'What is the Federal Reserve actually doing?'

Midterm Report

Our Dataset - Monthly & Quarterly Time series Data

Independent Variables - Labels	Dependent Variables - Labels
● FedFunds Effective Rate - FEDFUNDS ● Federal Reserve Balance Sheet liabilities ○ Total - RESPPNTNWW ○ Treasuries ■ All - TREAST ■ 10 year+ maturity - TREAS10Y ■ 5-10 year maturity - TREAS5T10 ■ 1-5 year maturity - TREAS1T5 ■ 91 day - 1 year maturity - TREAS911Y ○ Mortgage Backed Securities ■ Over 10 year maturity - MBS10Y ○ Federal Agency debt ■ All - FEDDT ■ 10+ year maturity - FEDD10Y ● Press Release Language ○ Positive - Positive ○ Neutral - Neutral ○ Negative - Negative ● M1 Money Supply - M1SL	● Asset Prices ○ SPY ETF S&P 500 index - SPY ○ US Home Prices - CSUSHPINSA ○ Wilshire 5000 company index - WILL5000INDFC ○ Wilshire small cap index - WILLSMLCAP ○ US Coporate Bond Total Return Index - BAMLCC0ACMTRIV ○ US Corporate Bond High Yield Total Return Index - BAMLHY0A0HYM2TRIV ● Macroeconomic Indicators ○ Personal Consumption Expenditures - PCEPI ○ Personal Consumption Expenditures excluding food and energy - PCEPILFE ○ Unemployment Rate - UNRATE ○ Gross Domestic Product - GDP ○ Share of total US wealth held by the top 1% - WFRBST01134 ● Interest Rates ○ 1 Year Treasury yield- GS1 ○ 5 Year Treasury yield- GS5 ○ 10 year treasury yield -GS10 ○ 30 year treasury yield - GS30 ○ 2 vs 10 Treasury yield - T202YM

Data Cleaning

Our data was available at various different frequencies, daily, weekly, monthly, and quarterly. We wanted all our data to be at a monthly frequency. This presented a complicated problem as some data needed to be differenced and weekly and quarterly data was not available for the first of every month. We decided we would want any missing data to be forward filled from the last available data point. If we forward filled, then took the difference and then took the value for the first of each month we would not end up with the actual difference for each month. If we took the differences from the original data and then forward filled and took the values for the first of each month this also wouldn’t be the actual change month to month. For our data where we decided to take the arithmetic change which had weekly wednesday data 'FEDDT','FEDD10Y','RESPPNTNWW', 'M1SL','TREAST','TREAS10Y','TREAS1T5', 'TREAS5T10', 'TREAS911Y', 'MBS10Y','WFRBST01134', we forward filled all the daily data with the prior totals and then took the monthly totals and differenced them. For data where we wanted the percentage change 'SPY','GDP','WILL5000INDFC', 'WILLSMLCAP','CSUSHPINSA', which either came at a quarterly or monthly interval we took the difference from the original data and forward filled it for the missing months, such as the case for GDP.

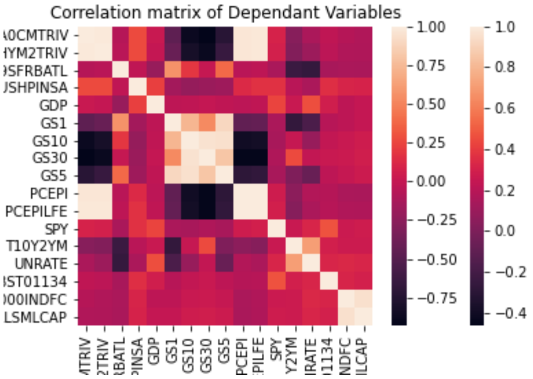
We had to read in CSV files from various data sources and standardize them into pandas datetime indexes, and outer merge them in order to maintain all our data and keep its time integrity. We used the above procedures before the merge in the case of the percentage change data or after merging and filling in the case of the arithmetic difference data.

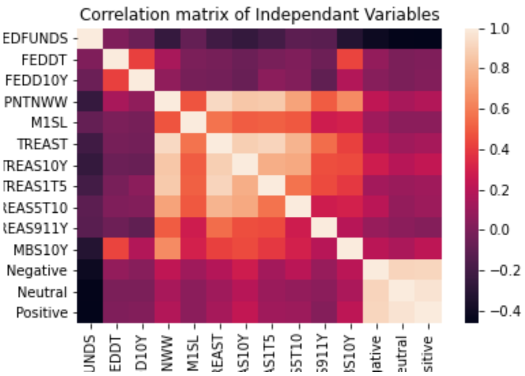
How will we avoid Over and Under fitting?

To prevent under fitting, we start with a model that uses all the available independent variables. We will then select and engineer features which are most explanatory and least correlated with our other available featurEs to reduce variance and increase explanatory power of the model.

To prevent overfitting, we would use k-folds cross validation to see if the coefficients of each feature would be stabilized under different sets of original data. We would also use both linear regression and tree models to analyze the data and see if we would receive similar results.

We are going to look for highly correlated independent and dependant variables using the correlation matrices below:


Correlation matrix of Dependant Variables


Correlation matrix of Independant Variables

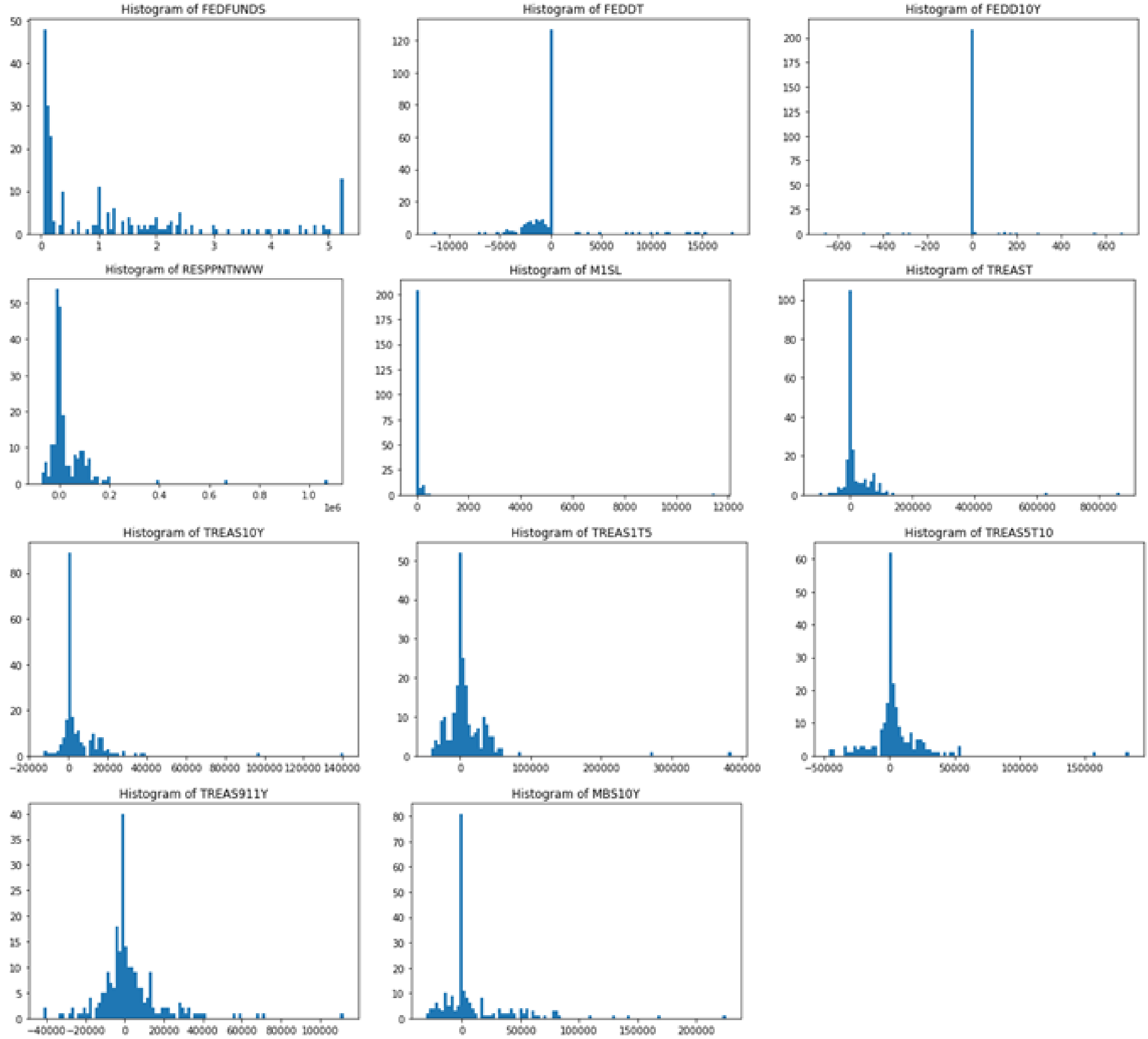
We see a high correlation between the total fed balance sheet and the individual assets on the balance sheet. We will likely exclude the total from our models.

Federal Reserve related numerical data:

Descriptive statistics:

	FEDFUNDS	FEDDT	FEDD10Y	RESPPNTNWW	M1SL
mean	1.2924	10.433	0.9509	32491.8348	83.1429
std	1.5877	3457.462	94.0477	99436.033	766.9199

	TREAST	TREAS10Y	TREAS1T5	TREAS5T10	TREAS911Y	MBS10Y
mean	21435.9375	5359.7634	8564.5804	4154.0893	2496.3036	10844.5
std	78032.9707	13547.1094	37251.7298	22375.6248	16882.0887	33021.9972

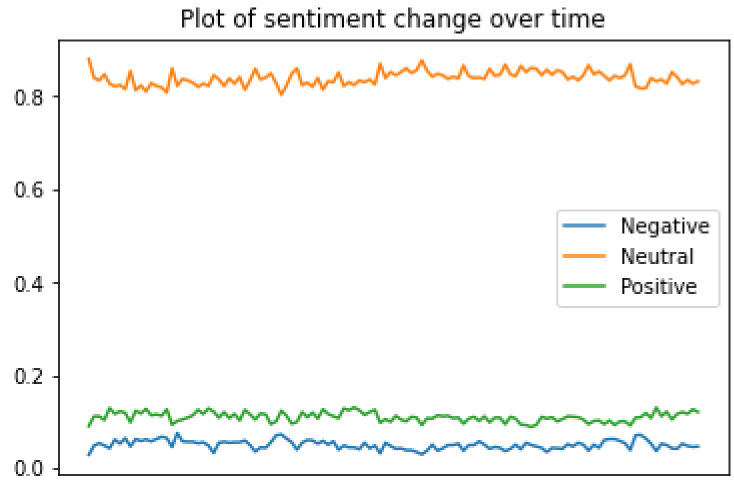


Press release language:

Press release language is processed through the use of NLTK's sentiment analysis. The resulting data point of each press release is a vector of three numbers each representing negative, neutral and positive sentiment.

Descriptive statistics:

	Negative	Neutral	Positive
mean	0.0506	0.8389	0.1105
std	0.0096	0.0153	0.0103



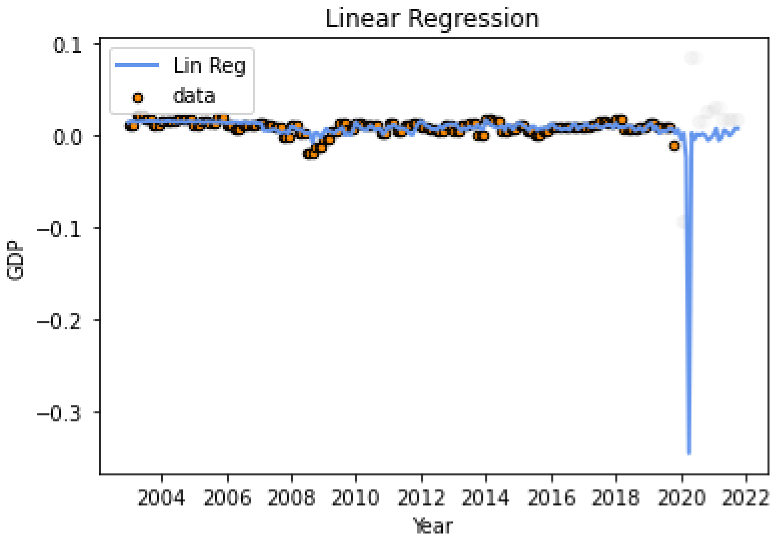
How many features and examples are present? How much data is missing or corrupted? How can we tell?

We have in total 12 features. Among the 12 features, the press language data points starts only from Jan 2007, and is recorded every one and half months, giving 118 data points in total.

The rest of the features are numerical data directly downloaded from the federal reserve. Some of them go back as early as the 1950s, while some of them are available at a daily frequency. However, since a lot of the features were not recorded before Jan 2003, we trimmed the data before January 2003 and kept data points only at a monthly frequency. This gives us 225 data points in each column.

Preliminary analyses of the data, including regressions or other supervised models, describing how you chose which features (and transformations) to use.

We first decided to use all our independent variables of Federal Reserve action predictive power for GDP.



This linear regression model clearly shows us that something the federal reserve did was enable the model to predict the collapse in GDP in 2020 and the quick recovery. This does not tell us whether the federal reserve caused the collapse, the recovery or neither, but at the least there is clearly a relationship between the datasets.

Finally, explain what remains to be done, and how you plan to develop the project over the rest of the semester.

We know the federal reserve cannot affect macroeconomic conditions in real time, but that their policies take time to take effect. For this reason we need to take lagged values of our time series and test correlations with our dependent variables of these various lags to get the best models. Exponential transforms, and variable interactions should also be tested. We will also test boosted decision trees and use the control burn and other coefficient reduction methods to help us understand what variables are most important in prediction as many of our variables are interrelated with each other. Our goal is to find the most important actions taken by the federal reserve and see what outcomes they are creating, not simply to create a model which can predict these macroeconomic conditions. Creating a valid model is the first step, but understanding our models will help us answer our ultimate question.