**Project Documentation – zhudanie CSC-324 Individual Project**

**Purpose**

The primary purpose of this project is to visualize regression data and fit a linear model to the data that predicts various stock performance indicators using changes 5 days in the past. The secondary purpose of this project is to visualize the financial data for over one thousand publicly traded companies. The aim of this project is to provide users with information available at a quick glance on metrics they care about. Visualizations are chosen to simplify data and present in a few seconds what may have taken hours of individual research.

**Data Description**

This project uses three separate datasets in its final implementation for different visualizations. The first dataset, compositeAugust.csv, contains daily stock performance and financial data on Apple (NASDAQ:AAPL) through the month of August 2021. Each observation is a day and contains variables like stock price, financial ratios, ESG performance scores, simple moving averages, and other data points. In addition to daily metrics, each observation contains changes in daily metrics 5 days prior to create lagged variables utilized in the regression modeling. A linear regression model and all regression analysis utilizes this dataset.

The second dataset, composite\_total2.csv, again contains daily stock performance and financial data but includes over a thousand publicly traded companies. Much like compositeAugust.csv, each observation is a day and each observation contains daily metrics as well as changes from 5 days prior. This second dataset is used in the industry-wide analyses in the “Industry Visualizations” and “3D Heat Map” tabs.

Finally, the third dataset, income\_statements.csv, contains quarterly information on over a thousand publicly traded companies. Each company appears only once in this dataset, and each observation is a snapshot of a company’s performance on the first day of August 2021. This dataset is used for the “Financial Visualization” tab which analyses trends on a quarterly level.

**How was the data collected?**

The data was scraped from financial reporting site Finviz.com over the course of a month. Daily information was saved using Google Sheets and built-in functionality. RStudio was used to merge daily datasets into the three datasets used that each span a month in time.

**Who are the users that this visualization was made for?**

This visualization was made for anyone with an interest in finance. Some parts of the visualization may be better suited for those with an intermediate understanding of stock movements and regression modeling. However, most of the visualizations are made to simplify complex data and are suitable for the general audience. Investors may find the project useful to help predict stock change in the future using current data.

**What questions are you trying to answer? What works?**

In my project the primary question I attempt to answer is if you can predict future stock movements using current data. I use a lagged variable that consists of data from 5 days prior to the observation date and fit a linear model predicting price as an outcome variable and the lagged data as the explanatory variable to address this. The secondary question my project tackles is how industry-wide financial trends present themselves visually. I plot income by sector and ESG performance by profitability and sector to provide insight into industry-level trends. Everything in the project works.

**What insights did you get from your data?**

In regards to my primary question, I find that for Apple in the month of August 2021, 5-day lagged price and simple moving average (SMA) change are statistically significant predictors in the simple linear model of current price. At face value, this means that the data tells us you can reasonably predict future movements using current data. However, the model is exceedingly basic and does not use multiple explanatory variables which means it almost definitely suffers from omitted variable bias.

For my secondary question, I find that income varies widely across different sectors. Financial companies seemed to perform the best in terms of gross revenue, and industrial companies the worst. Additionally, I find that there seems to be either no correlation or a slightly negative correlation between profit margin and ESG score across all companies. It is possible a quadratic function could be a better fit than a linear one to explain the relation between profit margins and ESG scores. Finally, there seems to be no relation between lagged variables and current prices on an aggregate level using data from all companies.

**What needs improvement?**

Future implementations could look to improving the regression model. This could be done in multiple ways. The regression model could be expanded to include multiple explanatory variables to provide a better understanding of the relationship between lagged metrics and current price. Additionally, explanatory and outcome variables could somehow be scaled so that they are comparable across companies and industries in a way that makes it possible to fit a model on the data. Currently, I use absolute values in variables which makes it difficult to include more than one company in the regression model.

Data could be improved to remove outliers in ways that are appropriate.

**Sources or References (with appropriate credits)**

*Building linear model and implementation in Shiny*

<https://www.analyticsvidhya.com/blog/2021/05/build-interactive-models-with-r-shiny/>

<https://statsandr.com/blog/a-shiny-app-for-simple-linear-regression-by-hand-and-in-r/>

<https://bookdown.org/paul/applied-data-visualization/example-a-simple-regression-app.html>

*Understanding linear models*

Fitting and interpreting models (R)

<https://www.youtube.com/watch?list=PLNUVZZ6hfXX1tyUykCWShOKZdIB0TIhtM&v=69U92Q3pwnA>

Modeling non-linear relationships (R)

<https://www.youtube.com/watch?v=j4MZ6ZdHnHg&list=PLNUVZZ6hfXX1tyUykCWShOKZdIB0TIhtM>

*Functions for GG Plot visualizations*

<https://www.maths.usyd.edu.au/u/UG/SM/STAT3022/r/current/Misc/data-visualization-2.1.pdf>

*Converting 2D models to 3D*

<https://www.rayshader.com/>

<https://github.com/adamBirenbaum/rocky_mt_models>

<https://github.com/tylermorganwall/rayshader/issues/78>

<https://www.maths.usyd.edu.au/u/UG/SM/STAT3022/r/current/Misc/data-visualization-2.1.pdf>

**Describing the process and development to make your work reproducible, for example, tidying data.**

To tidy data into an acceptable format for analysis, I compiled financial datasets of all days of August into one dataset by using a full join. I then applied functions to clean the data that converted characters to numbers, dropped duplicate rows, and dropped empty/unneeded observations. Finally, I utilized for loops for datasets compositeAugust.csv and composite\_total2.csv to create lagged variables used in regression and analysis for each dataset, respectively.

**Description of design decisions (encoding/mapping). Use the taxonomy of the what-why-how analysis framework presented by Tamara Munzner. Sections separated by tab.**

*Scatterplot*

|  |  |
| --- | --- |
| Idiom | Scatterplot |
| What: Data | Table: two quantitative value attributes |
| How: Encode | Express values with horizontal and vertical spatial position and point marks |
| Why: Task | Find trends, outliers, distribution, correlation; locate clusters; visualize linear regression model and best fit line. |
| Scale | Items: tens |

*Distribution*

|  |  |
| --- | --- |
| Idiom | Histograms |
| What: Data | Table: two quantitative value attributes |
| How: Encode | Express values with horizontal and vertical spatial position and bars to collect and sum data points in bins |
| Why: Task | Identify distribution of dependent and independent variables used in linear model, locate abnormal distributions |
| Scale | Items: tens |

*Industry Visualization*

|  |  |
| --- | --- |
| Idiom | Hexagonal heat maps |
| What: Data | Table: two quantitative value attributes faceted by sector |
| How: Encode | Express values with horizontal and vertical spatial position and bins with color representing frequency of occurrence |
| Why: Task | Identify densities in relationships between variables utilized in regression |
| Scale | Items: thousands |

*Financial Visualization*

|  |  |
| --- | --- |
| Idiom | Horizontal dot plot |
| What: Data | Table: one categorical variable and one continuous variable |
| How: Encode | Express ranges of income with horizontal spacing and dot intensity |
| Why: Task | Visualize total income of companies by sector and identify highest/lowest earning sectors |
| Scale | Items: thousands |

|  |  |
| --- | --- |
| Idiom | Scatterplot |
| What: Data | Table: two quantitative value attributes |
| How: Encode | Express values with horizontal and vertical spatial position and point marks with color representing sector |
| Why: Task | Identify trends between ESG scores and profitability, visualize performance by sector, identify most ESG-friendly sectors |
| Scale | Items: thousands |

*3D Heat Map*

|  |  |
| --- | --- |
| Idiom | 3D Color Density Plot |
| What: Data | Table: two quantitative value attributes |
| How: Encode | Express densities of relationships between two values using horizontal, vertical, and depth spatial position with color and height representing density |
| Why: Task | Visualize relationships between two variables of interest, identify specifically the density of most common occurrences of relationships |
| Scale | Items: thousands |