Assault Prediction From Stock Prices

Shlok Sethia, Shalaj Lawania, Dhivyadarsan GM, Chaitanya Varma K

ABSTRACT

The film '*Minority Report'* was one of the first mainstream ventures to introduce the futuristic idea of 'precrime': the ability to predict crime before it occurs. Our project aims to convert this futuristic idea into reality in order to create safe public spaces for the world. While we have a general idea of 'safe' and 'unsafe' neighborhoods in every city, currently there is a lack of clear insights on the number of anticipated assaults in an area in a given period of time, how assaults are affected by seasonality, weather conditions, unemployment rates and the stock prices of local companies. In particular, in this project, we intend to explore the relationship between market sentiment and the number of assaults, and the types of psychological behaviors that cause this difference.

OBJECTIVE

In this project, we explore the relationship between stock prices and the number of assaults in Boston. Since there is limited research on the relationship between stocks and assaults, we intend to build models that can improve crime prediction and resource allocation.

Based on the existing challenges in public safety, we have identified the following objectives for this project:

- Analyze historical crime data for the city of Boston and identify patterns in assaults
- Research external datasets for potential crime predictors like seasonality, historical data, economic indicators, and stock prices of companies headquartered in Boston.
- Develop Poisson models that can provide insights into heterogeneity in the number of assaults, along with a time-series-based crime prediction model.

Public safety institutions can use our models to predict the number of assaults in the future, and allocate resources for crime prevention and medical emergencies.

LITERATURE REVIEW

Predictive policing is one of the ways through which police departments in the United States have incorporated big data methods into their work in the last two decades. There is a considerable amount of research for predicting future crimes based on a given set of geographical and time-based features.

However, there is a lack of research when it comes to predicting crimes with the help of economic indicators. In this [2] paper, the authors find significantly positive effects of unemployment on property crime rates. Their estimates suggest that a substantial portion of the decline in property crime rates during the 1990s is attributable to the decline in the unemployment rate. The evidence for violent crime is considerably weaker and they have encouraged future work in that area.

There is similar research [3] in which the authors have used a random forest regressor to predict crime and quantify the influence of

urban indicators on homicides. Their approach has up to 97% of accuracy on crime prediction, and the importance of urban indicators is ranked and clustered in groups of equal influence, which are robust under slight changes in the data sample analyzed. Their results determine the rank of importance of urban indicators to predict crime, unveiling that unemployment and illiteracy are the most important variables for describing homicides in Brazilian cities.

There is research [4] that suggests that political uncertainty and crime rates are important determinants of stock market returns in Colombia. The findings indicate that stock market activity in Colombia partly depends on the crime rate.

With the help of the above-mentioned research papers, we were able to narrow down our interest focusing on predicting crime using economic indicators like stock market data of companies that are representative of the economy of the particular location and the unemployment rate of that location. Based on the results, we suggest resource allocation strategies that would help the police department.

DATA SCRAPING

We collected historical data on crime reports in Boston from 2018 to 2021. We grouped the data using `Occurred_On_Date` (the date the assault was recorded) and reduced the `Offense_Description` (type of crime) column from 800+ crime divisions into 15 main categories.

We scraped multiple sources to extract weather information (`Temperature`, `Snow`, `Precipitation`, `Humidity`, `Visibility`, `Cloud_Cover`), unemployment rate (`UNEMPLOY_RATE`), and created seasonality columns (`Year`, `Month`, `Weekend`, `Week`, `Day of Week`, `Day of Month`). In addition to this, we wanted to explore and build models using the `Location` column. Hence we extracted `ZIP CODE` using reverse geocoding and added features like `Total_lights` (Total number of street lights in that area) and demographics-related data like `Population`, `Median Age`, etc. Finally, we identified several companies with a major presence in the city of Boston. We used pandas DataReader to get stock-related information from YahooFinance.

For our study we selected 9 stocks:

John Hancock Investors (`JHI`): Finance
TJ Maxx (`TJX`): Retail
MassMutual Growth Fund (`MPGSX`): Finance
Thermo Fisher (`TMO`): Manufacturing
General Electric (`GE`): Energy
Blackstone Inc (`BX`): Finance
Boston Scientific (`BSX`): Manufacturing
Raytheon Technologies (`RTX`): Security
State Street (`STT`): Finance/Banking

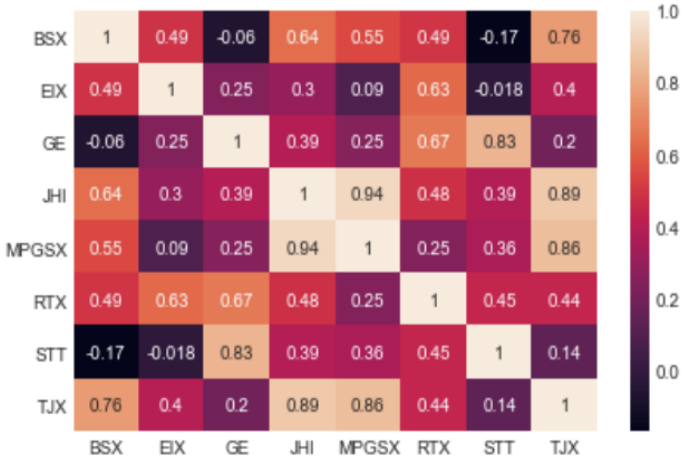
EXPLORATORY DATA ANALYSIS

This portfolio encapsulates trends in major economic sectors and serves as a good indicator of the general market sentiment in the city of Boston. *Appendix Table 1* is a high-level data dictionary of our work in scraping.

Correlation: *Graph 1* shows the correlation chart demonstrating the actual numerical values for the correlation between the stocks' daily return values by comparing the closing prices. We constructed correlation charts of stock prices of companies with a major presence in Boston. This helped us identify a

portfolio of non-multi-collinear and low-risk stocks

Graph 1: Correlation Chart, Stocks

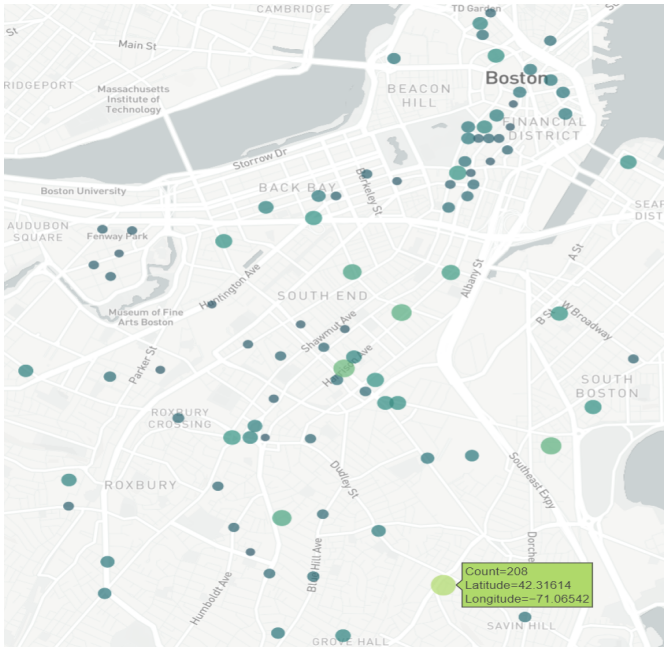


Fuzzy Clustering: Using assault crime data in Boston we identify red hot and cold blue areas for potential targeting and resource allocation. Clustering is performed to segregate data points into a number of clusters based on areas such that data points in the same area are in closer proximity to other data points in the same area than those in other areas. In simple words, the aim is to segregate groups with similar traits and assign them into clusters.

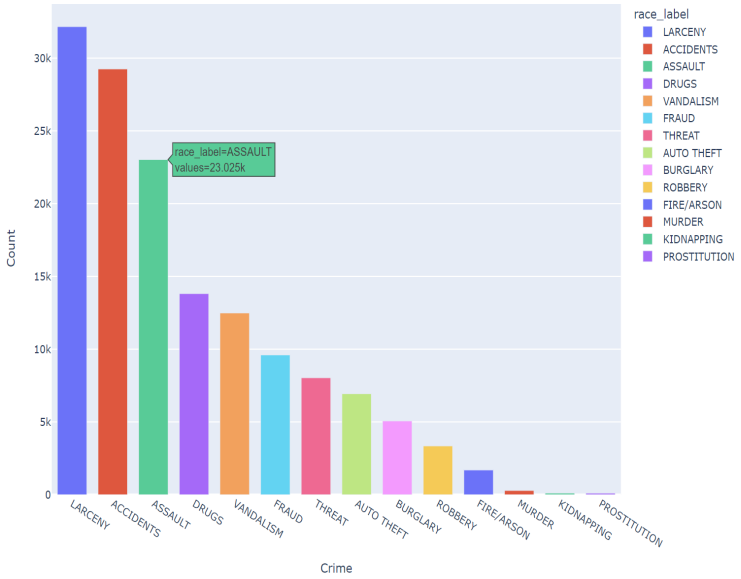
We performed fuzzy clustering using HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) to build an interactive dashboard that helps identify general trends and patterns in assaults over the years. HDBSCAN performs Density-Based Spatial Clustering over varying epsilon values and integrates the result to find a clustering that gives the best stability over epsilon. This allows HDBSCAN to find clusters of varying densities, and be more robust to parameter selection.

In practice, this means that HDBSCAN returns a good clustering straight away with little parameter tuning.

Graph 2: Fuzzy Clustering, Assaults In Boston



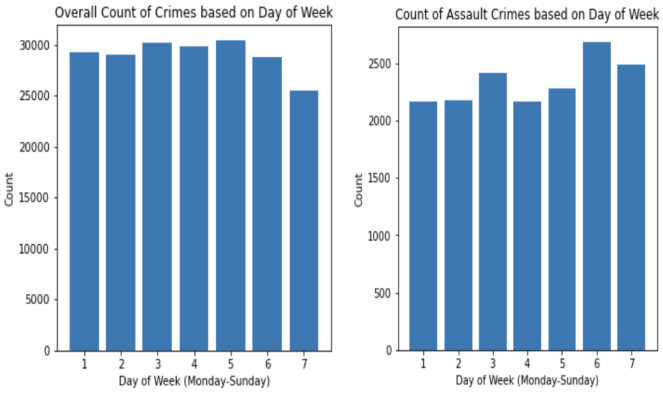
Graph 3: Distribution Of Types Of Crimes



Segregation: *Graph 3* describes the total number of various crimes that took place in Boston over the last three years. For our study, we focused on crimes pertaining to assault as we believed it would have a significant business impact.

Graph 4 depicts the number of assault crimes reported on each day of the week. The graph on the right depicts the total number of all crimes reported on each day of the week. We can see that while assault crimes peak over the weekend, the overall number of crimes reported on the weekend is lower compared to weekdays. This, with several other findings, helped capture the seasonality nature of these crimes. We even observed that in comparison to January, the number of assaults increases in May, June, July, October and November.

Graph 4: Average Assaults By Day



Feature Engineering: As seen in *Appendix Graph 1*, the number of assaults data is skewed with a right tail. We used log transformation to preserve the order of the data, improve readability and reduce the effect of outliers.

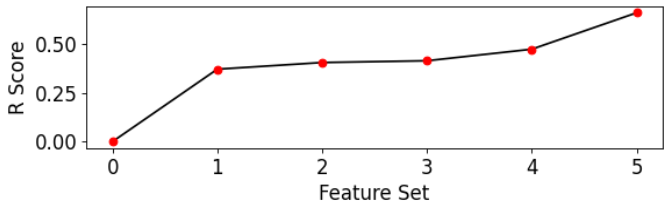
We created new columns with 7-day rolling averages of the stocks titled: `JHI_7`, `MPGSX_7`, `TJX_7`, `TMO_7`, `BCG_7`. Rolling averages reduced the noise from daily fluctuations and emphasized the long-term trends in stock prices (refer to *Appendix Graph 2*). Since our objective is only concerned with the long-term trends in the stock prices (bullish or bearish), daily volatility might make the model overfit.

LINEAR MODELS

Causal Inference: At the first step, we built a linear model to observe the correlation between the predictors. We used backtesting to identify the initial set of features that produce the highest R score. We also implemented an ElasticNet model to identify the ideal number of features to include in our models.

Graph 5 depicts the improvement in the R score as we increased the feature set with the addition of historical data, seasonality, weather, and stock prices. In particular, the jump in R-score from ~0.47 to ~0.65 was due to the addition of the five stocks.

Graph 5: R Score Improvement



Backward Stepwise: We built a backward elimination model to identify the significant features and removed those significant features that do not have a significant effect on the dependent variable. We identified the following variables as significant features:

1. `'cloud'`
2. `'Humidity'`
3. `'DJI'`
4. `'Precipitation'`
5. `'Weekend'`
6. `'Unemployment rate'`
7. `'Government Holiday'`
8. `'BSX'`
9. `'EIX'`
10. `'JHI'`
11. `'MPGSX'`
12. `'STT'`
13. `'TJX'`

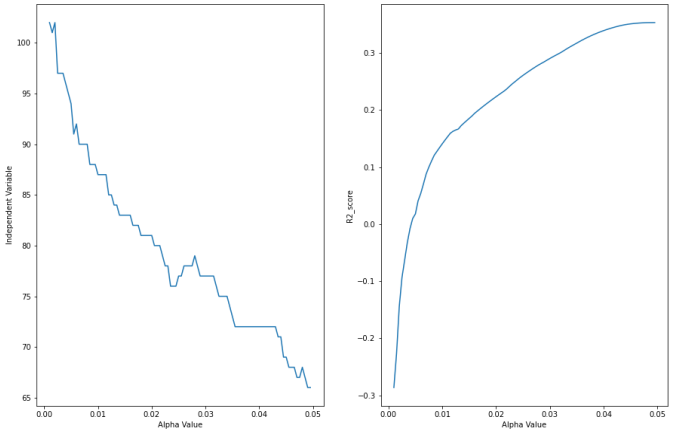
14. `'TJX_7'`
15. `'TMO_7'`
16. `'BSX_7'`
17. `'EIX_7'`
18. `'BCG_7'`
19. `'BCG_7'`
20. `'np.log(count_lag)'`
21. `'t'`
22. `'government holiday'`

and achieved an R2 score of 0.634.

ElasticNet: We also built an elastic net model with L1 = 0.1 after creating dummy variables and standardizing the independent variables. We iterated over multiple values of alpha and identified the alpha value with the maximum number of R2 score and also found the significant independent variables corresponding to the maximum R2 score.

We found 66 significant independent variables. However, the R2 score of 0.354 was not convincing.

Graph 6: ElasticNet Results

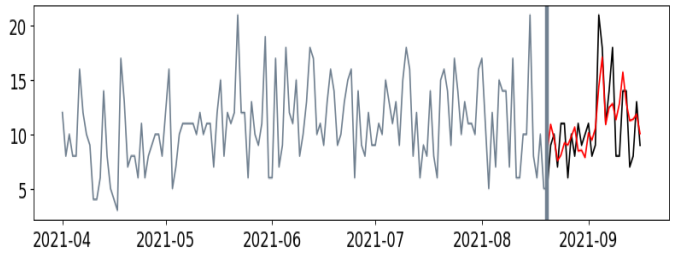


Appendix Graph 3 shows the linear regression results. All stock prices are statistically significant in our model, with positive coefficients for `MPGSX` and `BCG` and negative coefficients for the other three stocks.

Time-Series Prediction: We built a time-series prediction model for predicting the number of assaults three days into the future. All crime predictors were lagged by three days. The training and testing sets were split by date. Appendix Graph 4 shows the difference in correlation when the predictors are lagged. In particular, the weather variables (`Temperature`, `Snow`, `Precipitation`) lost correlation upon lagging. This lines up intuitively, as the number of assaults depends on the weather on a given day, rather than the weather in the past.

Graph 6 shows the comparison between our model predictions and actual data. The predictions capture the major trends really well however there is some loss in R score due to the lagged predictors, as the delayed effects of stock fluctuations on assaults are lower than the direct effect.

Graph 6: Predictions vs. Actual, Time-Series



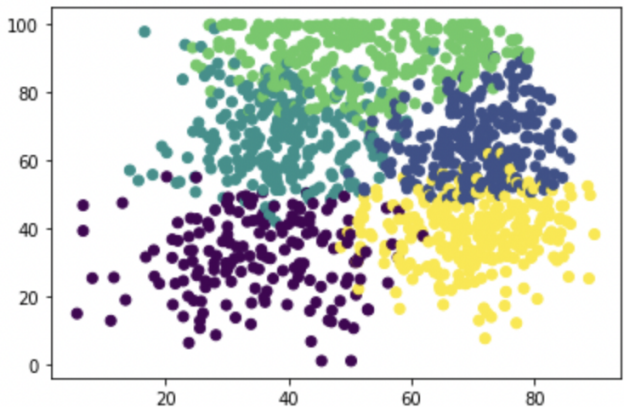
POISSON MODELS

We also built several Poisson models to identify the distribution of data. The baseline Poisson model did not fit the data well, as the data displayed clear heterogeneity.

Cluster Analysis: All attributes available in the data are analyzed to segment the crimes. Temperature(X-axis) and Cloud cover(Y-axis) turned out to be the major factors in forming

clusters and are shown below.

Graph 7: Cluster Analysis



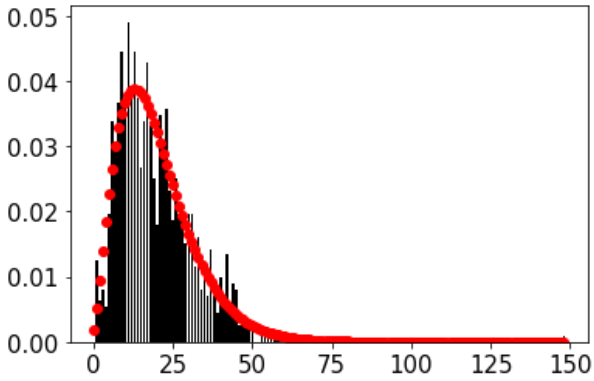
Clusters can be interpreted as crimes during 1) winters with low cloud cover 2) winters with high cloud clover 3) spring/fall 4) summers with low cloud cover and 5) summers with high cloud cover.

Appendix Graph 5 displays the cluster means for all the variables we identified.

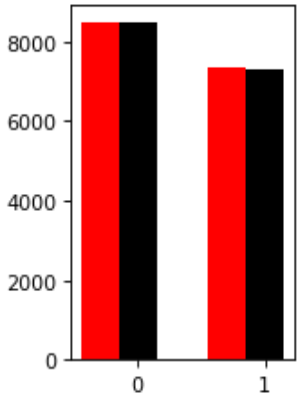
NBD Model: Since the data clearly contained multiple segments, the baseline Poisson model did not give a good fit. As a next step, we built an NBD model to identify the distribution of the lambda value.

Graph 8 shows the NBD distribution. Since the dataset only includes reported crimes, we added truncation to the model to reflect the actual likelihood of observing the data.

Graph 8: Cluster Analysis



Graph 9: AIC, BIC Comparison



We added the stock prices as covariates in our NBD model to further improve the fit. Upon adding the covariates, the joint log-likelihood increased from -4220 to -3652. Graph X depicts the difference in AIC and BIC between the NBD models.

IMPACT

In this section, we discuss the business impact that we can generate from this project.

Firstly, the improvement in R score upon using rolling averages instead of daily prices suggests that assaults are less influenced by daily shocks and disappointments, but rather by the overall trend of the market. In addition to our model, we can also restate our findings based on how stock market returns can impact an individual's psychology.
We also noticed that assaults tend to be higher if it's a sunny day on a weekend when compared to a rainy or snowy day in the winter or a weekday. However, they don't have to increase security in every part of the city impulsively. Our fuzzy clustering model using HDBSCAN will assist them to allocate resources based on the red hot and cold blue areas in the city. This can help allocate police

department resources efficiently depending on seasonality and crime heatmaps.

An interesting and counterintuitive observation was that increase in MPGSX stock in particular was linked with an increase in crime.

The positive relationship between market returns and crime may seem odd if one assumes that an increase in market activity is associated with a decrease in the number of poor people. However, most people do not invest in the stock market. For example, Malloy et al. (2009) [5] estimate that only 23% of households hold stocks (including retirement plans), so the majority of households do not directly benefit from market increases.

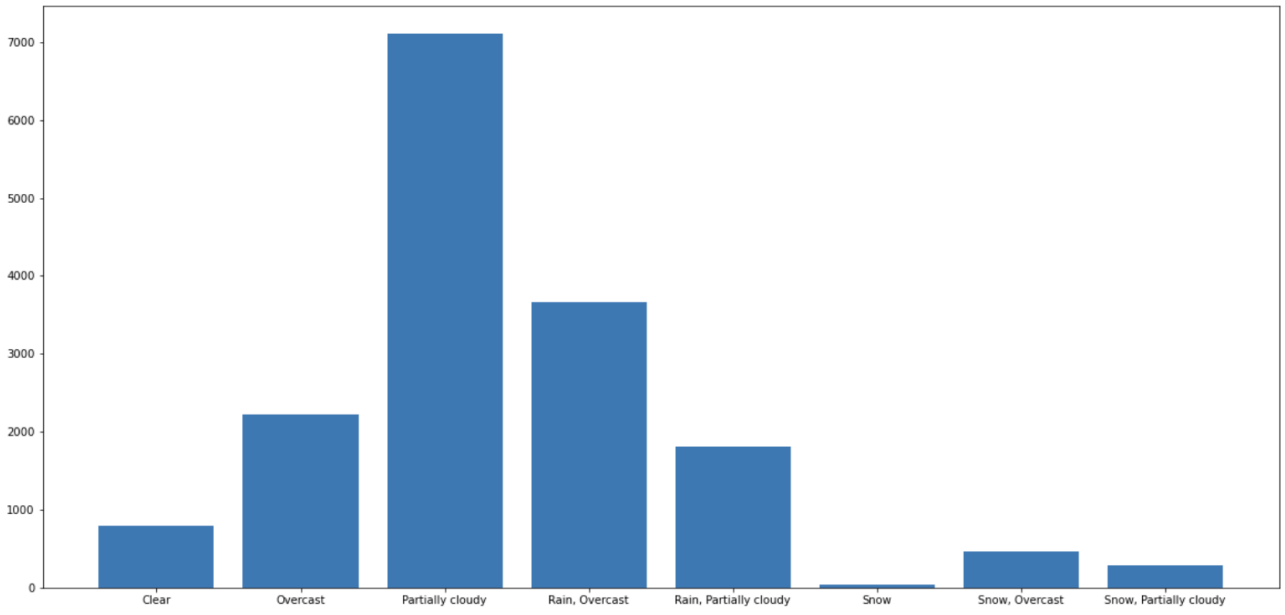
It is also likely that those more prone to commit crime (assaults) have no funds invested in the stock market and are more likely to be angered when the stock market goes up and they observe others benefiting, while their own relative wealth declines. This relative status effect is consistent with models such as Abel (1990) who posit that individuals care about their own consumption relative to others.

With this particular finding, we could suggest changes in monetary policy which is the most powerful weapon in the government's arsenal. Changes in regulations, subsidies, and taxes can have an immediate, and long-lasting impact not only on the companies but on the safety of society as a whole.

This report also adds to the literature on crime and the economy. This study differs in that individuals also appear to criminally react to future economic conditions that daily stock market returns may signal, rather than just to current economic conditions.

For example, crime has been linked to GDP consumer sentiment, growth, wages, unemployment, and income inequality.

Graph 10: Count of Assault Crimes based on different weather conditions



LIMITATION

We have seen from various sections of the report how economic conditions and stability of a location have a significant impact on crime counts. While we have tried to incorporate these economic condition features into our model with the help of features like the Unemployment rate and stocks with a major presence in these localized areas, we believe a lot more factors indicating the overall economy of these locations can be included.

Currently, the models we have implemented are altered to fit the crimes of Boston. A more robust and generalized model needs to be built to be suitable for all cities and towns.

FUTURE WORK

Gather city specific data to build a more robust model with a higher R2 score and perform cluster analysis on a more granular scale. Our current model only predicts crime 3 days into the future with a good accuracy, gathering more historic data and building more powerful models could help easily predict crimes further into the future.

Currently we only predict crimes which are of the assault nature, but in the future we'd like our model to predict any types of crimes that occur.

REFERENCES

[1] Engelberg, Joseph, and Christopher A. Parsons, 2016, Worrying about the stock market: Evidence from hospital admissions, Journal of Finance 71, 1227–1250.

[2] Raphael, Steven, and Rudolf Winter-Ebmer. "Identifying the Effect of Unemployment on Crime." *The Journal of Law & Economics*, vol. 44, no. 1, [The University of Chicago Press, The Booth School of Business, University of Chicago, The University of Chicago Law School], 2001, pp. 259–83, https://doi.org/10.1086/320275.

[3] Alves, Luiz & Valentin Ribeiro, Haroldo & Rodrigues, Francisco. (2017). Crime prediction through urban metrics and statistical learning.

[4] Franco-Laverde, Juan & Varua, Maria & Ozanne, Arlene. (2009). Understanding Crime, Political Uncertainty and Stock Market Returns. World Economics. 10. 109-116.

[5] Malloy, Christopher J., Tobias J. Moskowitz, and Annette Vissing-Jrgensen, 2009, Long-run stock-holder consumption risk and asset returns, Journal of Finance 64, 2427{2479.

APPENDIX

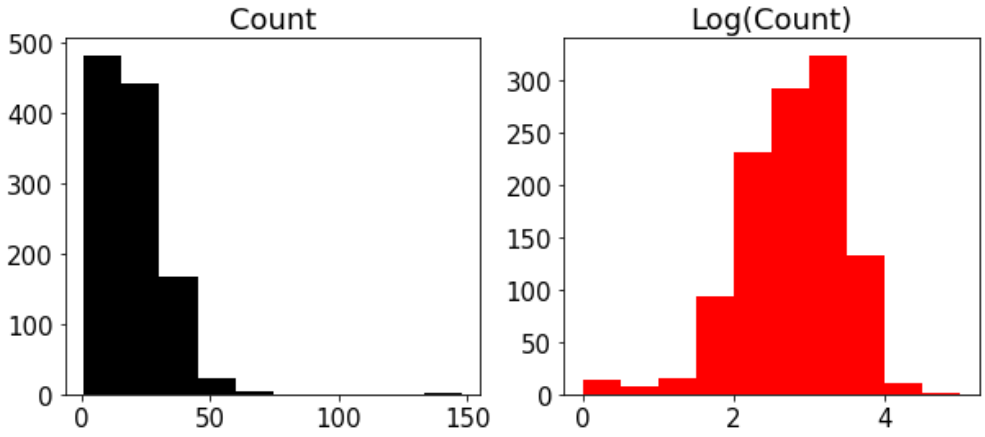
DATA

Appendix Table 1: Data Dictionary

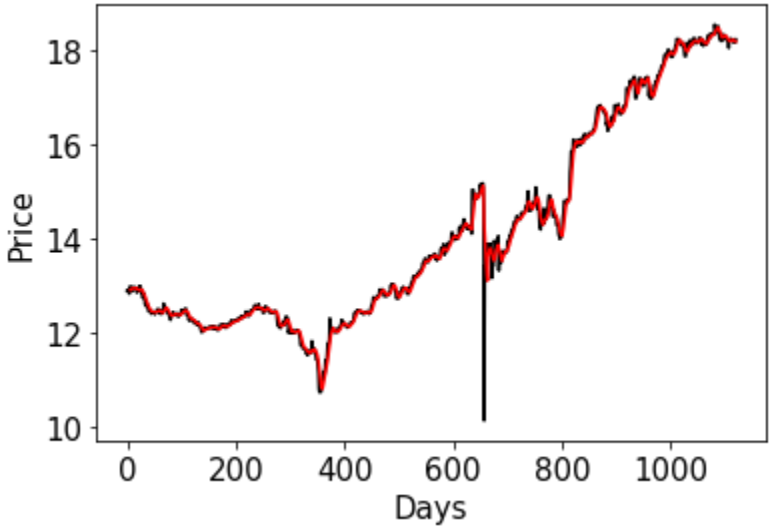
Source	Features Extracted
Visual Crossing	• Temperature • Precipitation • Snow • Snow Depth • Wind speed • Visibility • Cloud Cover • Relative humidity • Conditions.
Wall Street Journal	• Dow Jones
Boston Government	• Area Size of District • Number of street lights in each zip code
	• Zip Code Data
YCharts	• Unemployment Rate
Zip Atlas	• Population • Median Income • Gender • Total housing units • Median Age • Race
Office Holidays	• Federal • Government • Non-Public

	• Regional

Appendix Graph 1: Assault Data Distribution



Appendix Graph 2: Rolling Average Comparison (JHI Stock)



Appendix Graph 3: Causal Inference, OLS Summary

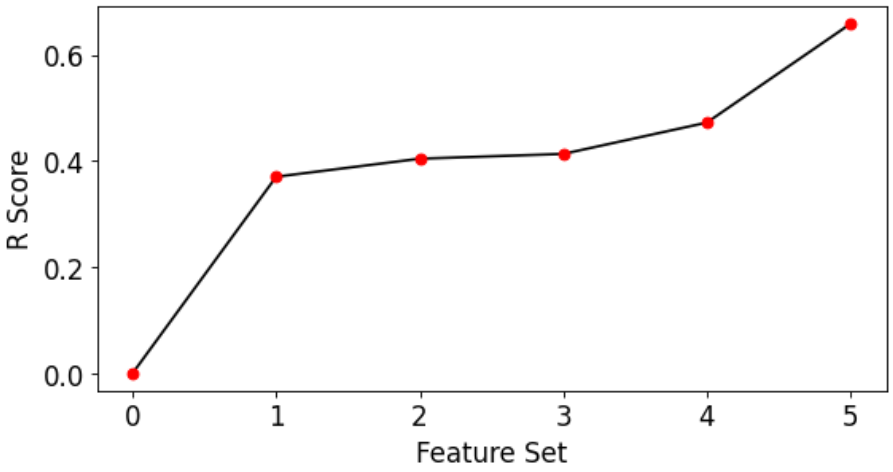
OLS Regression Results

Dep. Variable:	np.log(count)	R-squared:	0.645
Model:	OLS	Adj. R-squared:	0.638
Method:	Least Squares	F-statistic:	94.49
Date:	Wed, 29 Dec 2021	Prob (F-statistic):	3.22e-228
Time:	06:18:19	Log-Likelihood:	-600.95
No. Observations:	1115	AIC:	1246.
Df Residuals:	1093	BIC:	1356.
Df Model:	21		
Covariance Type:	nonrobust		

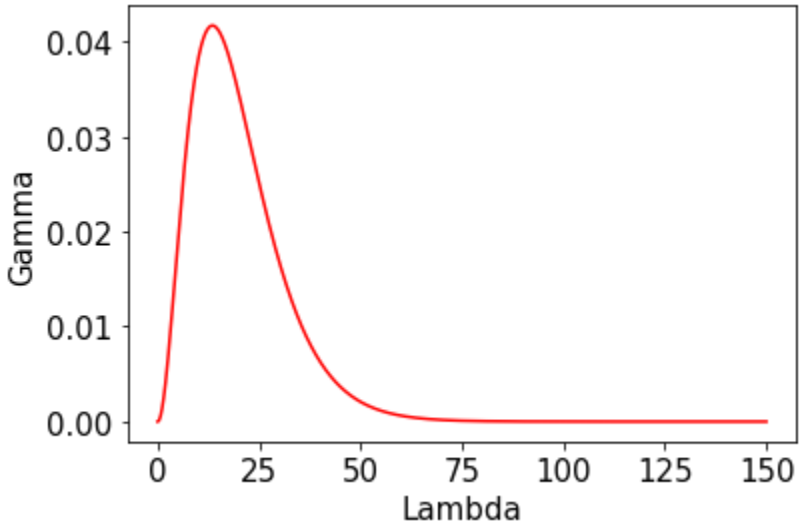
	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	3.4523	0.283	12.216	0.000	2.898	4.007
C(MONTH)[T.2]	0.0879	0.062	1.419	0.156	-0.034	0.209
C(MONTH)[T.3]	0.0363	0.061	0.591	0.555	-0.084	0.157
C(MONTH)[T.4]	-0.0036	0.066	-0.055	0.957	-0.133	0.125
C(MONTH)[T.5]	0.1099	0.076	1.444	0.149	-0.039	0.259
C(MONTH)[T.6]	0.1856	0.087	2.139	0.033	0.015	0.356
C(MONTH)[T.7]	0.1174	0.092	1.275	0.203	-0.063	0.298
C(MONTH)[T.8]	-0.0368	0.091	-0.404	0.687	-0.216	0.142
C(MONTH)[T.9]	0.2109	0.082	2.557	0.011	0.049	0.373
C(MONTH)[T.10]	0.2760	0.078	3.545	0.000	0.123	0.429
C(MONTH)[T.11]	0.1870	0.083	2.250	0.025	0.024	0.350
C(MONTH)[T.12]	-0.0164	0.081	-0.204	0.839	-0.175	0.142
C(WEEKEND)[T.1]	0.1742	0.028	6.256	0.000	0.120	0.229
np.log(count_lag)	0.2049	0.030	6.838	0.000	0.146	0.264
Temperature	0.0041	0.002	2.406	0.016	0.001	0.007
Precipitation	-0.0902	0.040	-2.229	0.026	-0.170	-0.011
I(UNEMPLOY_RATE)	-0.0825	0.010	-8.591	0.000	-0.101	-0.064
JHI_7	-0.0485	0.033	-1.483	0.138	-0.113	0.016
MPGSX_7	0.1590	0.034	4.685	0.000	0.092	0.226
TJX_7	-0.0184	0.003	-5.511	0.000	-0.025	-0.012
TMO_7	-0.0037	0.001	-6.022	0.000	-0.005	-0.003
Snow	-0.0555	0.019	-2.892	0.004	-0.093	-0.018

Omnibus:	328.875	Durbin-Watson:	1.407
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1405.248
Skew:	-1.334	Prob(JB):	7.15e-306
Kurtosis:	7.810	Cond. No.	8.46e+03

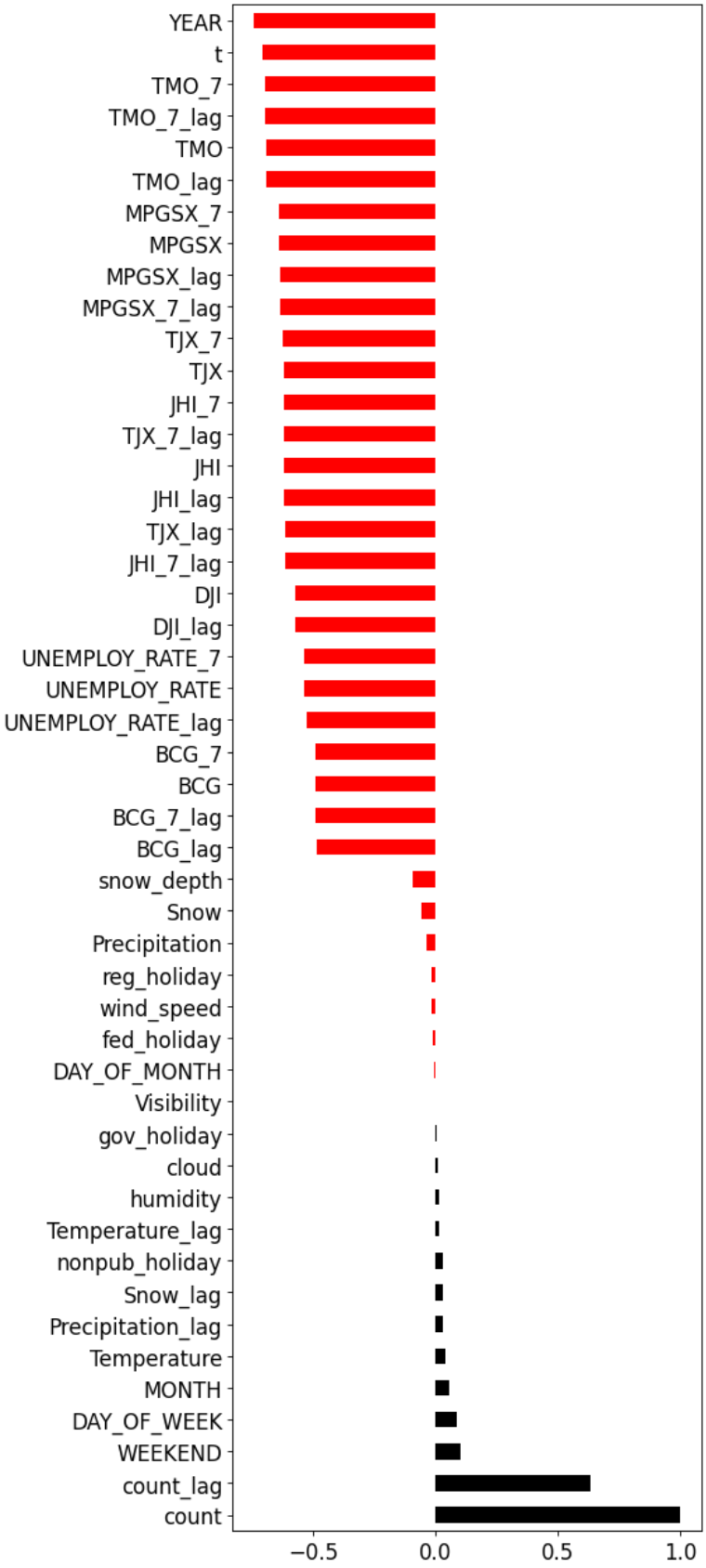
Appendix Graph 4: R-Score Improvement



Appendix Graph 5: Lambda Distribution, NBD Model



Appendix Graph 6: Correlation Comparison, Lagged vs. Non-Lagged

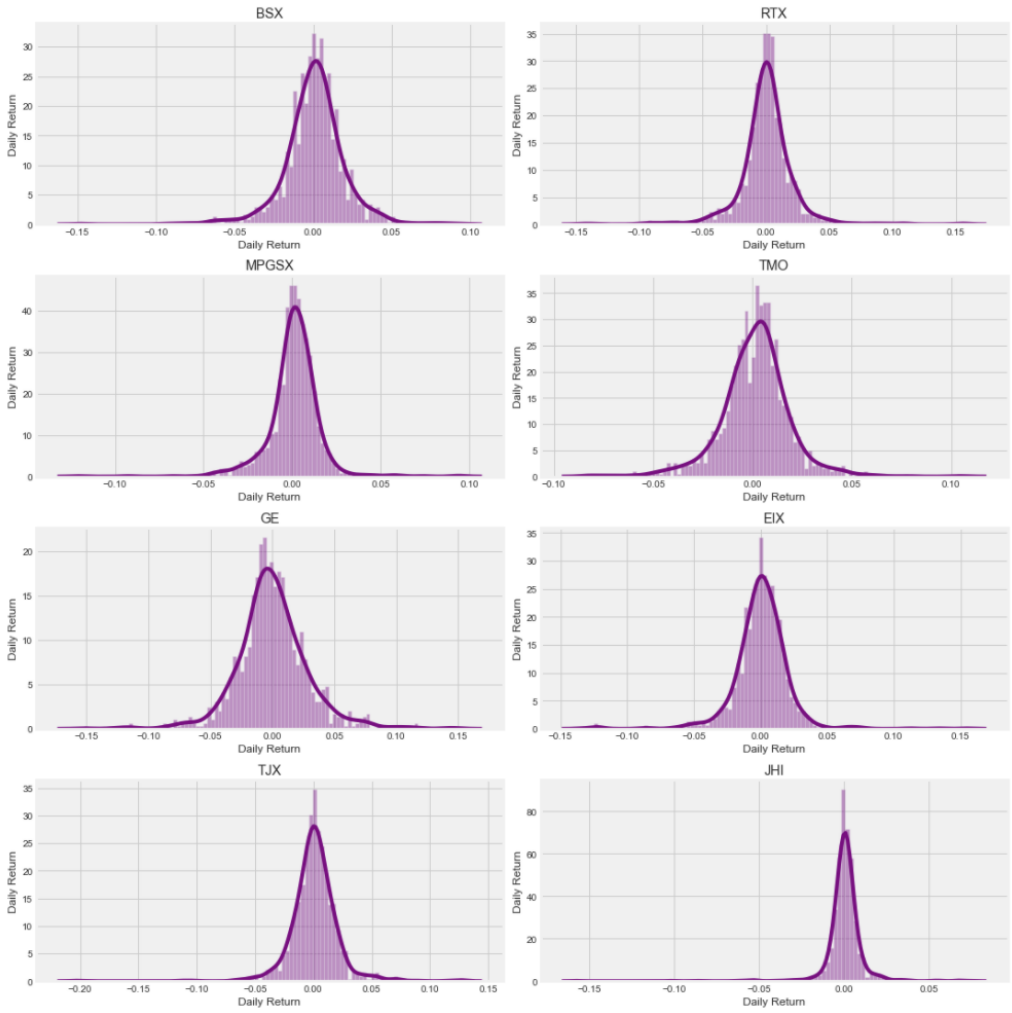


Appendix Graph 7: Cluster Analysis

clusters	Temperature	Precipitation	Snow	Snow_Depth	Wind_Speed	Visibility	Cloud_Cover	Relative_Humidity	TotalCrimes
0	71.932000	0.011200	0.000000	0.000000	16.316000	9.560000	45.308000	61.015200	361.760000
1	40.840000	0.241200	0.593800	1.615600	20.658000	7.512000	91.912000	80.440000	320.960000
2	29.059091	0.000000	0.040909	2.815909	19.609091	9.881818	24.709091	44.300909	297.772727
3	37.548077	0.076731	0.537500	2.264231	19.703846	9.336538	66.148077	62.129038	315.980769
4	69.394643	0.210000	0.000000	0.019464	16.898214	8.667857	80.683929	75.183929	364.142857

Appendix Graph 8: Stock trend for the various stocks considered

Appendix Graph 9: Distribution of stocks based on their daily return

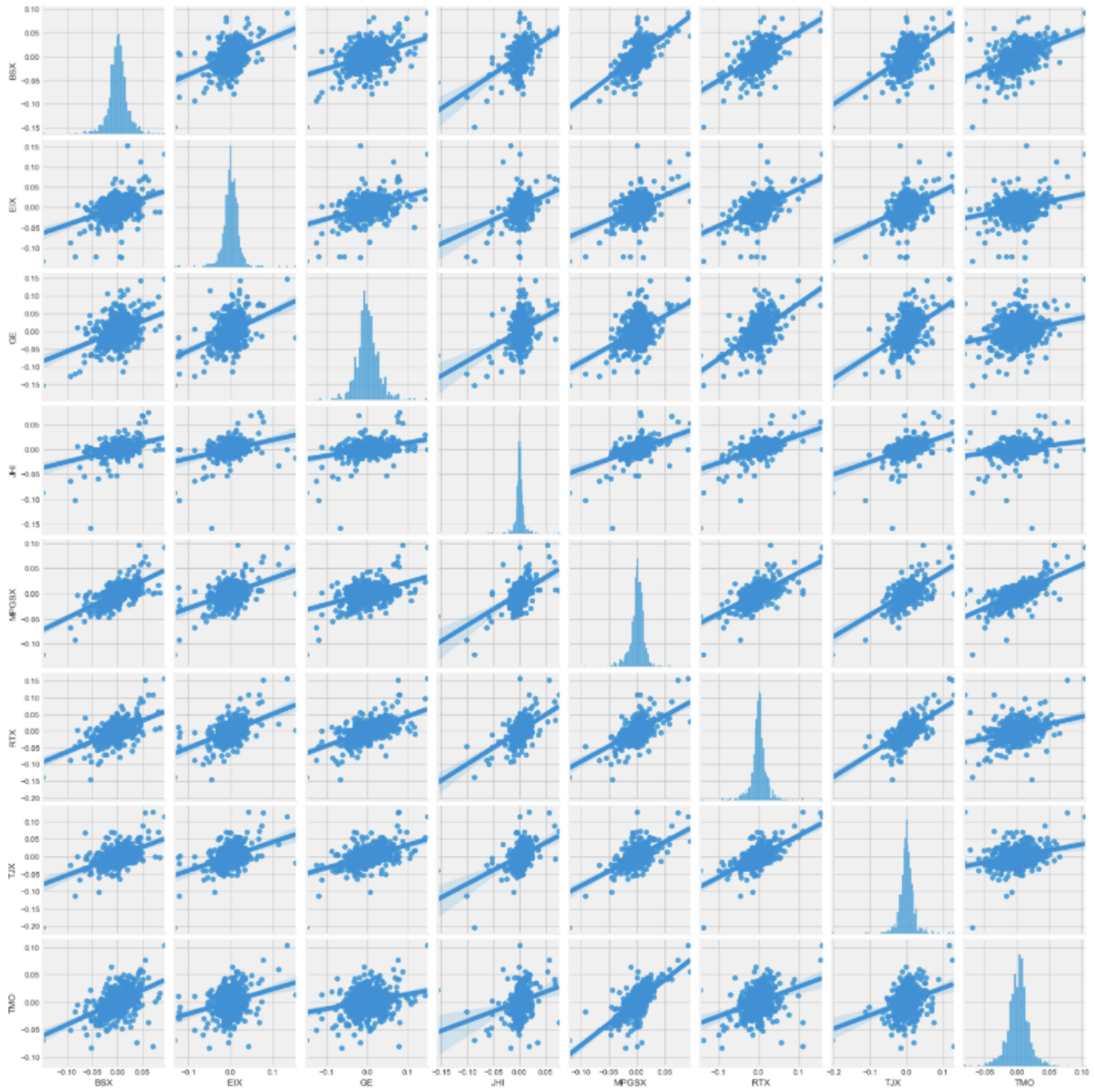


Appendix Graph 10: Percentage growth of stock since 2018

BSX	EIX	GE	JHI	MPGSX	RTX	TJX	TMO
63.014232	19.251488	-24.296771	42.496816	113.717009	18.699851	100.660347	233.605361

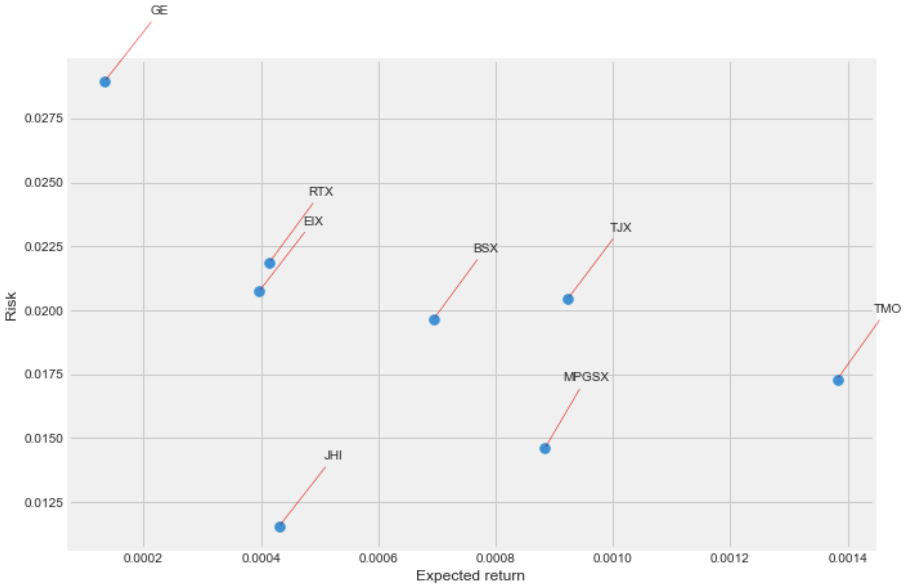
For example, if we see JHI it showed a growth of 42% in its stock price between this period.

Appendix Graph 11: Distribution and correlation between stocks



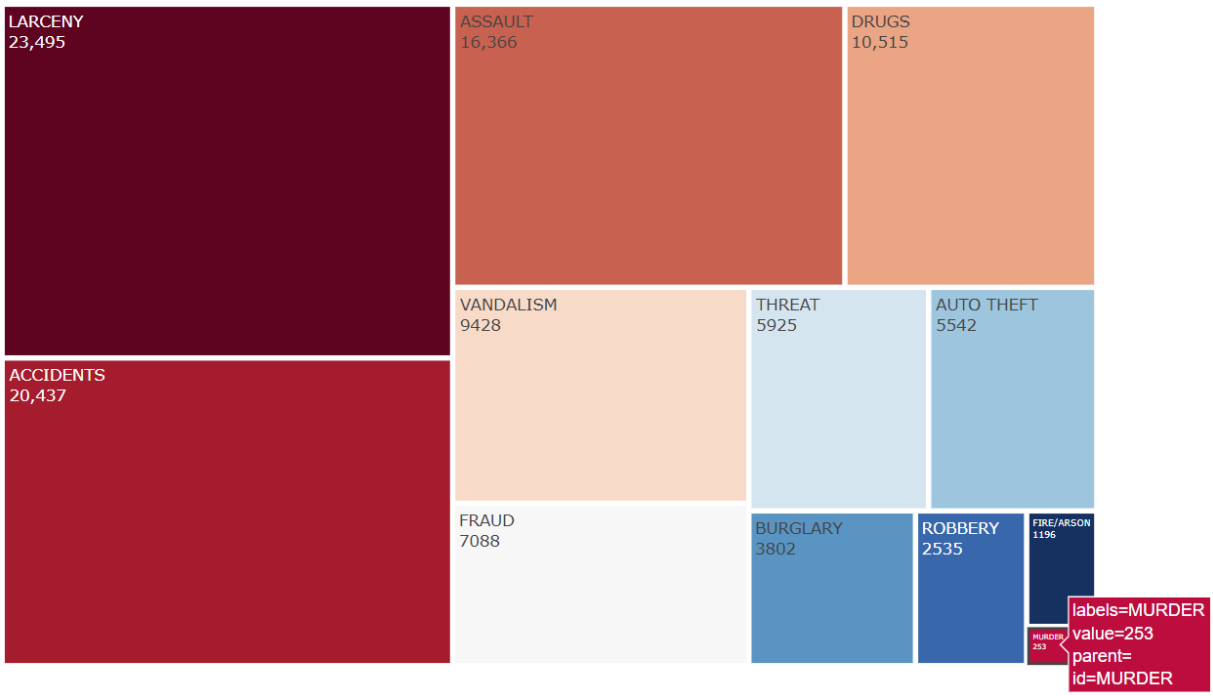
Describes a pair plot, mapping each variable in a dataset onto a column and row in a grid of multiple axes. Allowing us to see both distributions of single variables and relationships between two variables. Pair plots are a great method to identify trends for follow-up analysis.

Appendix Graph 12: Daily returns of each stock

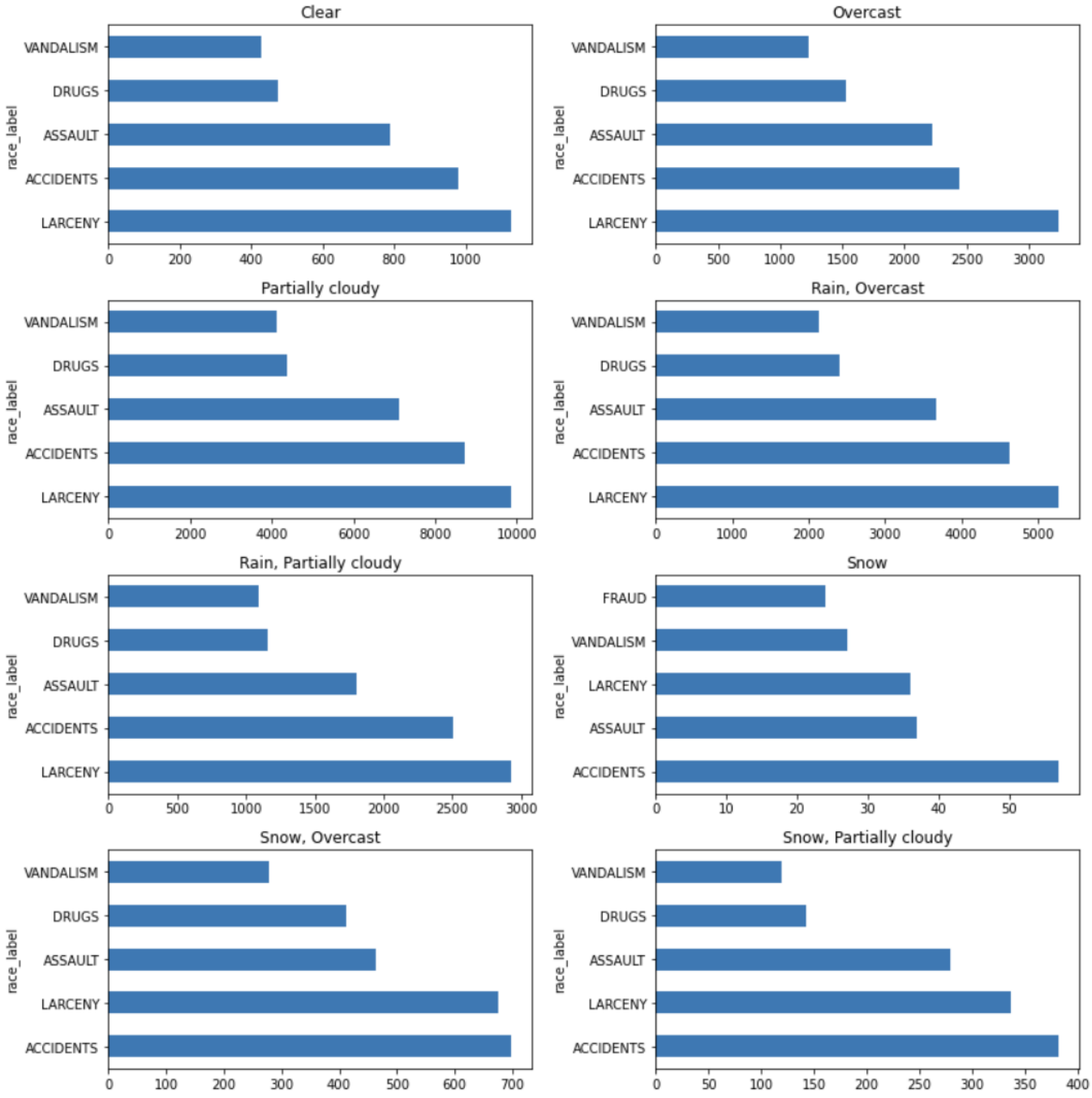


Gives us the daily return of each of the stocks we inspected plotted against the risk. We look at a few ways of analyzing the risk of a stock, based on its previous performance history, to get an idea of the volatility of the stock. We use standard deviation to estimate the risk

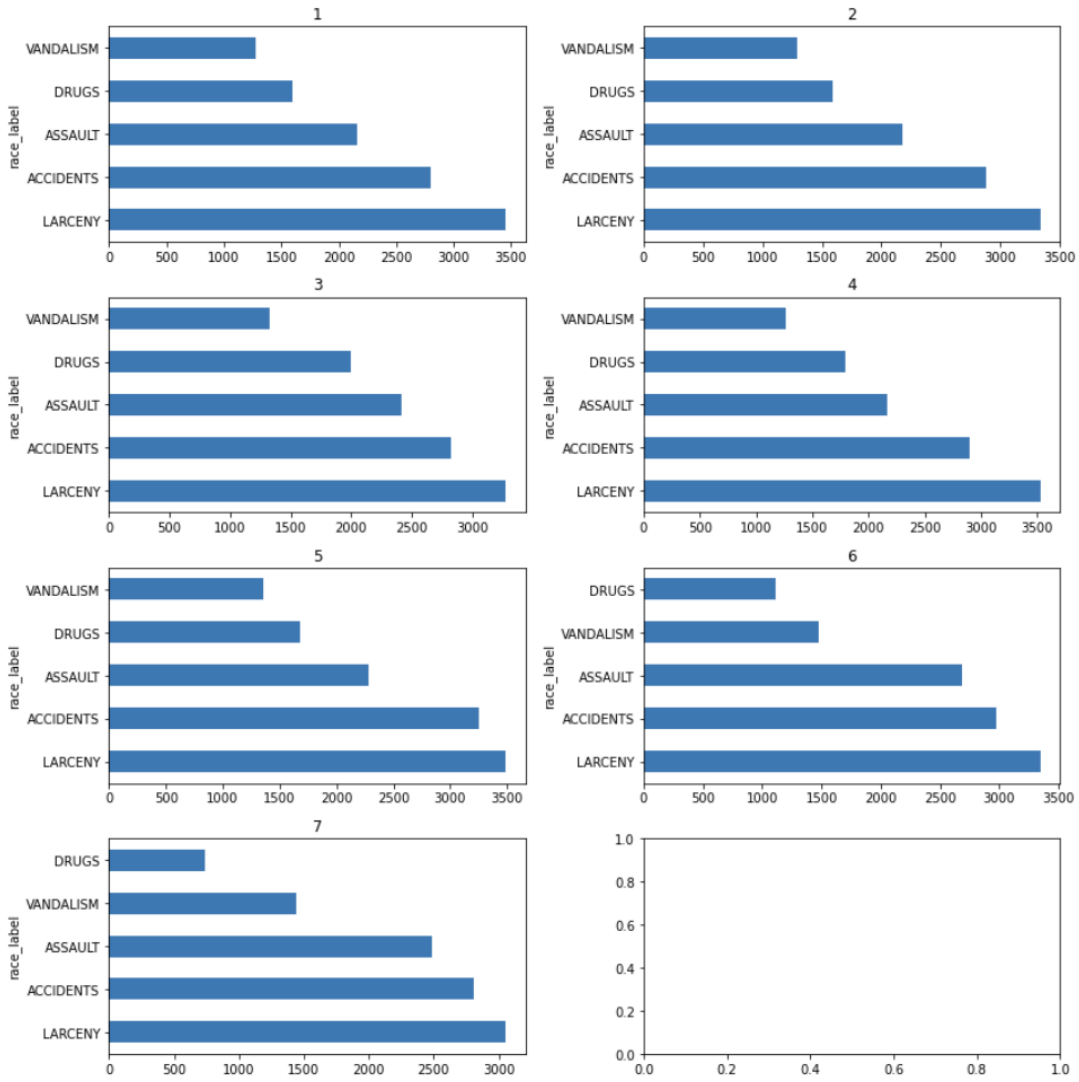
Appendix Graph 13: Treemap indicating the number of crimes based on category



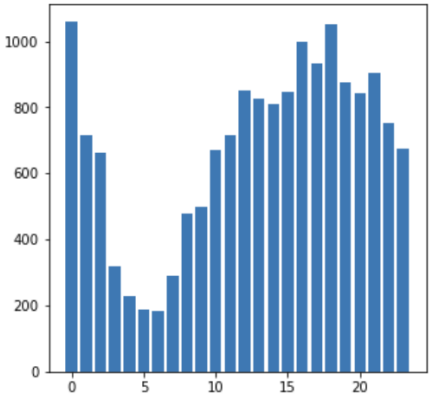
Appendix Graph 14: Top 5 categories of crimes that occur in different weather conditions



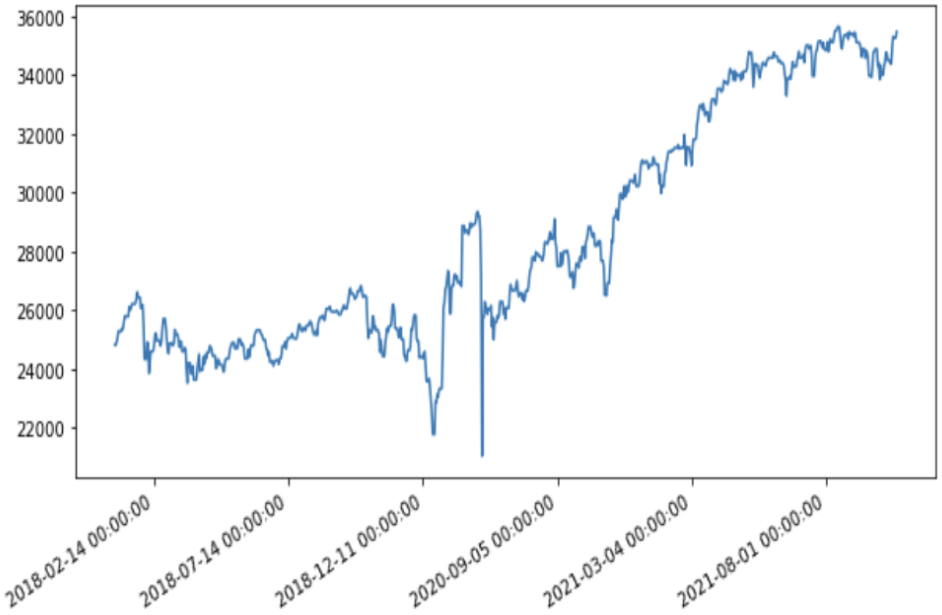
Appendix Graph 15: Top 5 categories of crimes that occur on different days of the week (Monday-Sunday)



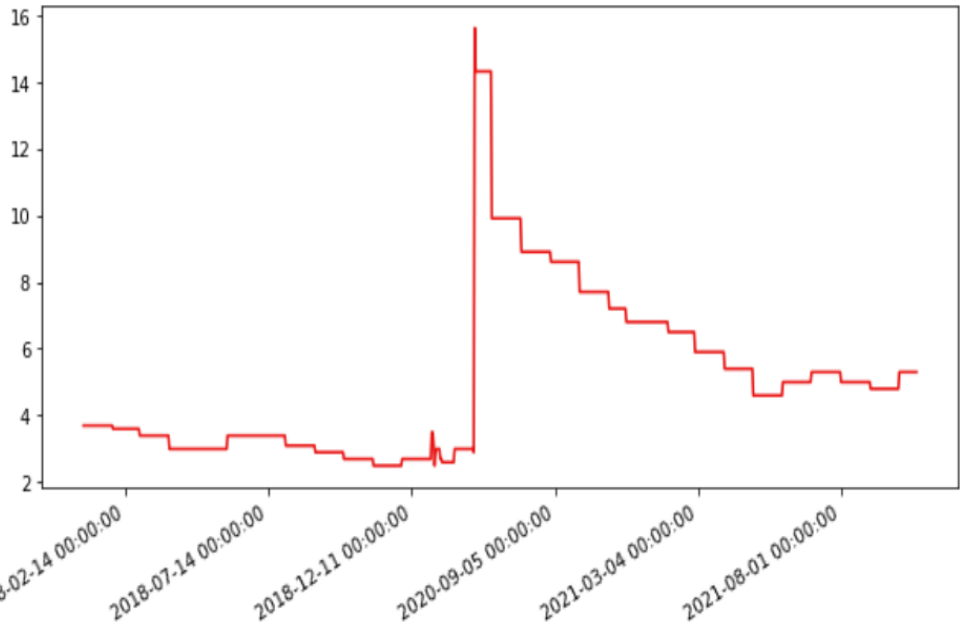
Appendix Graph 16: Count of Assault Crimes based on Hour of the day



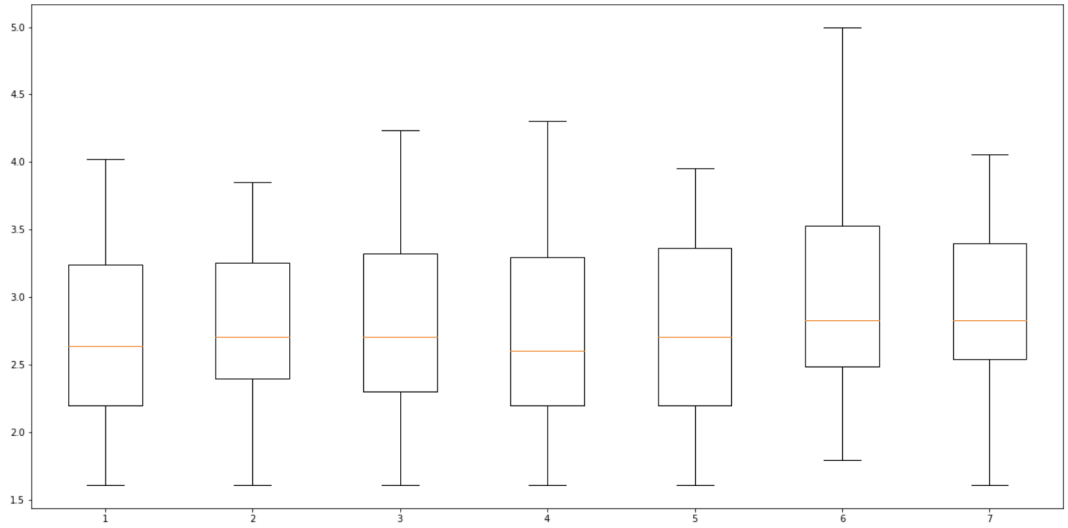
Appendix Graph 17: Dow Jones Price over time



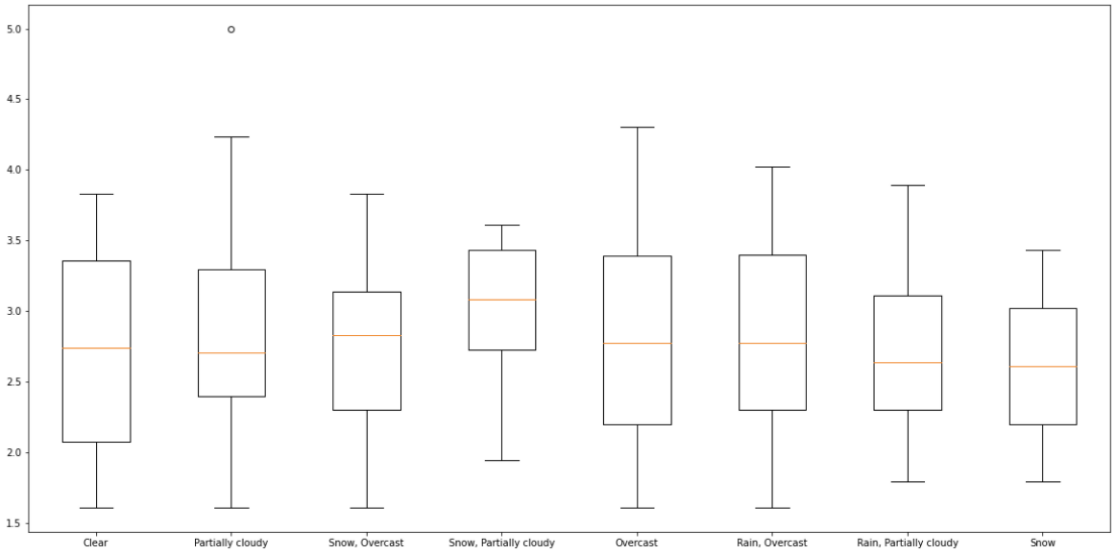
Appendix Graph 18: Unemployment Rate over time



Appendix Graph 19: Box plot to show Count of Assault Crimes based on Day of Week (Monday-Sunday)



Appendix Graph 20: Box plot to show Count of Assault Crimes based on different Weather Conditions



Appendix Graph 21: Top 5 categories of crimes that occur in different Districts