Milestone 3
David Bombara, Nathan Dennis, and Aditya Saxena

Table of Contents

Background

The Form 13F data sets are derived from EDGAR filings, a system managed by the U.S. Securities and Exchange Commission (SEC). Form 13F filings are required by institutional managers with over $100 million in assets under management to disclose their quarterly holdings. The goal of these filings is to provide transparency in the activities and holdings of large managers in the financial markets.

The data sets are extracted from XML files in EDGAR Form 13F and made available as tab-delimited text files on a quarterly basis. These files are generated from structured filings by institutional managers and are compiled into seven specific tables within each quarterly data set.

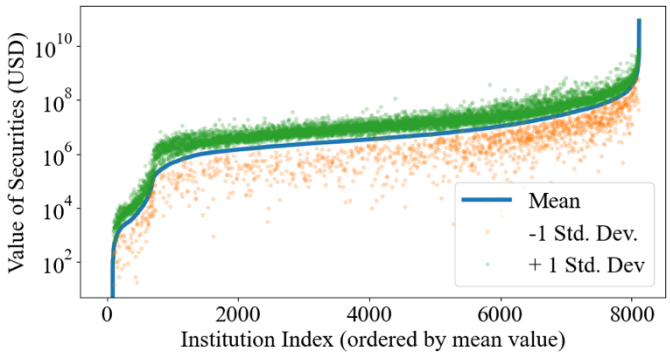
Structure and Organization of the Data

The data includes up to seven different tables, each representing different aspects of the filing information:

1. **COVERPAGE**: Provides cover page information like the filing manager's name, address, and report type. Key details include the report calendar quarter, amendment information, confidentiality status, and filing manager details.
2. **SUMMARYPAGE**: Offers a high-level summary of the filing, including counts of other included managers, the table entry total, and the total value of holdings. It also notes if certain information was confidentially omitted.
3. **INFOTABLE**: The primary data table containing detailed information about each holding, such as issuer name, class title, market value, share quantity, and voting authority. Key fields include *ACCESSION_NUMBER*, *INFOTABLE_SK*, and several columns providing identifiers for the financial instruments.

Finalized Research Question

The research question is: Can we predict the total value of the assets held by an institutional investor, given information about their location, number of shares held, and summary statistics of the individual holdings?



Data Cleaning and Augmentation

To clean the data, firstly I remove rows from *db_info* that report the "principal amount", instead of the number of shares. This makes the data easier to analyze while not reducing the insights we can gain from the data.

Next, I sort *df_info* by the *VALUE* column and notice two values whose holdings are clearly incorrect, based on my knowledge of the company share prices that these institutions reported. I thus remove these rows. I also converted the *SSHPRNAMT* column to a *float* datatype to ensure we can do numerical calculations on it.

Next, I compute summary statistics from *df_info*, grouped by the investor (given by *ACCESSION_NUMBER*). I compute the minimum, maximum, standard deviation, and the following percentiles: 1%, 10%, 25%, 50%, 75%, 90%, and 99%. I compute these statis

tics for two columns: *VALUE* (total value of individual security held) and *SSHPRNAMT* (number of shares).

I also sourced information on the United States (US) GDP by state from the U.S. Bureau of Economic Analysis (bea.gov). Therefore, we plan to include the 2024 Q2 state GDP as a feature. Another feature I added was an indicator variable indicating whether the investor was in the US.

Table 1: Raw Data Details

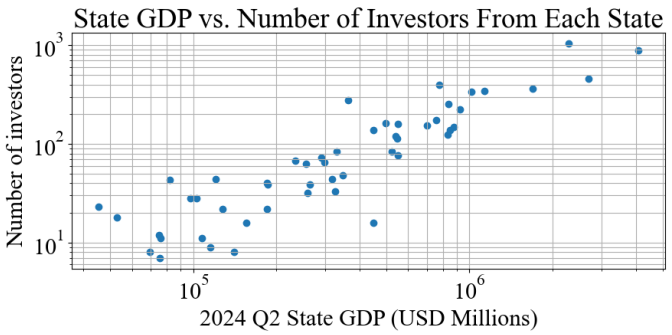
Table	Number of Rows	Number of Columns
COVERPAGE	10117	22
SUMMARYPAGE	8244	5
INFOTABLE	3278515	16

Data Visualization and Insights

The distribution of the value of securities by institutional investor is shown below. In this plot, each x-value represents a single investor. The investors are ordered by the mean value (in US dollars) of the securities they hold. To understand this statistic, consider the following example for a fictional investor: An investor holds $100 M in Apple shares, $50 M in Tesla shares, and $240 M in NVIDIA shares. This means the average security value of said investor is $130 M.

Investors are also required to report their office locations. Therefore, the figure below shows the correlation between state GDP and number of investors in the 2024-Q2 13F filings from each state. States with higher GDP tend to have more

investors submit Form13F filings. On a logarithmic scale, (taking the logarithm of both the x-axis and y-axis variables), the relationship is approximately linear.



Next, we study the correlations between features and predictors. To help produce a linear correlation, the logarithm of each feature will be used in the model implementation.

Correlation between Features and Prediction Variable

Firstly, the figure below 12 subfigures, each with *TABLEVALUETOTAL* on the y-axis. *TABLEVALUETOTAL* represent the total values of assets held by the investor. It doesn't exactly measure total assets under management, AUM, because some assets (like private equity) are not required to be reported on the 13F form. However, we can consider *TABLEVALUETOTAL* to be a lower-bound approximation of AUM. On the x-axis of each subfigure is a summary statistic of the *VALUE* column that were computed early. For example, "*10% VALUE*" is the 10% percentile of the value of securities held. The summary statistics are approximately positively correlated with *TABLEVALUETOTAL*. In this scatter plot, each dot represents a single institutional investor.

Similarly, we show the correlation between number of shares and *TABLEVALUETOTAL* in the next scatter plot. The number of shares is given in the column *SSHPRNAMT* from the dataset. Here, each dot again represents a single institutional investor. *"SHAMT"* also stands for "share amount" (number of shares).

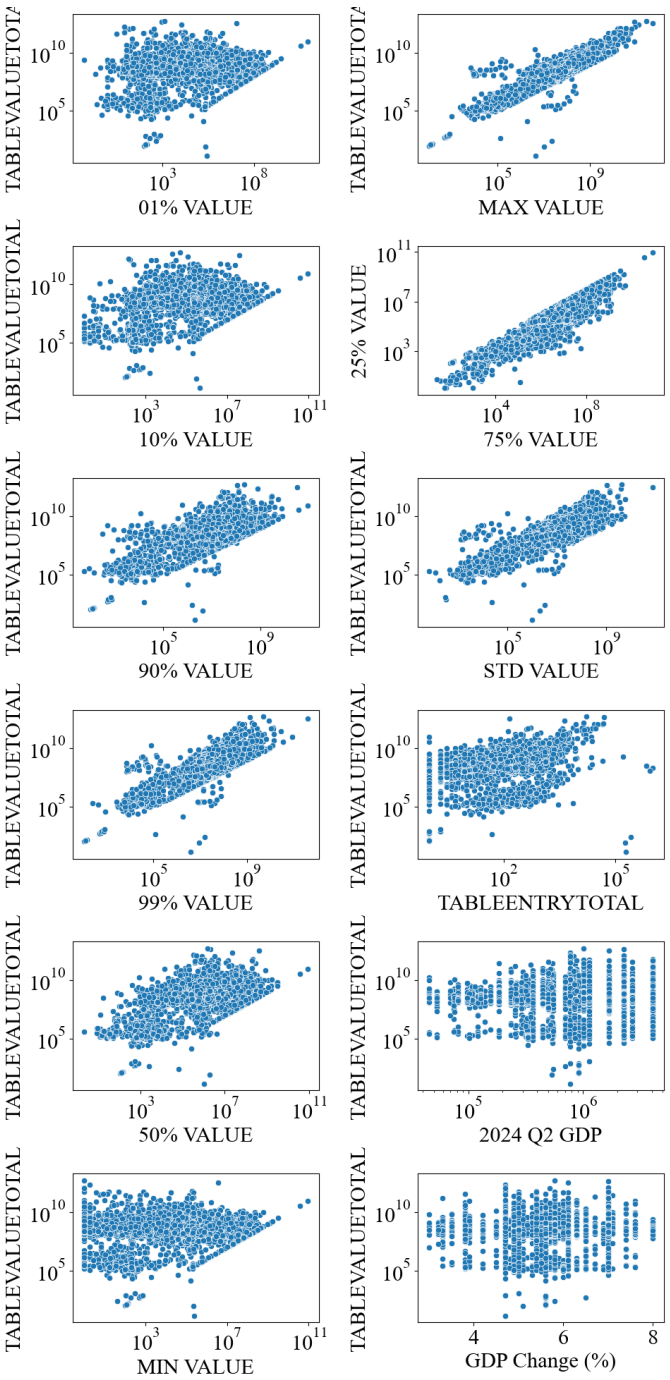
Thirdly, we plot a heatmap depicting the correlation matrix of the numerical columns in the dataset. We show heatmaps for the summary statistics of *VALUE* (securities value) and *SHAMT* (share amount). The top row is *TABLEVALUETOTAL*. Each cell of the heatmap lists the correlation value.

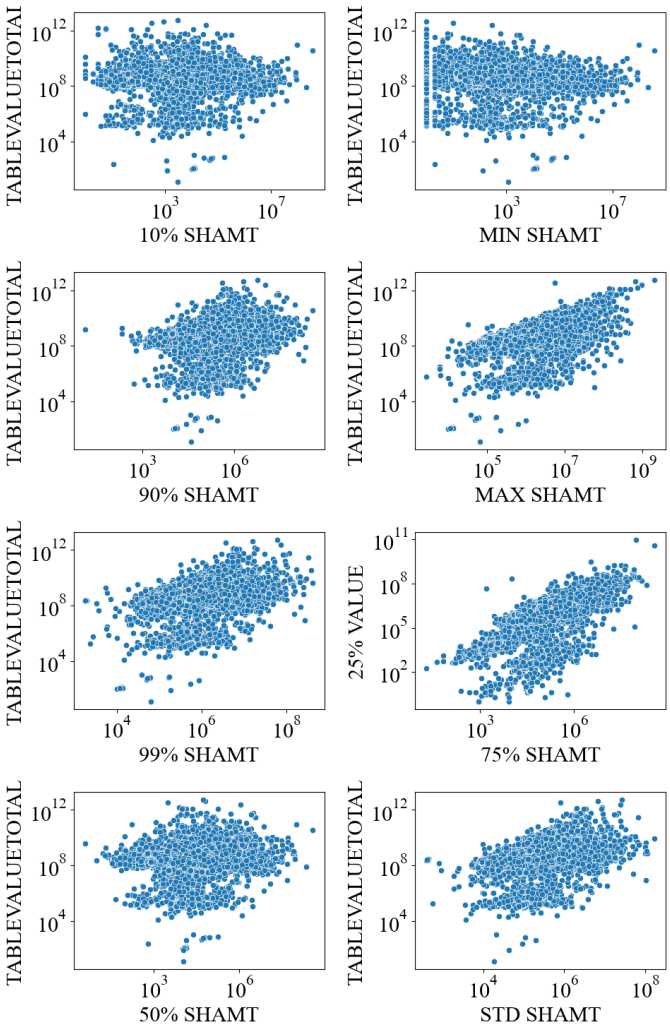
Noteworthy Findings

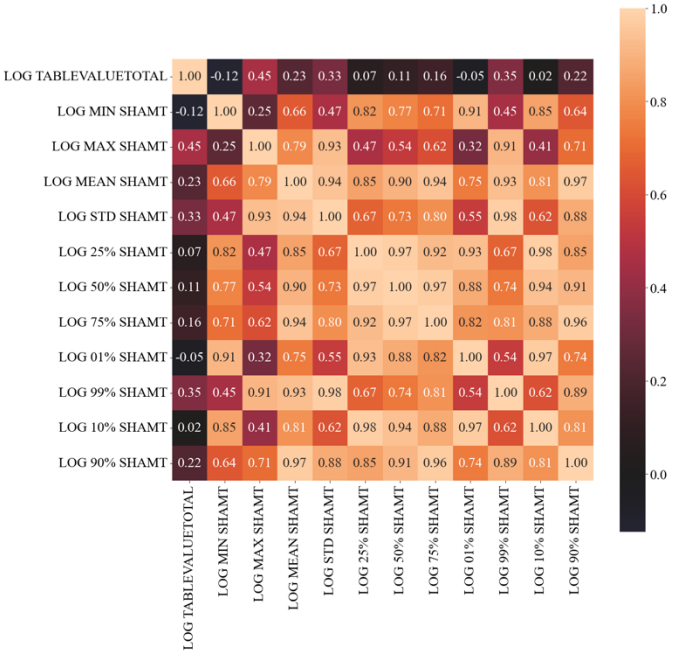
The numerical features should be transformed using by taking the logarithm. This makes the correlations between features appear to be somewhat linear. Additionally, the correlation between state GDP and number of investors from each state is quite interesting.

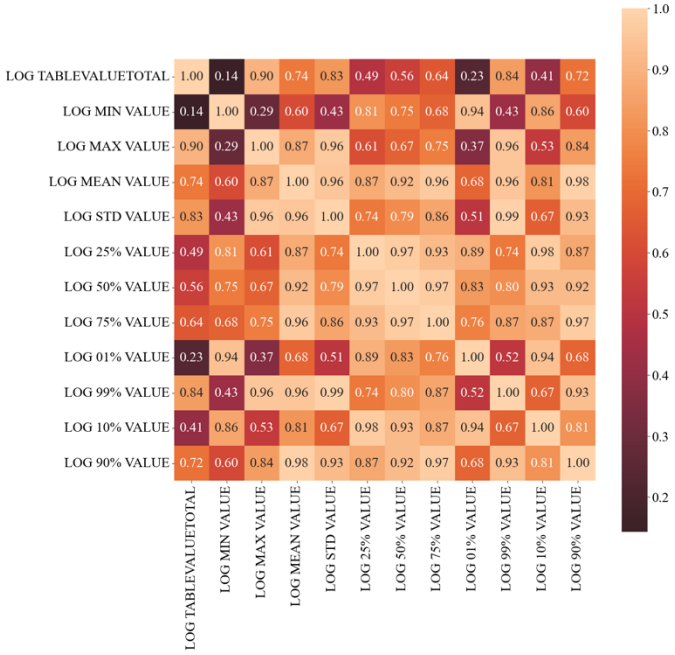
Baseline Model

For our baseline model, we implement multi-linear regression using a subset of features. Firstly, we scale the features to account for differences in value ranges. We use *Standard_Scaler* from *sklearn*. We choose the following predictors: *2024 Q2 GDP, LOG MAX SHAMT, LOG STD SHAMT, LOG STD VALUE, LOG MEAN VALUE, LOG MAX VALUE,* and *LOG MIN VALUE*. We then use *LinearRegression()* from *sklearn*. We are also planning to test the effectiveness of polynomial features. To access, the model performance, we will use mean squared error.




TABLEVALUETOTAL
10% SHAMT
MIN SHAMT
90% SHAMT
MAX SHAMT
99% SHAMT
25% VALUE
75% SHAMT
50% SHAMT
STD SHAMT


LOG TABLEVALUETOTAL 1.00 -0.12 0.45 0.23 0.33 0.07 0.11 0.16 -0.05 0.35 0.02 0.22
LOG MIN SHAMT -0.12 1.00 0.25 0.66 0.47 0.82 0.77 0.71 0.91 0.45 0.85 0.64
LOG MAX SHAMT 0.45 0.25 1.00 0.79 0.93 0.47 0.54 0.62 0.32 0.91 0.41 0.71
LOG MEAN SHAMT 0.23 0.66 0.79 1.00 0.94 0.85 0.90 0.94 0.75 0.93 0.81 0.97
LOG STD SHAMT 0.33 0.47 0.93 0.94 1.00 0.67 0.73 0.80 0.55 0.98 0.62 0.88
LOG 25% SHAMT 0.07 0.82 0.47 0.85 0.67 1.00 0.97 0.92 0.93 0.67 0.98 0.85
LOG 50% SHAMT 0.11 0.77 0.54 0.90 0.73 0.97 1.00 0.97 0.88 0.74 0.94 0.91
LOG 75% SHAMT 0.16 0.71 0.62 0.94 0.80 0.92 0.97 1.00 0.82 0.81 0.88 0.96
LOG 01% SHAMT -0.05 0.91 0.32 0.75 0.55 0.93 0.88 0.82 1.00 0.54 0.97 0.74
LOG 99% SHAMT 0.35 0.45 0.91 0.93 0.98 0.67 0.74 0.81 0.54 1.00 0.62 0.89
LOG 10% SHAMT 0.02 0.85 0.41 0.81 0.62 0.98 0.94 0.88 0.97 0.62 1.00 0.81
LOG 90% SHAMT 0.22 0.64 0.71 0.97 0.88 0.85 0.91 0.96 0.74 0.89 0.81 1.00
LOG TABLEVALUETOTAL
LOG MIN SHAMT
LOG MAX SHAMT
LOG MEAN SHAMT
LOG STD SHAMT
LOG 25% SHAMT
LOG 50% SHAMT
LOG 75% SHAMT
LOG 01% SHAMT
LOG 99% SHAMT
LOG 10% SHAMT
LOG 90% SHAMT
1.0
0.8
0.6
0.4
0.2
0.0


LOG TABLEVALUETOTAL 1.00 0.14 0.90 0.74 0.83 0.49 0.56 0.64 0.23 0.84 0.41 0.72
LOG MIN VALUE 0.14 1.00 0.29 0.60 0.43 0.81 0.75 0.68 0.94 0.43 0.86 0.60
LOG MAX VALUE 0.90 0.29 1.00 0.87 0.96 0.61 0.67 0.75 0.37 0.96 0.53 0.84
LOG MEAN VALUE 0.74 0.60 0.87 1.00 0.96 0.87 0.92 0.96 0.68 0.96 0.81 0.98
LOG STD VALUE 0.83 0.43 0.96 0.96 1.00 0.74 0.79 0.86 0.51 0.99 0.67 0.93
LOG 25% VALUE 0.49 0.81 0.61 0.87 0.74 1.00 0.97 0.93 0.89 0.74 0.98 0.87
LOG 50% VALUE 0.56 0.75 0.67 0.92 0.79 0.97 1.00 0.97 0.83 0.80 0.93 0.92
LOG 75% VALUE 0.64 0.68 0.75 0.96 0.86 0.93 0.97 1.00 0.76 0.87 0.87 0.97
LOG 01% VALUE 0.23 0.94 0.37 0.68 0.51 0.89 0.83 0.76 1.00 0.52 0.94 0.68
LOG 99% VALUE 0.84 0.43 0.96 0.96 0.99 0.74 0.80 0.87 0.52 1.00 0.67 0.93
LOG 10% VALUE 0.41 0.86 0.53 0.81 0.67 0.98 0.93 0.87 0.94 0.67 1.00 0.81
LOG 90% VALUE 0.72 0.60 0.84 0.98 0.93 0.87 0.92 0.97 0.68 0.93 0.81 1.00
LOG TABLEVALUETOTAL
LOG MIN VALUE
LOG MAX VALUE
LOG MEAN VALUE
LOG STD VALUE
LOG 25% VALUE
LOG 50% VALUE
LOG 75% VALUE
LOG 01% VALUE
LOG 99% VALUE
LOG 10% VALUE
LOG 90% VALUE
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2