Data Quality Report

Andres Camilo Viloria Garcia
aviloriagarcia00@mylangara.ca
Data Analytics
Langara College
Vancouver, BC

Diana Ortiz
dortizmontes00@mylangara.ca
Data Analytics
Langara College
Vancouver, BC

Summary:

This report summarizes the key insights from the Exploratory Data Analysis (EDA) and focuses on data quality assessment. It aims to acknowledge limitations and identify areas for improvement. The dataset used comprises 89 columns and over 53,000 rows. During the analysis, missing values were found in demographic variables, including P/E ratio ttm, P/B ratio mrq, D/E ratio mrq, free cash flow ttm, PEG Ratio 5 years, and ROE ttm. Additionally, non-historical data was identified in the dataset, and negative values were observed. These findings highlight the need for further investigation, imputation techniques for missing values, removal of non-historical data, and appropriate handling of negative values to enhance the dataset's quality and reliability for future analyses.

In the Results section, an in-depth analysis will be presented to elaborate on these findings. As part of this report, Python files were provided, which contains detailed information and further insights. The Jupyter notebook will provide details of the analysis and facilitate future improvements in the research.

Results:

Exploratory data analysis:

The code contained in this folder includes general analysis to help look at data before making any assumptions. It can help identify obvious errors, as well as better understand patterns within the data, detect outliers or anomalous events.
The data analyzed in this report corresponds to the dataset provided in Phase 4 of the project.

- **3.3 Understanding Final (Exploratory Data Analysis)**
  - **Demographics:**
    - Free cash flow: During the analysis, it was discovered that some stocks had extremely high values. For example, Ford (F) was found to have a value of 706. However, upon cross-referencing the data with values from Yahoo Finance, the maximum values found for Ford stock were 9.5 billion in 2022 and 5.5 billion in 2023. This suggests a discrepancy between the dataset's values and the actual values reported by Yahoo Finance,

indicating a potential data quality issue that needs to be addressed and validated for accurate analysis. reference 1 , reference 2.

- **Missing values:**
  It has been found that the only columns with empty values were located in the demographics variables.

  - The column " P/E Ratio ttm " has 13.31% elements nulls.
  - The column " P/B Ratio mrq " has 6.67% elements nulls
  - The column " D/E Ratio mrq " has 16.67% elements nulls
  - The column " Free Cash Flow ttm " has 10.00% elements nulls
  - The column " PEG Ratio 5 year expected " has 6.67% elements nulls
  - The column " ROE ttm " has 6.67% elements nulls

Missing values Key Findings: The dataset contains null values that can be categorized into two scenarios. The first scenario involves null values resulting from data unavailability during the scrapping process in April 2023. These null values can be considered acceptable and can be replaced with zeros since they represent periods where data was not accessible. However, in the second scenario, there are null values that correspond to variables with known representative numbers, such as the stock MCD having a value of 6.68 billion in April 2023 (reference1). This indicates an inconsistency or issue in accurately capturing and recording the data. To ensure data integrity and meaningful analysis, it is important to appropriately address these null values. Replacing null values with zeros can be suitable for the first scenario, but for the second scenario, needs to be replaced with actual values that can be accessible from other sources such as: ycharts.com and/or zacks.com. As this report provides a concise overview of the findings regarding data quality. For a more detailed analysis, you can access the comprehensive report in two formats: HTML or Jupyter Notebook.