
Análisis sobre precios de dispositivos móviles



Trabajo elaborado por Guido Droblas

Sobre la base de datos

Este estudio examina un conjunto de datos extraídos de publicaciones de ventas en Amazon, contiene información sobre los precios de los teléfonos móviles y diversas características (como el nombre, marca, modelo, capacidad de la batería, etc).





Introducción

El **objetivo** de esta investigación es analizar las **relaciones y los patrones** dentro del conjunto de datos para obtener información sobre los **factores** que pueden influir en los **precios** de los teléfonos móviles.

Al explorar las correlaciones entre las diferentes variables, podemos identificar los determinantes clave que impactan en la fijación de precios de los teléfonos inteligentes..



Preguntas a resolver

- ¿La **marca** de un teléfono móvil afecta su **precio**?
- ¿Existe una **correlación** entre el **tamaño** de la **pantalla** y la **capacidad** de la **batería**?
- ¿Qué impacto tiene la **resolución** de la pantalla en el **precio** de un teléfono?
- ¿Influye la **capacidad de almacenamiento interno** en la definición de ambas **cámaras (delantera y trasera)**?
- Además se intentará crear un modelo capaz de predecir el valor de un teléfono usando los parámetros del Data Frame.

Contexto

Una muy reconocida empresa de venta de celulares me solicita crear un modelo que prediga (basándose en ciertos datos) si la marca de determinado celular es tope de gama o no. Esto podría ser beneficioso ya que permite analizar las nuevas marcas de teléfonos, posibles competencias, que se están creando últimamente, como por ejemplo Nothing Phone.

Esta historia es solo para fines ilustrativos.



Este sería un posible ejemplo aplicado a una situación verosímil.



EDA – Exploratory Data Analysis

- Análisis Univariado
- Análisis Bivariado
- Análisis Multivariado



Valores nulos y duplicados

```
Conteo nulos:
unnamed:_0      0
name            0
brand           0
model           0
battery_capacity_(mah)  0
screen_size_(inches)  0
touchscreen     0
resolution_x    0
resolution_y    0
processor       0
ram_(mb)        0
internal_storage_(gb)  0
rear_camera     0
front_camera    0
operating_system  0
wi-fi          0
bluetooth      0
gps            0
number_of_sims 0
3g             0
4g/_lte        0
price          0
dtype: int64
```

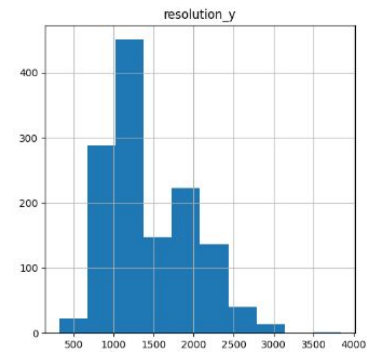
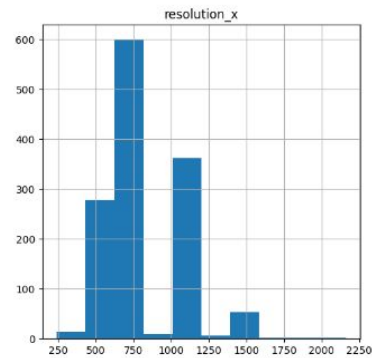
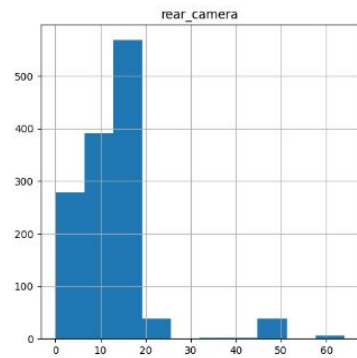
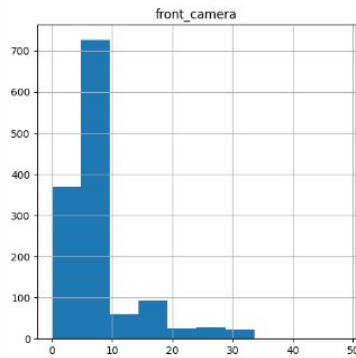
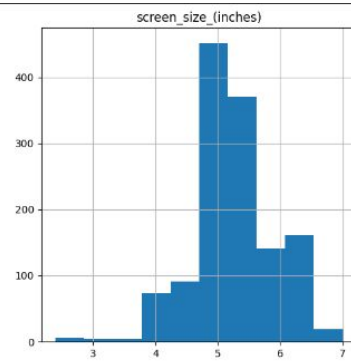
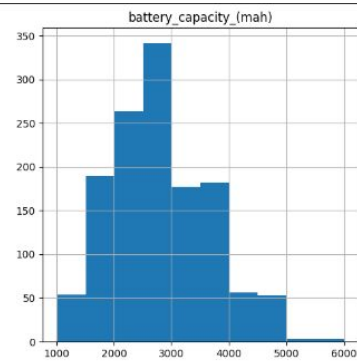
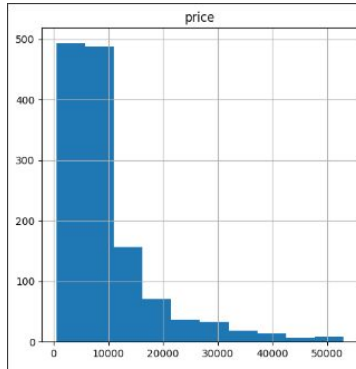
```
Conteo duplicados:
False    1359
dtype: int64
```

El Dataset no cuenta con valores nulos ni duplicados.

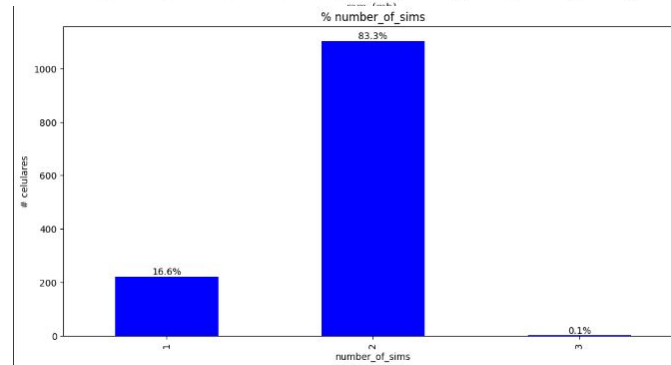
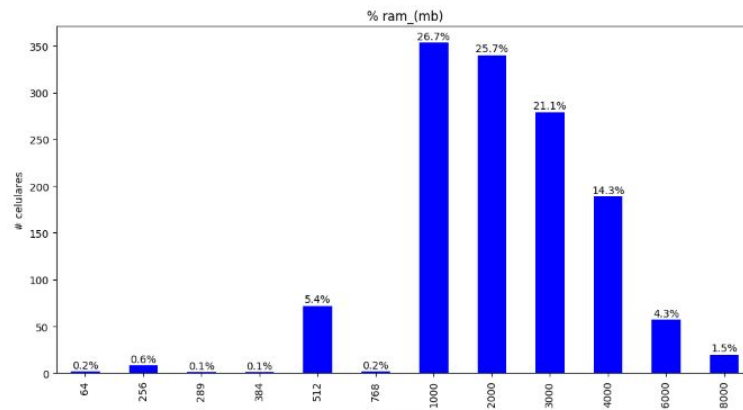
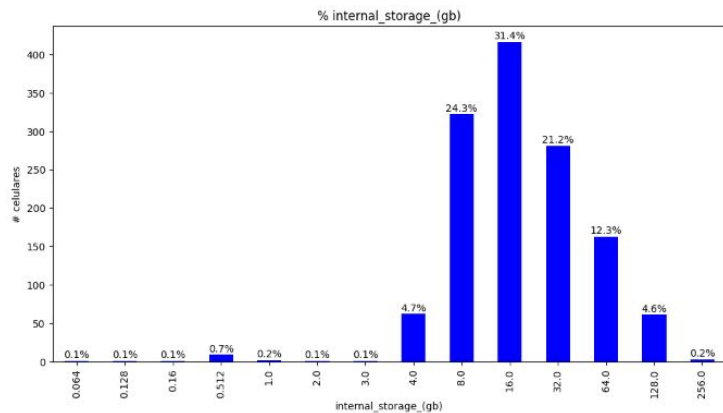
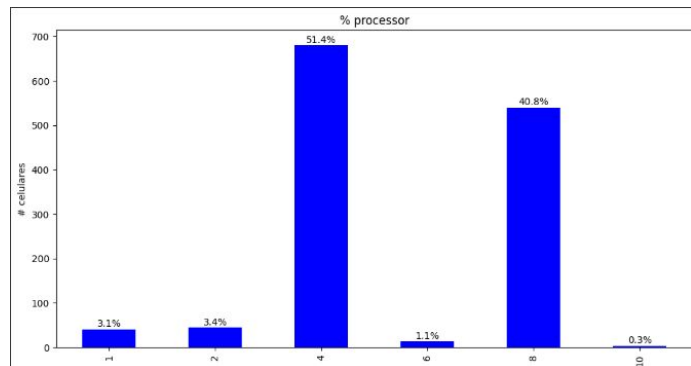
```
Información:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1359 entries, 0 to 1358
Data columns (total 22 columns):
#   Column                Non-Null Count  Dtype
---  -
0   unnamed:_0            1359 non-null  int64
1   name                  1359 non-null  object
2   brand                 1359 non-null  object
3   model                 1359 non-null  object
4   battery_capacity_(mah) 1359 non-null  int64
5   screen_size_(inches)  1359 non-null  float64
6   touchscreen           1359 non-null  int64
7   resolution_x          1359 non-null  int64
8   resolution_y          1359 non-null  int64
9   processor             1359 non-null  int64
10  ram_(mb)              1359 non-null  int64
11  internal_storage_(gb)  1359 non-null  float64
12  rear_camera           1359 non-null  float64
13  front_camera          1359 non-null  float64
14  operating_system      1359 non-null  object
15  wi-fi                 1359 non-null  int64
16  bluetooth             1359 non-null  int64
17  gps                   1359 non-null  int64
18  number_of_sims        1359 non-null  int64
19  3g                    1359 non-null  int64
20  4g/_lte               1359 non-null  int64
21  price                 1359 non-null  int64
dtypes: float64(4), int64(14), object(4)
memory usage: 233.7+ KB
None
```

Información sobre las columnas del Dataset.

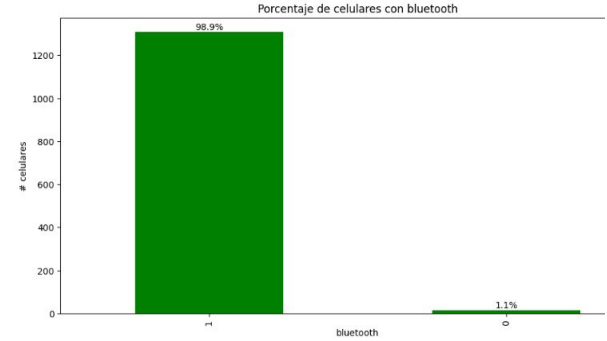
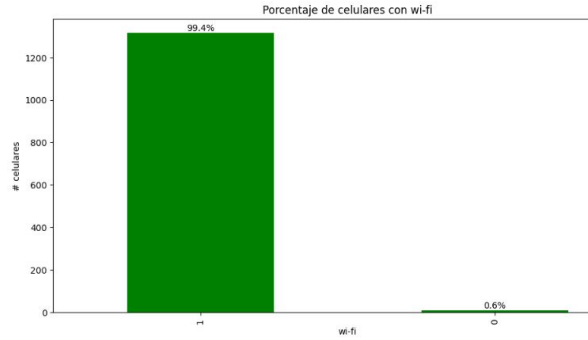
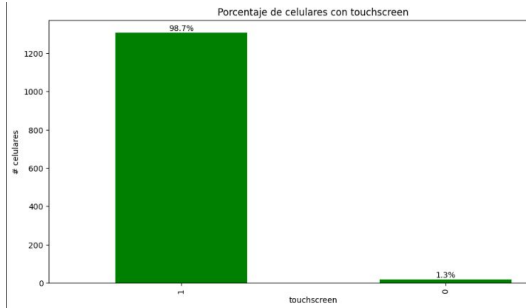
Análisis Univariado



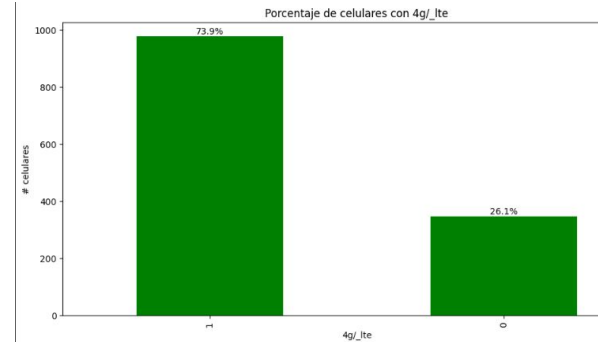
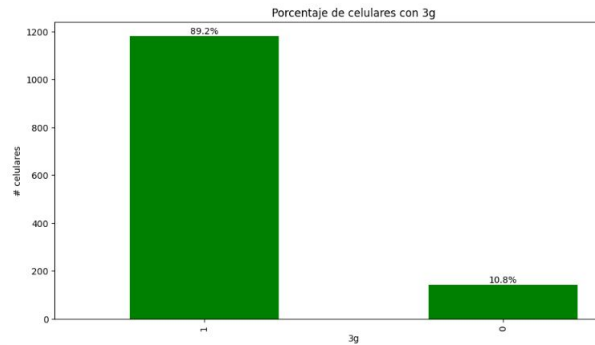
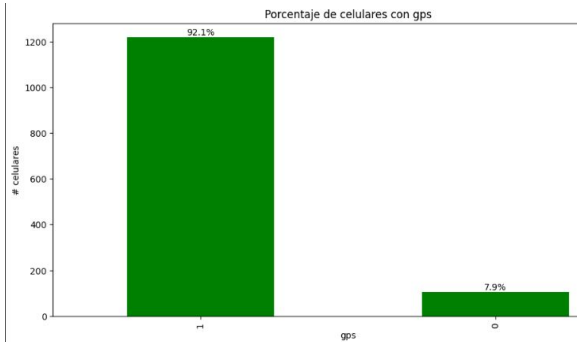
Análisis Univariado



Análisis Univariado



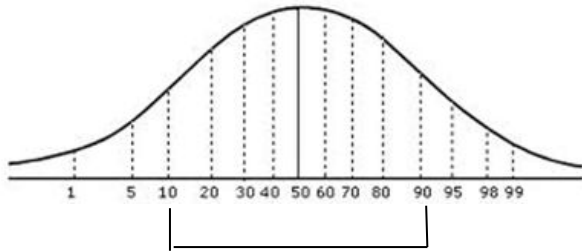
Observando estos gráficos podemos concluir que casi todos los datos corresponden a celulares inteligentes (smartwatch), en su mayoría cuentan con pantalla táctil y bluetooth.





Valores Atípicos

En este punto, con la ayuda visual de los gráficos de análisis univariados, se eliminaron algunos valores atípicos que causaban ruido.



Se tomaron como válidos los datos correspondientes a valores entre los cuantiles 10% y 90%

Se eliminaron 35 datos, un 2.58 % del total.



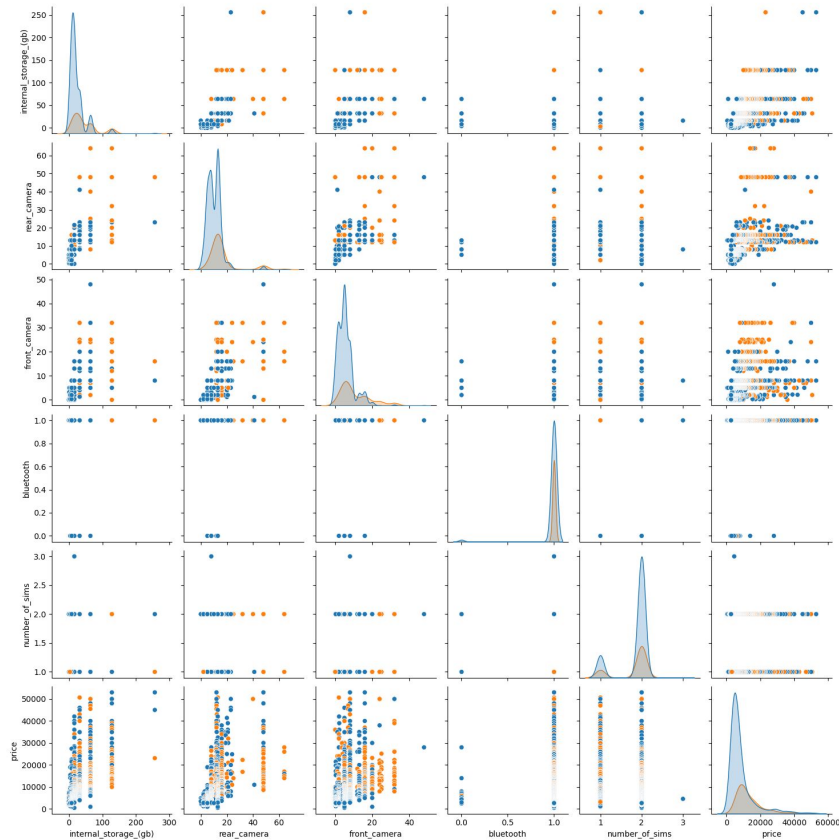
Ingeniería de variables

Se agregaron las columnas "brand_type" la cual tiene valores: "Famosa" o "No famosa" según la marca, tomado de un dato extra estadístico sacado de un de un estudio realizado por Omdia en el 2Q2022, dividiendo las marcas entre "reconocidas mundialmente" y "pequeñas".

Además, he agregado la columna "gama" de valores: "Alta" o "Baja", la cual toma el promedio de precio de la marca, si el valor del celular es mayor a este lo considera alto será de gama alta y si está por debajo, gama baja.

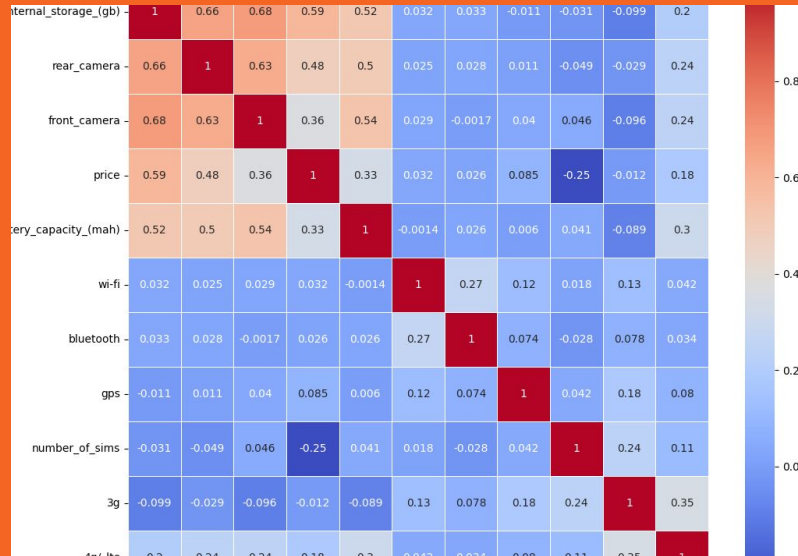
Por último, agregué la columna "top", donde separo en 3 grupos todos los celulares según su precio. Los tamaños de los grupos pueden ajustarse.

Análisis Bivariado



Vemos las relaciones
1 a 1 de las variables

Análisis Multivariado



Vemos las correlaciones entre las variables

Siendo las más fuertes:

- front_camera y internal_storage (0.68)
- rear_camera y internal_storage (0.66)
- front_camera y rear_camera (0.63)
- price y internal_storage (0.59)

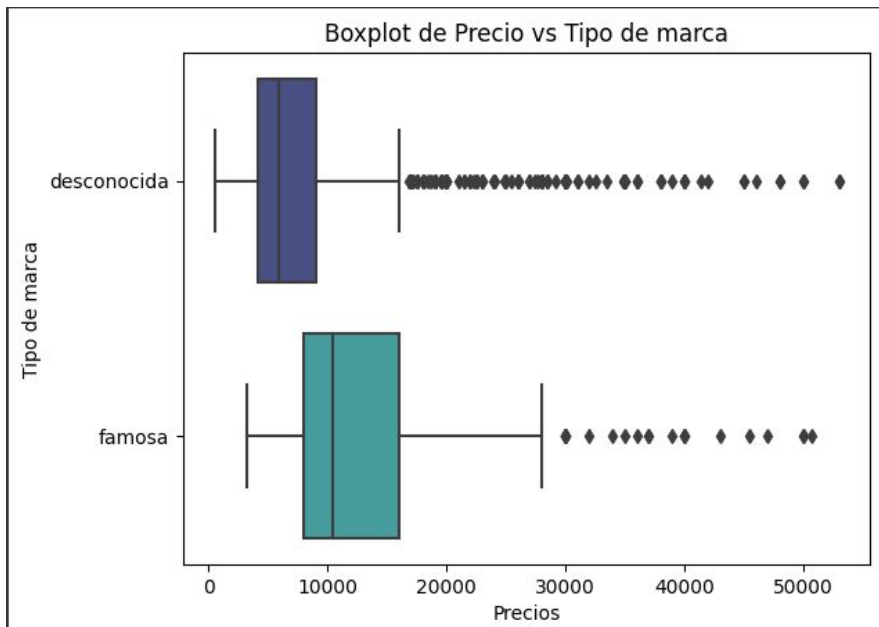


Solución a preguntas planteadas

¿La marca de un teléfono móvil afecta su precio?

Hipótesis: Las marcas premium tienden a tener precios más altos en comparación con las marcas menos conocidas o de bajo presupuesto.

SAMSUNG

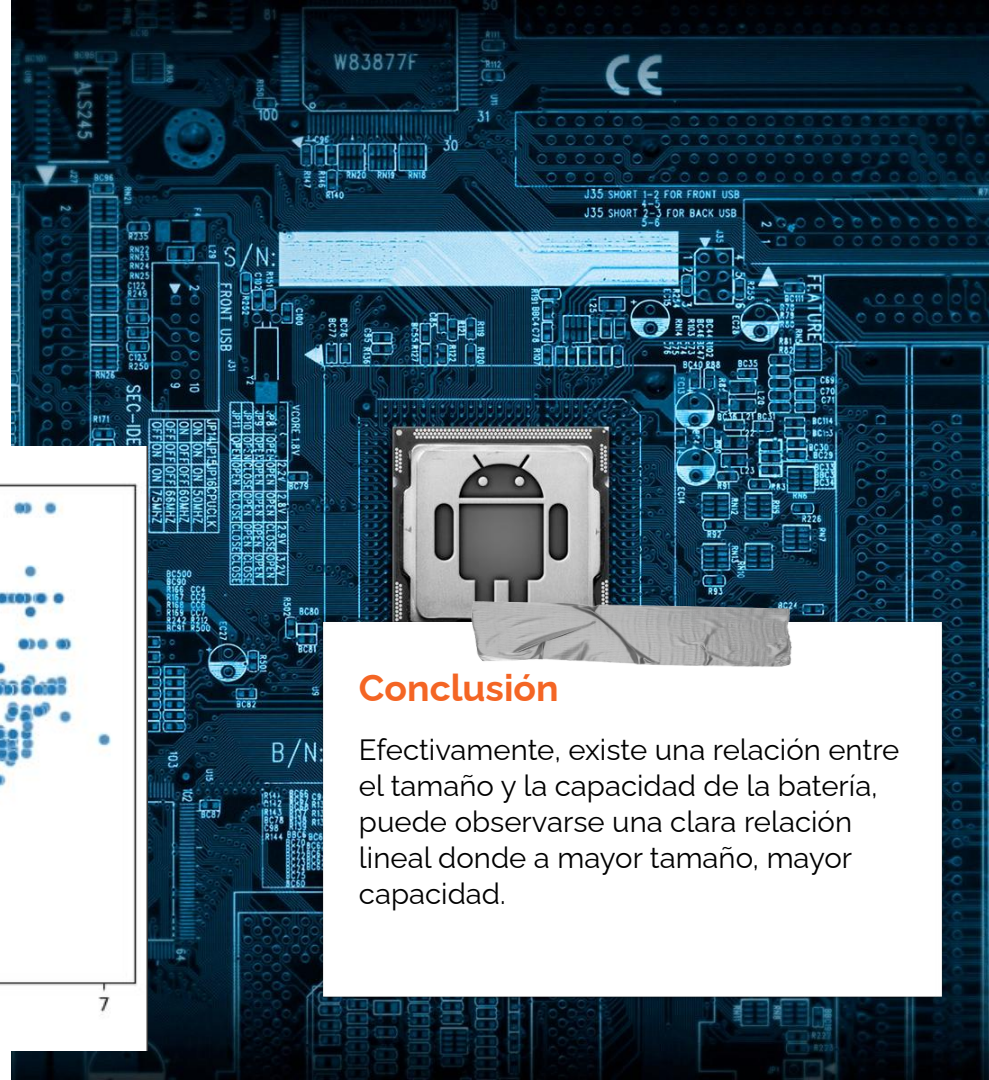
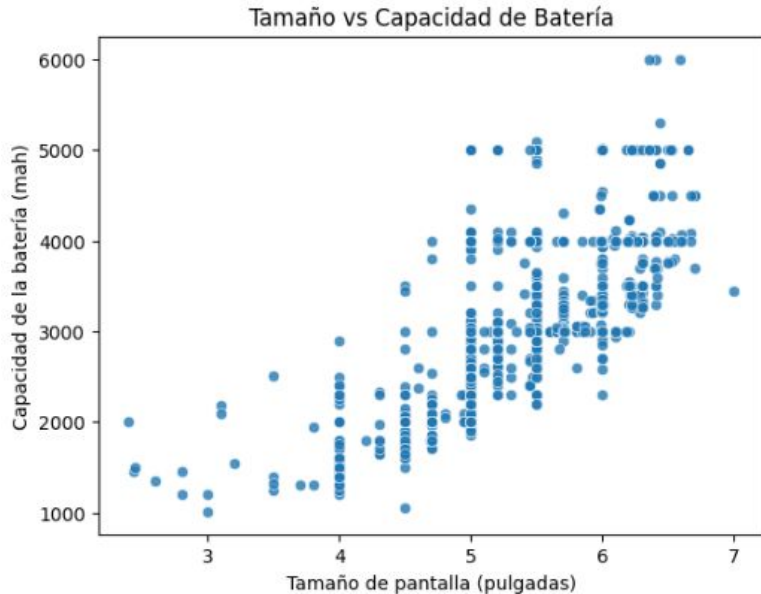


Conclusión

Se observa cómo las marcas desconocidas tienen mucha mayor cantidad de outliers de precio más alto que el promedio, en comparación con las conocidas y cómo estas tienen una menor media.

¿Existe una correlación entre el tamaño de la pantalla y la capacidad de la batería?

Hipótesis: Los tamaños de pantalla más grandes se asocian con mayor espacio para la batería.

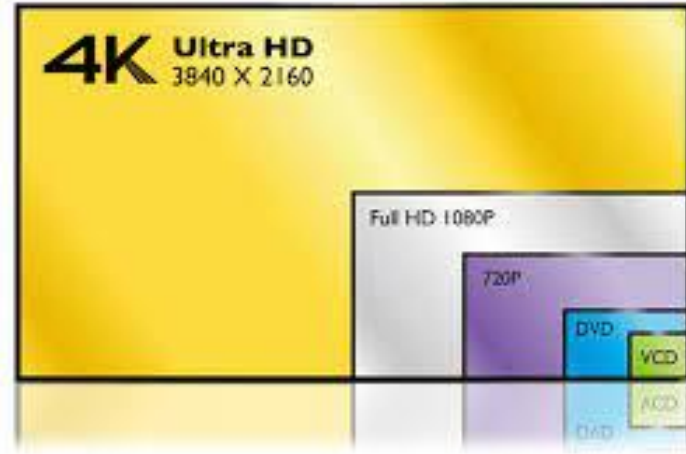


Conclusión

Efectivamente, existe una relación entre el tamaño y la capacidad de la batería, puede observarse una clara relación lineal donde a mayor tamaño, mayor capacidad.

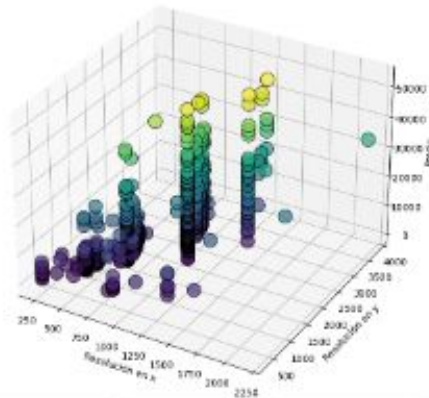
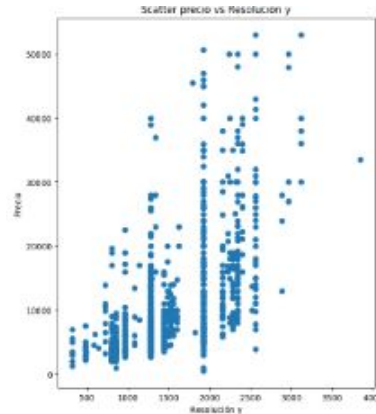
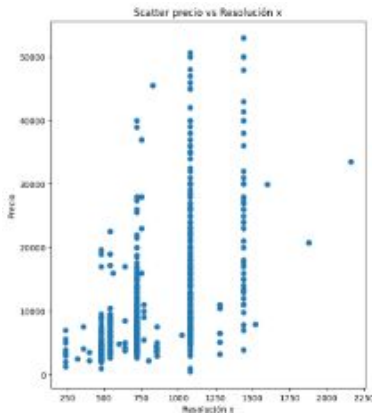
¿Qué impacto tiene la resolución de la pantalla en el precio de un teléfono?

Hipótesis: Los teléfonos con pantallas de alta resolución es probable que tengan precios más altos, ya que ofrecen una mejor calidad visual.



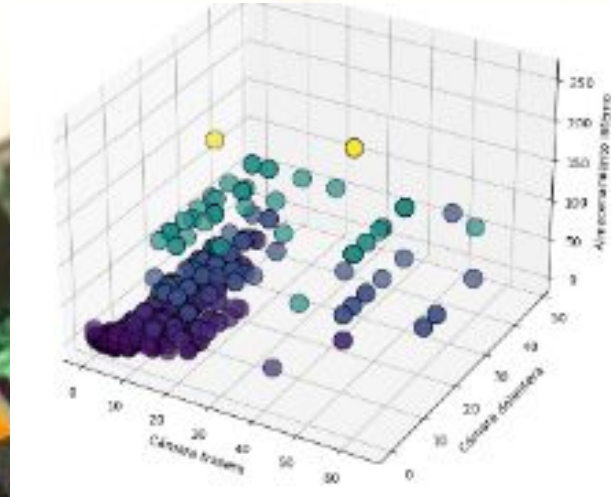
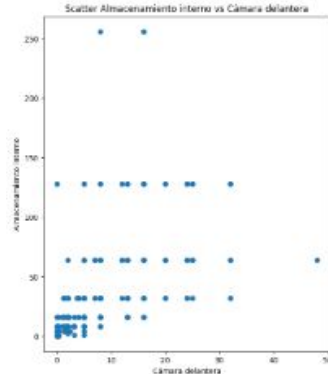
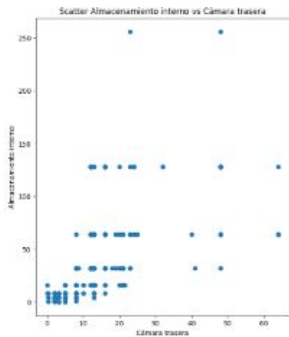
Conclusión

No existe una clara relación entre estos factores. Supongo se debe a que las resoluciones tienen pre establecidos (HD+ 720×1280 píxeles, Full HD+ 1080×1920 píxeles y WQHD+ 1440 x 2560 píxeles).



¿Influye la capacidad de almacenamiento interno en la definición de ambas cámaras (delantera y trasera)?

Hipótesis: Los teléfonos móviles con mayores capacidades de almacenamiento interno deben a tener más capacidad de guardar fotografías más pesadas, ya que ofrecen más espacio para aplicaciones, medios y archivos.



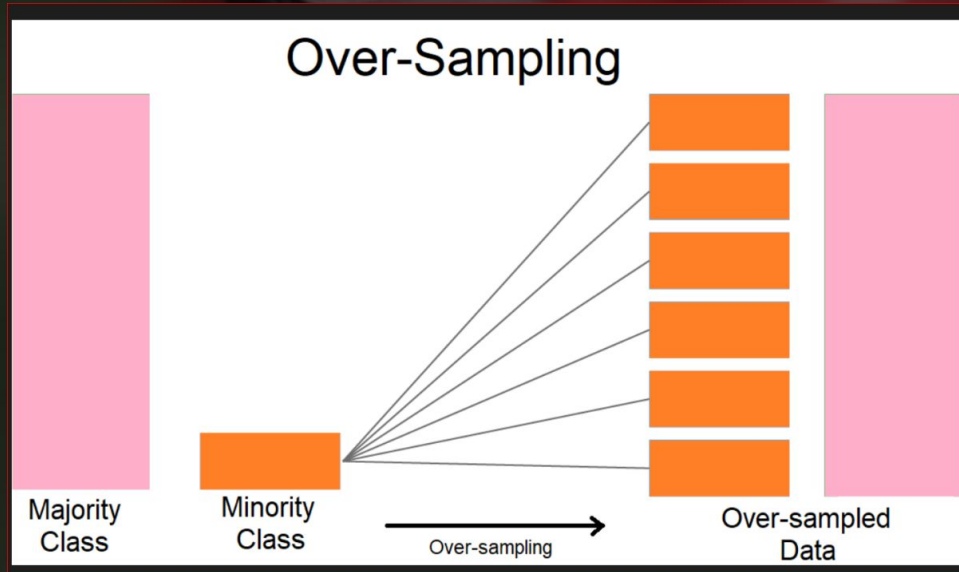
Conclusión

Si bien en este caso hay mayor dependencia entre las variables que en el caso anterior, no es lo suficientemente contundente como para considerarla una relación fuerte.

Modelos de ML



Over-Sampling



Al presentarse un desbalance de 73.3% a 26.7%, a favor de las marcas “desconocidas”, se utilizó un procedimiento de Over-Sampling, el cual iguala ambas cantidades.

Modelos de ML probados :

- **Decision Tree Classifier**
 - Con selección de variables Forward y Floating
- **Support Vector Machine**
 - Probando valores de c de 0.1, 0.2, 0.3
- **Random Forest**
 - Con Cross Validation: Stratified KFold

Métricas

Modelo	Parámetros	Score
Tree Classifier	['Ramas ', 17, 'Variables ', 13]	['F1 Score', 0.87625]
Support Vector Machine	['c ', 0.1]	['F1 Score', 0.73071]
Random Forest	n_estimators=30, criterion='gini',max_depth=4	['F1 Score', 0.7907]

Para el modelo de predicción de precio:

Mejor modelo: Tree Classifier

Accuracy: 87,63%

Método: Forward Selection

Ramas (17):

Variables tenidas en cuenta (13):

- 'battery_capacity_(mah)'
- 'screen_size_(inches)'
- 'touchscreen'
- 'resolution_x'
- 'resolution_y'
- 'processor'
- 'ram_(mb)'
- 'internal_storage_(gb)'
- 'rear_camera'
- 'front_camera'
- 'operating_system'
- 'wi-fi'
- 'price'

Gracias por la atención

Para ver el trabajo completo dirigirse a

[Colab](#)

o

[Github](#)