


PROYECTO. EVENTO SMART SHOPPING

Equipo 15. Integrantes:

- Ana Laura López --> any.lom.01@gmail.com**
- Alma Violeta Alvarez --> almavioletaalvarez@gmail.com**



EVENTO SMART SHOPPING

Planteamiento: Una tienda americana realiza una temporada de ofertas durante el año, en donde sus clientes pueden obtener sus productos a un menor precio. Los clientes pueden usar métodos de pago como lo son tarjetas de crédito o débito y pago en efectivo. Tomando en cuenta que es posible que algunos productos puedan obtener mínimos montos de compra. La tienda valora mucho los comentarios de los clientes, y recibe una calificación de acuerdo a la experiencia de compra del cliente.

PREGUNTAS



Con la situación presentada podemos plantearnos algunas preguntas que podrían resolverse mediante el análisis y la ciencia de datos. Las preguntas obtenidas son las siguientes:

- 1. ¿Qué método de pago es el más utilizado por los clientes?**
- 2. ¿Qué producto es el más consumido?**
- 3. ¿Cuál es el promedio de la satisfacción del cliente después de consumir en la tienda?**
- 4. ¿Cuál es el promedio en valor de los artículos vendidos?**
- 5. ¿En qué día hubo más ventas?**
- 6. ¿Si se pudiera relacionar la calificación del cliente con el método de pago, cuál sería el método de pago preferido?**

```
df = pd.read_csv('/content/drive/MyDrive/DataSetsAny/Fashion_Retail_Sales.csv')
df.head()
```

	Customer Reference ID	Item Purchased	Purchase Amount (USD)	Date Purchase	Review Rating	Payment Method
0	4018	Handbag	4619.0	2023-02-05	NaN	Credit Card
1	4115	Tunic	2456.0	2023-07-11	2.0	Credit Card
2	4019	Tank Top	2102.0	2023-03-23	4.1	Cash
3	4097	Leggings	3126.0	2023-03-15	3.2	Cash
4	3997	Wallet	3003.0	2022-11-27	4.7	Cash

ANÁLISIS GENERAL DEL DATASET

```
#Obtenemos dimensiones del dataset  
df.shape
```

```
(3400, 6)
```

+ Código

+ Texto

```
[5] #Obtenemos el nombre de las columnas contenidas en el dataset  
df.columns
```

```
Index(['Customer Reference ID', 'Item Purchased', 'Purchase Amount (USD)',  
      'Date Purchase', 'Review Rating', 'Payment Method'],  
      dtype='object')
```

descripcion

Analizando el nombre de las columnas, podemos observar que tenemos datos de compras como es el ID del cliente, el producto adquirido, la cantidad que se pagó por el producto, la calificación del cliente, y el método de pago.

```
#Verificamos si tenemos valores nulos en nuestras columnas  
df.isna().sum()
```

```
Customer Reference ID      0  
Item Purchased             0  
Purchase Amount (USD)    650  
Date Purchase             0  
Review Rating            324  
Payment Method            0  
dtype: int64
```

LIMPIEZA DEL DATASET

```
[ ] # Llenamos los valores nulos de la columna Purchase Amount (USD) con ceros, pues es un dato numérico
df_drop['Purchase Amount (USD)']=df_drop['Purchase Amount (USD)'].fillna(0)
#df_drop = df_drop.dropna(subset=['Purchase Amount (USD)'])
```

```
[ ] # Llenamos los valores nulos de la columna PReview Rating con ceros, pues es un dato numérico
#Despues de análisis estadístico se decide borrar los registros con calificación NaN, porque produce sesgos
#Que esten en 0 puede afectar negativamente a la estadística de la tienda, cuando podríamos pensar que simplemente no emitieron su voto
#Al ser una variable que determina el desempeño de la experiencia, si es importante
#df_drop['Review Rating']=df_drop['Review Rating'].fillna(0)
df_drop = df_drop.dropna(subset=['Review Rating'])
```

```
[ ] # Volvemos a verificar nuestros datos nulos en todo el dataset
df_drop.isna().sum()
```

```
Customer Reference ID    0
Item Purchased           0
Purchase Amount (USD)    0
Date Purchase            0
Review Rating            0
Payment Method           0
dtype: int64
```

PRIMERA VARIABLE (PURCHASE_AMOUNT)

```
Estimados de locacion
purchase_amount
Media: 120.96228868660599
Mediana: 86.0
Media truncada: 84.60926076360683
Desv. estandar: 354.9204878895212
```

Con los datos obtenidos de nuestro primer analisis de estimados de locación podemos observar que tenemos una media elevada comparando con la mediana y la media truncada, aquí podemos hacer la hipótesis de que es muy probable que en nuestro dataset existan datos atípicos que estén elevando la media.

**Antes de
eliminar datos
atípicos**

```
Estimados de variabilidad
purchase_amount
Valor minimo: 0.0
Percentil 10: 0.0
Percentil 25: 24.0
Percentil 50: 86.0
Percentil 75: 145.0
Percentil 90: 180.0
Valor maximo: 4932.0
```

Con los resultados obtenidos del análisis de los estimados de variabilidad podemos deducir que efectivamente tenemos datos elevados, debido a que es mucha la diferencia entre el percentil 90 y el valor máximo. Es probable que sean estos datos los que aumenten la media calculada anteriormente.

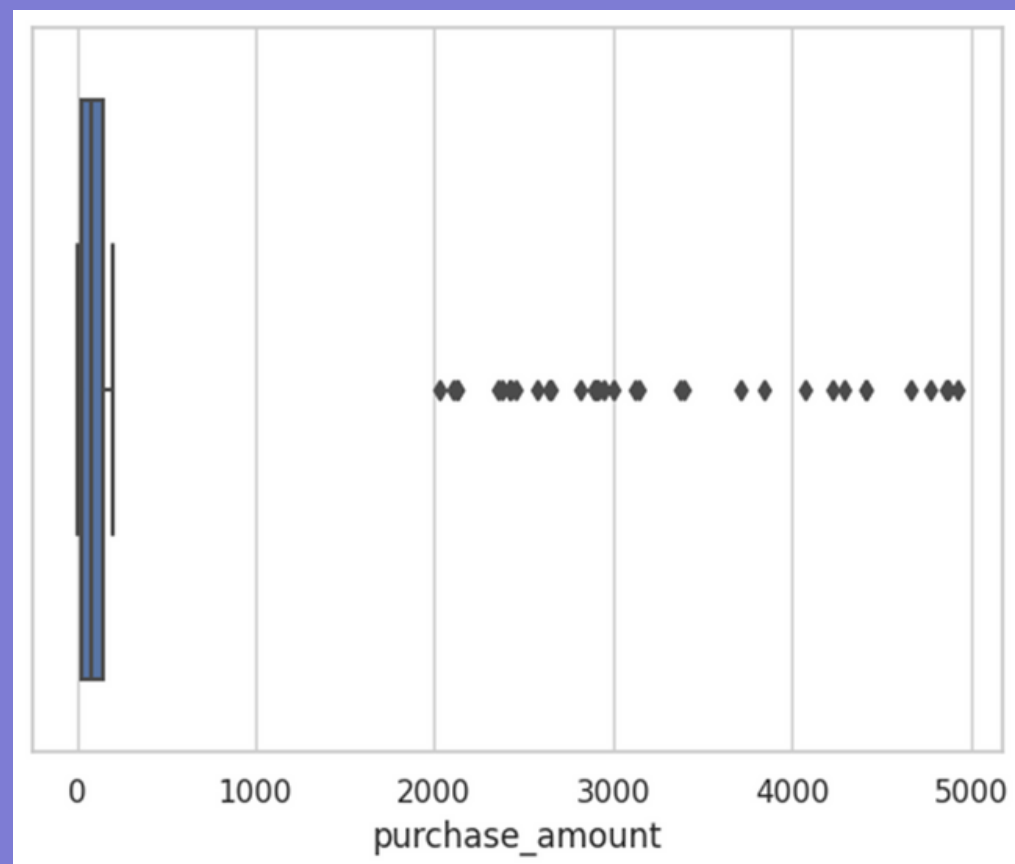
PRIMERA VARIABLE (PURCHASE_AMOUNT)

Después de
eliminar
datos atípicos

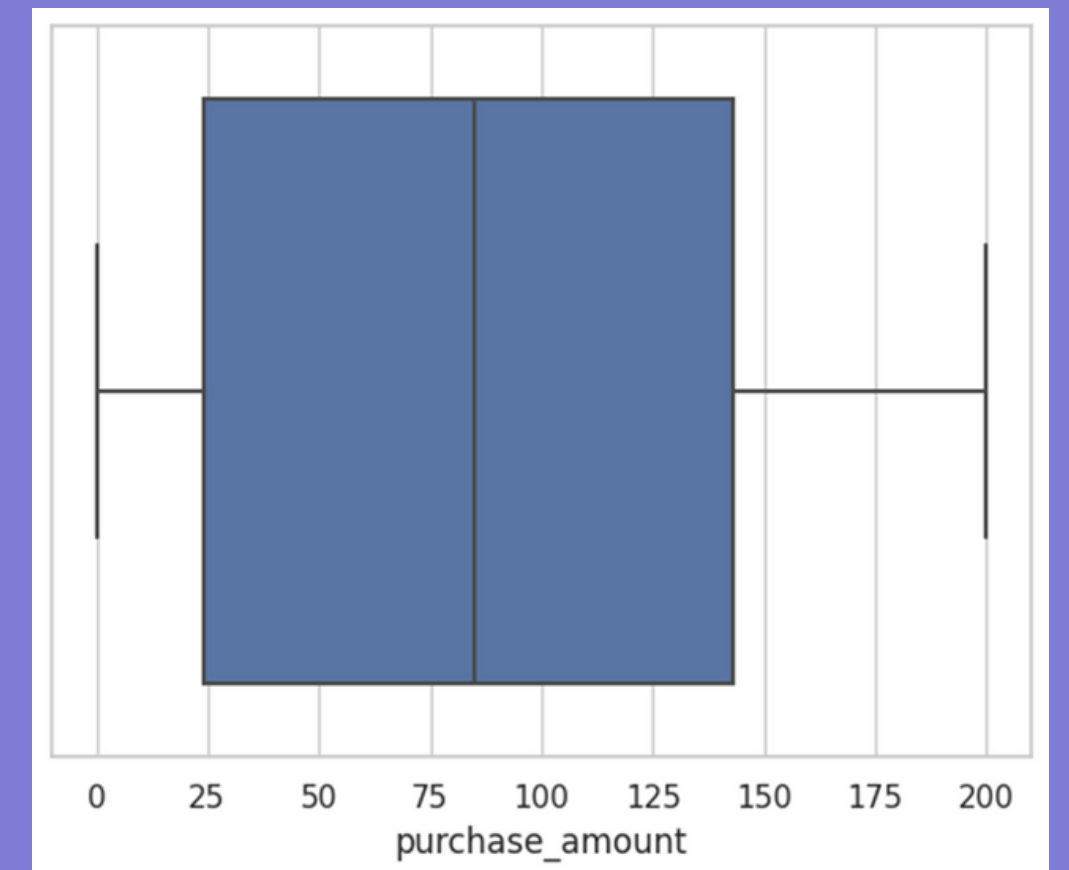
```
Estimados de locacion
purchase_amount
Media: 120.96228868660599
Mediana: 86.0
Media truncada: 84.60926076360683
Desv. estandar: 354.9204878895212
Comparacion
Datos filtrados por IQR
Media: 85.54043392504931
Mediana: 85.0
Media truncada: 83.31059983566146
Desv. estandar: 65.01935948926558
```

Analizando los datos después del filtrado podemos observar que las medidas de locación se ven mejor y más parecidas entre ellas.

Antes

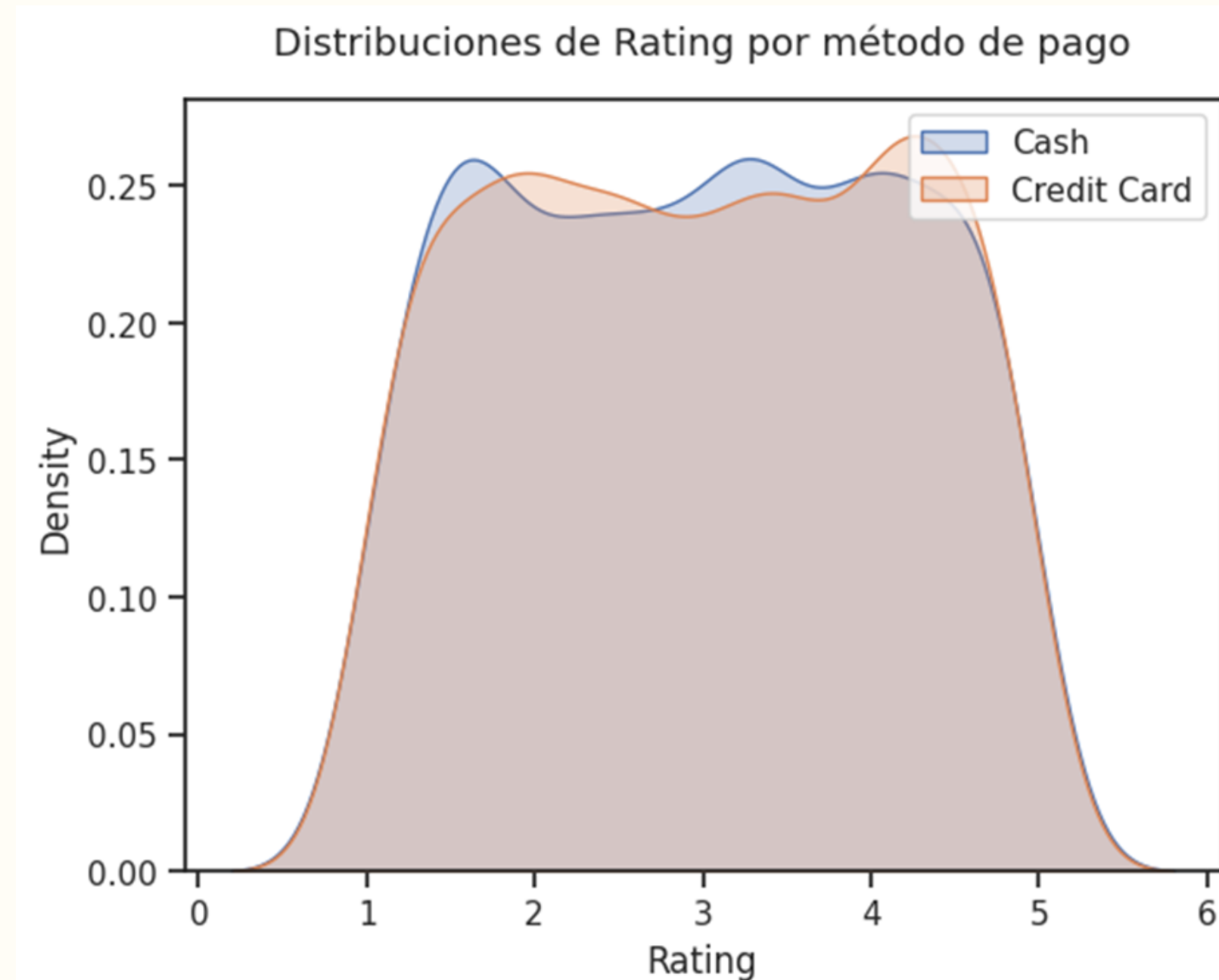


Después

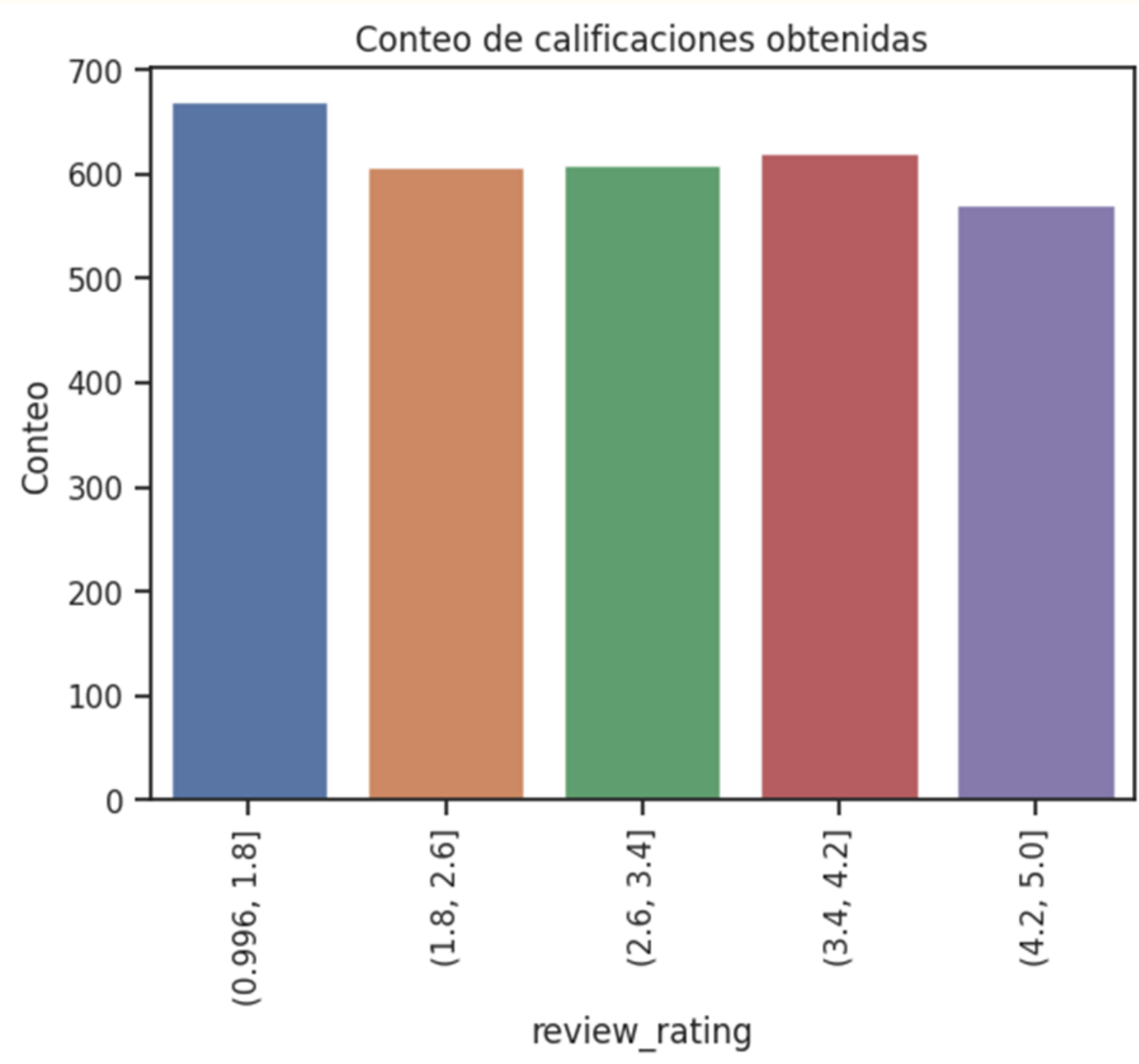


GRÁFICA DE DENSIDADES

Realizamos grafica de densidad para relacionar las variables review_rating con payment_method, payment_method es una variable categorica del método de pago (efectivo o con tarjeta) asociamos ambas variables para ver la relacion del tipo de pago con la calificación



EXPLORACIÓN DE VARIABLES CATEGÓRICAS

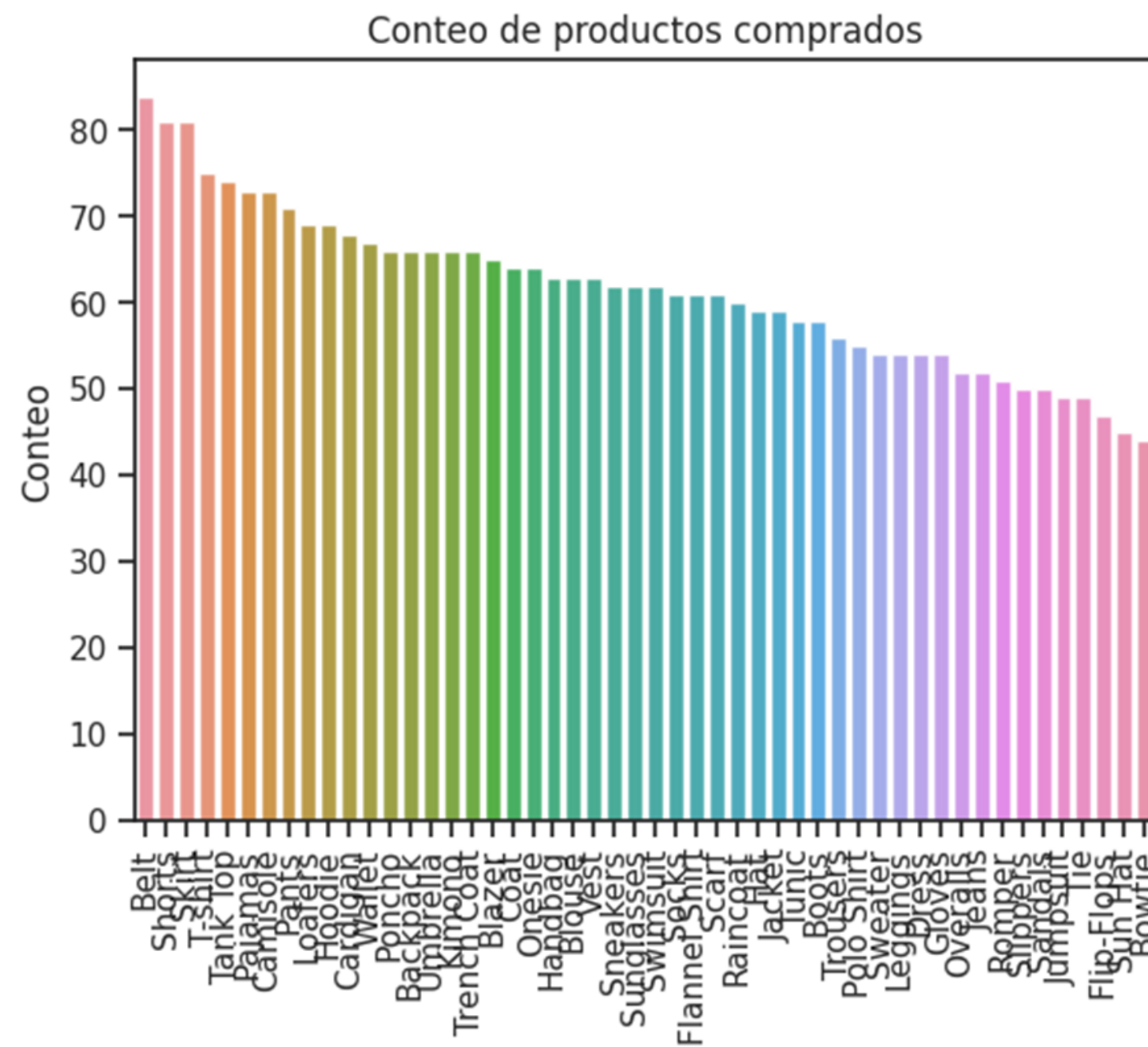


Realizando la gráfica de barras de nuestras variables correspondientes, lo que podemos observar es que tenemos un mayor de datos en calificaciones que son un poco bajas, pero el segundo intervalo donde hay más datos es uno de los más altos, todos son muy proporcionados, pero en esta encuesta lo ideal es que hubiera minoría de datos en los primeros intervalos. Sin embargo esto puede servir como referencia para la mejora de la tienda.

Primera variable:
item_purchased

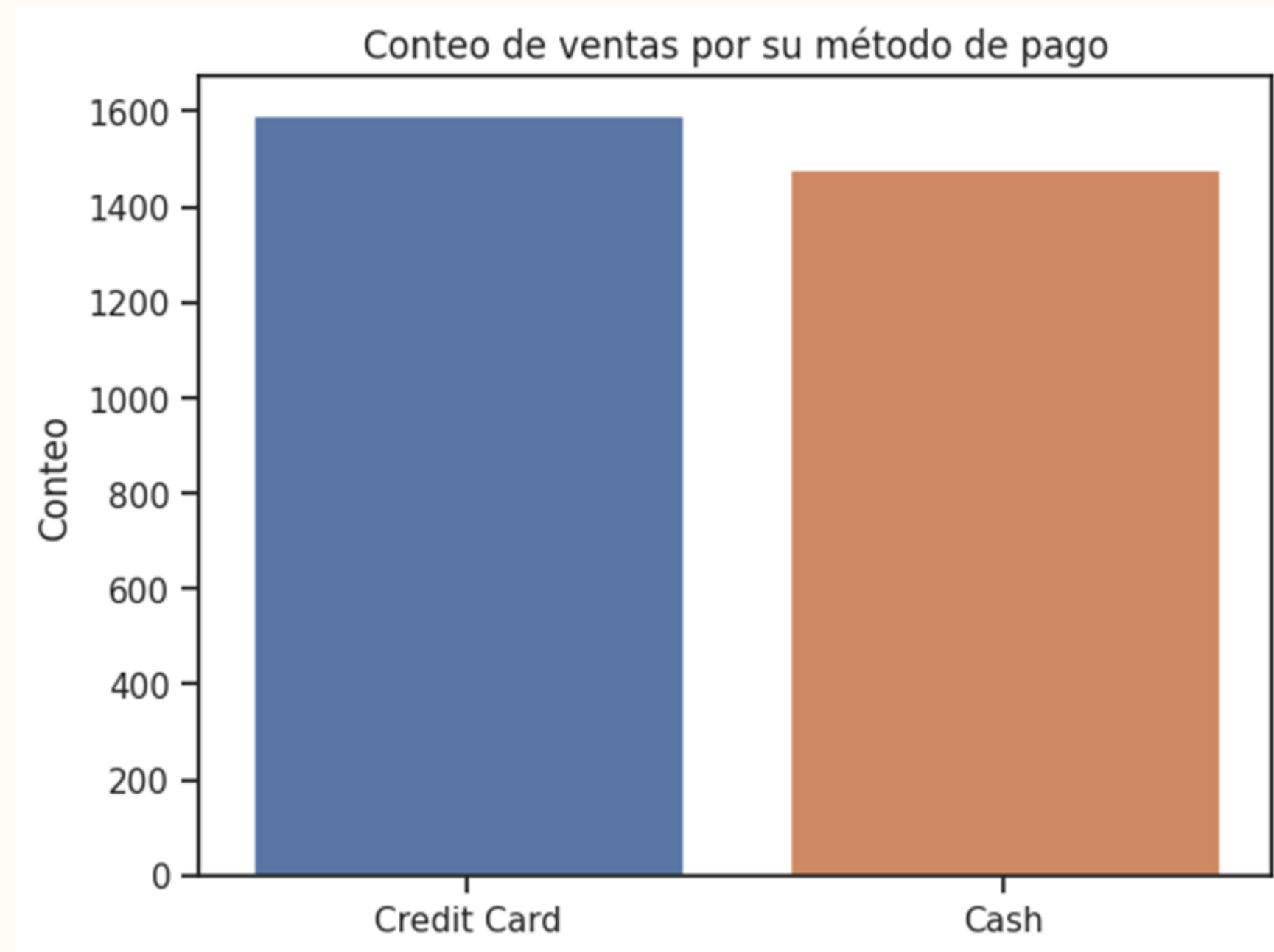
PRODUCTOS VENDIDOS VS EL CONTEO DE CADA UNO

Decidimos realizar una grafica de barras para saber cómo están distribuidas las ventas de acuerdo al producto.



VARIABLE PAYMENT_METHOD

Gráficamente se visualiza cuál es el método de pago que más se utilizó, en esta ocasión, aunque los resultados son muy parejos, el más usado es mediante tarjeta de crédito.



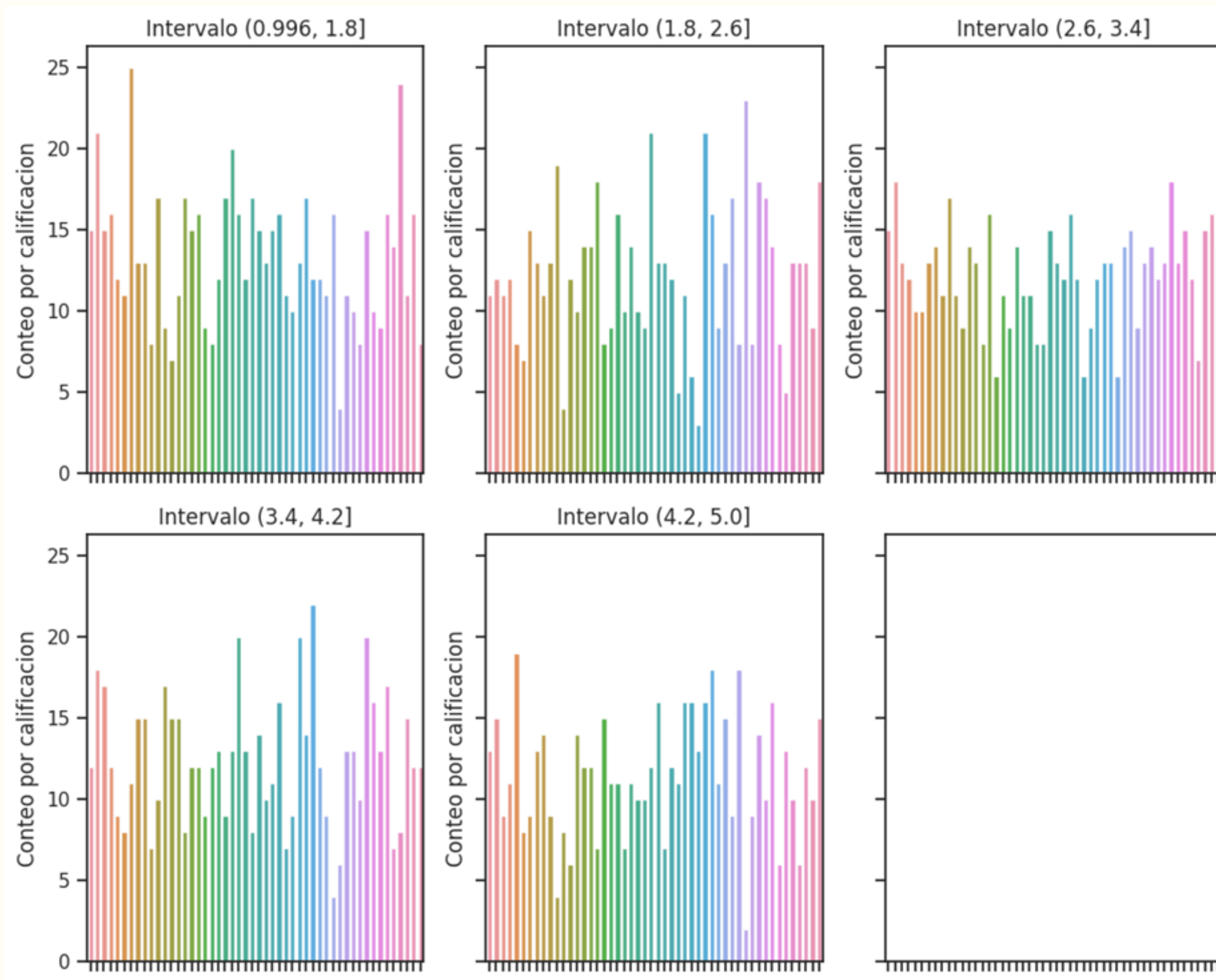
TABLAS DE CONTINGENCIA

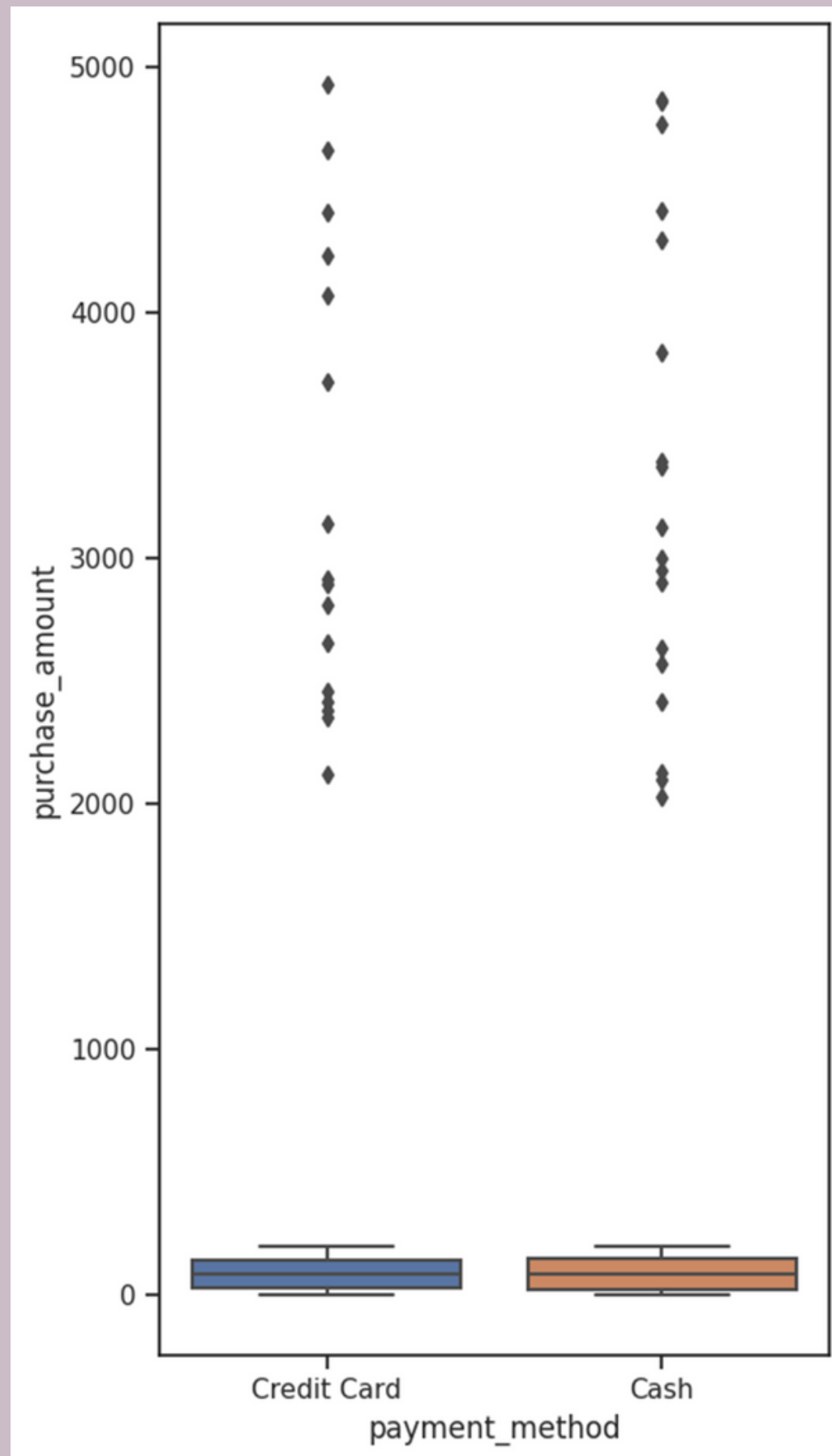
```
[ ] #Se hace tabla de contingencia para agrupar por item y calificacion
crosstab_review_item = pd.crosstab( df2['interval_rating'], df2['item_purchased'])
crosstab_review_item
```

item_purchased	Backpack	Belt	Blazer	Blouse	Boots	Bowtie	Camisole	Cardigan	Coat	Dress	...	Swimsuit	T-shirt	Tank Top	Tie	Trench Coat	Trousers	Tunic	Umbrella
interval_rating																			
(0.996, 1.8]	15	21	15	16	12	11	25	13	13	8	...	8	15	10	9	16	14	24	11
(1.8, 2.6]	11	12	11	12	8	7	15	13	11	13	...	18	17	14	8	5	13	13	13
(2.6, 3.4]	15	18	13	12	10	10	13	14	11	17	...	12	13	18	13	15	12	7	15
(3.4, 4.2]	12	18	17	12	9	8	11	15	15	7	...	10	20	16	13	17	7	8	15
(4.2, 5.0]	13	15	9	11	19	8	9	13	14	9	...	14	10	16	6	13	10	6	12

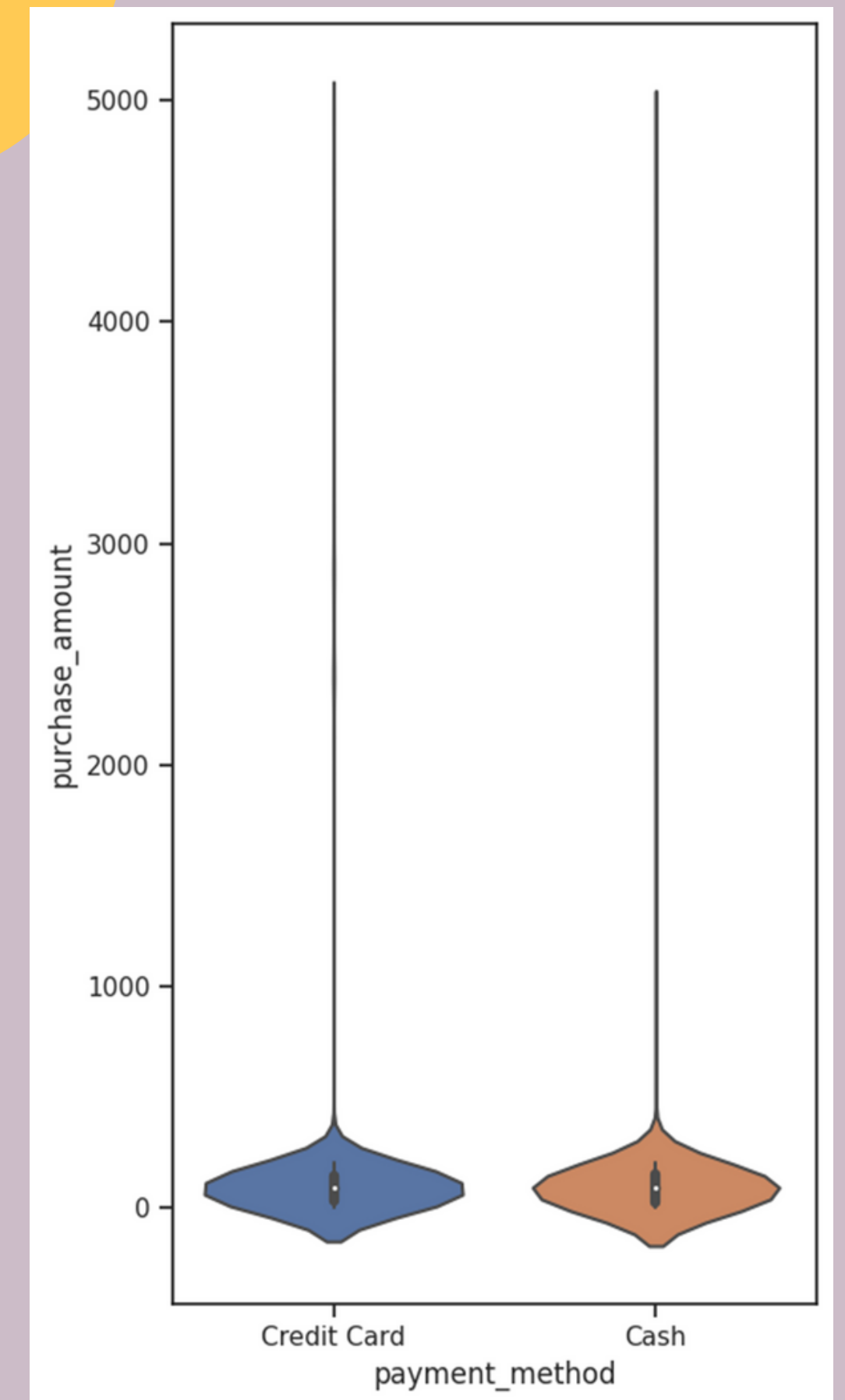
5 rows x 50 columns

En este caso decidimos realizar una tabla de contingencia para agrupar los datos del tipo de producto y cuántas unidades de ese producto fueron clasificadas en los intervalos de calificación.





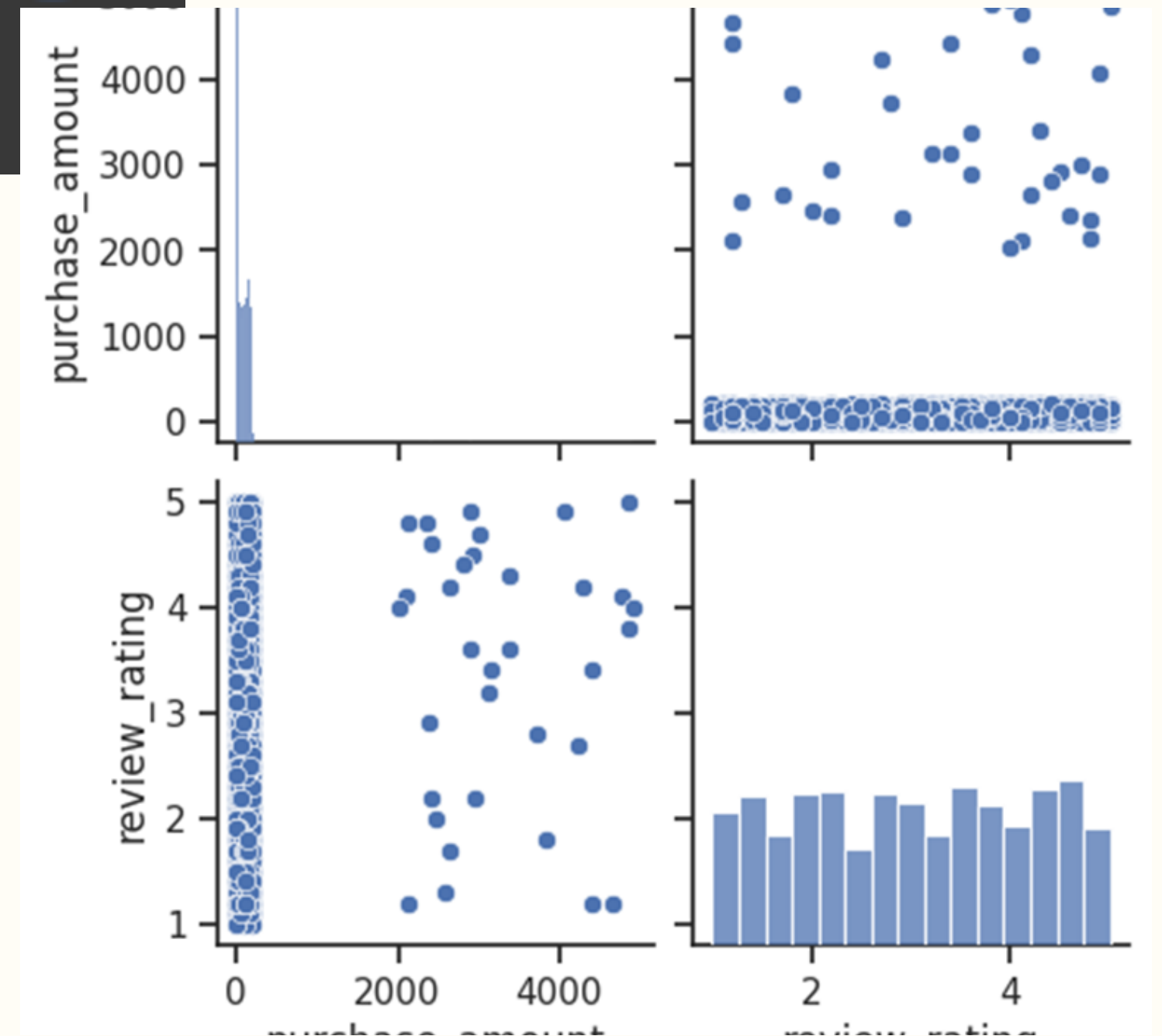
Analizando los resultados obtenidos podemos ver gráficamente que nuestros datos están concentrados en compras menores a 1000 USD y ambas están distribuidas proporcionalmente en las dos opciones de pago. También es posible observar algunos datos atípicos en lo que se categorizan como precios altos.



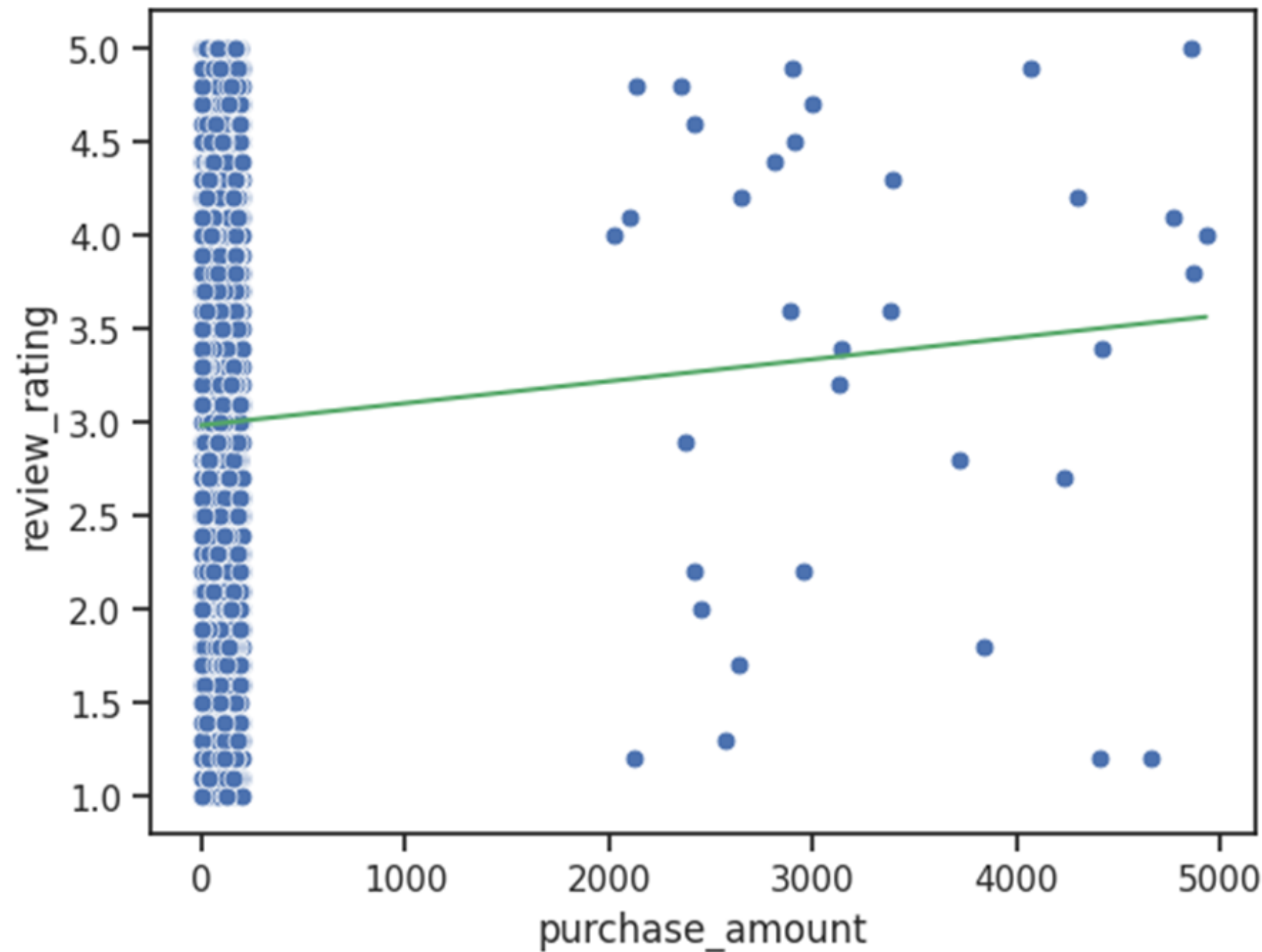
MATRIZ CORRELACIONES

	purchase_amount	review_rating
purchase_amount	1.000000	0.036044
review_rating	0.036044	1.000000

Se intentó limpiar el dataset para intentar obtener mejores resultados en cuanto a correlación, sin embargo estos no fueron buenos, así que solo se optó por eliminar las calificaciones nulas del dataset.

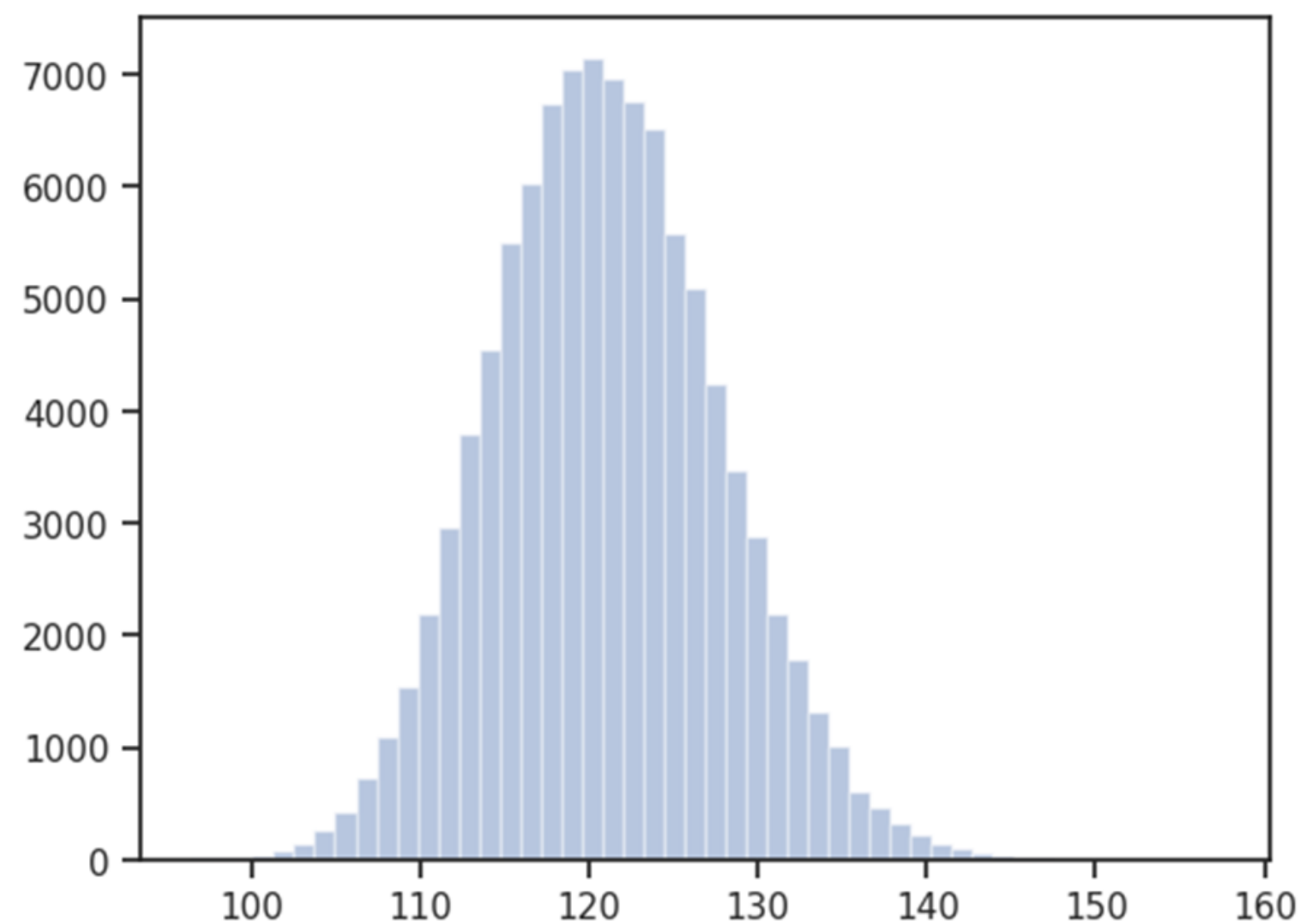


ENTRENAMIENTO REGRESIÓN LINEAL COMO PRÁCTICA



**Al realizar el ejercicio,
podemos darnos cuenta
que sería muy
complicado predecir
alguna de nuestras dos
variables numéricas en
base a la otra, entonces
no sería un método
efectivo en este caso.**

BOOTSTRAP VARIABLE PURCHASE_AMOUNT



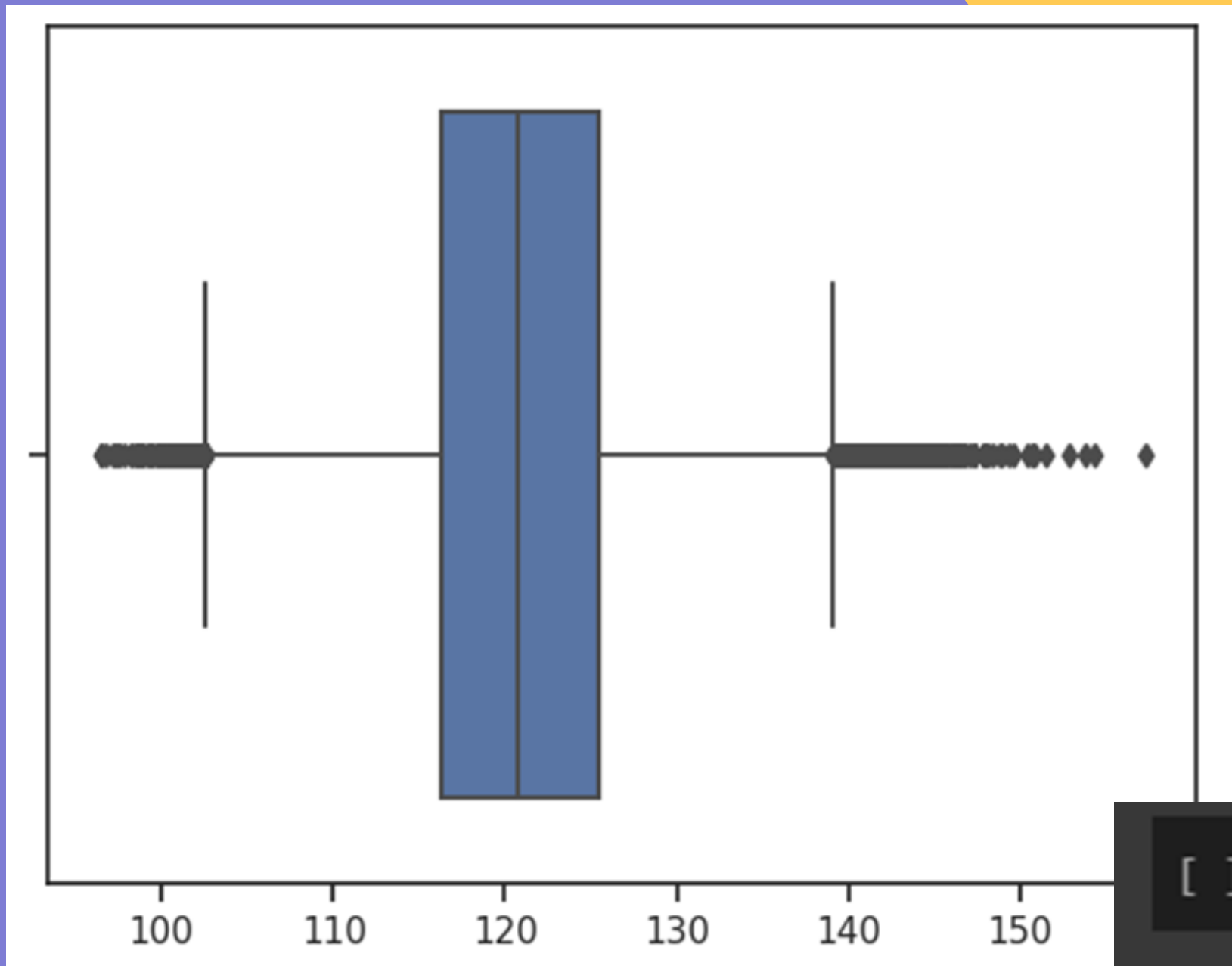
Se hizo Bootstrapping con muestras del 90% para obtener la media, al igual que la mediana.

```
[ ] serie_means_amount.skew()
```

```
0.18918847020944585
```

```
[ ] serie_means_amount.kurtosis()
```

```
0.04428218677668605
```



ERROR ESTÁNDAR

```
[ ] print('Error estándar: ', serie_means_amount.std())
```

```
Error estándar: 6.747222978655318
```

```
[ ] print('Valor minimo: ', serie_means_amount.min())  
print('Valor Maximo: ', serie_means_amount.max())  
print('Rango: ', serie_means_amount.max()-serie_means_amount.min())
```

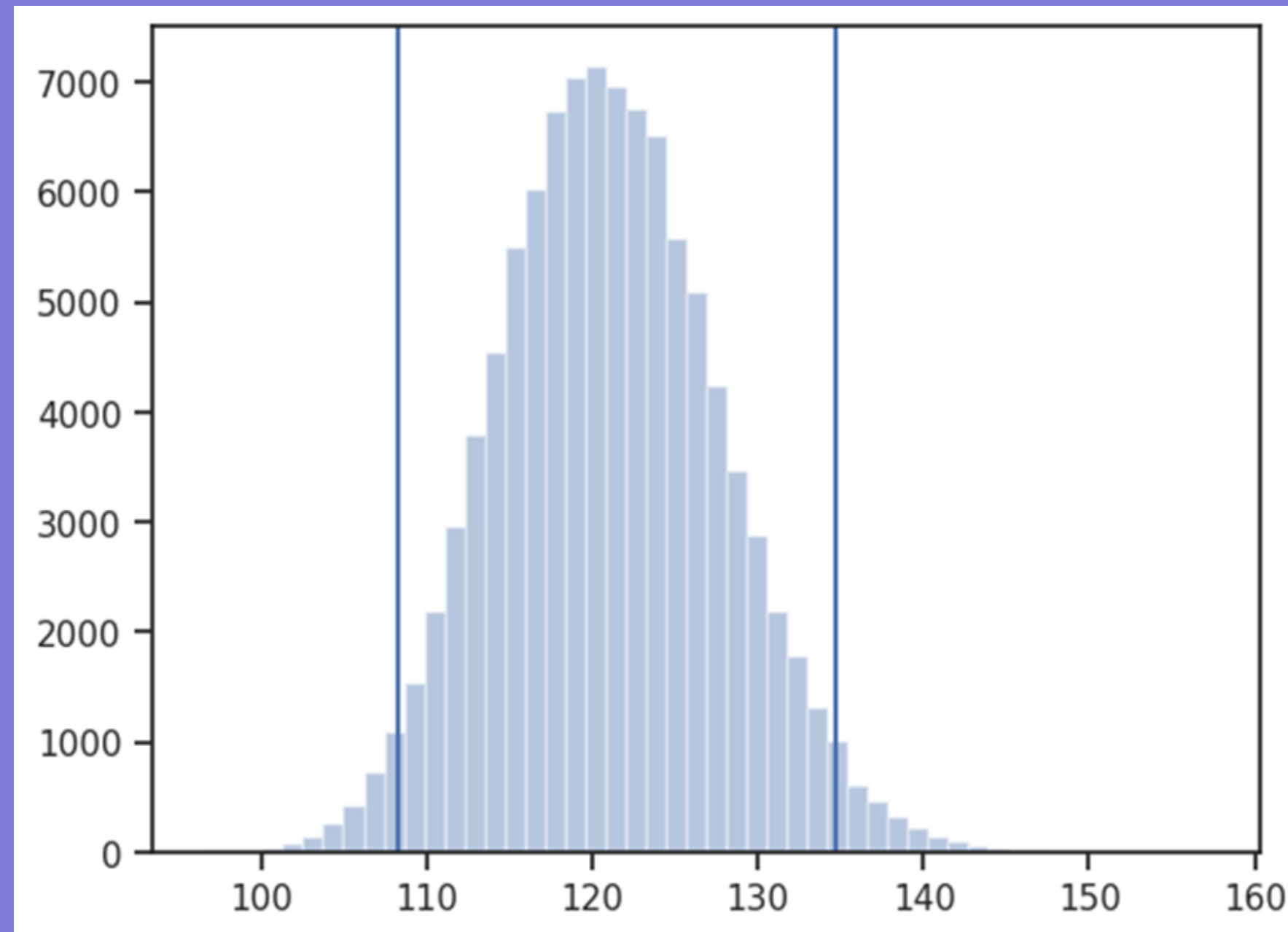
```
Valor minimo: 96.46170520231213
```

```
Valor Maximo: 157.25541907514452
```

```
Rango: 60.793713872832384
```

INTERVALO DE CONFIANZA

Intervalo de 95% de confianza de la media: $108.3279985549133 < 120.96228868660599 > 134.70705382947975$



Intervalo de 95% confianza de la media: $120.96228868660599 \pm 13.189527637283227$

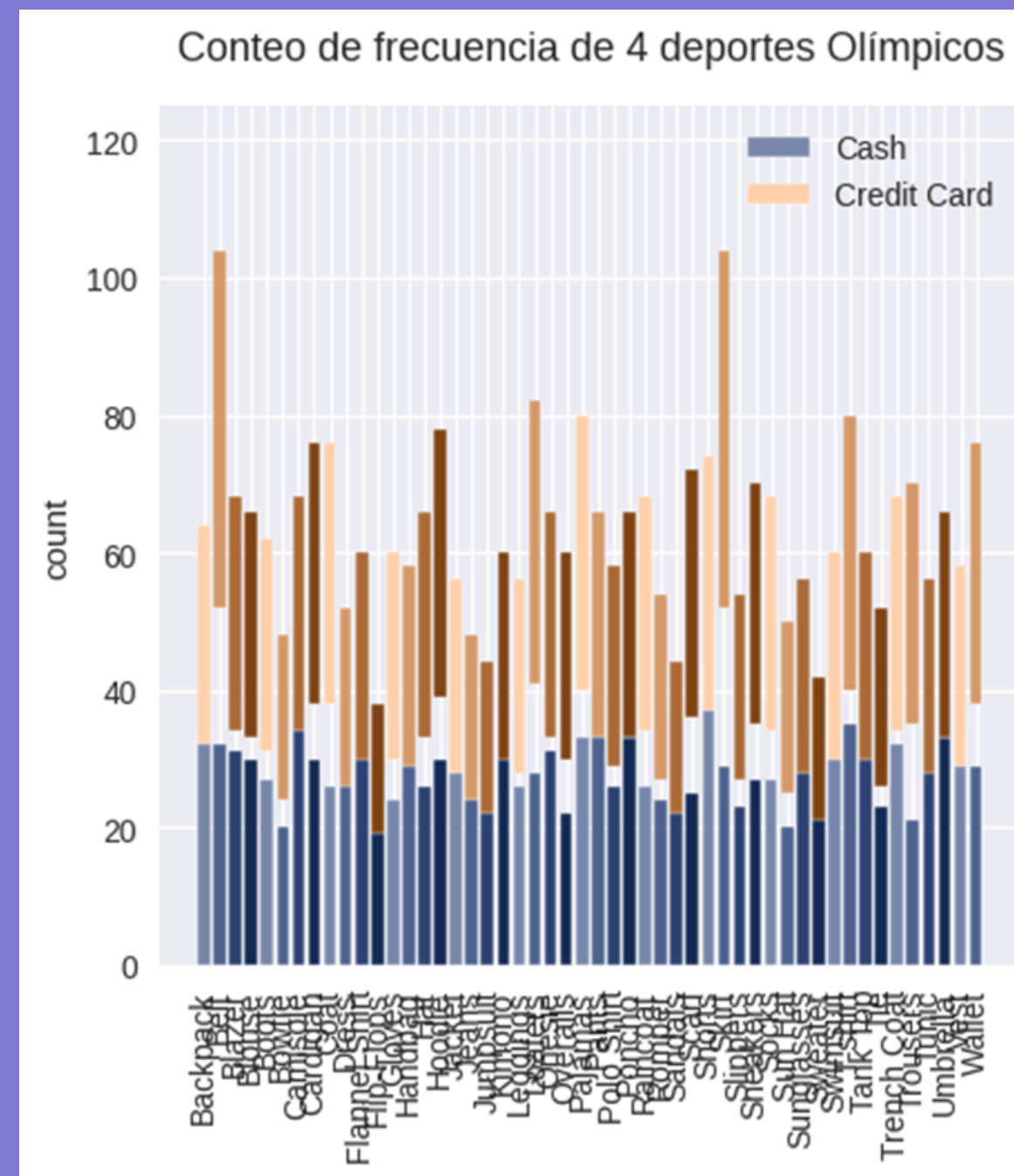
TREEMAP

Se opta por hacer un treemap, que esté categorizado por el tipo de producto obtenido y a su vez, sus hijos son la proporción de los métodos de pago con los que fueron adquiridos



GRÁFICA DE BARRAS APILADA

De igual forma, para ejemplificar este tipo de gráfico, se hace una categorización por le tipo de producto apilando los metodos de pago para cada uno.



NLP

Para abarcar el tema de procesamiento de lenguaje natural, se usa un dataset específico. Este dataset clasifica un fragmento de texto y está clasificado en cuál es la emoción que lo representa. Debido a que es un análisis de texto que está enfocado en los sentimientos, decidimos enfocar nuestro análisis en la palabra 'feel'.

```
[128] all_text.concordance('feel', lines=20)
```



Displaying 20 of 1361 matches:

```
i feel like i am still looking at a blank c
blank canvas blank pieces of paper i feel like a faithful servant i am just fe
r if i am feeling festive i start to feel more appreciative of what god has do
be able to take care of this baby i feel incredibly lucky just to be able to
cky just to be able to talk to her i feel less keen about the army every day i
less keen about the army every day i feel dirty and ashamed for saying that i
dirty and ashamed for saying that i feel bitchy but not defeated yet i was dr
ess the slaughter of others i didn't feel abused and quite honestly it made my
e i also loved that you could really feel the desperation in these sequences a
ust know to begin with i am going to feel shy about it i feel try to tell me i
th i am going to feel shy about it i feel try to tell me im ungrateful tell me
worst daughter sister in the world i feel that it is something that will never
will never really be resolved i just feel like all my efforts are in vain and
ts are in vain and a waste of time i feel absolutely foolish for allowing myse
know if anybody will ever be able to feel how i feel or at least relate when e
body will ever be able to feel how i feel or at least relate when everything i
lf missing and longing for it them i feel as if i am the beloved preparing her
ring herself for the wedding i would feel i missed out on a wealth of treasure
i did not read i finished the film i feel kind of regretful that i wasn't able
le to catch this on the big screen i feel like im caring about my body not in
```

DATOS ESTADÍSTICOS DE LA VARIEDAD LÉXICA DE NUESTRO DATASET

```
[130] #Total de palabras
      len(all_text)

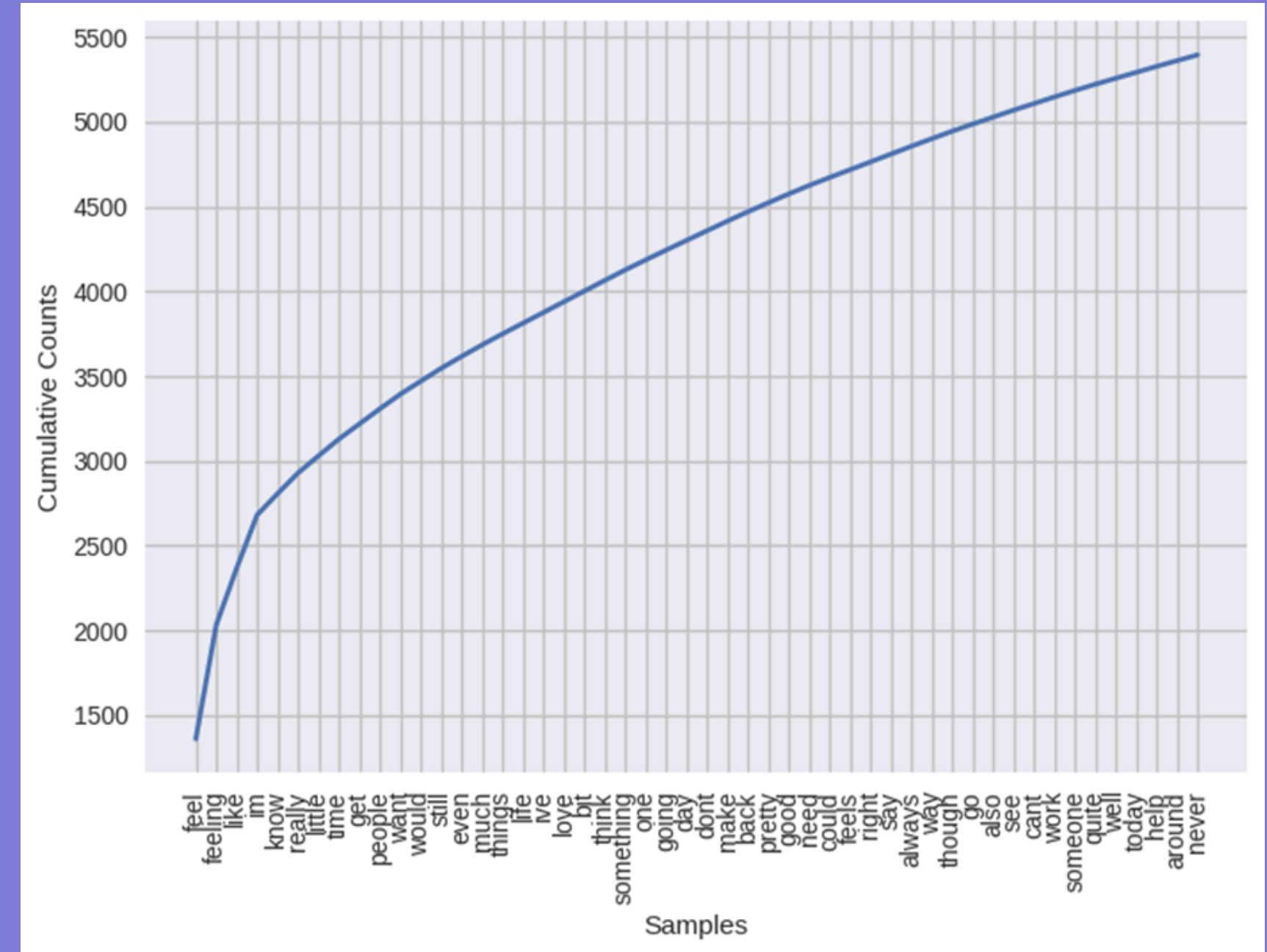
      37683

[131] #Total de palabras distintas
      len(set(all_text))

      4791

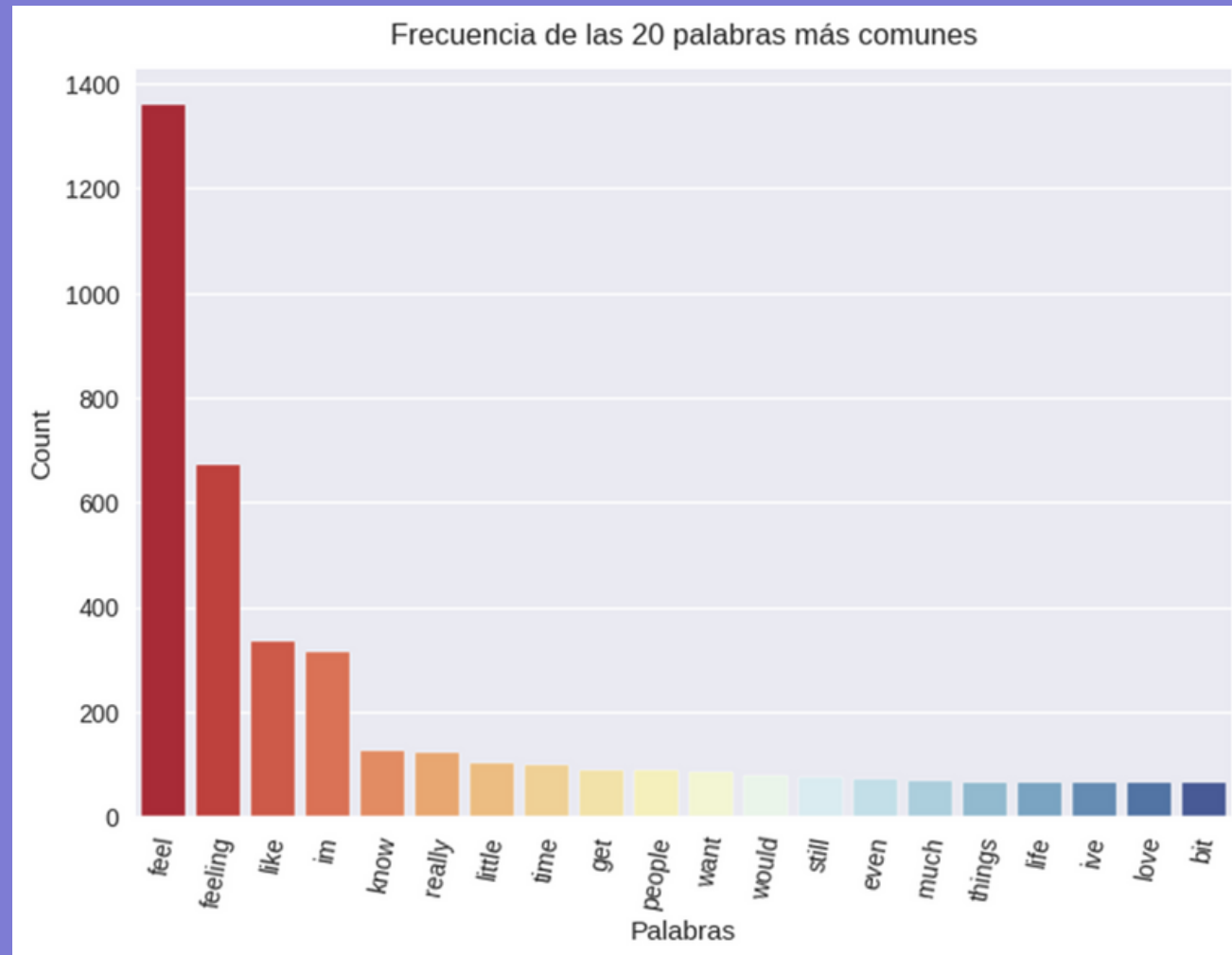
[132] #riqueza lexica
      len(set(all_text)) / len(all_text)

      0.12713955895231271
```

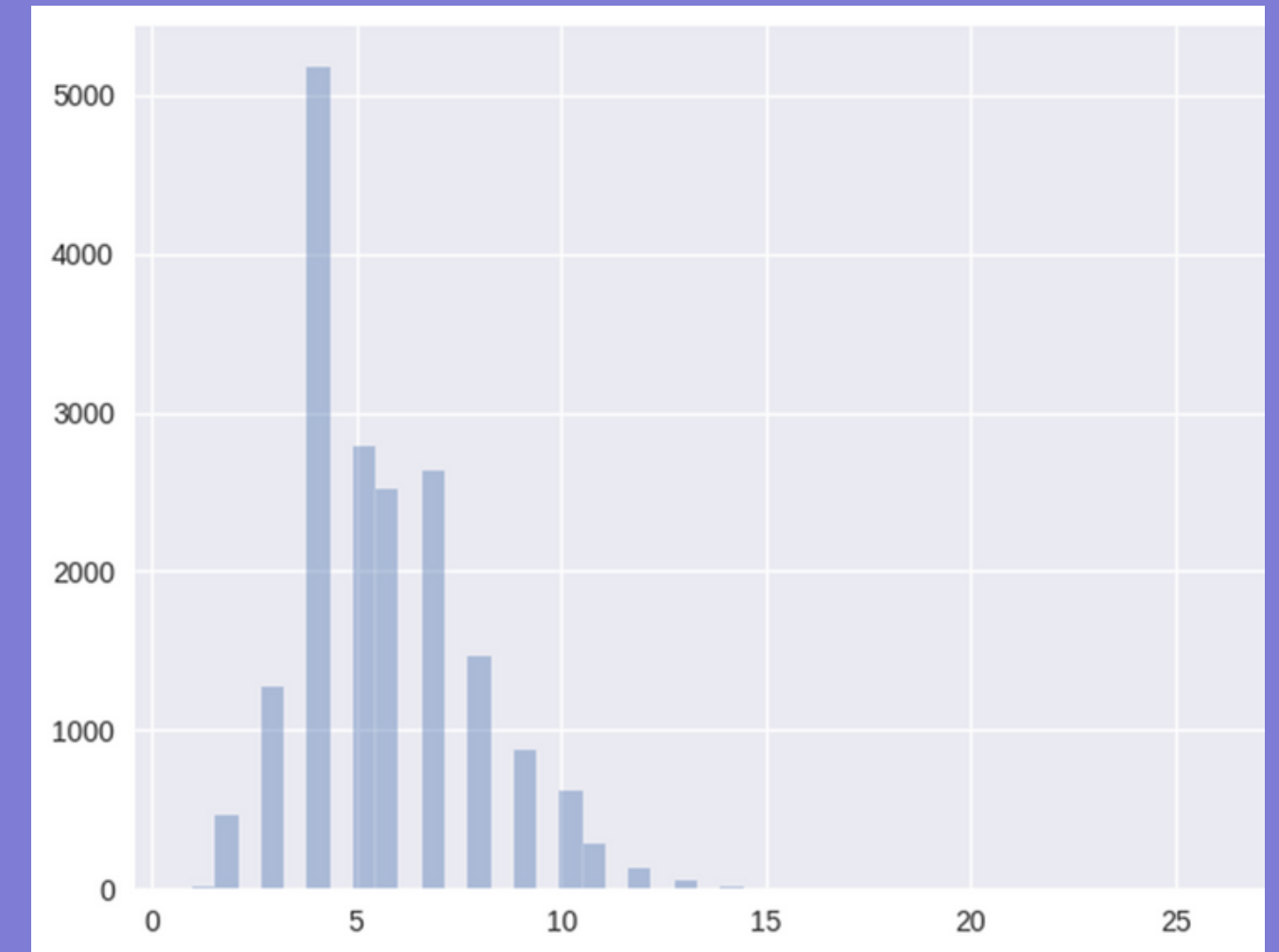


Gráfica de frecuencias de las 50 palabras más usadas en el dataset.

DATOS ESTADÍSTICOS DE LA VARIEDAD LÉXICA DE NUESTRO DATASET



Histograma para visualizar la frecuencia de longitudes de palabras y oraciones



Gráfica de barras de frecuencia de las 20 palabras más comunes



Nubes de palabras para identificar temas importantes

CLASIFICACIÓN SUPERVISADA Y NO SUPERVISADA

Para efectos de practicar la regresión logística, importamos un dataset más que es acerca de una predicción de un ataque al corazón, evaluando ciertos parámetros médicos que influyen en el padecimiento y su categorización.



	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	output
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

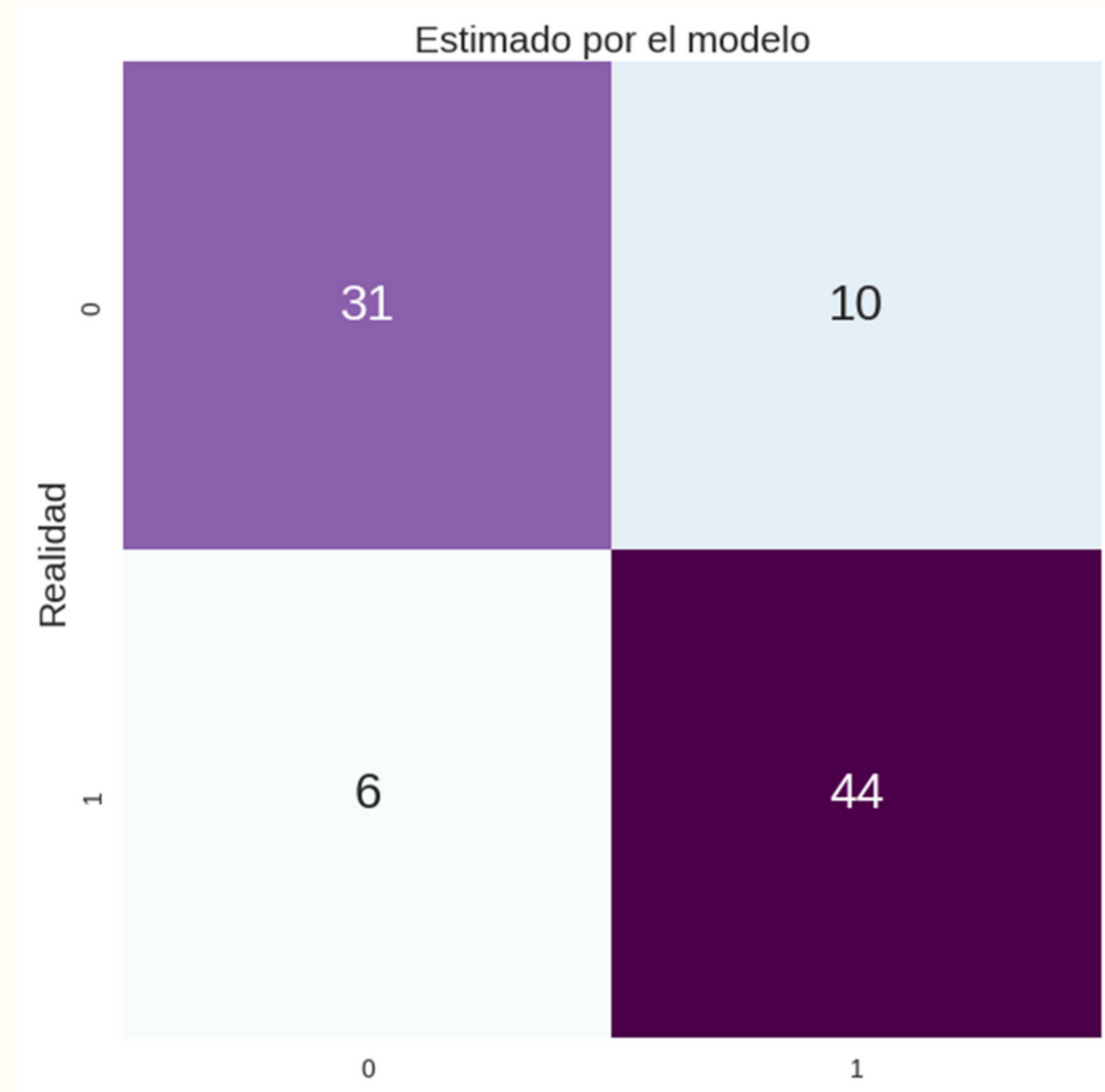
SCORE DE LA REGRESIÓN LOGÍSTICA

```
✓ [167] logreg.score(X_test, y_test)  
0.8241758241758241
```

El score obtenido de la aplicación de la regresión logística en nuestra variable 'target' que es el indicador del ataque al corazón, fue muy alto, es decir que nuestro conjunto de datos tiene una buena clasificación.

MATRIZ DE CONFUSIÓN

Obtuvimos los siguientes
resultados de nuestra
matriz de confusión:
Verdadero Positivo: 44
Falso Negativo: 6
Falso Positivo: 10
Verdadero Negativo: 31



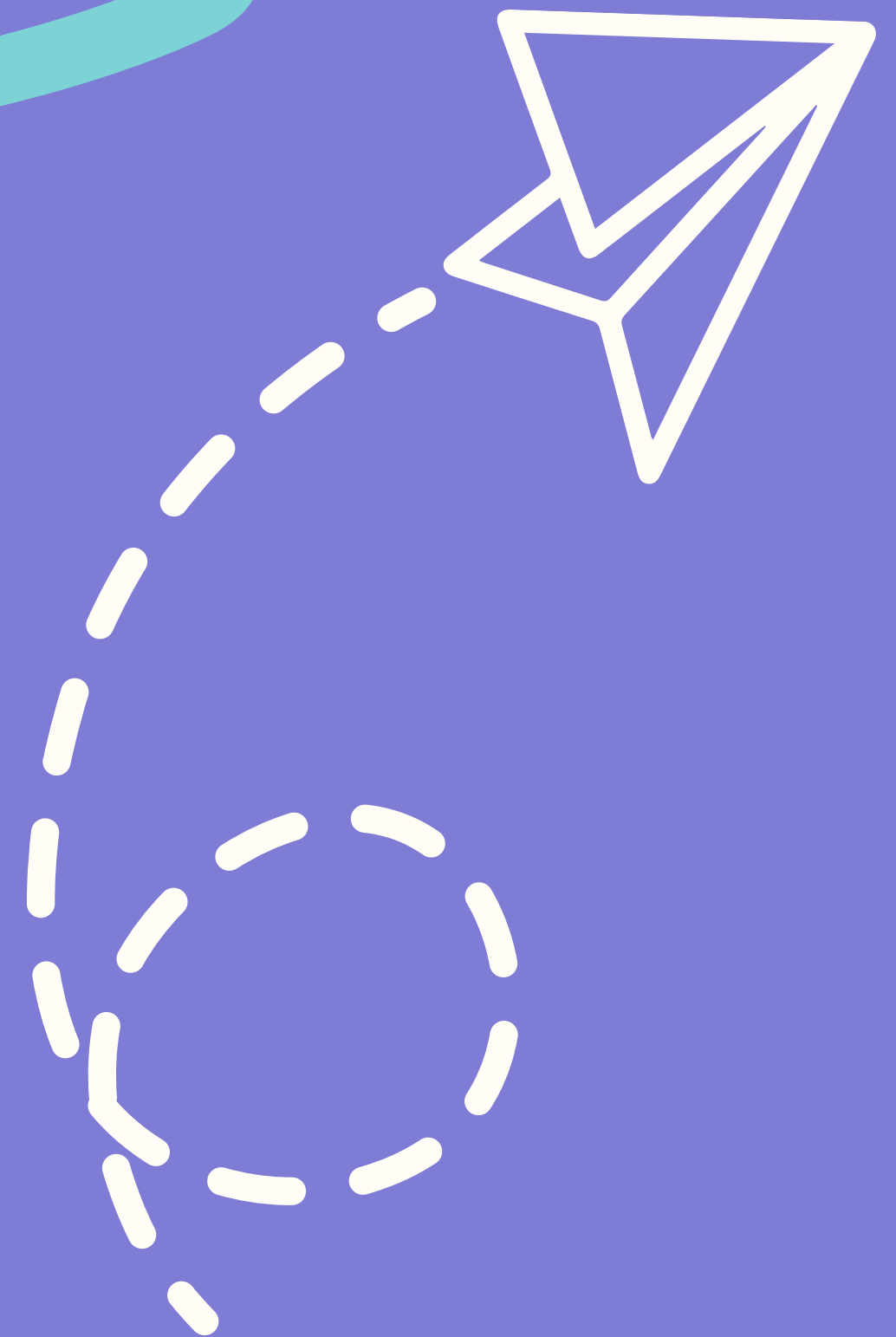
MEDIDAS DE PRECISIÓN, EXACTITUD, SENSIBILIDAD Y ESPECIFICIDAD

Precisión: 0.8148148148148148

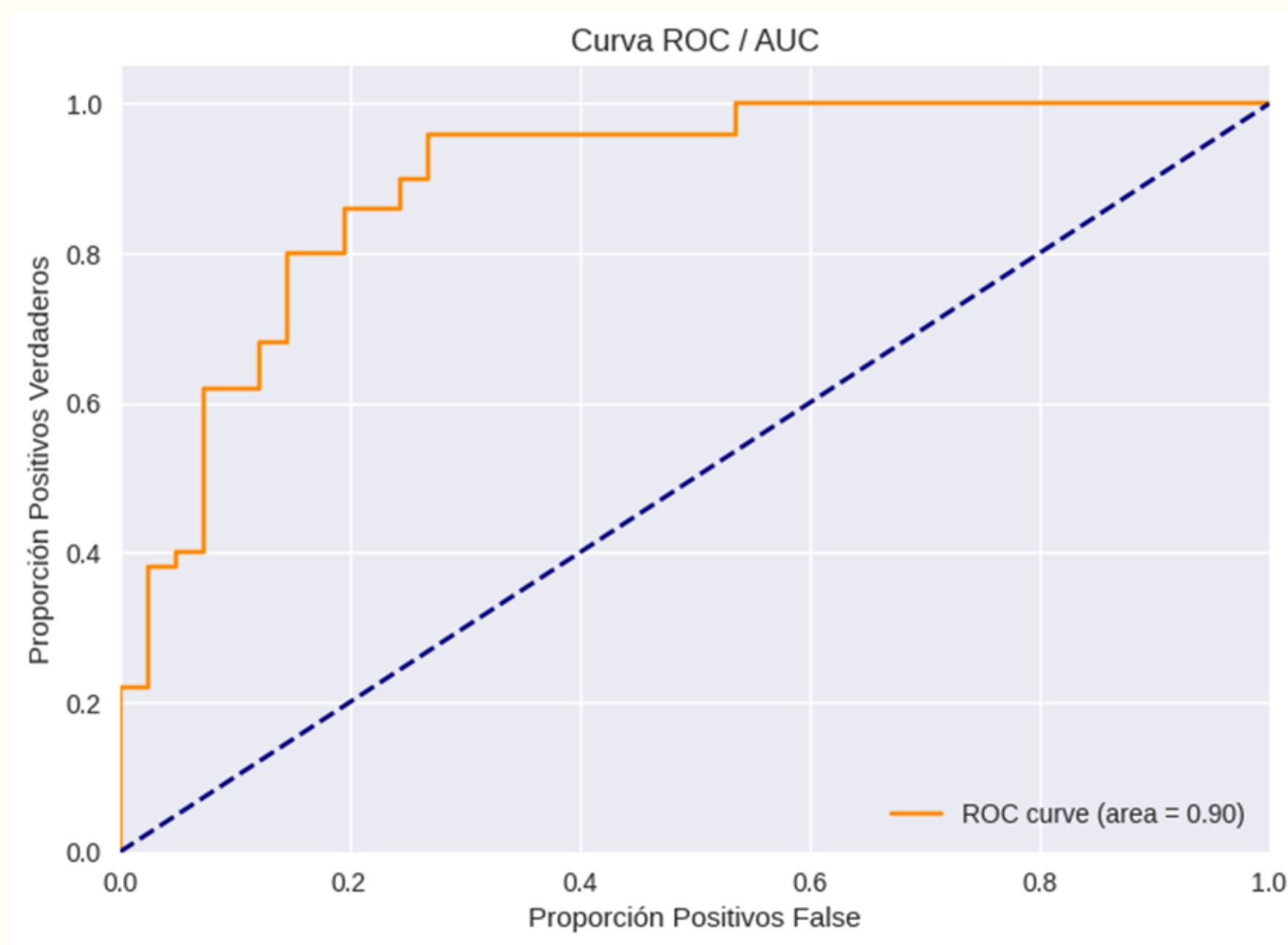
Exactitud: 0.8241758241758241

Sensibilidad: 0.88

Especificidad: 0.7560975609756098



CURVA ROC / AUC



FUENTES

- Dataset Smart Shopping:
<https://www.kaggle.com/datasets/zeesolver/fashion>
- Dataset Feelings (NLP):
<https://www.kaggle.com/datasets/praveengovi/emotions-dataset-for-nlp>
- Dataset Heart Attack (Regresión Logística):
<https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset>