Cervantes Rubí Brandon

Minería de Datos 2023-1, G.3, 17/09/2022
Practica No.3 Análisis Exploratorio de Datos (EDA)

Objetivo.

Hacer un análisis exploratorio de datos con base en información obtenida de Yahoo Finanzas. Por ejemplo, datos de Spotify, Facebook, Amazon y Aeroméxico.

Características.

El análisis exploratorio de datos es una etapa fundamental para los proyectos que se desarrollan en la actualidad relacionados con la IA actual y todos los subcampos que esta abarca. Se trata de dar un acercamiento a cómo es que están compuestos estos datos; el tipo, su distribución, las dimensiones de las matrices de valores.

Yahoo Finance ofrece una amplia variedad de datos de mercado sobre acciones, bonos, divisas y criptomonedas. También proporciona informes de noticias con varios puntos de vista sobre diferentes mercados de todo el mundo, todos accesibles a través de la biblioteca `yfinance`.

Fuente de datos

- High: Precio máximo de la acción.
- Low: Precio mínimo de la acción.
- Open: Precio con el que abre la acción de un determinado periodo.
- Low: Precio con el que cierra la acción en un determinado periodo.
- Volume: Cantidad de actividad comercial.
- Dividends: Cantidad de dividendos.
- Stock Splits: Cantidad de splits.

Desarrollo.

Durante toda la práctica se utilizaron las siguientes bibliotecas; `pandas` para el manejo de los datos, `numpy` para trabajar con vectores/matrices más cómodamente, `pyplot` de `matplotlib` para graficar los resultados del análisis y por último `seaborn` para manejar aún más estilos y graficas en la visualización de datos, `pandas` que nos proporciona estructuras para la manipulación y análisis de datos. Además de que por buenas prácticas se utilizó un alias para cada una de estas bibliotecas con la palabra reservada `as`.

En esta ocasión los datos con los que se trabajó se consumen a través de la librería `yfinance` la cual funciona como un API que nos comunica con el portal de Yahoo Finance que contiene información de las acciones de empresas que cotizan en las diferentes bolsas de valores al rededor del mundo.

La forma en la que podemos obtener los datos es con el método `Ticker` de `yfinance` el cual recibe como parámetro el identificador de la empresa que estamos buscando, con el podemos analizar sus precios a lo largo del tiempo con el método `history` definiendo una fecha de inicio, fin y el intervalo de tiempo el cual será de un día para todos los ejemplos de la práctica, nos devuelve el siguiente dataframe cuyos valores se detallaron anteriormente.

```
SpotifyHist = DataSpotify.history(start = '2019-1-1', end = '2022-9-8', interval = '1d')
```

Date	Open	High	Low	Close	Volume	Dividends	Stock Splits
2019-01-02	111.660004	115.629997	110.360001	113.739998	861100	0	0
2019-01-03	112.080002	113.345001	108.589996	109.019997	1082300	0	0
2019-01-04	112.059998	121.470001	111.500000	118.510002	2484800	0	0
2019-01-07	115.040001	123.865997	113.279999	119.360001	2516200	0	0
2019-01-08	121.440002	122.769997	114.699997	117.480003	1257100	0	0
...	...	...	...	...	...	...	...
2022-08-31	109.500000	110.080002	107.349998	108.150002	1050400	0	0
2022-09-01	106.059998	107.089996	102.180000	106.519997	1570600	0	0
2022-09-02	107.550003	108.290001	103.360001	104.419998	1405300	0	0
2022-09-06	104.059998	104.370003	100.620003	102.589996	962800	0	0
2022-09-07	102.349998	106.279999	102.349998	105.860001	652300	0	0

A partir de este punto se comenzó a trabajar con la caracterización de los datos, como primera función se reviso `head`, la cual nos regresa las n primeras líneas que especifiquemos, su contraparte seria `tail`, la cual nos regresa las n ultimas líneas del `dataframe`.

Al igual que con la primera práctica, se dividió el análisis exploratorio de datos en 4 puntos principales. El primero de ellos se centra en definir la estructura de los datos que regresa Yahoo Finance en cuanto a sus dimensiones y sus tipos de dato.

- `shape`: Este atributo nos regresa cual es la dimensión del dataframe que puede ser visto como una matriz, por ejemplo, en esta ocasión las dimensiones son de 928 x 7 lo que quiere decir que se tienen 928 renglones o registros de las acciones y por cada uno de estos registros se tienen 7 valores los cuales corresponden a las variables iniciales.

- `dtypes`: Este atributo nos regresa el tipo de dato de cada una de las variables o columnas en el dataframe, en este caso regresa 7 valores los cuales varían entre float64 e int64, dado a que la mayoría son indicadores financieros todo el dataframe está compuesto de números.

El segundo paso dentro del análisis de datos corresponde a identificar datos faltantes, dependerá de cómo venga el dataframe, sin embargo, es muy común que haya registros sin valor en alguna de sus variables, y estas al ser importante para el análisis pueden afectar el desempeño del modelo.

Una forma de identificar los datos faltantes es con el método `isnull` en conjunto con `sum`, agrupando primero todos aquellos valores que son nulos para después sumarlos y devolver una lista con cada una de las variables numéricas con la cantidad de valores nulos en ellas

```
SpotifyHist.isnull().sum()
```

```
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   Open          928 non-null    float64
 1   High          928 non-null    float64
 2   Low           928 non-null    float64
 3   Close         928 non-null    float64
 4   Volume        928 non-null    int64
 5   Dividends     928 non-null    int64
 6   Stock Splits  928 non-null    int64
dtypes: float64(4), int64(3)
memory usage: 58.0 KB
```

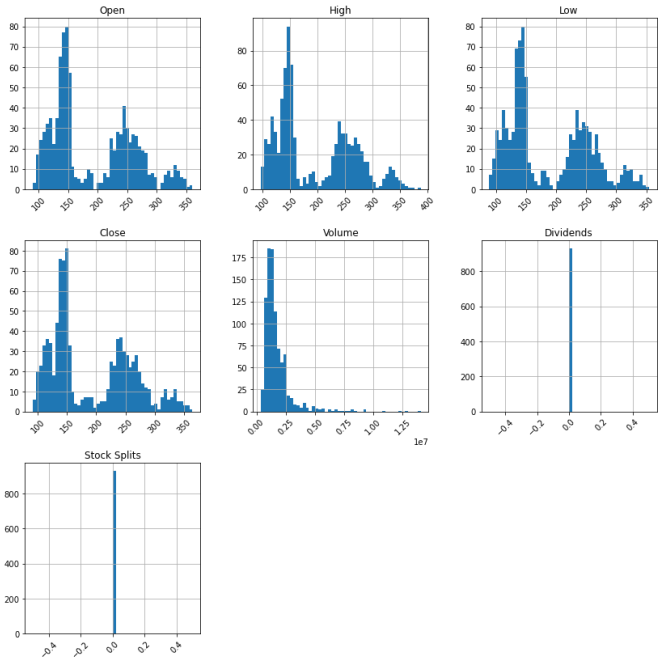
En ella podemos ver que en esta ocasión ninguno de los registros cuenta con valores nulos en su estructura, es decir que son datos completamente transparentes por parte de las empresas.

También se revisó el método `info`, el cual es similar a `dtype`, con la diferencia de que además de regresarnos el tipo de dato, nos regresa la cuenta de los valores que no son nulos.

El tercer paso se centra en la detección de valores atípicos, es decir, valores que no hacen mucho sentido cuando se comparan con la mayoría de los valores que corresponden al conjunto de datos, valores que se alejan mucho de la media.

Podemos ayudarnos de diferentes tipos de gráficas, como histogramas o diagramas de caja para detectar estos posibles errores, por ejemplo, un porcentaje mayor a 100, o valores que se alejen demasiado de los demás. Para cada variable se graficaron histogramas los cuales nos dan una idea de sus valores y su rango.

```
SpotifyHist.hist(figsize=(14,14), xrot=45, bins = 50)
plt.show()
```



Para los parámetros se utilizó `figsize` y `xrot` para modificar el tamaño de los histogramas y con `bins` podemos hacer aún más preciso el histograma que estemos creando pues `bins = 50` subdivide el histograma con 50 barras.

En primera instancia los valores atípicos también se pueden tratar de errores al momento de tomar la medición, sin embargo, hay que identificar cuales son para corroborar si en efecto se trata de un error o de un valor que nos pueda indicar que los datos están corruptos. En la mayoría de los casos simplemente se elimina para que no afecte mucho al modelo que se desarrolle, o en el mejor de los casos se puede corregir por el valor correcto.

En estos datos, el histograma de `Volume` presenta valores sesgados hacia la izquierda, sin embargo, es normal que haya días atípicos en los que la cantidad de transacciones suba por cuestiones de tendencias. Además, observamos que los `dividendos` y `splits` presentan la mayoría de sus datos en 0.

El método `describe` de pandas nos muestra un resumen de las estadísticas de cada una de las columnas indicando la cuenta de los registros, la media, la desviación estándar, el valor mínimo, los cuartiles que también indican la mediana al 50%, y el valor máximo.

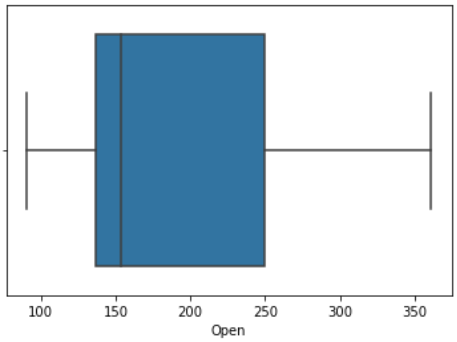
	Open	High	Low	Close	Volume	Dividends	Stock Splits
count	928.000000	928.000000	928.000000	928.000000	9.280000e+02	928.0	928.0
mean	191.646390	195.605927	187.637602	191.667758	1.738217e+06	0.0	0.0
std	68.571980	70.010612	66.978685	68.475777	1.317898e+06	0.0	0.0
min	90.440002	97.070000	89.029999	91.940002	3.945000e+05	0.0	0.0
25%	136.515003	139.740002	134.285000	137.150002	1.050350e+06	0.0	0.0
50%	153.595001	156.855003	151.012497	154.339996	1.383300e+06	0.0	0.0
75%	249.719994	254.449997	245.337498	249.709995	1.988100e+06	0.0	0.0
max	360.910004	387.440002	354.178009	364.589996	1.404930e+07	0.0	0.0

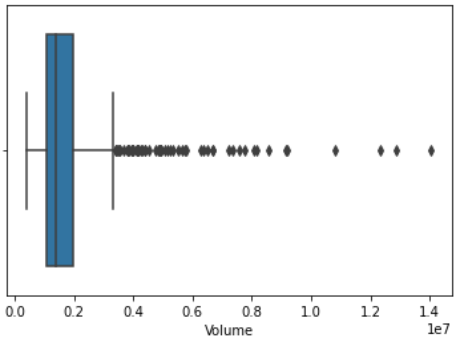
En este resumen podemos apreciar lo inestable que pueden ser los valores de las acciones pues en la mayoría de las ocasiones los valores de la media y la mediana se ven prácticamente duplicados por los valores máximos que se presentan, a su vez, cuando los comparamos con los valores mínimos podemos observar que la media ahora es la que duplica a los índices mínimos.

Otra forma muy común de detectar los valores atípicos son los gráficos de caja o boxplots, este tipo de grafica se basa en los cuartiles pues la caja abarca los cuartiles 1, 2 y 3 (25%, 50% y 75%). El siguiente código muestra este tipo de grafica para cada una de las variables sobre las cuales se sospecha la existencia de valores atípicos.

```
VariablesValoresAtipicos = ['Open', 'High', 'Low', 'Close', 'Volume']
for col in VariablesValoresAtipicos:
  sns.boxplot(col, data=SpotifyHist)
  plt.show()
```

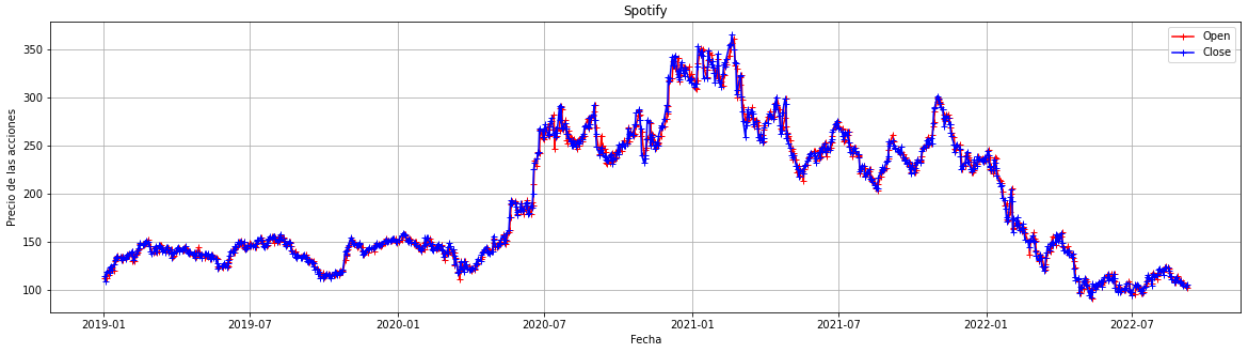
A continuación, se muestran las gráficas de caja.





A modo de ejemplo se consideraron las variables `open` y `volume`, en estos podemos ver como para el primer caso si bien hay valores que se alejan de la media no es tan notorio el cambio por lo cual se puede considerar normal la subida. En cambio, para la variable de `volume` volvemos a observar como hay una gran cantidad de valores que se alejan de la media y por mucho lo cual lo podemos atribuir a que existen días en los que la actividad financiera es mucho mayor que en la mayoría.

Podemos además hacer un grafico que nos relacione el comportamiento de las variables `Open` y `Close` considerando su valor a través del tiempo el cual se ve de la siguiente manera.



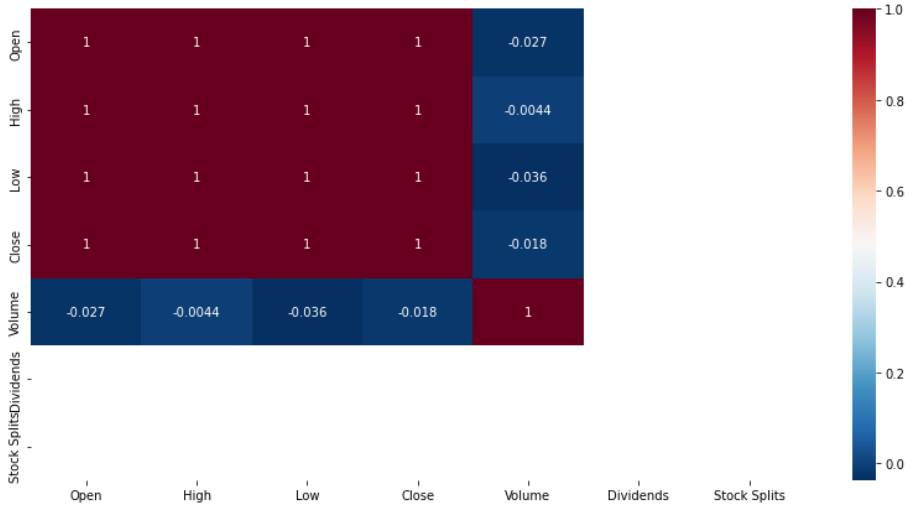
En el podemos ver en rojo los valores de apertura y en azul los valores de cierre, estas graficas prácticamente se sobreponen pues al tener un periodo de cambio tan pequeño es normal que las acciones cierren muy cerca del valor con el que abrieron.

El último paso dentro del análisis de correlación de variables tiene que ver con la identificación de correlaciones para medir el grado de similitud entre pares de variables, es decir que tan parecido se comportan. Estas medidas se pueden obtener en una matriz de correlaciones la cual tiene índices que van de -1 a 1, donde los más cercanos a 1 o -1 indican un valor de correlación fuerte y los mas cercanos a 0 indican un nivel de correlación débil.

En primera instancia podemos obtener esta matriz de correlaciones con el método D, el cual nos regresa la siguiente matriz.

	Open	High	Low	Close	Volume	Dividends	Stock Splits
Open	1.000000	0.998372	0.998514	0.996621	-0.026878	NaN	NaN
High	0.998372	1.000000	0.998185	0.998535	-0.004356	NaN	NaN
Low	0.998514	0.998185	1.000000	0.998477	-0.036423	NaN	NaN
Close	0.996621	0.998535	0.998477	1.000000	-0.017530	NaN	NaN
Volume	-0.026878	-0.004356	-0.036423	-0.017530	1.000000	NaN	NaN
Dividends	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Stock Splits	NaN	NaN	NaN	NaN	NaN	NaN	NaN

En ella podemos observar a simple vista que la mayoría de estas variables tienen un índice de correlación muy alto, casi de 1. Podemos apreciar esto de mejor manera si además graficamos el mapa de calor para esta matriz el cual pinta cada uno de los espacios dentro de la matriz de colores muy rojos y azules cuando el índice se acerca a sus valores máximos y mínimos pudiendo de esta manera distinguir mejor cuales son las correlaciones fuertes.



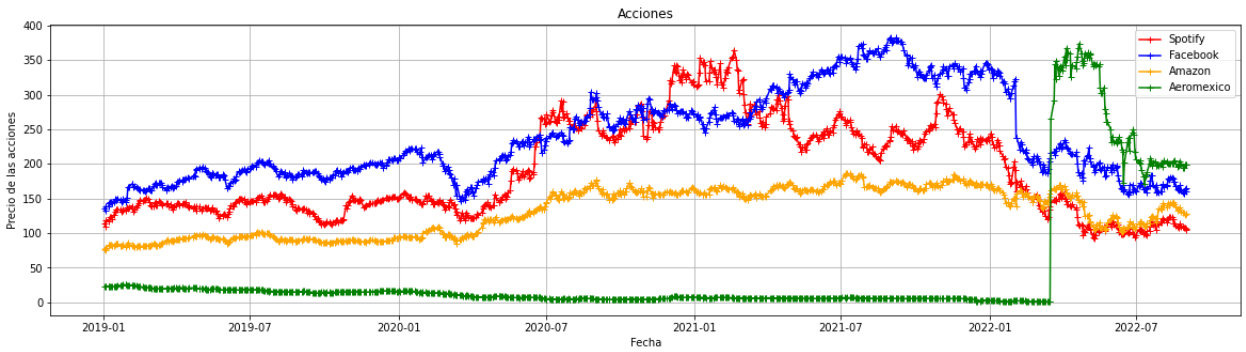
Como podemos ver la mayoría de las variables se encuentran fuertemente correlacionadas a excepción de volumen, pues al tener un intervalo de tiempo muy corto estos índices prácticamente no varían con respecto a sus valores, por ejemplo, el precio de apertura es prácticamente el mismo que el de cierre variando por muy pocas unidades. En cambio, volumen si llega a cambiar bastante entre los días pues hay días con mucha mas actividad que otros, esto dependerá de las tendencias del momento.

El siguiente paso dentro del análisis fue observar el comportamiento de algunas empresas importantes durante la pandemia. Tal es el caso de Facebook, Amazon, Spotify y Aeroméxico de igual manera se obtuvieron sus datos a través de yahoo finance desde el 2019 a la actualidad con intervalos de un día.

Puesto que el valor de cierre suele ser más importante para los análisis financieros se decidieron eliminar todas las demás columnas para observar solamente el comportamiento de cierre. Esto se hizo para todas las empresas, después se concatenaron haciendo uso de un `join` con `inner` es decir que solamente se obtuvo la intersección de estos valores para eliminar fechas en las que por ejemplo Amazon si tenia registro de sus acciones, pero Facebook no y de esta manera no tener valores nulos desde la concatenación de las empresas.

```
# Se integran los cierres del precio de las acciones
Acciones = pd.concat([SpotifyClose, FacebookClose, AmazonClose, AeromexClose],
                     axis = 'columns', join = 'inner')
Acciones
```

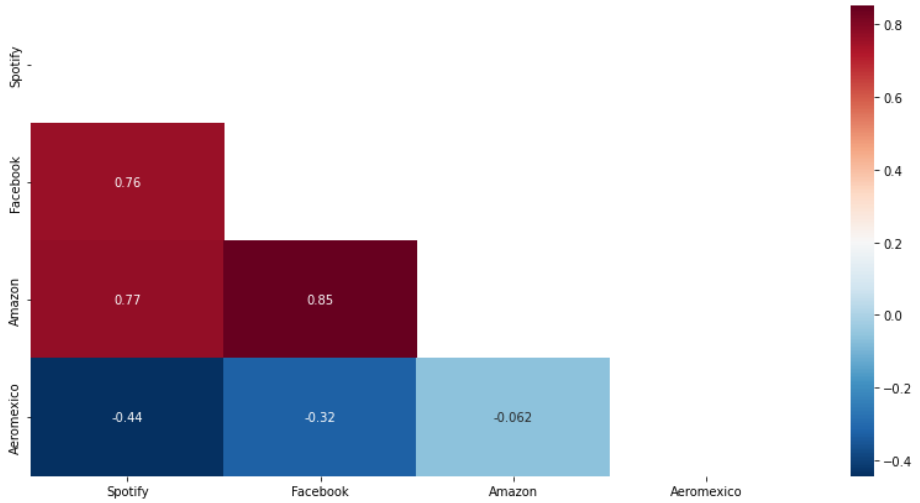
Posterior a esto se grafico el comportamiento de sus precios de cierre a lo largo de los 3 años del análisis, en ella podemos ver a Spotify en rojo, Facebook en azul, Amazon en naranja y Aeroméxico en verde.



En esta grafica podemos ver que las 3 empresas centradas en tecnología no solo mantuvieron sus precios, sino que inclusive fueron subiendo estos índices a lo largo de la pandemia observando como la que mas genero fue Facebook, cosa que probablemente tenga que ver con el anuncio de su metaverso para futuros años. Además, es importante destacar el comportamiento de Aeroméxico la cual durante muchísimo tiempo estuvo inactiva con valores muy bajos hasta que se regularizan los vuelos por cuestión de la pandemia maso menos a principios del año 2022 cuando tiene una subida en sus acciones bastante considerable.

Además, podemos obtener el mapa de calor de la matriz de correlación para identificar como es que varia el comportamiento de estas empresas considerando los 3 años de análisis.

A continuación, se muestra el mapa de calor.

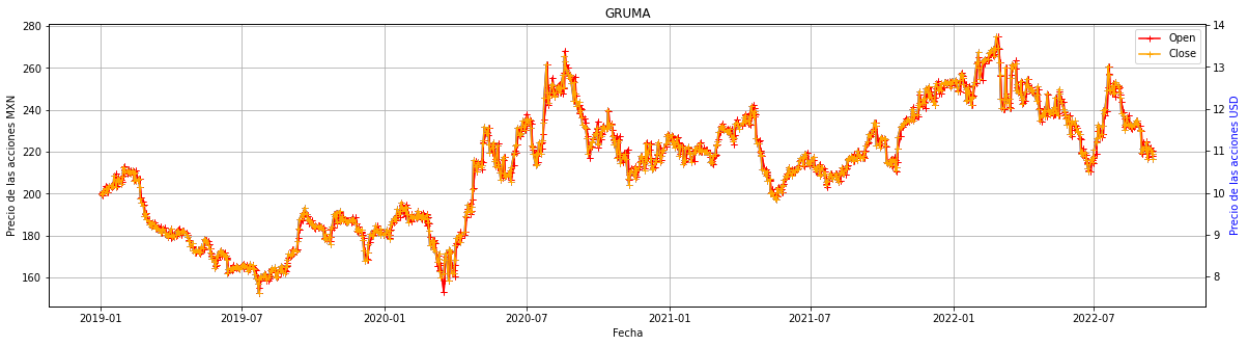


Como se veía en la grafica anterior las empresas que tienen un mayor nivel de correlación son las 3 relacionadas directamente con la tecnología Amazon, Facebook y Spotify teniendo una correlación fuerte, entre ellas las que mayor correlación tienen son Amazon y Facebook. A diferencia de Aeroméxico la cual tiene correlaciones moderadas o muy débiles con respecto a las otras empresas.

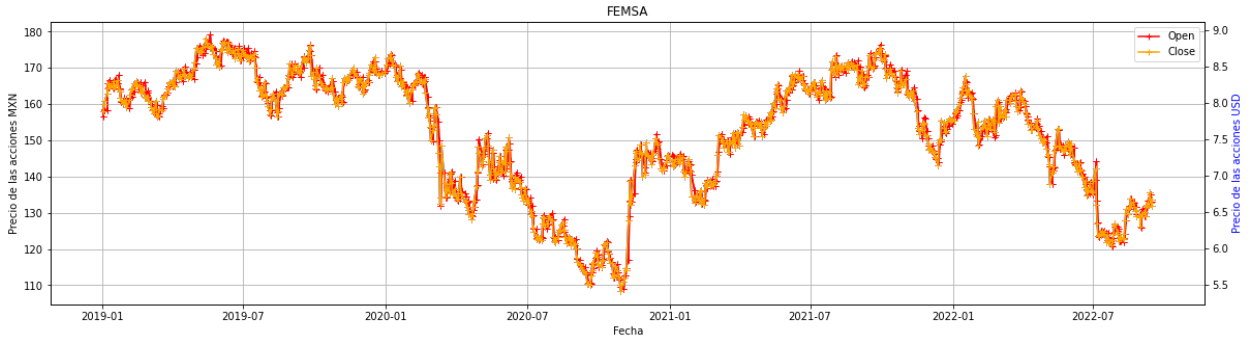
Como ultimo paso dentro de la práctica se analizó el comportamiento de 4 empresas mexicanas; Aeroméxico, Grupo Aeroportuario del Pacifico, GRUMA y FEMSA. Las últimas dos son empresas dedicadas a la industria de los alimentos.

A continuación, se muestran las graficas del comportamiento de cierre de estas empresas en los 3 años de análisis destacando en el eje izquierdo el precio de las acciones en MXN y del lado derecho su precio en USD, además de que se muestra en naranja su valor de cierre y en rojo su valor de apertura.

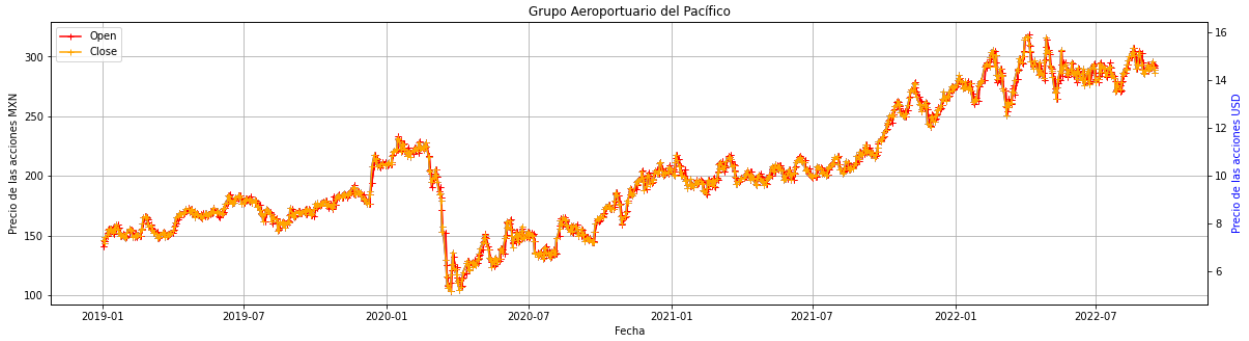
GRUMA



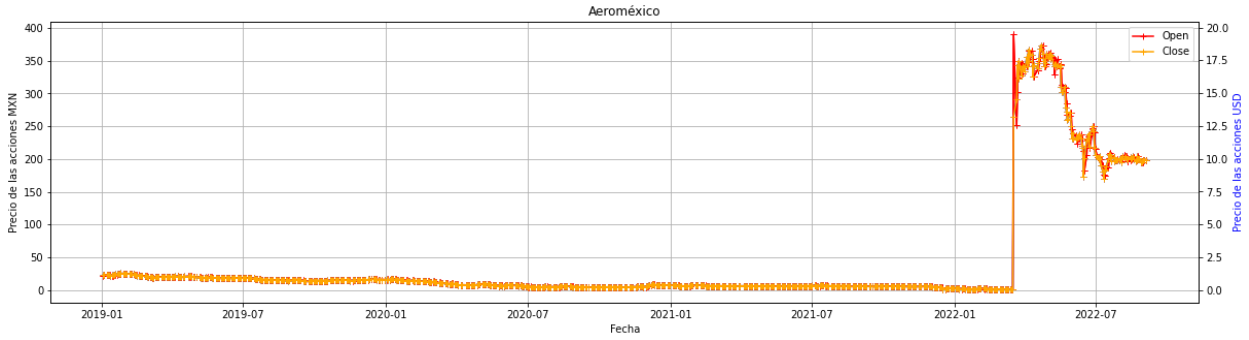
FEMSA



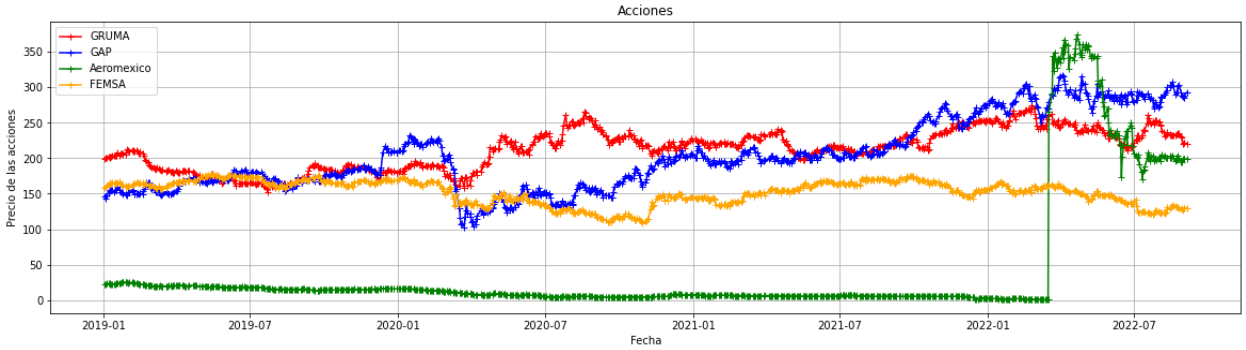
Grupo Aeroportuario del Pacifico



Aeroméxico

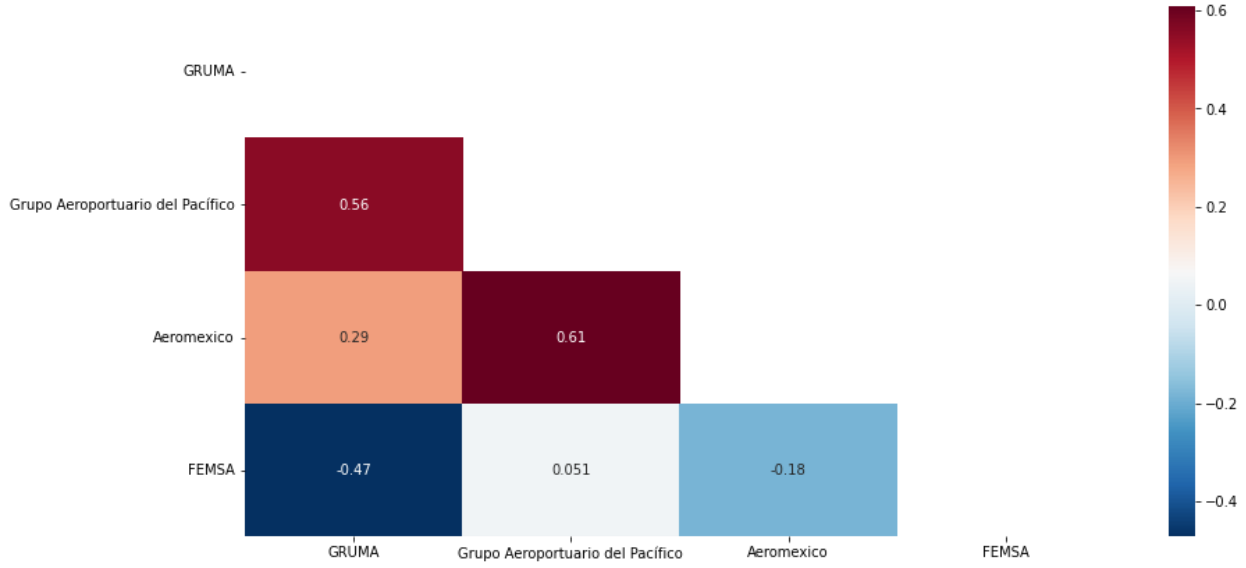


A continuación, se muestra la grafica donde podemos ver la comparativa de como es que se comportaron las acciones de las empresas de la pagina anterior durante la pandemia.



En ella podemos destacar que, si bien el grupo aeroportuario se maneja en la misma línea que Aeroméxico, supo llevar mejor la pandemia, esto debido a que al contar con mayores lugares de operación tiene un negocio más diversificado el cual es muy poco probable que se venga abajo como paro con Aeroméxico, si bien tuvo sus dificultades en 2019 se supo mantener y se repuso. Además, podemos ver que con respecto a las empresas de alimentos la pandemia prácticamente no tuvo influencia en el precio de sus acciones pues la comida es un negocio que siempre va a ser rentable.

Podemos además graficar el mapa de calor para conocer que tan parecido es el comportamiento en las acciones de estas empresas mexicanas.



En el se observa que la mayor correlación es moderada dándose entre Aeroméxico y el Grupo Aeroportuario del Pacifico pues ambos negocios se llevan sobre la misma línea, además existe otra entre el mismo Grupo Aeroportuario del Pacifico y GRUMA por como se manejaron en pandemia y por último entre FEMSA y GRUMA pues a ambos les afecta lo que pase en la industria de los alimentos.

Conclusiones

Durante esta práctica se logró aplicar el análisis exploratorio de datos a otro ejemplo real el cual podría tener impacto si se desarrolla a profundidad pues se trata del comportamiento de la bolsa en meses anteriores lo cual se puede extrapolar a un comportamiento a futuro con datos obtenidos de Yahoo Finance.

Sin embargo, el análisis exploratorio de datos es el primer paso en este tipo de proyectos, en el solamente se pretende conocer cual es la estructura de los datos, su tipo, sus posibles errores o valores atípicos, sus correlaciones, etc. para de esta manera saber si es que nos van a servir para el tipo de problema al que nos vayamos a enfrentar o si por otro lado tendremos que obtener mas datos o darle formato a los que ya tenemos.