Statistical Analysis

There are a large number of exploratory data analysis techniques to be drawn from the stock market data and the news data. Here a statistical analysis is focused on drawing conclusions between correlations between different features in the data set and reporting on statistical measurements for individual companies. Before exploring these data it would be of interest to understand how short term returns vary day-to-day. Graphing stock market opening or closing prices over time makes data look as if the short term returns or random and independent. This can be checked by performing a ljung-box test and comparing the p-value to a threshold value. If the p-value is lower than the threshold then the null hypothesis fails and therefore the data is not considered random noise.

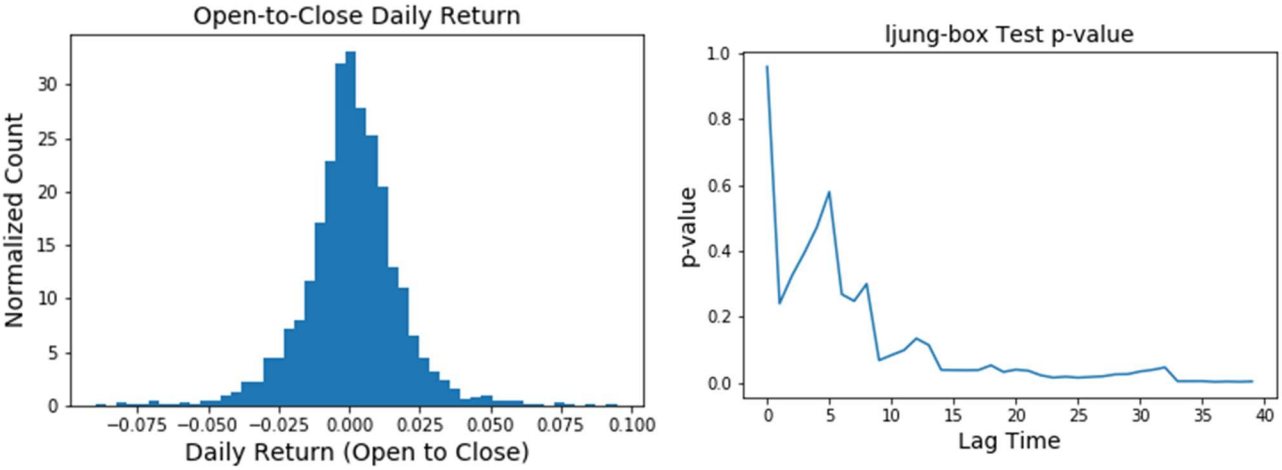
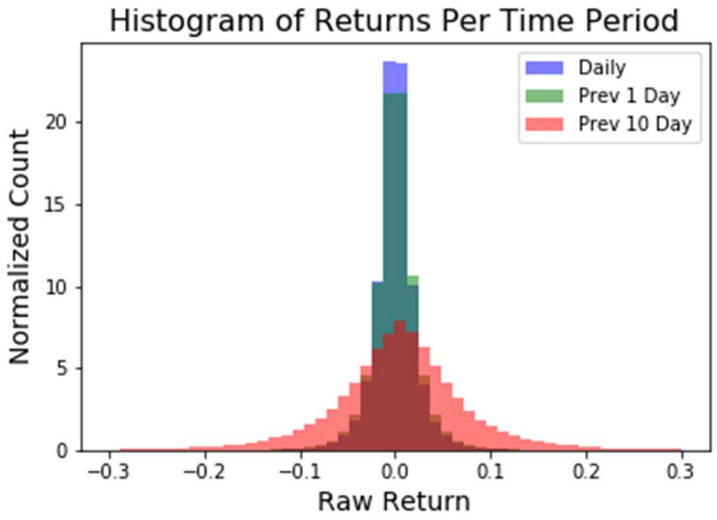


Figure 1. Left graph: Histogram of daily returns for Agilent Technologies Inc. Right graph: Ljung-box test p-values versus lag time.

In the figures above is a summary of the daily returns and the p-values for the ljung-box test performed on daily returns for Agilent Technologies Inc. Because of computational limitations performing the test on all companies would have been impractical hence a single company was chosen. Notice, at short lag times <10 the p-value is significant and therefore the null hypothesis fails to be rejected and therefore daily returns at these lag times can be considered random noise. However, at larger time lags the p-value drops which indicates that the null hypothesis will eventually be rejected for large enough lag times. Hence, at long lag times the daily returns do not show random behavior. This trend is statistically interesting and shows a threshold for a timeframe in which daily returns are no longer random and independent.

Based on the analysis above it suggests that there is a non-random change in the returns over longer lag times. We can plot the returns for different time periods on the same histogram to get a sense as to what the difference is with different return time periods. The histograms below show the differences in returns using daily pricing, the return over the previous one day and the return over the previous 10 days. Notice, as the return time increases the spread of returns increases. Also, in this case, the average return steadily increases from 1.98e-4 daily, 4.93e-4 previous 1 day and 4.40e-3 for the previous 10 days.



When modeling a complex system of features it is important to understand what, if any, correlations exist between features. Here a series of features was chosen and then pair-wise correlations were calculated between each set of features. Several features from the original market and news data sets were chosen along with calculated features day_diff and log_r which are the difference between the closing and opening daily price and the daily log return defined as log10(Close Price/ Open Price). A heat map with these pair-wise correlations is shown for

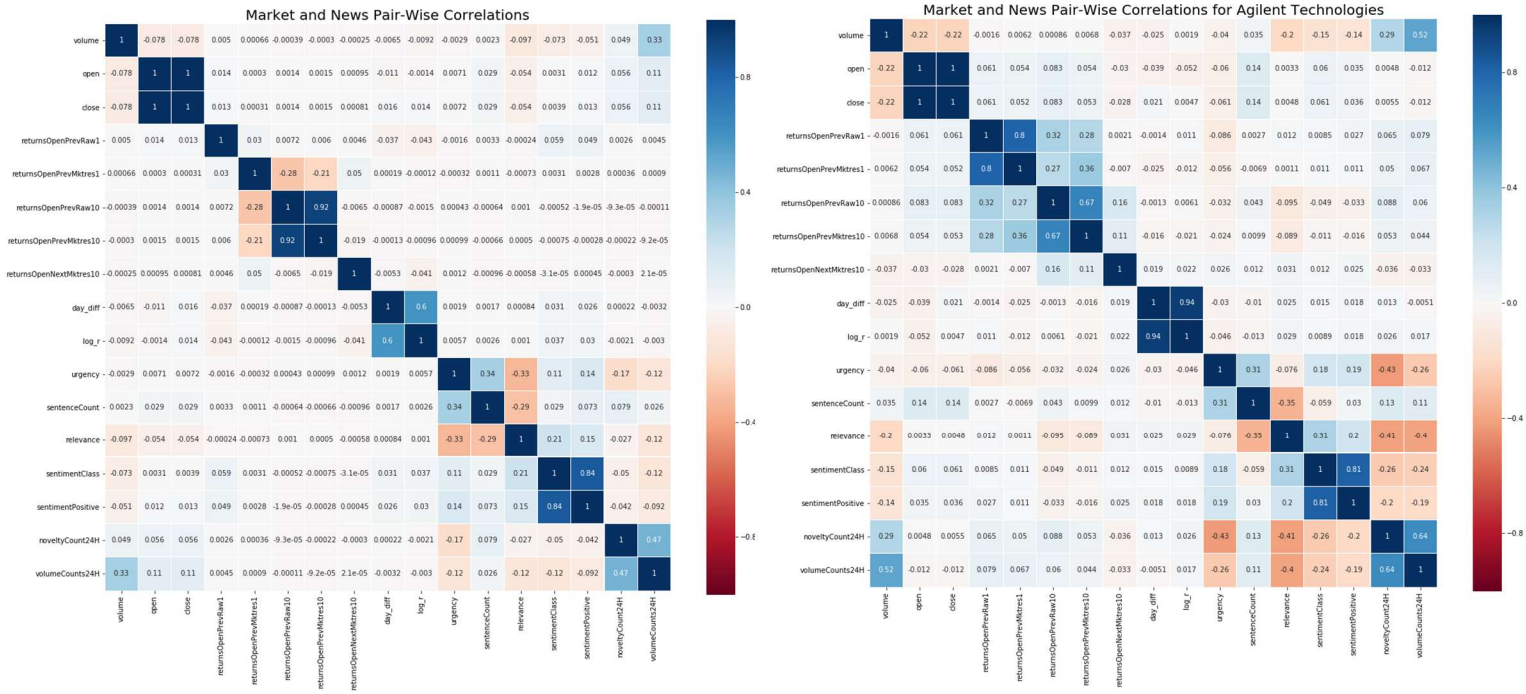


Figure 2. Left graph: Heat map of pair-wise correlations for entire data set. Right graph Heat map of pair-wise correlations for Agilent Technologies Inc.

It is apparent from the heat map that there are not strong (either negatively or positively) correlations between different feature pairs. The features that are closely correlated are logically related in the first place. For instance, returnsOpenPrevRaw1 and returnsOpenPrevMktres1 are highly correlated which makes sense. The returnsOpenPrevMktres1 is simply the raw value except it has be corrected to take into account a standardization term for the market. Hence, the exact values in these heat maps do not in themselves provide deep insight. However, it is interesting that there seems to be stronger correlations for the individual company (Agilent) compared to the correlations with the entire data set. This may suggest that how company features are related is dependent on the type of company and therefore to properly model the system these difference may have to be accounted for.