Goals:

Plan an Exploratory Data Analysis (EDA): Use a public financial dataset from a source like Kaggle to perform an EDA. Analyze things like GDP data, company financial reports, or historical stock data to create visualizations and uncover initial insights - 30 Points

Draft and Start an outline with code for this project - 70 Points

Notes:

Purpose: Produce a focused EDA using a public financial dataset (e.g., World Bank GDP, Kaggle company financials, or historical stock prices) to surface descriptive statistics, trends, and initial insights.

Objective: Understand structure, quality, and key patterns in the chosen dataset; identify trends, seasonality, outliers, and relationships.

Selected Focus Areas: time-series trend analysis, distributional analysis, correlation and drivers, volatility/returns (for stocks), and sector/category breakdowns (for company financials or GDP components).

Example Datasets & Sources:

- World Bank / IMF GDP data: country-level GDP (real & nominal) — World Bank API or Kaggle mirror.
- Historical stock prices: S&P 500 constituents or single-company histories — source: Yahoo Finance (via yfinance) or Kaggle “S&P 500 Historical Data”.
- Company financial statements: Kaggle datasets like “Financial Distress” or “Fundamentals & Financial Statements” (income statements, balance sheets, cash flows).
- Choose one dataset for this EDA. Example choice: S&P 500 historical prices + company sector mapping (for cross-sectional and time-series analyses).

Data Acquisition & Preparation:

- Acquire: use Kaggle CLI (kaggle datasets download -d <dataset>) or yfinance for stocks (example: pip install yfinance then yfinance.download(tickers, start, end)).
- Ingest: read CSV/Parquet into pandas DataFrame (pd.read_csv / pd.read_parquet).
- Clean: standardize date formats (pd.to_datetime), ensure numeric types, strip currency symbols, handle duplicates.
- Missing data: quantify (df.isna().sum()), visualize gaps (heatmap), impute or drop depending on context.
- Resample/align: for time-series, set Date index and resample to daily/weekly/monthly as needed.

EDA Methodology

1. Structure & Quality Checks: data types, row/column counts, unique keys, missingness, and duplicates.
2. Univariate Analysis: summary stats, histograms, KDEs, boxplots to assess distribution and outliers.
3. Time-Series Patterns: line charts of levels and growth rates, rolling means, decomposition (trend/seasonality/residual) using statsmodels.
4. Returns & Volatility (stocks): compute daily returns, cumulative returns, volatility (rolling std), drawdowns.
5. Cross-sectional Analysis: group-by sector/category for totals, medians, growth rates; pivot tables.
6. Correlation & Drivers: correlation matrix, heatmap, scatterplots; for macro data, examine GDP vs. unemployment/ inflation if available.
7. Anomaly Detection: flag extreme values (z-score, IQR) and investigate data-source errors vs. true anomalies.

Ideas for Visualization - probably pick 3-4

- Time-series line plot: levels and growth rates (with confidence intervals).
- Histogram & KDE: distribution of returns or ratios.
- Boxplot: distribution by sector or year.
- Heatmap: correlation matrix.
- Bar charts / Treemaps: top contributors by category (e.g., top countries by GDP or top sectors by market cap).
- Rolling volatility chart: volatility vs. time.
- Event overlays: annotate major events (market crash, policy change) on time series.
  Statistical Tests & Quick Models
- Stationarity tests: ADF or KPSS for time-series stationarity.
- Auto-correlation: ACF/PACF plots to inspect autocorrelation structure.
- Simple regression checks: e.g., GDP growth ~ unemployment rate (if available) or returns ~ market index for beta estimation.
- Clustering/segmentation (optional): cluster companies by financial ratios (P/E, ROE) for exploratory grouping.