Information Visualization

Final Report

Group 7: Julie Chea, Tianyu Gao

Table of Contents

Context of consultancy & Problem Statement **2**

Initial Data Quality **2**

Data preparation process **2**

Data Discovery 2
Data compilation and joining 3
Grouping, formatting 3
Final Data 3
Descriptive Summary Statistics (not yet) 4

Reflections and speculations on data preparation tools (not yet) **4**

Data quality assurance process (NA) 4
Harnessing the power of software combinations (NA) 4

Wonders for collaborative workflows (NA) 4

Context of consultancy & Problem Statement

The aim of this consultancy project is to provide an understanding three of the most popular market index funds: the S&P500, Dow Jones, and FT100. Harnessing the power of Tableau and R, this project will include forecasting of future stock prices, as well as an observation of the relationship between the prices and returns of these indices with commonly associated economic variables such as unemployment rate, CPI and interest rates. We will also explore the effect of seasonality on these returns. Overall, this project will compare and contract these indices as well as explore their relationship with other factors. It might also help establish for beginner investors which index is best to unilaterally invest in given specific consideration and constraints. This consultancy is not aimed mostly at professional investors although it may certainly be helpful in giving them an overview of the market and their interconnections with economic factors, but at people trying to better understand these indices as they enter the finance field.

Initial Data Quality

The dataset that we are using is compiled from stock price and trade volume data from yahoo finance. The three indexes put in focus through this project are three amongst the most popular index funds. We have added data downloaded from FRED which is a widely trusted and credible source on macroeconomic data about unemployment and CPI in the United States. Lastly, we have used data from the OECD about short- and long-term interest rates, which is also a very credible source. Overall, the initial data was credible and relatively complete although there were some cleaning processes which we had to go through and difficulty in compiling the data.

In terms of overall initial data quality, each dataset that were used are individually valid and credible, although there is work, we needed to do to make sure that the datasets were consistent with each other. Overall, the initial datasets can be said to have integrity as well as timelessness. Although there are some missing values in each dataset which were not easy to notice, the datasets were relatively complete with little null values.

We can therefore consider that the **initial data quality is relatively high.**

Data preparation process

Data Discovery

The initial data discovery process started with the problem statement and the aim of this consultancy. Since we aimed to analyze market indices and their relationships with other commonly associated variables, we went to the most trusted sources for these datasets and made sure that they were monthly so they could be joined and compiled together through a mixture of excel and tableau. Monthly data was decided on instead of weekly data which is

normally used in stock price analysis because most economic data was available in a monthly format. We downloaded the economic data using the Samir Khan yahoo downloader excel.

Data compilation and joining

Compilation was done through excel and Tableau. The initial market indices data was compiled together with unemployment and CPI data from FRED using excel. Data was downloaded for the last ten years, and copy pasted on one same sheet on Excel. The data for short and long-term interest rates were joined with the dataset through Tableau. In order to observe seasonality, we have also created an excel sheet where we created a column to assign months and joined it through Tableau PREP. Finally, we get our final dataset, as shown as "Output 3" in exhibit 1 below.

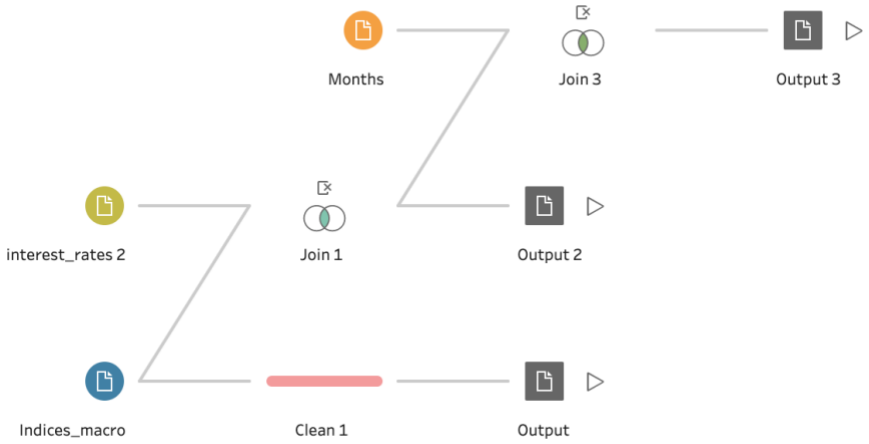

Exhibit 1: Tableau capture of joining process

There were some issues in the initial process as a month was entirely missing from the dataset without mentioning that it was a null value. We needed to check for this small discrepancy in order to make sure that there would not be big errors down the line. Meticulous rigor was put into making sure that these small details would not hinder data analysis down the line.

Grouping, formatting

In terms of formatting, it was initially impossible to join the datasets together as the dates in the time series were not formatted at dates but rather as strings. We needed to go through the process of formatting these dates to be uniform across multiple datasets so they could be joined. We also used the categorization of months in order to make interesting observations across month groups. Some datasets were in numbers instead of percentages and we made sure that the formatting for percentages was consistent as this could lead to big mistakes down the line.

Final Data

The final data was of high quality. It is complete, comes from credible sources, timeless and consistent. It has multiple columns that are of value and can all be used to make interesting observations through the visualization process.

Descriptive Summary Statistics (not yet)

Reflections and speculations on data preparation tools (not yet)

Data quality assurance process (NA)

Tableau and other data preparation tools can be extremely useful in modern organizational context. In a single glance using Tableau, we can see exactly how much of the data is composed of null values, therefore helping organizations know early the extent to which their data is complete. It also helps in terms of formatting, in terms of knowing the type of the data: numbers, strings, dates or other data type. Down the line, without using tableau or data preparation tools, if no conscious effort is made to check the types, codes might not work for intended purposes. For example, calculations cannot be made for strings. These are not obvious on csv files if we skip directly to data analysis.

Tableau is also useful to make sure that the data is consistent. For example, by joining datasets instead of purely copy pasting on excel, we could avoid making mistakes due to a null value that wasn't mentioned, making sure the data is correct and datasets are consistent with one another and could be joined.

Harnessing the power of software combinations (NA)

We can use a combination of Tableau prep and MySQL in order in order to elevate the data not only to check for data quality, but also to make sure that the data is fit to answer the questions posed. Tableau prep can be used to make sure that all the data falls into the proper needed data types. It can also be used to join data in a way that might be more visually intuitive than SQL. Tableau prep allows the user to just join two tables without mention of specific columns and allow the user to manually delete unnecessary or redundant columns.
MySQL can then be used for categorization. It can look at strings and other types and categorize.

Wonders for collaborative workflows (NA)