UNIVERSIDAD SAN FRANCISCO

FUNDAMENTOS DE CIENCIA DE DATOS

PROYECTO #1
CRISP-DM, ELT, EDA, DATA WRANGLING

GABRIELA ZUMARRAGA – 345769

QUITO 28 DE MARZO 2025

PROYECTO #1
CRISP-DM, ELT, EDA, DATA WRANGLING

Tema: Análisis del desempeño de las empresas en el S&P (Standard and Poor's) 500 utilizando datos históricos del mercado bursátil

1. BUSINESS UNDERSTANDING

El índice S&P 500, (Standard & Poor's 500), es un índice que indica el rendimiento de las 500 mayores empresas que cotizan en Estados Unidos (por capitalización de mercado). Es un indicador clave de la salud económica de la bolsa estadounidense ya que abarca distintos tipos de industrias, como tecnología, atención médica, servicios financieros, bienes de consumo y energía, seleccionadas por su tamaño, liquidez y representatividad económica.

El objetivo principal de este proyecto es analizar las tendencias y patrones históricos de precios y volúmenes de las acciones de las empresas del S&P 500 para proporcionar insights accionables.

a. Contexto

Hipótesis principal: Los movimientos históricos de precios y volúmenes pueden revelar tendencias repetitivas o correlaciones significativas que permitan prever cambios futuros en el mercado.

Qué necesitas comprobar o resolver: Identificar patrones relevantes en los precios y volúmenes que puedan ayudar a entender mejor el comportamiento del mercado.

Pregunta de investigación: ¿Qué patrones de precios y volúmenes se observan en las acciones de las empresas del S&P 500, y cómo pueden usarse para anticipar movimientos futuros?

b. MVP

Se realizarán el análisis de los datos y se crearán dashboards, gráficos o reportes que:

- Identifiquen las tendencias más destacadas en los precios y volúmenes históricos para sugerir posibles tendencias futuras utilizando modelos de regresión.
- Proporcionen visualizaciones claras que permitan a los inversionistas tomar decisiones informadas.
- Permitan entender el comportamiento histórico del mercado para quienes buscan optimizar sus estrategias de inversión.

2. DATA UNDERSTANDING

a. Origen

Los datos que se utilizaran provienen de Kaggle y contiene precios históricos de acciones para todas las empresas que actualmente se encuentran en el índice S&P 500.

- Enlace**:** S&P 500 Stocks (daily updated)

b. Estructura de datos

El dataset contienen 3 archivos con la siguiente estructura inicial. El detalle del análisis se realizó en el notebook 01_EDA.ipynb.

- sp500_stocks

Campo	Tipo de dato	Descripción
Date	float	Fecha del registro en formato yy-mm-dd
Open	float	Precio de la acción al inicio del mercado (esta es información del NYSE, por lo tanto, todo en USD)
High	float	Precio más alto alcanzado del periodo
Low	float	Precio más bajo alcanzado del periodo
Close	float	Precio de cierre del mercado.
Volume	int	Cantidad de acciones negociadas.

Adj Close	float	Similar al precio de cierre del mercado, pero incluye acciones de la empresa como dividendos y divisiones de acciones.
Symbol	float	Símbolo de la empresa

Los campos que contiene este archivo son importantes para el análisis ya que permitirá ver patrones históricos de las empresas y explorar relaciones significativas que permitan predecir el comportamiento de los precios y actividad en el mercado.

- sp500_companies

Campo	**Tipo de dato**	**Descripción**
Exchange	Object(string)	Intercambio donde se negocian sus acciones.
Symbol	Object(string)	Símbolo de la acción.
Shortname	Object(string)	Nombre corto de la empresa.
Longname	Object(string)	Nombre completo de la empresa.
Sector	Object(string)	Sector en el que opera la empresa.
Industry	Object(string)	Industria, dentro de un sector, en la que opera la empresa.
Currentprice	Float	Precio actual de la acción.
Marketcap	int	Capitalización de mercado actual.
Ebitda	Float	Ganancias antes de intereses, impuestos, depreciación y amortización.
Revenuegrowth	Float	Crecimiento de ingresos.
City	Object(string)	Ciudad matriz de la empresa.
State	Object(string)	Estado matriz de la empresa.
Country	Object(string)	País matriz de la empresa.
Fulltimeemployees	int	Número de empleados a tiempo completo.
Longbusinesssummary	Object(string)	Descripción general de la compañía.
Weight	float	Porcentaje de participación en el índice S&P 500 (según su capitalización de mercado).

Estos datos nos permitirán relacionar los patrones y predicciones encontradas, no solo por empresa sino también por sectores (industriales, geográficos). Igualmente permitirán evaluar el rendimiento para realizar predicciones (current price), analizar diferencias de comportamiento de las acciones (Exchange), ver índices de impacto (marketcap) e influencia de las empresas (weight).

- sp500_index

Campo	Tipo de dato	Descripción
Date	Datetime	Fecha del registro.
S&P 500	float	Valor del índice S&P 500.

Estas variables permiten analizar tendencias del índice S&P 500 relacionadas con el mercado bursátil.

c. Posibles desafíos:

- Datos faltantes: se debe analizar que tipo de estrategia se debe abordar en caso de que haya muchos datos faltantes ya que pueden comprometer el análisis. Si el porcentaje es pequeño se puede considerar la imputación de valores según corresponda.
- Datos duplicados: se debe analizar si existen duplicados, ya que la redundancia podría distorsionar los resultados.
- Inconsistencias: el tipo de datos y formatos tanto en campos de tipo numérico como strings se deben normalizar para evitar dificultades a la hora de procesar los datos.
- Precisión de los datos: el dataset fue sacado de Kaggle y según la fuente fue actualizado hace 3 meses. Si bien los datos no están al día, podríamos encontrarnos con posibles inconsistencias si la data no fue bien actualizada como se menciona. Esto podría influir en los modelos predictivos.

Estos posibles desafíos se analizarán dentro del EDA y se procesarán en el Data Wrangling según corresponda.

3. DATA PREPARATION

Para la limpieza y transformación de los datos se tomará en cuenta el análisis realizado en el EDA. Para los 3 archivos se realizará los siguiente:

- Análisis de los valores duplicados: de existir valores duplicados, se eliminarán las filas.
- Análisis de los valores nulos: dependiendo el porcentaje de valores nulos se eliminarán o se imputarán valores considerando medianas, medias o valores cercanos para no afectar el análisis.
- Análisis de strings: en columnas descriptivas se validará que no existan posibles duplicados (en supuestos valores únicos) por ejemplo en sectores, industrias, países, empresas.
- Análisis de los tipos de datos: se verificarán los tipos de datos con los que se trabajarán en las distintas columnas.
- Se considerarán posibles variables derivadas para análisis futuros.

Cabe recalcar que el detalle se encuentra en el notebook DataWrangling.