UNIVERSIDAD DEL SINÚ

Elías Bechara Zainúm

Seccional Cartagena

Análisis de datos a Zomato.csv

Kennet Morales

Sara Jiménez

Juan Mosquera

Nestor Ospino

Luis Mario Mercado

Jhesus Mulford

04 / 03 / 2024

Universidad del Sinú seccional Cartagena de Indias

Sede Plaza Colón

Docente

Cristian Cuadrado Beltrán

CONTENIDO:

-Introducción

-Objetivos

-Descripción de los datos

-Resumen estadístico inicial

-Relaciones entre variables

-Identificación de outliers

-Transformación de datos

-Exploración temporal

-Conclusiones y hallazgos

-Código y metodología

- **Introducción:**

En este informe podrás encontrar la manipulación y el análisis de la data “Zomato”, siendo necesario decir que no tenemos un contexto previo ni el conocimiento del objetivo que actualmente están teniendo estos datos, en el que inicialmente identificamos como “precios” en fechas específicas, relacionando estos datos como “ventas”; sin embargo, lo que se pretende a continuación es realizar un análisis descriptivo con el cual se pueden identificar patrones y a su vez los posibles riesgos.

- **Justificación**

- **Objetivos:**

- Realizar el análisis exploratorio de datos EDA.
- Comprender y dar un sentido a los datos.

- **Descripción de los datos:**

Empezando desde cero explorando la data Zomato se comprende su estructura y el significado de sus datos; la data está compuesta por 631 filas y 7 columnas llamadas “Date”, “Open”, “High”, “Low”, “Close”, “Adj Close”, “Volume”; siendo *Date* un dato temporal y todos los demás datos numéricos (*Open, High, Low, Close* y *Adj Close* float64 y *Volume* int64).

Identificando que los datos de las columnas *Open, High, Low, Close* y *Adj Close* son equivalente a un registro de precios en determinada fecha según su categoría; considerando a la columna *Open* como un precio inicial, a la columna *High* como el precio más alto, a la columna *Low* como el precio más bajo, a la columna *Close* como el precio finalmente vendido, a la columna *Adj Close* como un ajuste en el precio final; todo estos datos registrados referente a una fecha en la columna *Date,* esta data cuenta con un registro casi diario a partir del 23-07-2021 hasta el 07-02-2024, es decir, en un lapso de tiempo de 913 días se registraron 631 ventas.

Con lo anterior podemos decir que la data zomato es el registro minucioso de los precios de las ventas que se realizaron ciertos días en un determinado rango de tiempo.

- **Resumen estadístico:**

Posteriormente hemos identificado que no se presentan datos faltantes en este conjunto de datos por ende no es necesario recurrir a estrategias para el tratamiento los datos faltantes la data y se inicia el análisis estadístico.

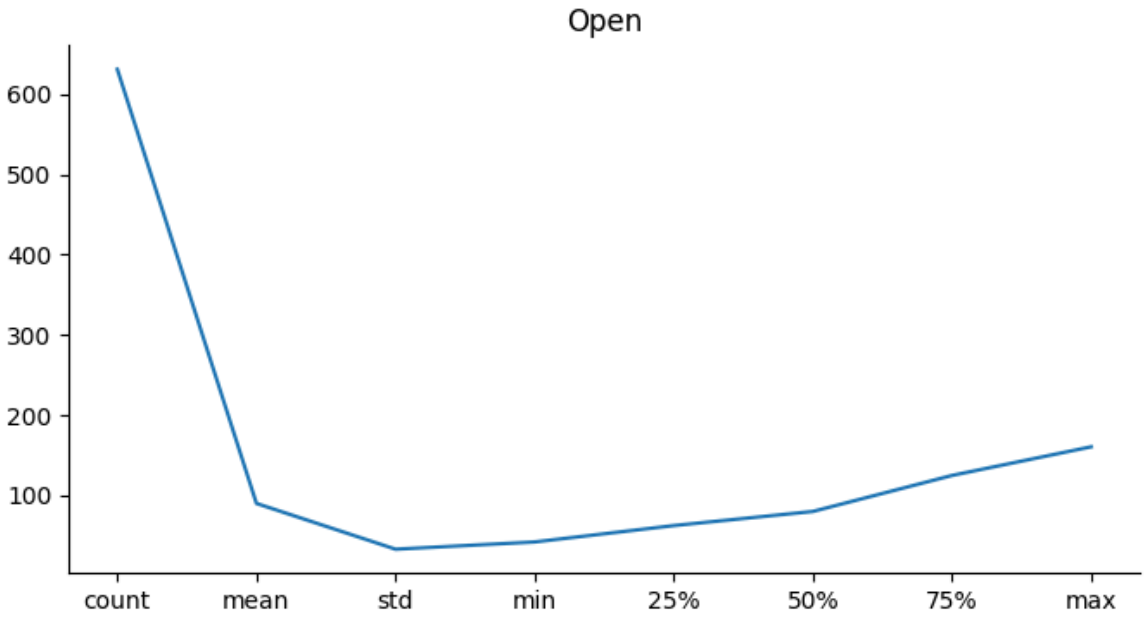
La imágen que se presenta a continuación muestra la estadística descriptiva de la data:

index	Open	High	Low	Close	Adj Close	Volume
count	631.0	631.0	631.0	631.0	631.0	631.0
mean	90.01125203803487	91.83573672107765	87.86830447226625	89.70768627099841	89.70768627099841	67317368.60697305
std	32.75763864854208	33.31954509282716	31.900848676831284	32.62189990889618	32.62189990889618	74610303.18581134
min	40.849998	44.400002	40.599998	41.650002	41.650002	0.0
25%	62.549999	63.450001	61.125	62.0749985	62.0749985	28007875.0
50%	80.0	81.0	78.099998	79.699997	79.699997	47597101.0
75%	124.4749985	126.75	121.5250015	124.5999985	124.5999985	75254391.0
max	161.149994	169.0	154.25	160.300003	160.300003	694895290.0

Columna Open

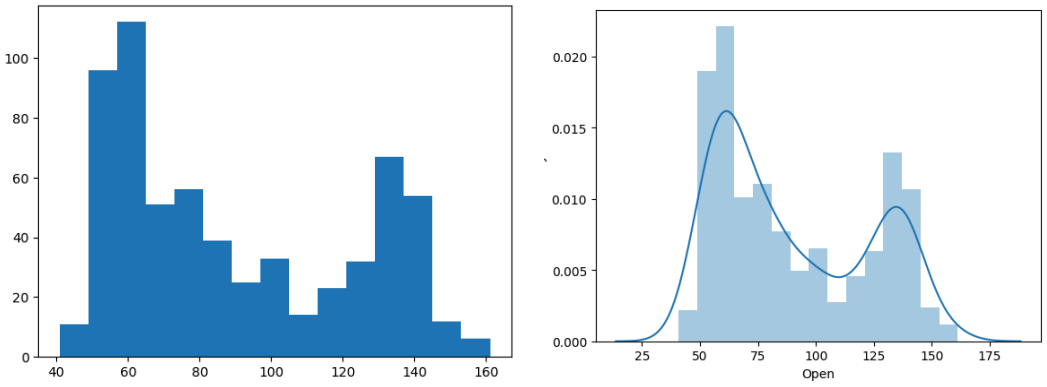
(Columna de los precios iniciales)

A continuación se muestra el resumen estadístico de la columna open:



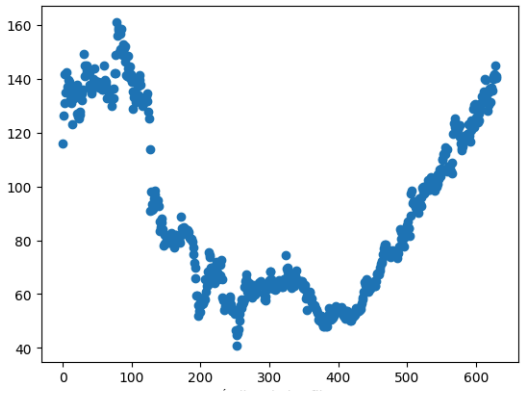
- Mediana: 90.011252
- Desviación Estándar: 32.75
- Mínimo: 40.849998
- Percentil 25 (Q1): 62.549999
- Percentil 50 (Q2): 80.000000
- Percentil 75 (Q3): 124.474998
- Máximo: 161.149994

- **Histograma**



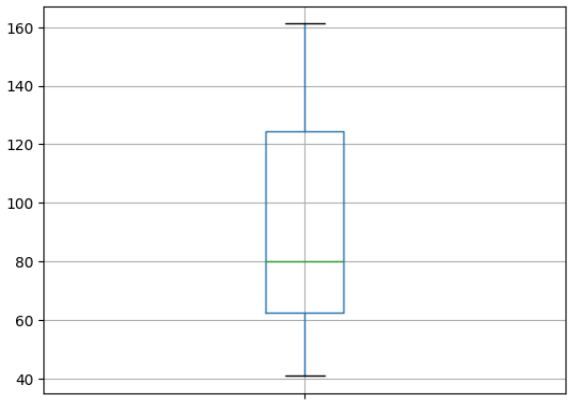
La anterior gráfica muestra la distribución simétrica con una concentración de los datos en un rango muy bajo, es decir, los precios de ventas iniciales son muy bajos en su mayoría permitiendo considerar como valores atípicos a los precios altos ya que hay menor frecuencia de ellos.

- **Diagrama de dispersión**

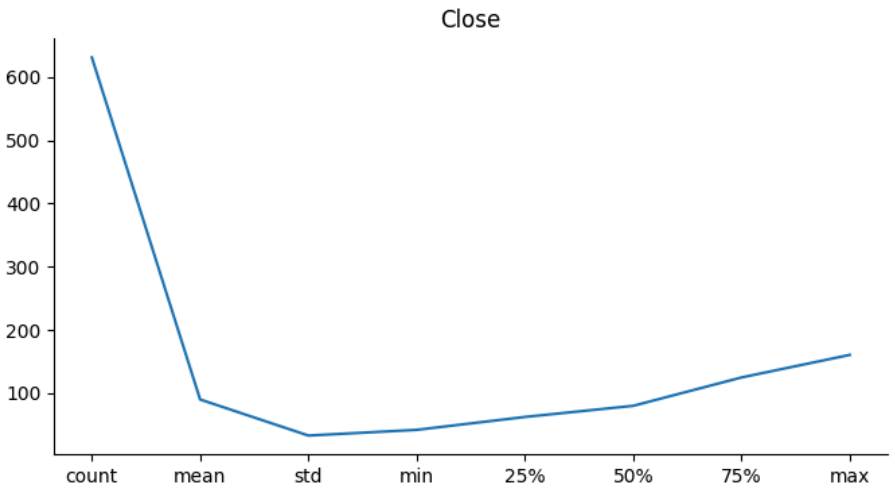


En este diagrama observamos que los puntos forman líneas ascendentes o descendentes que nos indican un patrón o tendencias fuertes en los precios iniciales de venta, por otro lado identificamos también como valores atípicos a puntos que están lejos de las líneas tendencia.

- **Boxplot**



Columna Close



(Columna de precio finalmente vendido)

Mediana: 89.707686

Desviación estándar: 32.621900

Valor mínimo: 41.650002

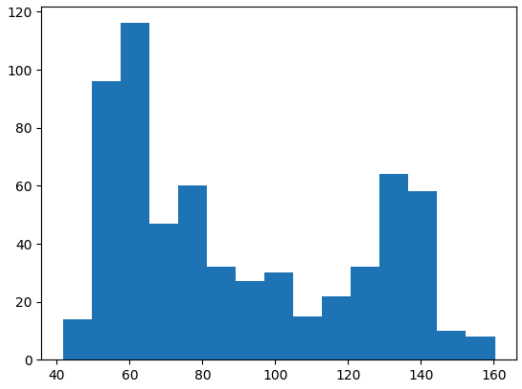
Percentil 25(Q1): 62.074998

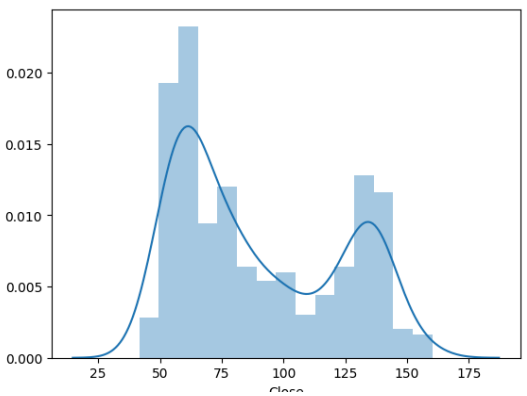
Percentil 50(Q2): 79.699997

Percentil 75(Q3): 124.599998

Valor máximo: 160.300003

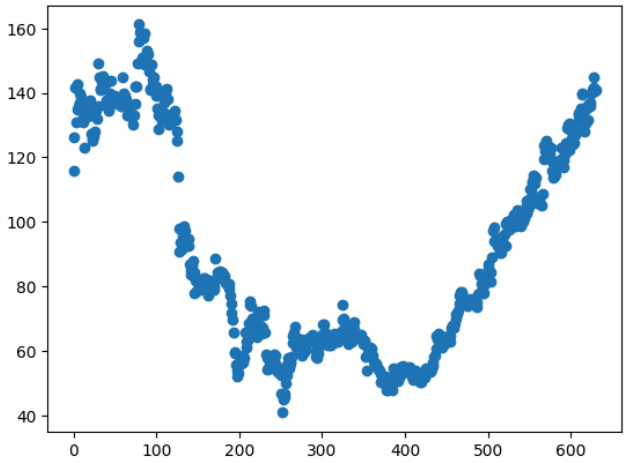
- **Histograma**





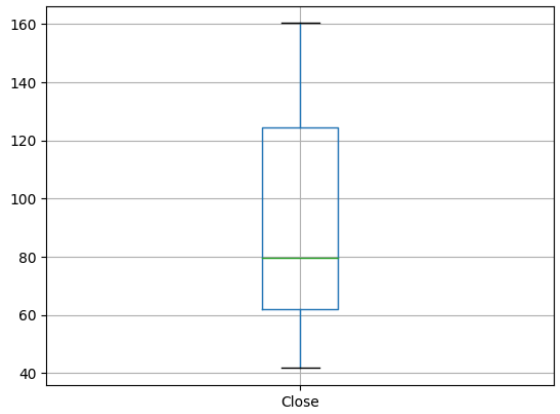
La anterior gráfica muestra la distribución simétrica con una concentración de los datos en un rango muy bajo, es decir, los precios de ventas finales en su mayoría son muy bajos considerando a los precios altos como valores atípicos ya que hay menor frecuencia de ellos.

- **Diagrama de dispersión**

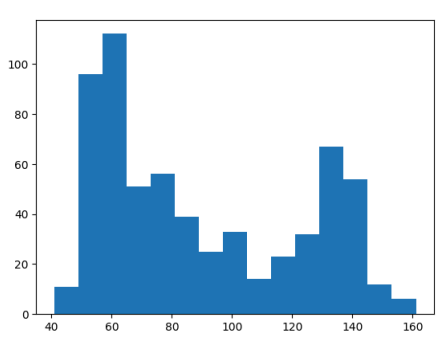


En este diagrama observamos que los puntos forman líneas ascendentes o descendentes que nos indican un patrón o tendencias fuertes en los precios finales de las ventas, identificando a su vez como valores atípicos a puntos que están lejos de las líneas tendencia.

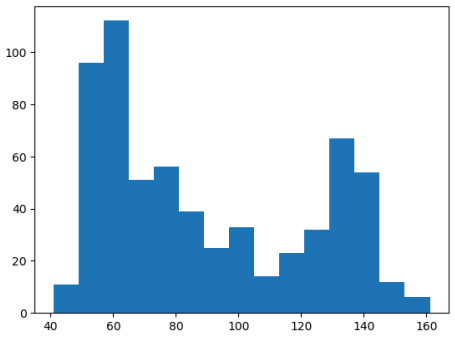
- **Boxplot**



- **Relación entre variables**



Open



Close

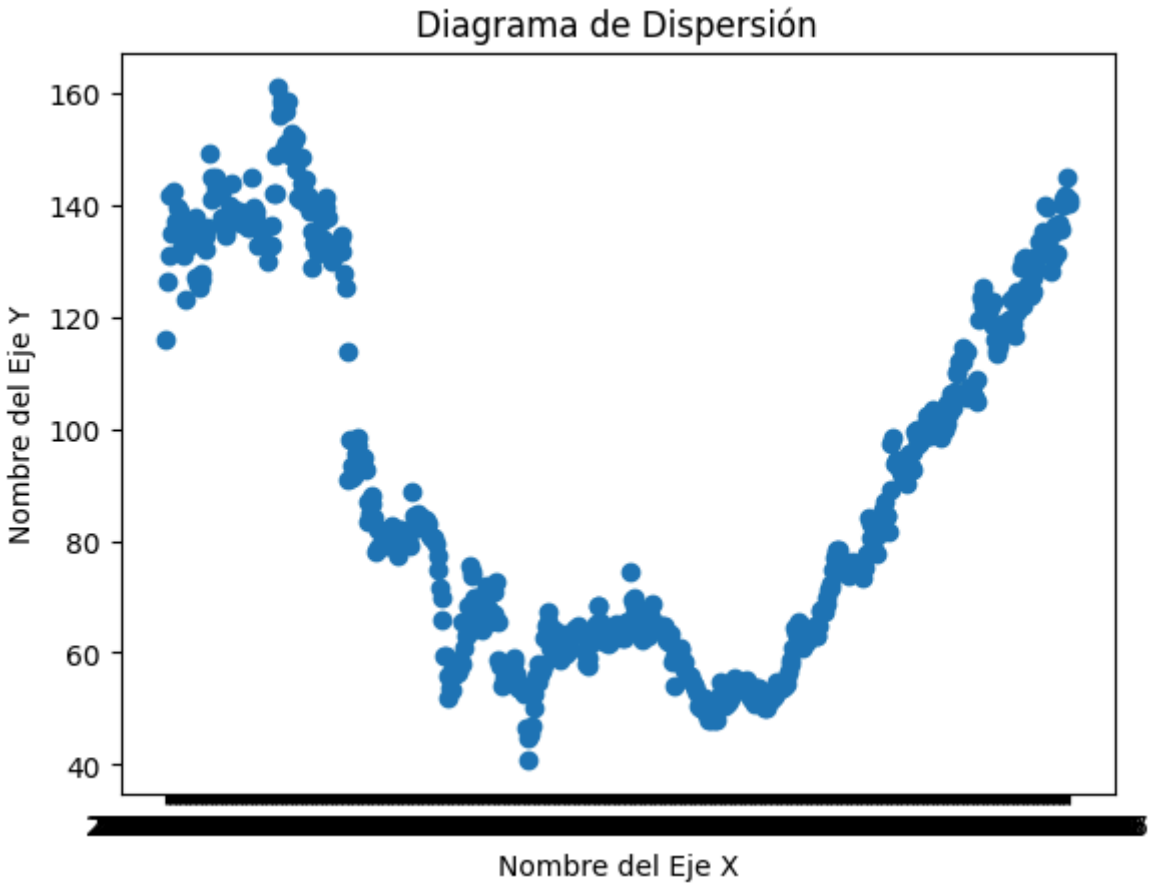
Teniendo en cuenta los gráficos anteriores y el resumen estadístico de la columna *Open* y la columna *Close* se observa que la mediana de la columna Open (Precio inicial de la venta) es un poco más alta que la columna Close (Precio final de la venta), lo que nos permite decir que las ventas fueron más bajas referente a su precio inicial

Con respecto a la desviación estándar, se puede decir que los datos de *Open* (con desviación estándar 32.75) y *Close* (32.621900) tienen una diferencia mínima.

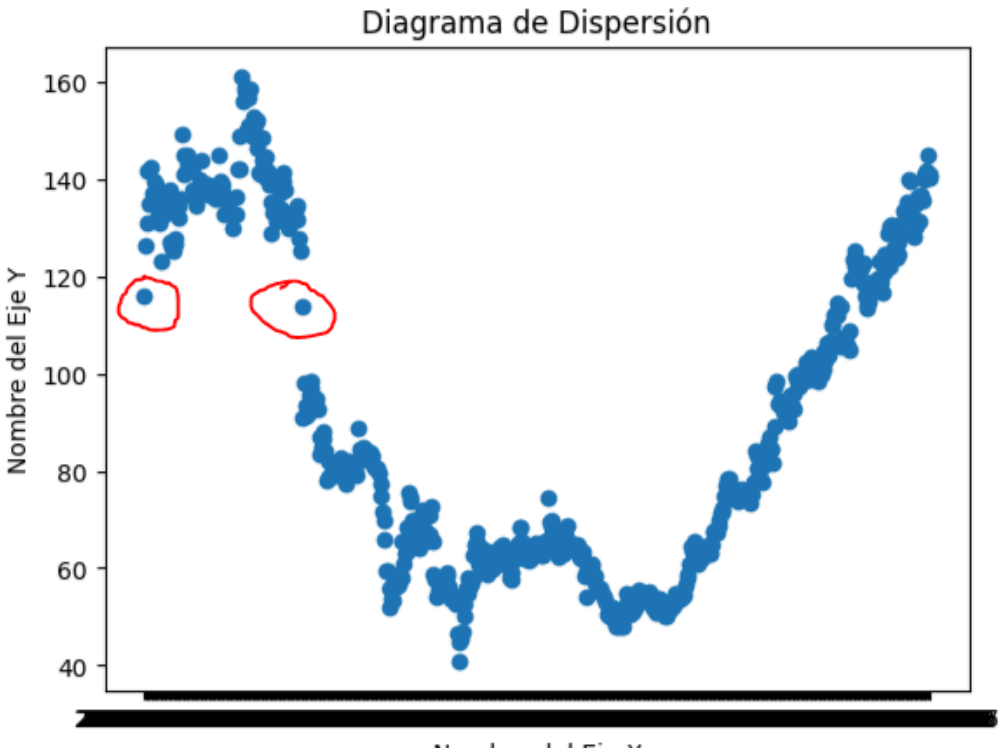
Se puede decir que los datos de *Open* están más cerca de la mediana que los datos de *Close*.

- **Identificación de outliers:**

Mediante un diagrama de dispersión que abarca todos los datos, se identificaron varios outliers.

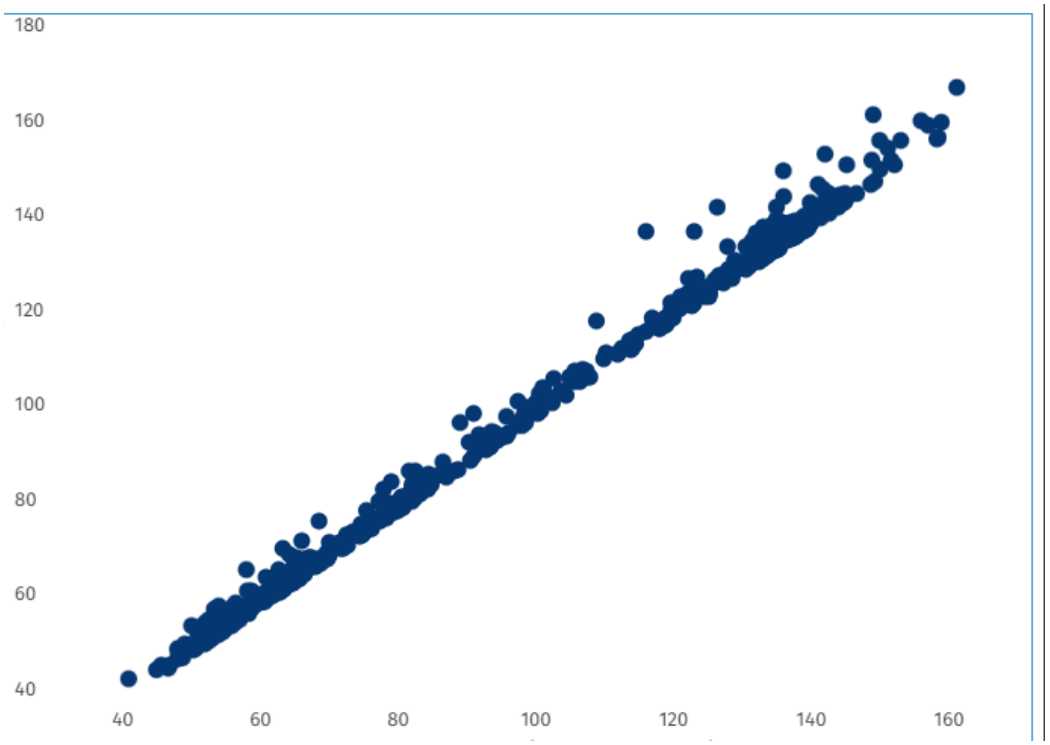


A continuación los outliers encontrados:

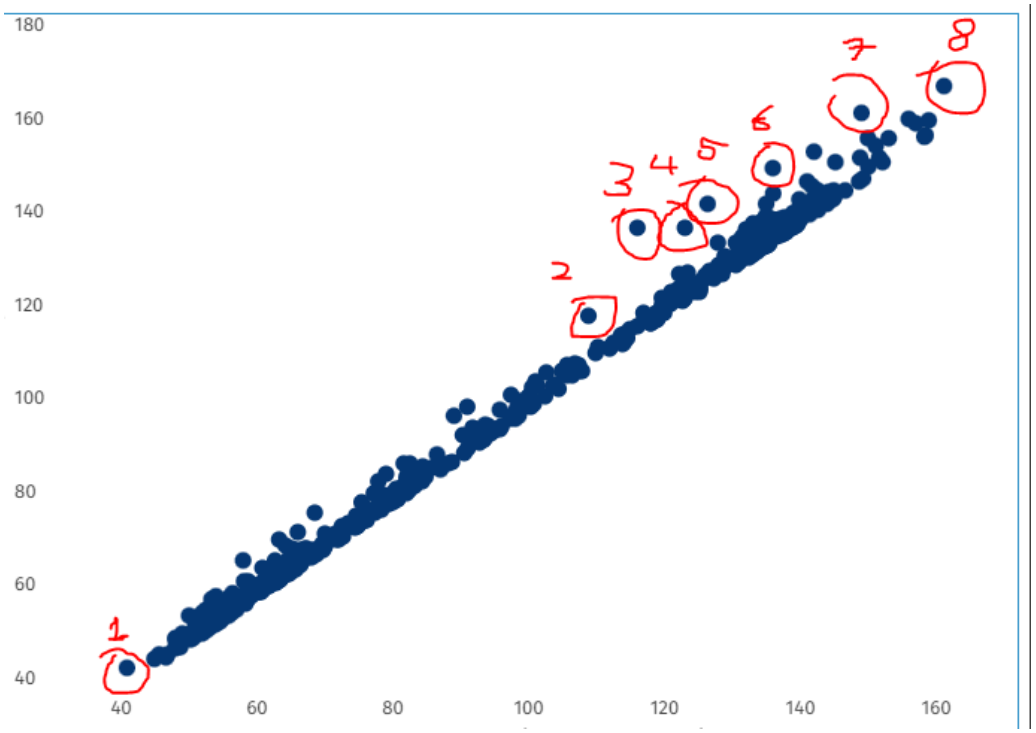


Como podemos ver, estos 2 datos aislados están entre el rango de 100 y 120.

Otro diagrama que muestra mejor y más completa la presencia de outliers es:

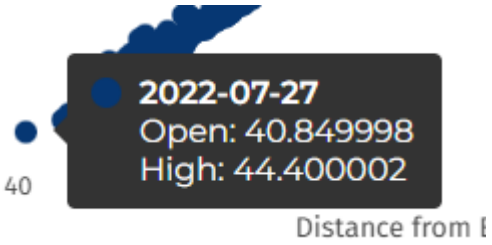


Outliers identificados:

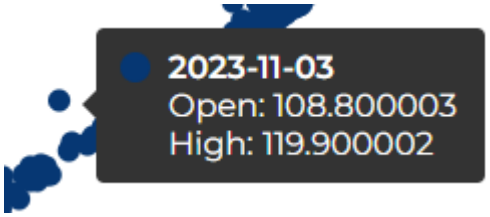


Corresponden a:

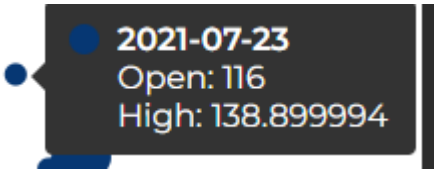
Para el número 1:



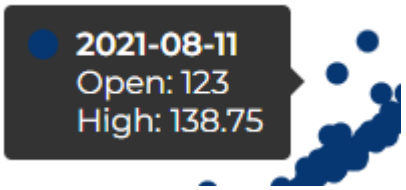
Para el número 2:



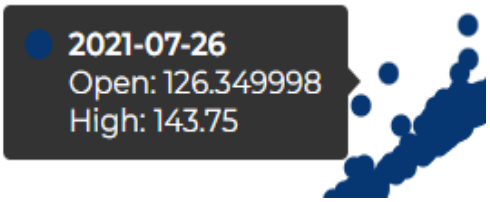
Para el número 3:



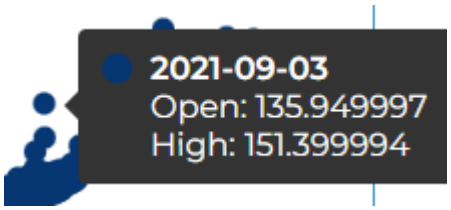
Para el número 4:



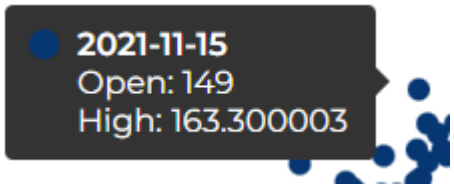
Para el número 5:



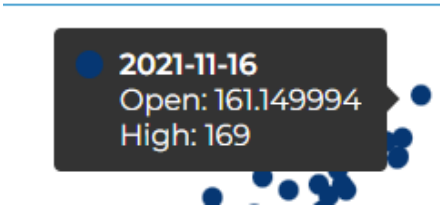
Para el número 6:



Para el número 7:



Para el número 8:

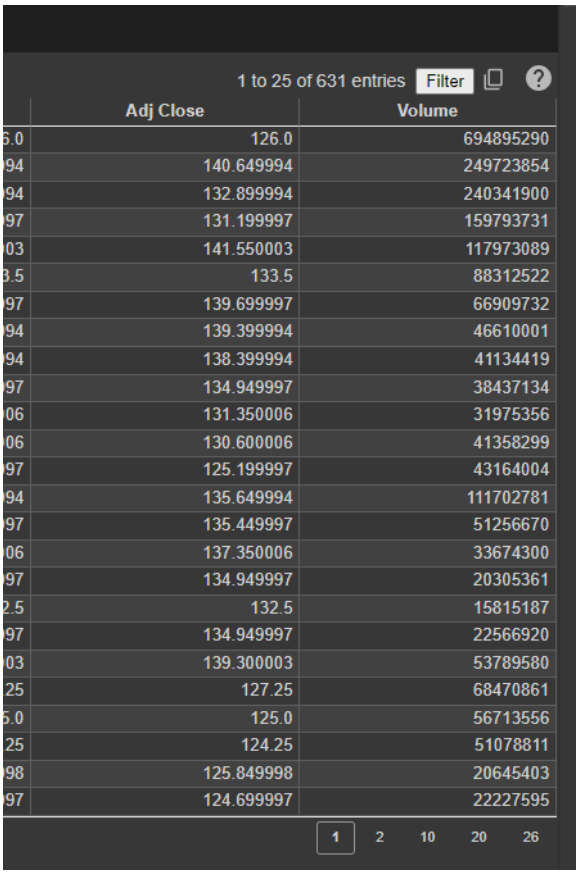


Posteriormente a la identificación, se procederán a eliminar, aislar o tratar los outliers o datos anómalos para una mejor simetría en la data.

- **Transformación de datos**

Para la transformación de datos utilizamos la columna volume, para poder trabajar mejor con esos datos lo que se hizo fue sacarle raíz cuadrada a dicha columna para simplificar los datos y poder manipularlos mejor.

Así lucían los datos antes de simplificarlos:



1 to 25 of 631 entries Filter

	Adj Close	Volume
6.0	126.0	694895290
94	140.649994	249723854
94	132.899994	240341900
97	131.199997	159793731
03	141.550003	117973089
8.5	133.5	88312522
97	139.699997	66909732
94	139.399994	46610001
94	138.399994	41134419
97	134.949997	38437134
06	131.350006	31975356
06	130.600006	41358299
97	125.199997	43164004
94	135.649994	111702781
97	135.449997	51256670
06	137.350006	33674300
97	134.949997	20305361
2.5	132.5	15815187
97	134.949997	22566920
03	139.300003	53789580
25	127.25	68470861
5.0	125.0	56713556
25	124.25	51078811
98	125.849998	20645403
97	124.699997	22227595

1 2 10 20 26

Este fue el resultado de simplificar (transformar) los datos obtenidos en la columna ‘volume’, se redondean los decimales para tener una mejor visión de estos.

1 to 25 of 631 entries Filter

Adj Close	Volume
126.0	26360.87
140.65	15802.65
132.9	15502.96
131.2	12640.95
141.55	10861.54
133.5	9397.47
139.7	8179.84
139.4	6827.15
138.4	6413.61
134.95	6199.77
131.35	5654.68
130.6	6431.04
125.2	6569.93
135.65	10568.95
135.45	7159.38
137.35	5802.96
134.95	4506.15
132.5	3976.83
134.95	4750.47
139.3	7334.14
127.25	8274.71
125.0	7530.84
124.25	7146.94
125.85	4543.72
124.7	4714.62

1 2 10 20 26

Para poder simplificar los datos sacando raiz cuadrada se utilizo la libreria panda y la siguiente línea de código:

```
[7]  import pandas as pd

     df = pd.read_csv('zomato.csv')

     # Transformar los datos (para calcular la raiz de la columna volume)
     df['Volume'] = df['Volume'].apply(lambda x: x**0.5)

     df.to_csv('zomato1.csv', index=False)
```

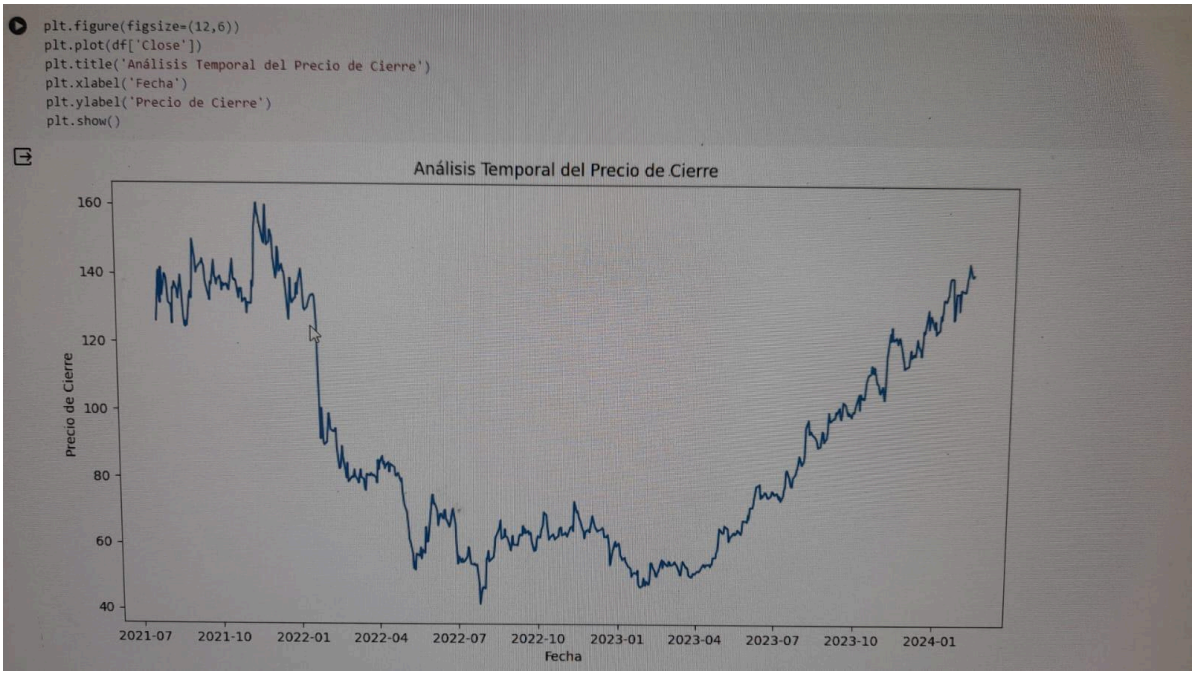
Todo esto era guardado en un formato csv, para poder mostrarlo como una tabla más adelante

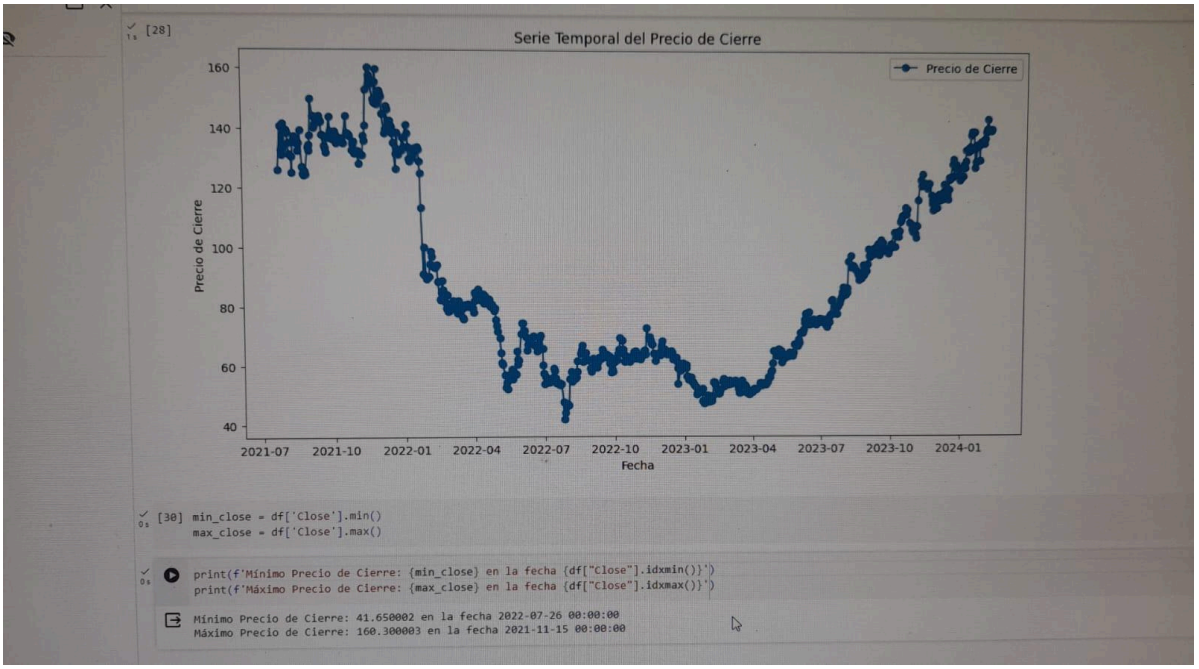
Para poder redondear a dos decimales los resultados obtenidos, se utilizo la libreria panda, además de la función df.round(2) para poder realizar el redondeo, como saldra a continuación:

```
[10] import pandas as pd

     df = pd.read_csv('zomato1.csv')

     # Redondear todos los valores numéricos a dos decimales
     df = df.round(2)


     df.to_csv('zomato1.csv', index=False)
```

- **Análisis temporal:**

```
plt.figure(figsize=(12,6))
plt.plot(df['Close'])
plt.title('Análisis Temporal del Precio de Cierre')
plt.xlabel('Fecha')
plt.ylabel('Precio de Cierre')
plt.show()
```





La gráfica muestra el análisis temporal del precio de cierre de un restaurante, utilizando datos del archivo "zomato.csv". El eje X representa la fecha, mientras que el eje Y representa el precio de cierre.

Elementos de la gráfica:

- **Línea azul:** Muestra la tendencia del precio de cierre a lo largo del tiempo.
- **Puntos grises:** Representan el precio de cierre para cada fecha individual.
- **Eje X:** Muestra las fechas en formato YYYY-MM.
- **Eje Y:** Muestra el precio de cierre en unidades monetarias..

Información que se puede obtener de la gráfica:

- **Tendencia general:** La línea azul indica si el precio de cierre ha tendido a aumentar, disminuir o mantenerse estable a lo largo del tiempo.
- **Variabilidad:** Los puntos grises muestran la variabilidad del precio de cierre a lo largo del tiempo.
- **Patrones:** Se pueden observar patrones estacionales o cíclicos en la gráfica.
- **Valores atípicos:** Los puntos grises que se encuentran alejados de la línea azul pueden indicar valores atípicos.

- **Conclusiones y hallazgos:**

En conclusión, después de realizar un exhaustivo Análisis Exploratorio de Datos (EDA) en la base de datos de Zomato, se puede concluir que hay una serie de insights valiosos que pueden ser extraídos para comprender mejor el comportamiento de los datos y las tendencias en la data. Desde datos como precios, precios mínimos y máximos, el análisis revela patrones significativos que pueden ser utilizados por los propietarios, los inversores y los responsables de marketing para tomar decisiones informadas y estratégicas. Este análisis proporciona una base sólida para la toma de decisiones empresariales, destacando áreas de oportunidad y potenciales desafíos que pueden abordarse para mejorar la experiencia del cliente y maximizar el éxito comercial en la plataforma de Zomato.

- **Código y metodología:**

Código fuente: CodigoFuente.ipynb

Metodología:

Se usó google colab para estructurar los bloques de código pertinentes para analizar la data.

Se usaron diferentes herramientas para graficar, las cuales son: Google colab y infogram.