**CS109B Project: What makes an asset class attractive?**

Group 16: Yang Liu, Ellen Sun, Hayden Mckee, Qi Chu

1. **Abstract**

This project is to quantify what makes an asset class attractive from perspective of human visual, technical analysis and time series analysis. It aims to predict the forward 1 week return of each individual asset class on secondary market including equity and bond etc. Based on the attributes of their own past security price and trading volume pattern, we utilized different sets of models from simple multiplayer perceptron to deep neural networks (CNN, RNN, LSTM) to formulate a multi-asset trading strategy by longing/shorting the predicted top/ bottom asset classes.

1. **Background**

In US secondary market, there are plenty of asset classes being traded on daily basis for investor to access different market exposures, including equity, bond, commodity, and foreign assets. Traders and portfolio managers are making predictions and bets on these asset classes on an ongoing basis, based on the analysis of macro, fundamental, technical, or sometimes even just simple visuals.

The commons visual and technical tools could range from moving average of price, candle chart, long term momentum, short term reversal, relative strength indicator, moving average convergence/divergence etc. However, these tools work in different manners for different asset classes in various time periods. What’s worse is that analyst could cherry pick the one that works best for history, but does not generalize well for future. Therefore, a systematic machine learning tool is desperately needed to assist humans to make technical and visual decisions in a more timely and efficient manner.

In this project, we are aiming to build a neural network to extract the essential visual and technical features from history and make systematic predictions for future prices. The wish list feature of neural networks would be a combination of feature extraction and time series/context sensitivity.

1. **Data Description**

The dataset is daily pricing and trading volume data of 16 major asset classes for the period of 1970 - 2022 from Bloomberg. Below are the asset classes including 6 bonds, 7 equities and 3 commodities.



The pricing data are mainly price and volume based. Below are the 6 data fields downloaded from Bloomberg.



1. **Task**

This project proposes to use past return data and visual pattern to predict forward 1 week return of different asset classes. There are many different settings we can try to formulate the research question. Considering past literature findings, we propose preliminary steps as follows.

4.1 **Input transformation**

There are different ways of translating time series data of asset class returns into images. Below are the techniques we plan to use for input transformation purpose.

4.1.1 Black and white graphs with open/close/high/low prices, daily trading volume and moving average of past returns

4.1.2 Wang and Oates (2015) proposed another two methods:

* 1. Graiman Angular Fields (GAF). In short, this methodology tries to map normalized time series into polar coordinates with mathematical operations. It gives a relationship between every data point and every other data point over time.
  2. Markov Transition Fields (MTF). Advantage of this method is that it extracts positional information of data points and conclude how related to two data points are in the time series.

4.2 Lookback window

Another parameter we can try to vary is the look back window when forming images. We can certainly test the effect of different lookback window. For starters, we would like to test:

* 1. Short term return information, i.e. lookback window of 5 days.
  2. Medium term return information, i.e. lookback window of 21 days.
  3. Longer term return information, i.e. lookback window of 63 days.

4.3 Label

Between a classification problem and regression problem, we lean toward using regression i.e. continuous labels at this point. The reason is since the nature of this project is to forecast returns, having a continuous estimation is more convenient to form backtest strategies.

* 1. Predicting forward 1 week return
  2. Winsorized at 95% and 5% percentiles

4.4 Model structure

4.4.1 First we would establish the baseline model of regression with the help of PCA analysis.

4.4.2 On top of baseline model, we would want to try FFNN model as a secondary baseline model for more complicated neural networks.

4.4.3 With 2 baseline models, we plan to build and test relatively simpler CNN structures.

* 1. Preferably at this step we should’ve narrowed down the most promising specs list before (image translation method, lookback window etc.)

4.4.4 Next we can try to test preferred specs on more sophisticated models. A natural question to ask is can we use pre-trained models to add value besides simple CNN structure.

* 1. AlexNet
  2. GoogleLetNet

4.4.5 Lastly, we think it would also be interesting to try Long short term memory (LSTM) network since it is generally seen as a good tool for time series prediction.

4.5 Training and testing split

Full Data: 53 years of daily pricing data from 1970 to 2022

Training Data: first 48 years from 1970 to 2007

Testing Data: last 15 year to test the model from 2008-2022.

* 1. Model Evaluation

Since it is a regression problem, naturally we want to examine common metrics such as MSE. More importantly, since the nature of the project is to generate future returns, we will implement a backtest strategy to observe the performance of long-short portfolios formed based on predicted result.

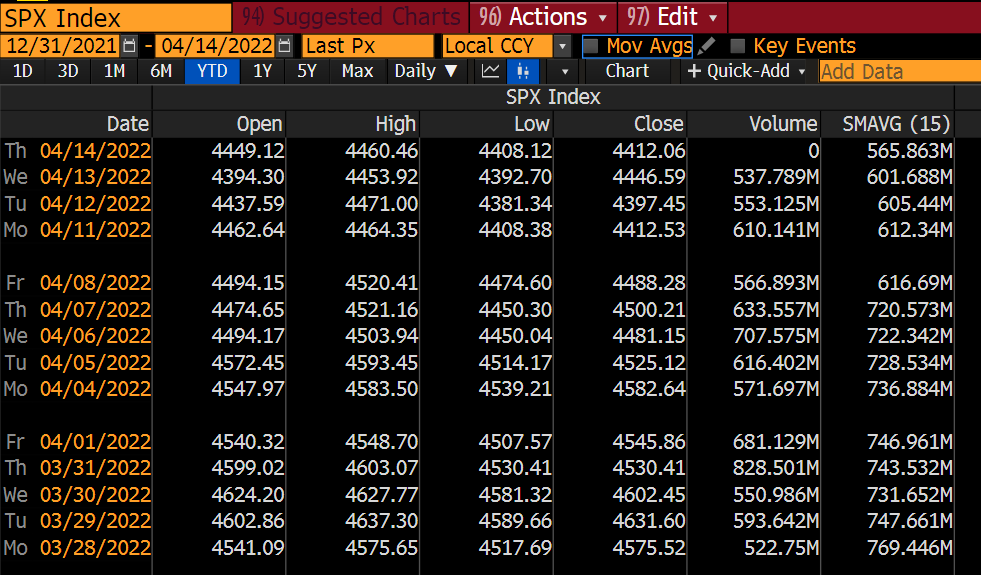
1. **Exploratory Data Analysis**

5.1 **Basic dataset information**

Below is an example of daily return chart of LRG (US Large Cap Core Equity), with the top panel as pricing and bottom as trading volume. Daily prices of this asset class are represented as a candle chart, with one candle per day. Top and bottom edges of the candle represent the open and close price or vice versa. Sticks on the top and bottom represents the intraday high and low prices.

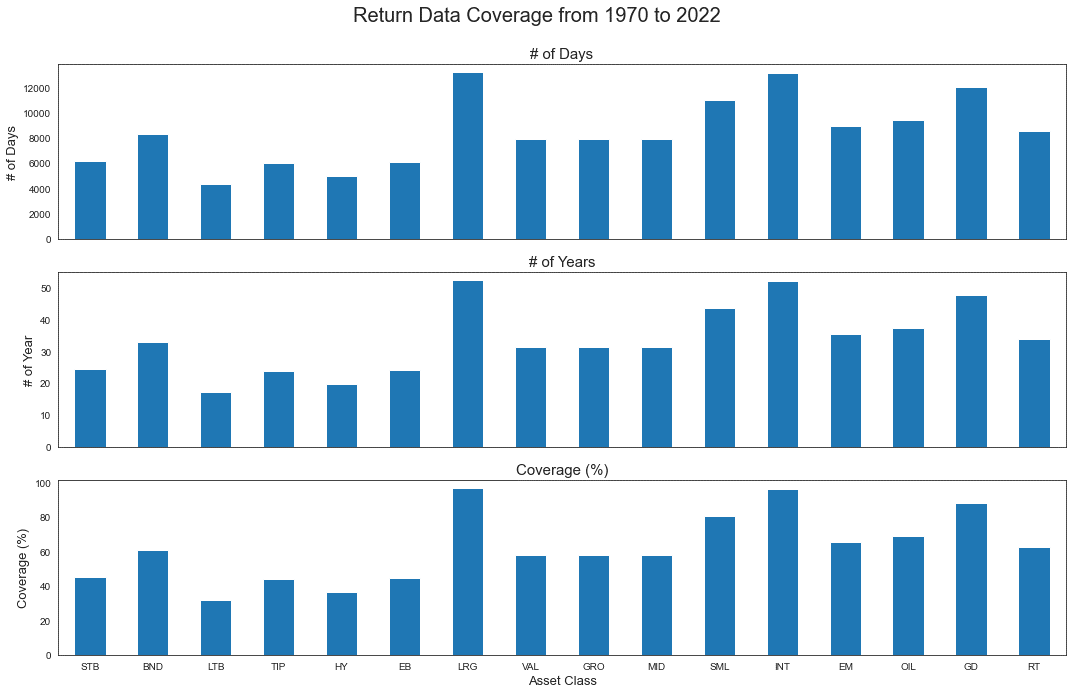


In this project, it would take tremendous and meaningless efforts to download this chart for each asset class every day for 53 years. Instead, the data behind the chart is downloaded and will be converted back to chart format using the techniques mentioned in section 4 above.

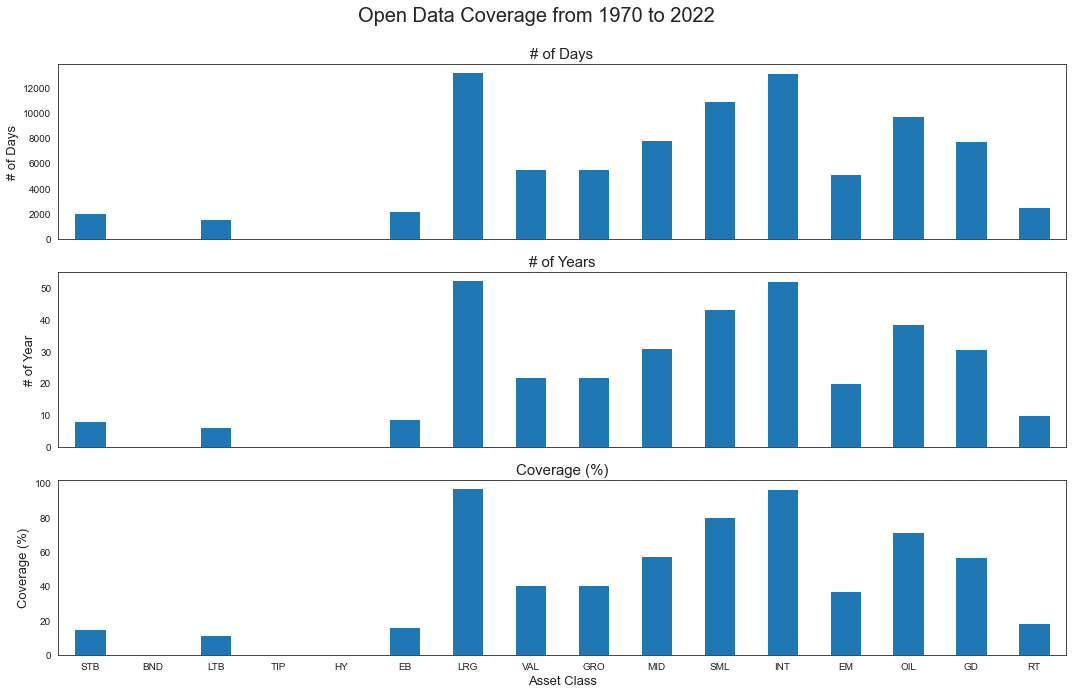


5.2 **Data coverage**

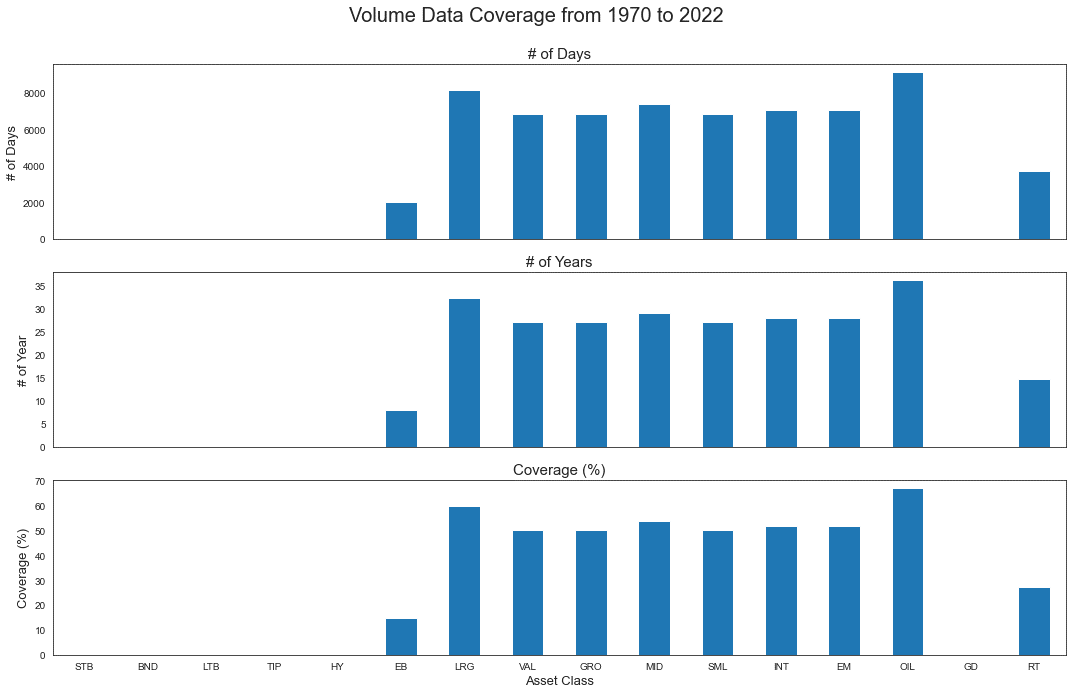
The dataset has good coverage overall. All asset classes have more than 17 years of data, with most having more than 30 years. LRG (US Large Cap Core Equity) and INT (International Equity) have the full coverage of 53 years. For those with shorter history, we will use data before 2008 as training data and same window as others as testing data.



For many of the fixed income asset classes, there is just close price without open, high and low. To align the data fields, we make those 3 fields the same as close price.

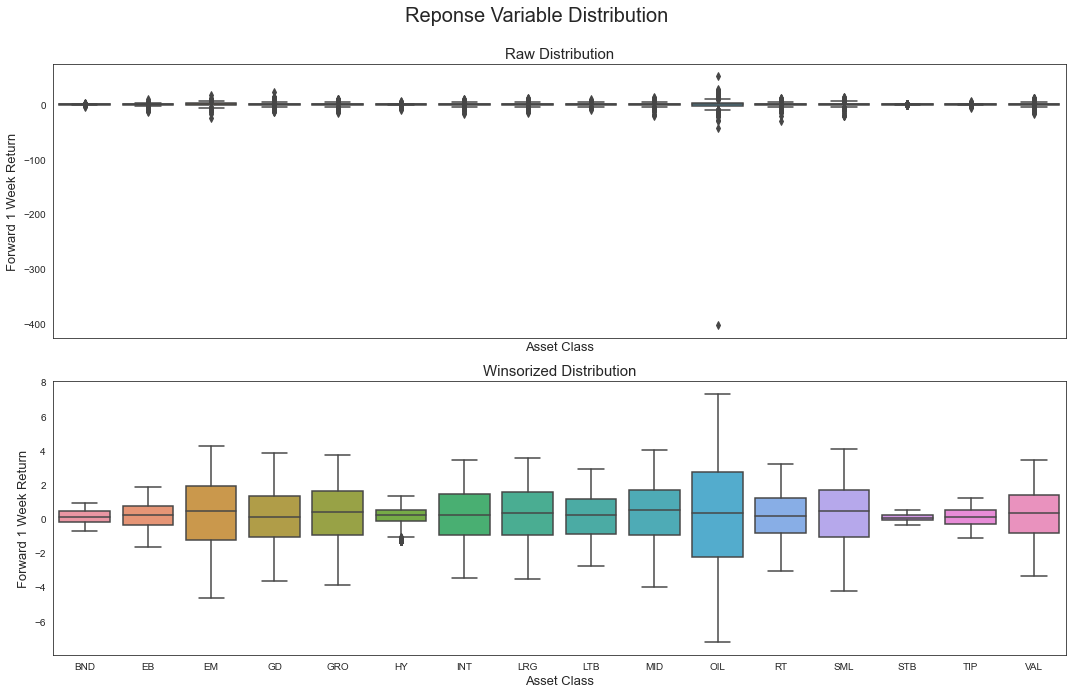


One thing we also notice is that most asset classes does not have volume data. Therefore, we are dropping volume data and will only focus on pricing data, including open, close, high and low prices.



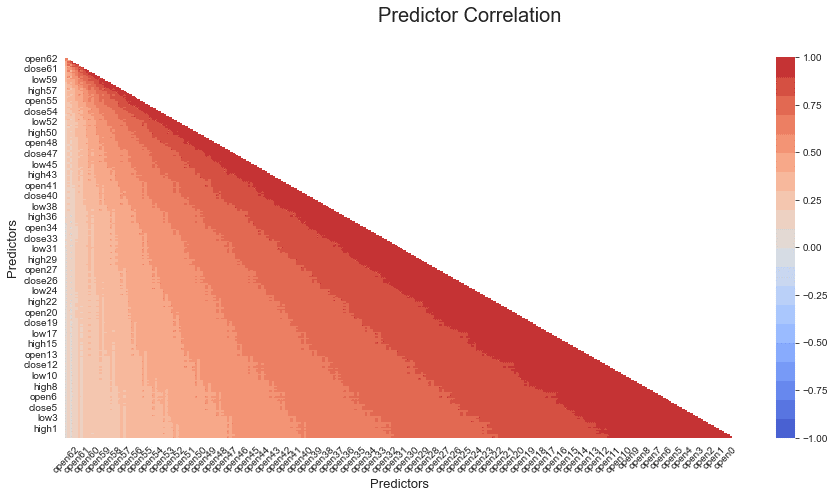
**5.3 Distribution of response variable - forward 1 week return**

The distribution of forward 1 week return could be off the chart, for example OIL in the top panel of chart below. To mitigate the noise there, we winsorize the response variable at 5% and 95% based on their own distribution. In the. After winsorization, the distribution of response variable looks more reasonable and lies between +/-8%.



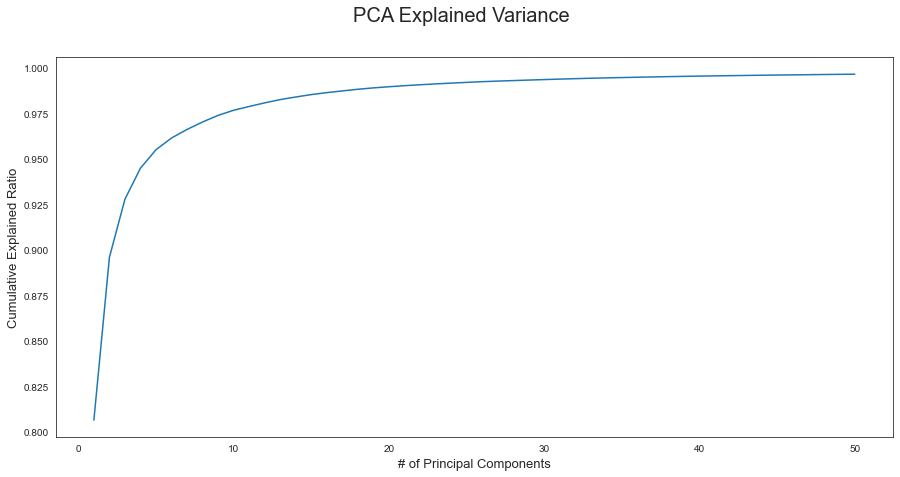
**5.3 Correlation of predictors**

Below is the correlation plot of predictors, with plenty of them highly correlated. And this does make intuitive sense, since predictors from the same day (high, low, open and close) are supposed to be closer to each other than other days. However, this high correlation would create the multi-collinearity issue for regression, which we will use as baseline model. Instead, we will do a PCA analysis first to extract the major features of the predictors and then run regression of response variables against the important principal components.



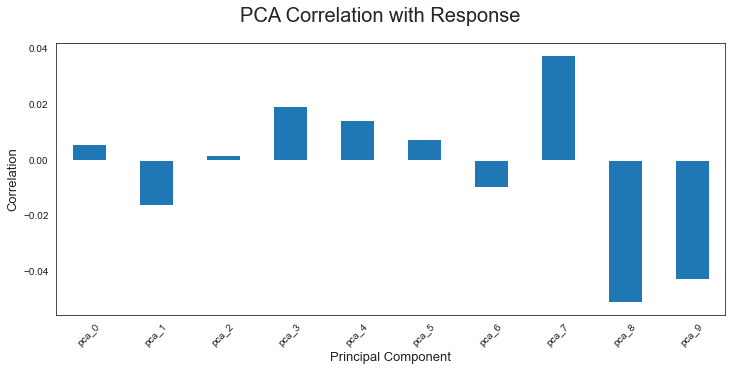
**5.5 PCA importance**

Below is the cumulative explained variance of PCA analysis on 252 predictors. As we can see, the first 10 principal components play very important roles in explaining the variance. Therefore, we will use the first 10 principal components for baseline model.



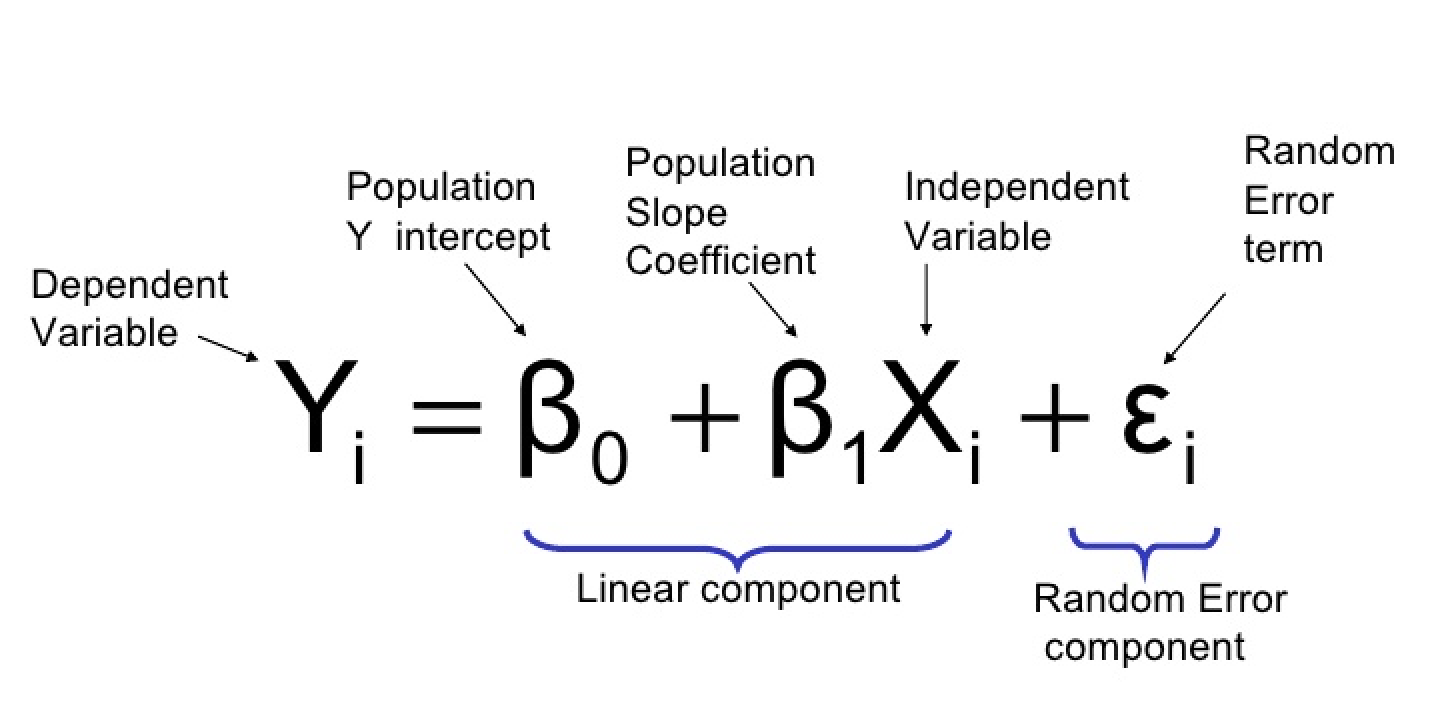
**5.5 Correlation with the response variable**

We do see that very little correlations of principal components with response. With these components uncorrelated with each other, there would likely very few features to be picked up by regression.



1. **Baseline Model**

A simple multi-linear regression model is used here as the baseline model, with the number of followers of a playlist as the response variable and all eligible numeric and categorical predictors included.



We split the data into training and test data with the date of 2007/12/31. All models in this project are fitted with training data and evaluated with test data.

Initially, solely looking at the relationship matrix above between the predictors and the response variable, it was clear that there is no linear relationship between these two. We suspect that PCA focus too much on capturing the noise without the intention to extract informative features for future predictions. Our multi-linear model produced the following scores.

***Model Performance***:

Train MSE: 4.63

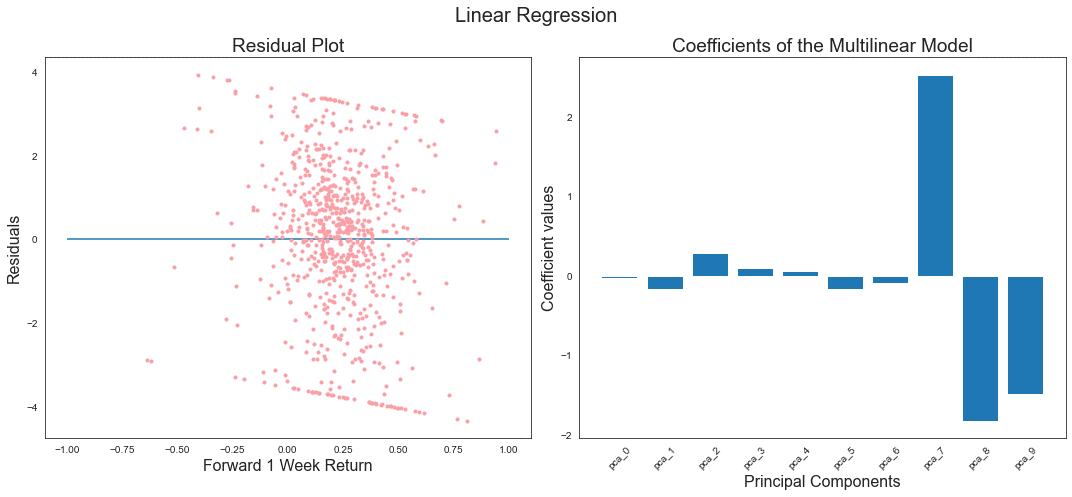
Test MSE: 5.83

Train R2: 1.07%

Test R2: -1.33%

A decent deterioration in the R2 score is observed, but both train and test are close to zero.

Based on the residual plot, residuals are mostly scattered evenly above and below the horizontal zero line. This means that our model does not capture enough information features from predictors. Based on this bar plot of regression coefficients, feature 7 seems to contribute positively for prediction and vice versa for feature 8 and 9. Unfortunately, the top features extracted by PCA does not seem to help much.



The conclusion here is that more enhancement is needed to improve the prediction. The wish list could cover the spectrum of visual feature extraction, time series awareness and more complicated approximation. We believe that FFNN, CNN, RNN and LSTM are perfect candidates to explore and improve the model.

**References:**

[1] **Wang, Z., T. Oates.** 2015. Encoding Time Series as Images for Visual Inspection and Classification

Using Tiled Convolutional Neural Networks, Trajectory-Based Behavior Analytics: Papers from the

2015 AAAI Workshop.[2] **Jiang, J., Kelly, B., and Xiu, D.** 2020. (Re-)Imag(in)ing Price Trends (December 1, 2020). Chicago Booth Research Paper No. 21-01