Luke A. & James M.
Data Science
Professor Culotta
10/11/2023

Data Science Milestone 1

Our group will consist of Luke Albright and James Manzer for the data science project and presentation. We are exploring Fama-French and equity markets as well as Major League Baseball statistics as two possible datasets for our project, which are explained further below.

The first dataset we are interested in studying involves the Fama-French Data Library. James and Luke are both finance majors and have a strong interest in understanding financial markets and the instruments traded in them. Fama-French improved the historically significant CAPM asset pricing model, which is able to explain approximately 70% of portfolio returns. By including more variables that adjust the model based on the company's market capitalization and book-to-market ratio, the update model is able to explain approximately 90% of portfolio returns. We believe a good starting point to pull data from is Fama-French's own website, which includes numerous datasets that include current and historical calculations of the different research factors in the model as well as portfolio data. We can then compare these datasets to the historical performance of the S&P 500 to see what the correlation is between the returns of a large cap index and the different factors in the Fama-French model. There are many ways to expand upon this research by including different equity or economic data to compare the research factors to. Exhibit 2 is a scatterplot comparing the SP500 index level and the market risk premium for a given week beginning in 2015. There is little correlation between these variables.

The second dataset we are interested in studying involves Major League Baseball statistics. Both of us have a history of playing baseball and are now fans of the Chicago Cubs and Texas Rangers. James is even a member of the club baseball team here at Tulane. Given our passion for baseball and its vast records of statistics, we feel compelled to take a deep exploration into a number of possible variables. We have identified two websites in the public domain that store a trove of information pertaining to the MLB. The two sites are the Lahman Database and Retrosheet (links to both are included below). The Lahman Database has data such as pitching, hitting, fielding, team statistics, and more from 1871 to the 2022 season. Retrosheet has data covering play-by-play, salary, team statistics, and even a few unique datasets such as ejections. Retrosheet data typically spans from 1913 to the 2022 season. There are many variables to explore but a preliminary idea was to investigate the "keys" of success for a team. One of the most straightforward views would be to compare the total amount a team spends on player salaries and total wins for that season. Exhibit 3 is a scatterplot of salaries and wins from the 2015 season (y-axis represents 100 millions). As you can see, there is not a strong correlation between total spending and wins in 2015. This data may change if it is compared year over year or team to team.

We are roommates which makes collaboration throughout this project much easier. We plan to store any pertinent data and code through GitHub and Google CoLab. Our schedules align best on Monday and Wednesdays (as well as weekends) which is when we plan to complete any necessary meetings or collaborative work. However, noting again that we are

Luke A. & James M.
Data Science
Professor Culotta
10/11/2023

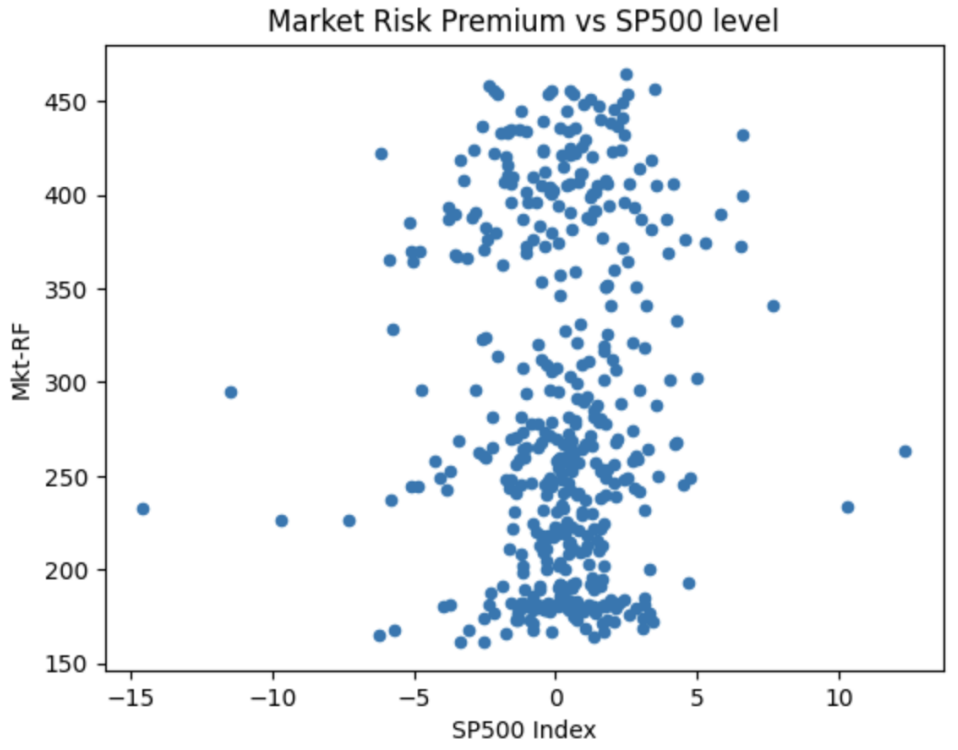
roommates, we expect our schedule to be much more fluid. Our GitHub site can be found at this link: https://github.com/lalbright22/DataScience2023.git

Exhibits:

Exhibit 1

Database	**Link to Database**
Fama-French Data Library	https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html
Lahman Database	Download Lahman Baseball Datbase (seanlahman.com)
Retrosheet	Retrosheet

Exhibit 2



Luke A. & James M.
Data Science
Professor Culotta
10/11/2023

Exhibit 3