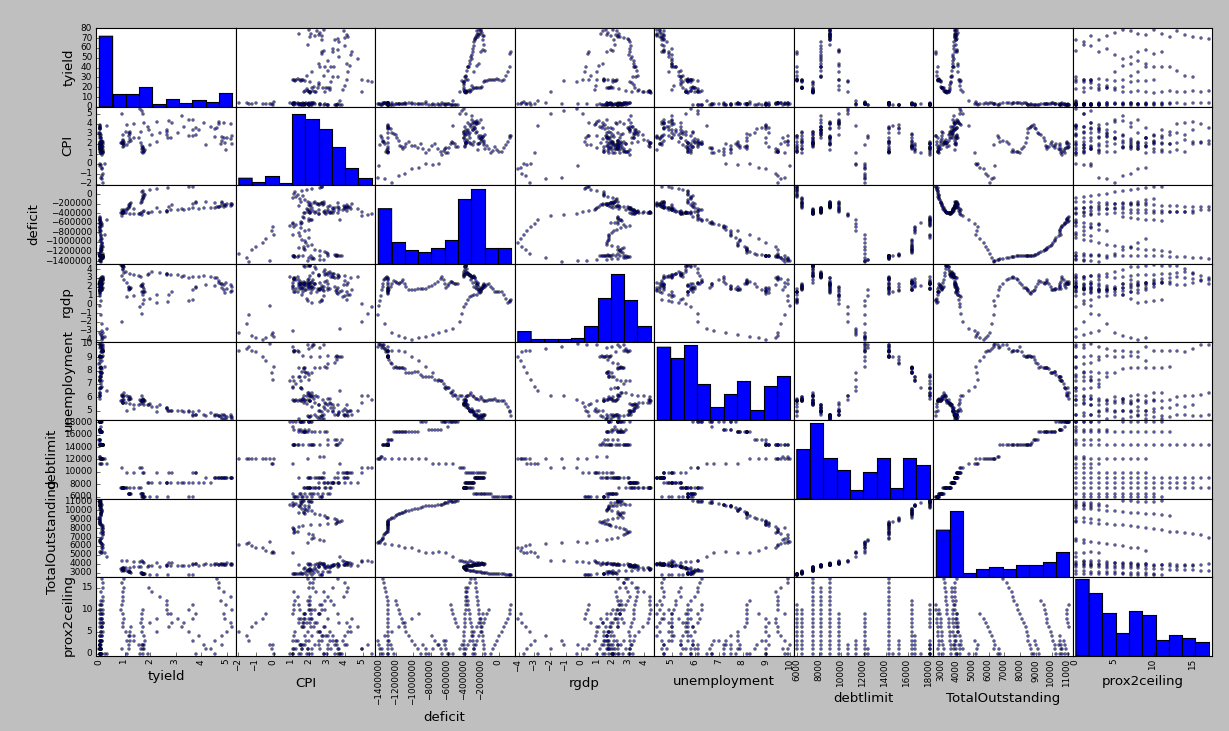
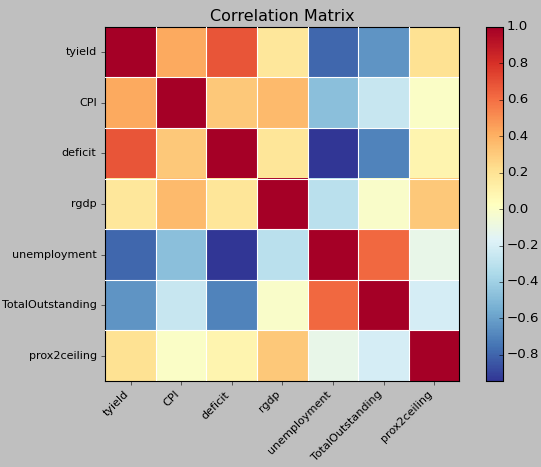
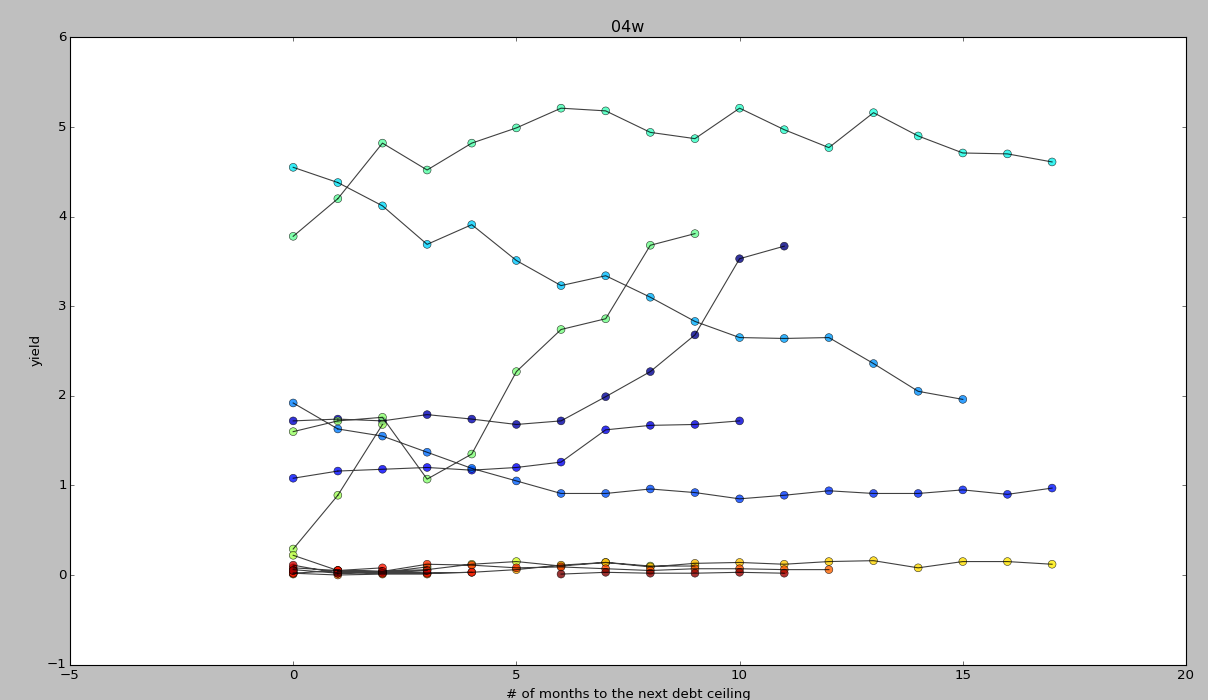
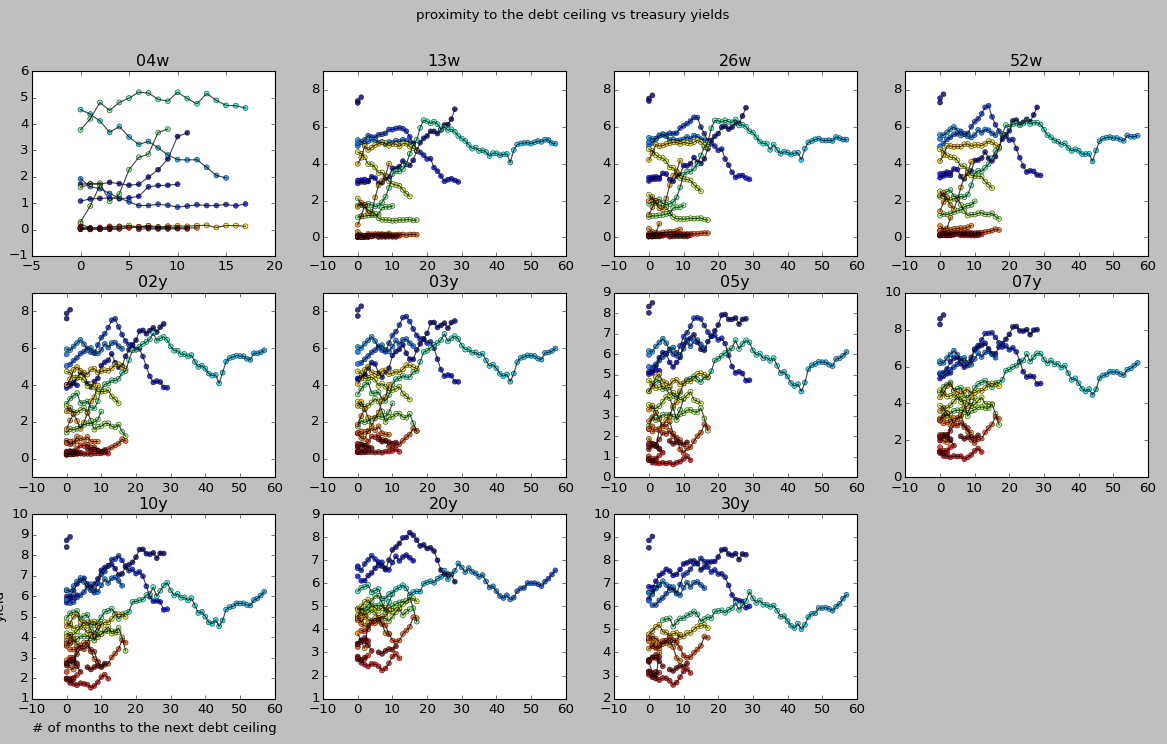
* **Problem statement and hypothesis**
* Analyze/predict the US treasury yields given the macroeconomic factors GDP, CPI, unemployment rates, amount outstanding, proximity to the debt ceiling and deficit levels.
* Background info: US treasuries consist of different types of securities ranging from 4 weeks to 30 years. This project includes bills [4 13 26 52] weeks and coupons [2 3 5 7 10 30] years and focuses on the 4-week bills/10-year bonds. Yields negatively correlates to prices – lower the yield, higher it sells (which means the Treasury is raising more money)
* **Description of your data set and how it was obtained (ts = time series)**
* Treasury yields (target, monthly ts) – constant maturity data, monthly averages <http://research.stlouisfed.org/fred2/series/DGS10>
* Unemployment rate (monthly ts) – represents the number of unemployment as a percentage of the labor force <http://research.stlouisfed.org/fred2/series/UNRATE>
* CPI (monthly ts) – measure of the average monthly change in the price for goods and service paid by urban consumers between any two time periods <http://research.stlouisfed.org/fred2/series/CPIAUCSL>
* Debt limit (irregular ts) – history of debt limit containing date/amounts <http://www.whitehouse.gov/sites/default/files/omb/budget/fy2013/assets/hist07z3.xls>
* Deficit (annual ts) – deficit at the end of the fiscal year <http://research.stlouisfed.org/fred2/series/FYFSD/>
* RGDP (quarterly ts) – inflation adjusted value of the goods and services produced by labor and property located in the United States <http://research.stlouisfed.org/fred2/series/GDPC1>
* Amount outstanding (monthly ts) – amount outstanding for each US treasury security. Data obtained using Federal Reserve (spreadsheet format)
* **Description of any pre-processing steps you took**
* I converted all the raw excel files to csv files, and made sure that files contain only headers and data. For the debt limit file, I had to manually process the information to come up with dates/values
* I created an import function to read in the files
* GDP/CPI are usually measured using year over year growth rates, so I had to convert the raw numbers to percentages
* I created a DataFrame that concatenates all the explanatory variables and indexed the df on a monthly basis
* Because the frequency of each time series was different, I had to use interpolate/bbfill to fill the NaN’s
* I calculated the proximity to the debt ceiling using index arrays (details in code)
* I truncated the df to analyze: 1990-9-1 to 2014-9-1 (data is more accurate/complete during this time)
* What you learned from exploring the data, including visualizations
* A lot of the explanatory variables exhibit multicolinearity – from scatter matrix
  + Total debt outstanding has a positive correlation with deficit
  + Debt limit positively correlates with debt outstanding
  + CPI year-over-year growth rate seems to have no relationship with the yield, unless it is very high or very low (GDP exhibits a similar behavior, but only when it is on the lower side)
  + When unemployment rate is high, the yield is substantially lower



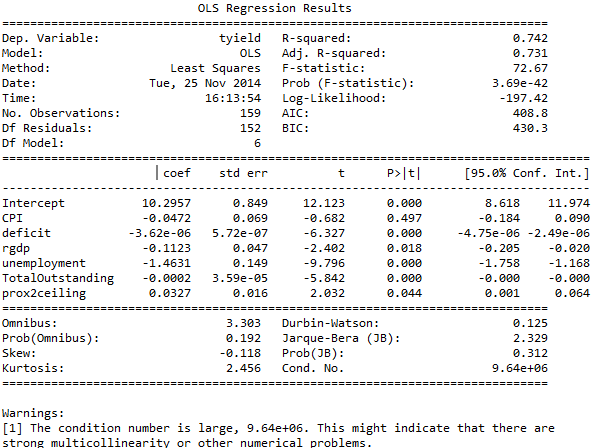


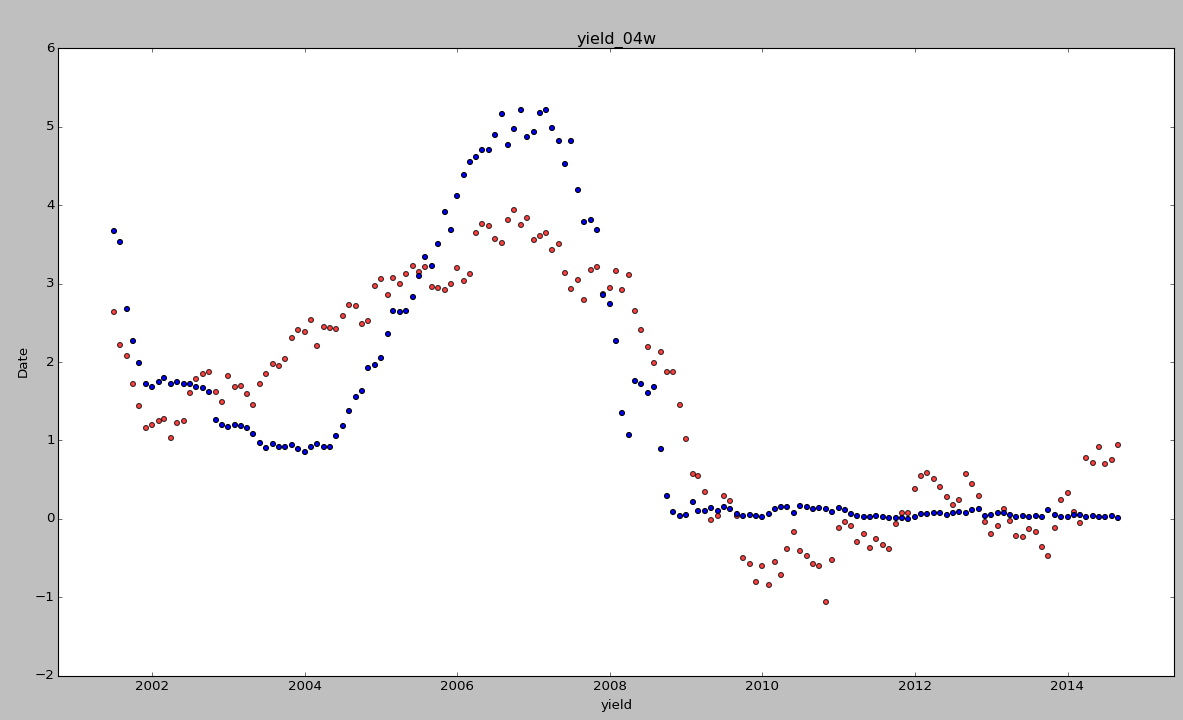
* Out of curiosity, I wanted to see if proximity to the debt ceiling had any effects on the yield (closer to debt ceiling = more likely the US will default = lower price = higher yield). I graphed the 4 week yields and saw no patterns (same applies to the other terms). This may indicate that the debt ceiling is simply a rubber stamp – the markets are not spooked by debt ceiling threats.



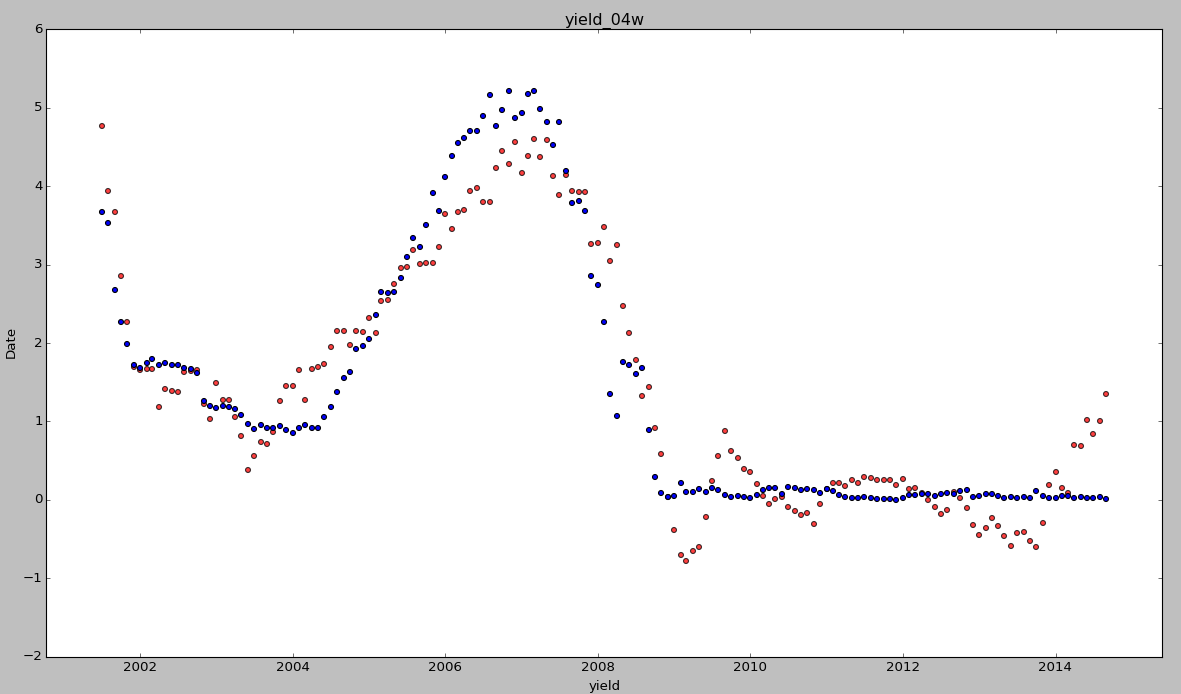


* **How you chose which features to use in your analysis**
* Bond buyers often use these features to evaluate the overall health of the market and the yield/worth of US Treasuries. These factors are also commonly discussed in the Federal Reserve minutes
* **Details of your modeling process, including how you selected your models and validated them**
* I ran a regression model on all the explanatory variables and the 4 week yield as my target variable. I also plotted my predictions (red) vs actual observations (blue)





* The R-squared value 0.742 is not too bad. From the graph above, it seems like the predicted yields are hovering around the true observations. However, it failed to capture micro trends: especially between 2003 and 2008. The 4 week yield also stabilized in late 2008, which the model failed to capture (<- maybe adding a categorical flag indicating whether data is prior to 2008 can improve the model?)
  + The results also indicate that there is strong multicolinearity – I decided to add more interaction terms. Because the p value for CPI is greater than the significance level of 0.05, I’ve decided to throw it out. The prox2ceiling variable is also pretty big, but I decided to keep it since it’s less than 0.05.
* I ran the new model: tyield ~ deficit:unemployment + deficit:TotalOutstanding + deficit + rgdp + unemployment + TotalOutstanding + prox2ceiling
  + The new model has a better R^2 value of 0.893; interaction term deficit:TotalOutstanding, prox2ceiling and TotalOutstanding now are above the significance level.
* I ran the new model with even fewer terms: tyield ~ deficit:unemployment + deficit + rgdp + unemployment
  + The new model still has multicolinearity problems; the R^2 value improved to 0.889, and all of the explanatory variables are now above 0.05. Here’s the plot of predicted (red) vs actual (blue)



* **Your challenges and successes**
* Successes: data concatenation – lining up the explanatory variables in a meaningful way, calculating proximity to the debt ceiling, etc.
* Challenges: unable to fix multicolinearity problems, need to figure out how ‘time’ affects the model (python time series tools?)
* **Possible extensions or business applications of your project**
* Extend yield analysis to longer termed bonds and see if the same factors/combination of factors affect the yields
* Separate dataset into pre-2008 and post-2008 to better fit the curves
* I might use k-means clustering to further examine the explanatory variables
* Categorize the yields (<0%, 0-1%, 1-2%, etc) and use a decision tree (hoping there are algorithms for non-binary trees) to determine what are the most important factors and corresponding cutoff values
* **Conclusions and key learnings**
* Not sure yet…