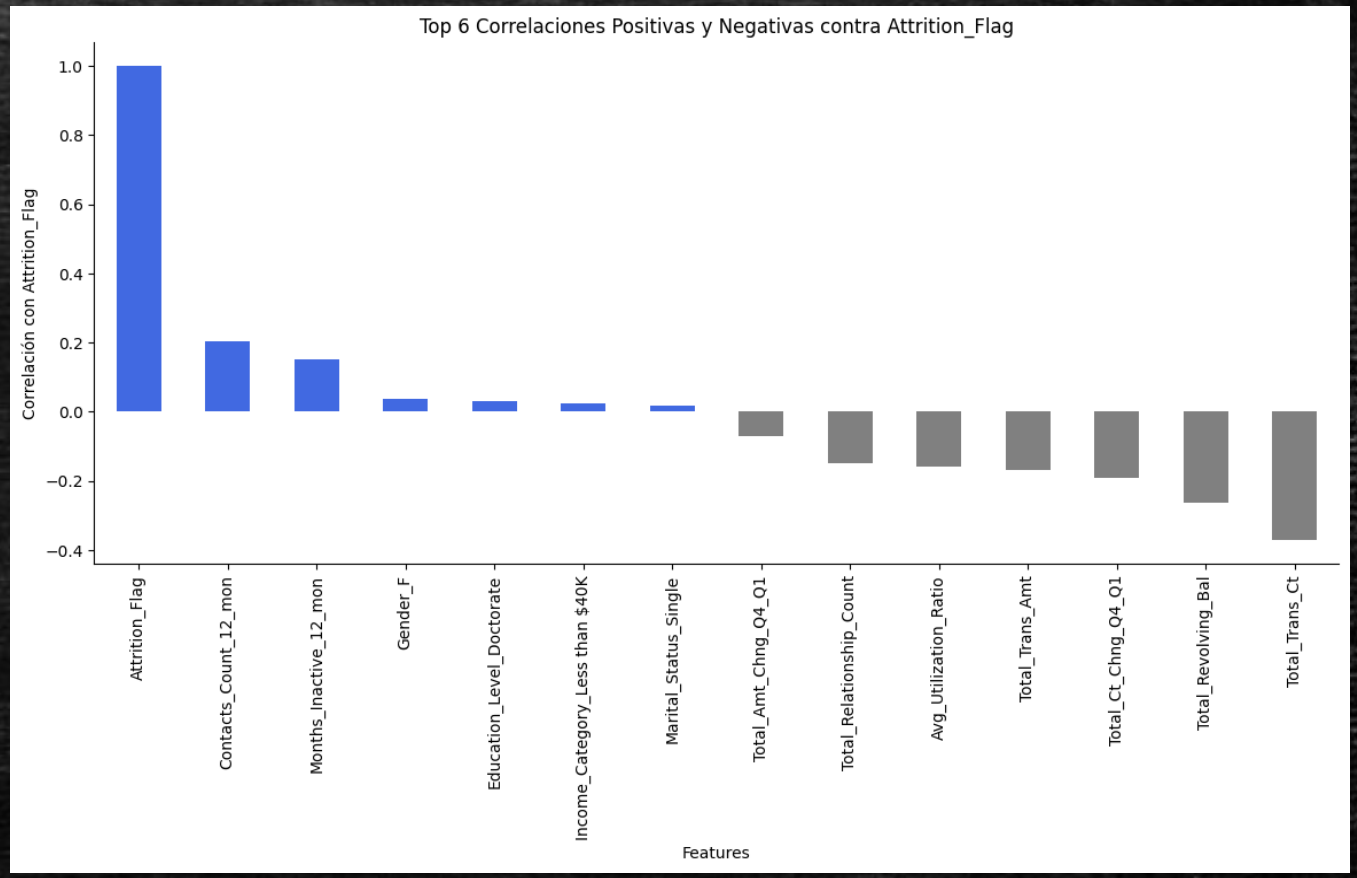


ANÁLISIS MULTIVARIADO



Top Correlaciones Positivas y Negativas contra Attrition_Flag

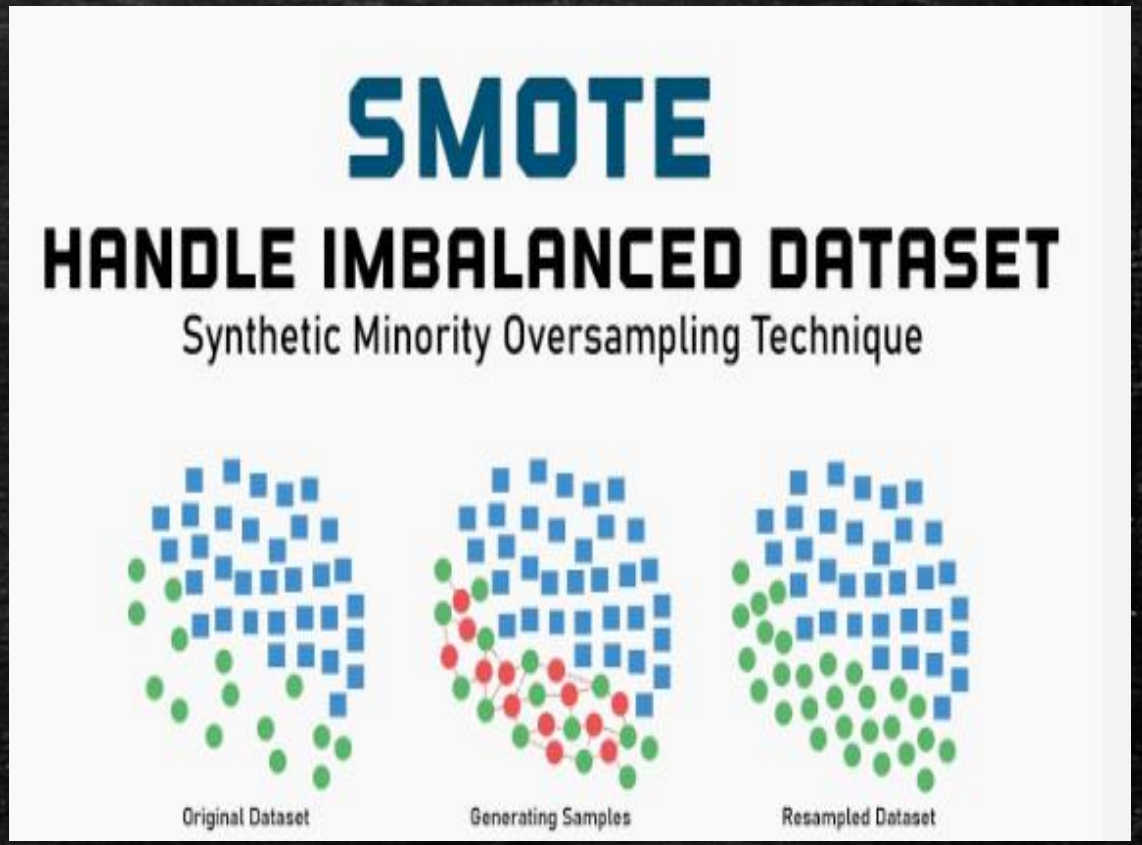
- El gráfico sugiere que los clientes que han tenido menos contactos en los últimos 12 meses entre 2 a 4 meses, son más propensos a abandonar dentro de ese periodo de tiempo
- La correlación positiva implica que los clientes que han tenido más contactos con la empresa en los últimos 12 meses tienen menos probabilidades de abandonar



IMPLEMENTACIÓN DE ALGORITMO DE MACHINE LEARNING PARA CLASIFICACIÓN

FEATURE ENGINEERING Y SELECCIÓN DE ALGORITMO

- De acuerdo a las variables que tenían una correlación alta o se encontraron influyentes. Se hizo una selección con ellas para dejarlas como set entrenamiento.
- Se aplicaron técnicas de encoding a las variables categóricas, así como también se normalizaron las variables numéricas y dejaron como Float para una convergencia más rápida.

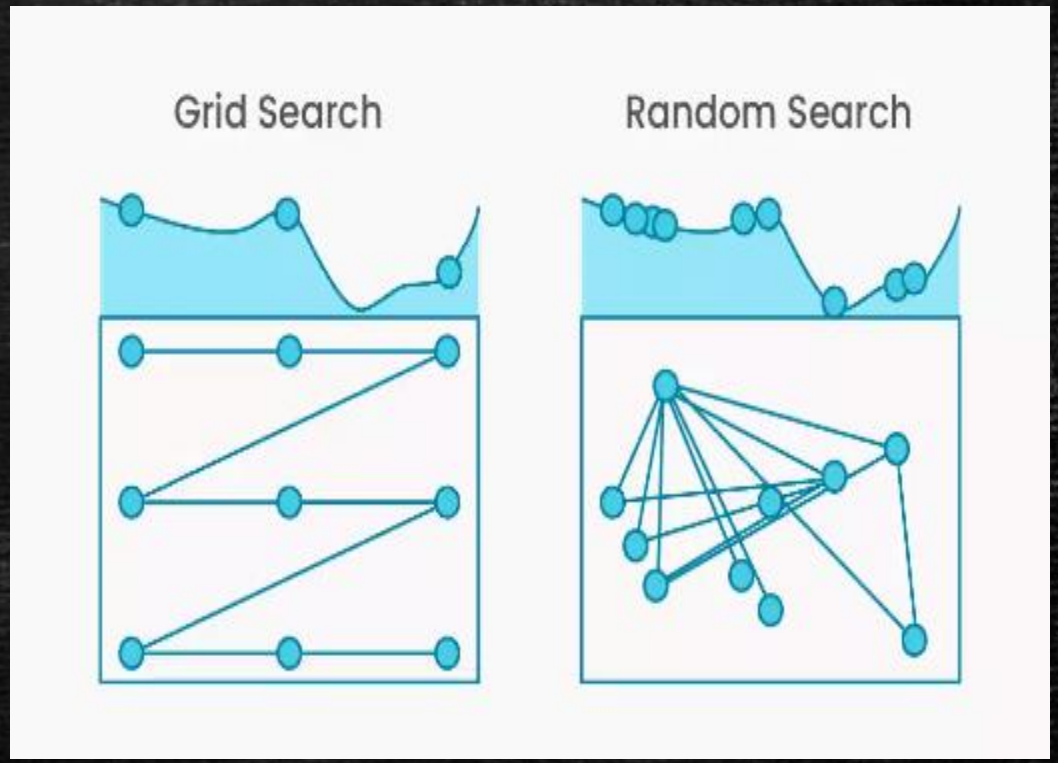


FEATURE ENGINEERING Y SELECCIÓN DE ALGORITMO

- Se probaron varios algoritmos de clasificiación como Regresión Logística, Random Forest y Stochastic Gradient Boosting.
- Se aplicó SMOTE para contrarrestar la naturaleza desbalanceada del set de datos (Oversampling).
- Por cuestiones de desempeño se eligió XGBoost

FEATURE ENGINEERING Y SELECCIÓN DE ALGORITMO

- Se utilizó GridSearchCV para realizar selección de hiperparámetros óptimos teniendo como objetivo "f1" buscar un balance entre Precisión y Recall.
- Se fijó como estimador XGBoost.
- Luego de aplicarlo se escogieron los mejores parámetros



VIENDO EL DESEMPEÑO DE XGBOOST

- Luego de obtener los hiperparámetros de GridSearchCV se definieron en XGBoost.
- Obteniendo un Accuracy de 87% para el Set de Prueba.
- Un roc_auc_score de 91% para los casos de Deserción.
- Se imprime un Reporte de Clasificación

VIENDO EL DESEMPEÑO DE CROSS VALIDATION

- En el reporte de Clasificación se puede observar la naturaleza imbalanceada del Set de Datos, que se corrigió con SMOTE.
- El resultado de STRATIFIED K-FOLD CV SOBRE XGBOOST.
- Obtuvo:
- CV MSE: 0.09
- Train MSE: 0.09
- Test MSE: 0.13

```
# Imprimir Reporte de Clasificación
print(classification_report(y_test, y_pred))
```

	precision	recall	f1-score	support
0	0.94	0.90	0.92	1701
1	0.59	0.72	0.65	325
accuracy			0.87	2026
macro avg	0.77	0.81	0.79	2026
weighted avg	0.89	0.87	0.88	2026

CONCLUSIÓN PARTE 1

- A través de preguntas formuladas y diferentes tipos de análisis se llegó a la conclusión que las variables que más influyen en la deserición son:
- 'Contacts_Count_12_mon' la cantidad de contactos que tuvo el cliente con la compañía en los últimos 12 meses
- 'Months_Inactive_12_mon' la cantidad de meses que el cliente ha estado inactivo durante los últimos 12 meses
- Teniendo estos detalles se puede tratar de generar una estrategia para estar más en contacto con los clientes o darle un enfoque tipo "Customer Success" mostrando a los clientes los beneficios que tienen y cómo sacar partido a lo que no han utilizado aún.

CONCLUSIÓN PARTE 2

- Notar también que en tercer lugar influyen con menor proporcción el Género Femenino, aunque la cantidad de dependientes (hijos, familiares) no sugiere ser un factor de influencia

- En cuarto lugar el nivel de educación con Doctorado, quizás al tener un nivel de educación más alto tienen mejores herramientas o medios para buscar diferentes compañias y servicos

- En quinto lugar ingresos inferiores a 40 mil dólares al año y en sexto lugar clientes Solteros, podría ser que estas deserciones estén asocidas a malos hábitos crediticios o el nivel de ingresos

- También se pueden fortalecer las variables que tienen correlación negativa, para hacer que los clientes permanezcan más tiempo en la compañía

CONCLUSIÓN PARTE 3

- En lo que respecta al modelo de Machine Learning tiene un buen desempeño parar identificar datos que no ha visto.
- Se trató de armonizar el resultado de Validación Cruzada a través de la definición de Hiperparámetros
- Buscando siempre que el intercambio entre ajustar un parámetro y otro no llevara ni a un Overfitting ni a un Underfitting.