Analysis of Crime

Analysis of Crime Database for the city of San Francisco with various Machine Learning Algorithms.

Sonam Malhotra & Sooraj Shetty

Abstract

Crime Analysis is an extensive concept which is applied by the policing community. Over the last few years' crime analysis has become a general term that includes intelligence analysis, criminal investigative analysis, tactical crime analysis, strategic crime analysis, operation analysis and administrative crime analysis. This helps us to better understand the occurrence of crime and what steps the police could take to avoid incidence of such crime in the near foreseeable future. For our project, we are considering the crime occurrence data for the city of San Francisco, California. As one of the iconic cities in the US, this would be a landmark for in-depth analysis of crime. Our project identifies the various methods, describes what is required to use them, and assesses how accurate they are in predicting future crime incidence. Factors such as data requirements and applicability for law enforcement use will also be explored, and our project will close with recommendations for further research and a discussion of what the future might hold for crime forecasting.

The current project intends to use machine learning algorithms to predict the incidence of crime for the city of San Francisco based on the data gathered regarding various crime reports from the city.

Introduction

San Francisco is the most densely settled large city (population greater than 200,000) in California and the second-most densely populated major city in the United States after New York City. A popular tourist destination, San Francisco is known for its cool summers, fog, steep rolling hills, eclectic mix of architecture, and landmarks, including the Golden Gate Bridge, cable cars, the former Alcatraz Federal Penitentiary, Fisherman's Wharf, and its Chinatown district. San Francisco is also the headquarters of five major banking institutions and various other companies such as Levi Strauss & Co., Gap Inc., Salesforce.com, Dropbox, Reddit, Square, Inc., Dolby, Airbnb, Weebly, Pacific Gas and Electric Company, Yelp, Pinterest, Twitter, Uber, Lyft, Mozilla, Wikimedia Foundation, and Craigslist. It has several nicknames, including "The City by the Bay", "Fog City", "San Fran", and "Frisco", as well as older ones like "The City that Knows How", "Baghdad by the Bay", "The Paris of the West", or simply "The City". As of 2016, San Francisco is ranked high on world liveability rankings.

Golden Gate Bridge

Motivation

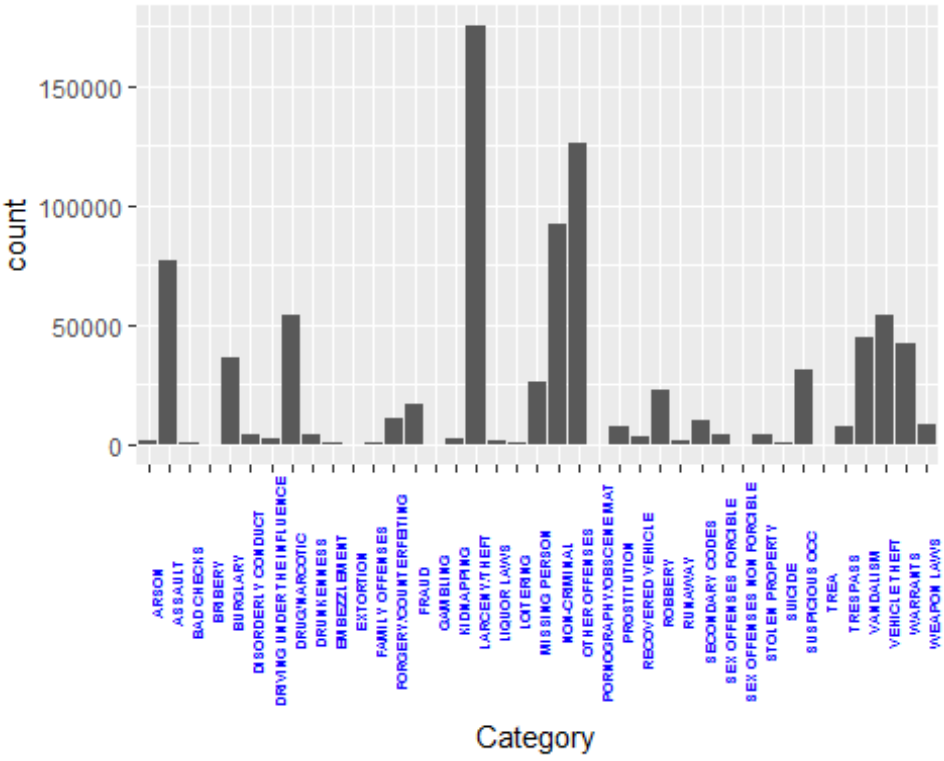
Over the last few years a new worldwide socio-economical order lead to an increasing number on crime rates and raised the need to find new ways to handle information about criminality. To better understand its causes, local, regional and national security authorities turned to new decision support tools such as crime analytics and other information technologies to help them in finding better solutions to forecast crime and take preventative steps to prevent these crimes from occurring.

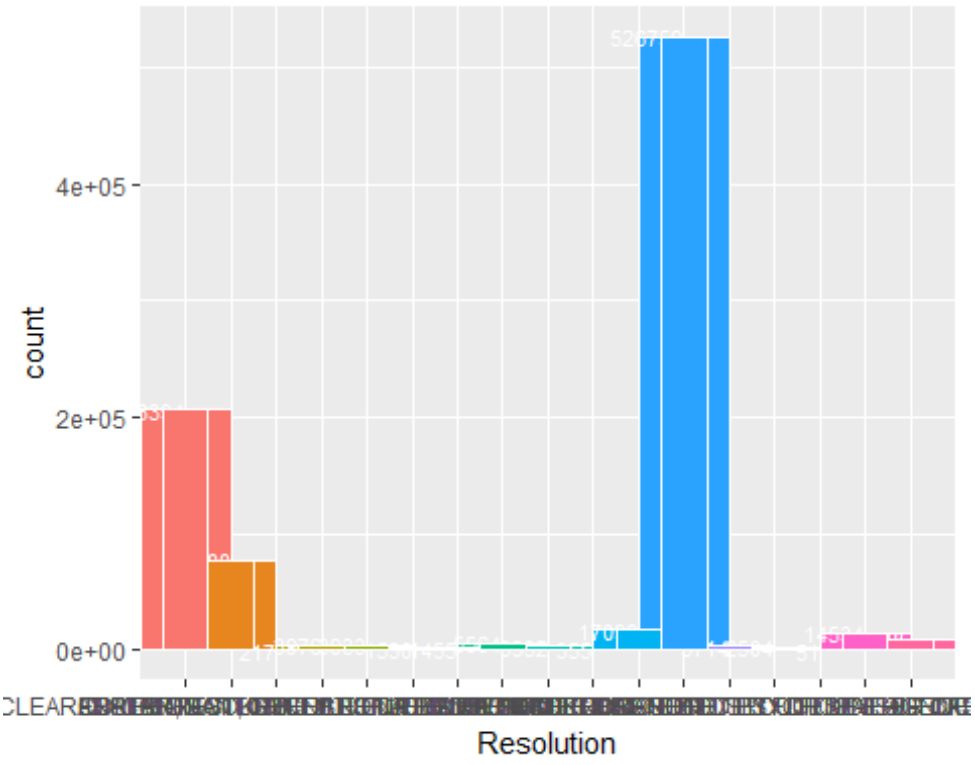
Background

Statistics are an important tool in crime analysis and police forces are using it in a more effective way to discover useful information, reduce crime rate by being proactive through the analyzed data. Statistics help strategic decisions and turn vast amounts of meaningless numbers into a general picture of crime events which can be used to predict the model for further years utilizing the resources fully. If every crime point has a geographical location (in space) and that every point has a considerable amount of information added to it, we can relate all the individual information (point) with all the others to construct a complete analysis.

Solution

For our project, we are analyzing the dataset which consists of information pertaining to the crime incidence for the city of San Francisco from 2003 to 2015. This information also includes details of the reported crime such as what was the resolution of the crime, location of the crime and type of crime. The extensive data provided from the dataset will enable us to analyze crime patterns and thus allow us to forecast the crime in an area. As per our project we used few models from supervised and unsupervised learning, time series analysis and data mapping for better visualization.

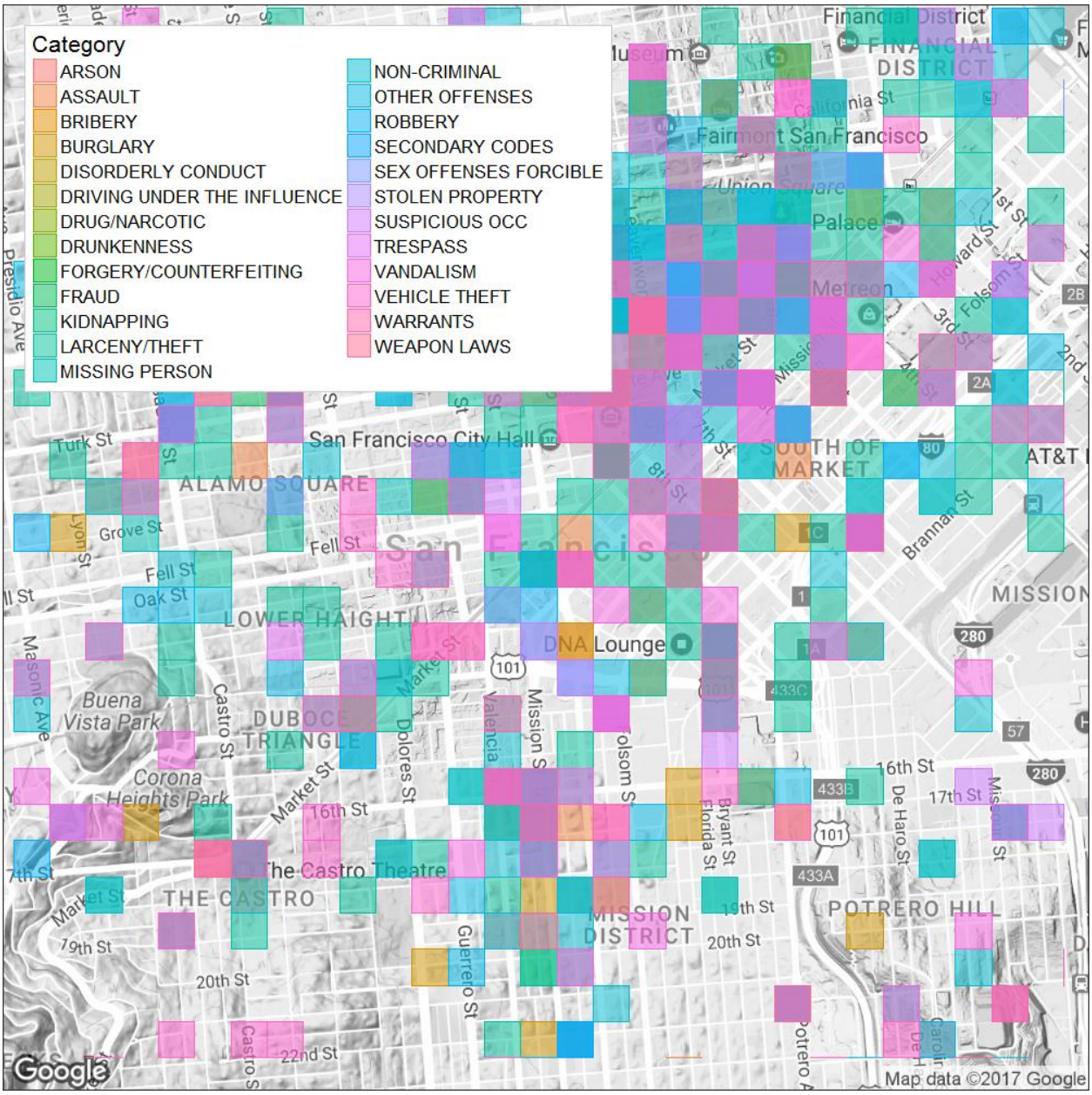




Description

Our database is extracted from San Francisco Police department. The sudden growth in the population has brought an inequality in terms of living, housing shortages leading to no scarcity of crime in the city This dataset has the incidents reported from various counties of San Francisco. To protect the privacy of victims, other information like address is provided at the block level. The objective of the project is to predict the requirement of police officers on a specific time in a county where the percentage of crime is high in a day with the help of historical data. This data can also be used to define the zones in the county where it is safe to live or any additional need of police patrolling to prevent crimes. Below are some methods which we used to analyze the data.

Following image give a view of mapping of categories of crime in San Francisco city.



Methods

Decision Trees

Decision tree is one of the important classification technique and a key step in data mining. A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their outcomes. The decision tree can be linearized into decision rules, where the outcome is the contents of the leaf node, and the conditions along the path form a conjunction in the if clause.

Each node in the tree is a decisions/tests. Each path from the tree root to a leaf corresponds to a conjunction of attribute decisions/tests. The tree itself corresponds to a disjunction of these conjunctions.

Attribute selection is one of the fundamental step to construct a decision tree and these attributes indicate the variables in our dataset which can be referred as a class or classifier. These attributes can be used to plan.

Information gain

Heuristic: choose the attribute that produces the "purest" nodes. That is, the most homogeneous splits. A popular impurity criterion is information gain. Information gain increases with the average purity of the subsets. The idea is to choose the attribute that gives greatest information gain as the root of the tree.

Association Rules

Association rule learning is a unsupervised method of generating "If this then that" type rules based on statistical associations in transactional data. A rule like {strawberries, chocolate}⇒{ice cream}{strawberries, chocolate}⇒{ice cream} for if somebody buys strawberries and chocolate together, they are likely to also buy ham ice cream meat. It is intended to identify "strong rules" that is, predictive rules.
An Association Rule: is an implication of the form $X \Rightarrow Y$X⇒Y, where $X, Y \Rightarrow I$X,Y⇒I

$P(A \cap B) = P(A)P(B) \Leftrightarrow P(B) = P(B|A)$P(A∩B)=P(A)P(B)⇔P(B)=P(B|A)

That is, two random variables are statistically in dependent if the occurrence of B does not affect the probability of A, and vice versa. Two random variables are statistically in dependent if the occurrence of B does affect the probability of A, and vice versa.

The term "association" is closely related to the term correlation and to the term mutual information.

Support, Confidence, Lift and Conviction

To select interesting rules from the set of all possible rules, constraints on various measures of significance and interest are used. The best-known constraints are minimum thresholds on support and confidence.

Support

It is defined as the proportion of transactions in the database which contains the data set. That is, the fraction of transactions that contain the data set.
The support count is the frequency of occurrence of different categories in our dats set

Confidence

The confidence value of a rule is the proportion the transactions, that can be interpreted as an estimate of the conditional probability

$P(EY|EX)$P(EY|EX),

the probability of finding the RHS of the rule in transactions under the condition that these transactions also contain the LHS.

Lift

The lift of a rule is defined as the ratio of the observed support to that expected if both variable were independent.

Conviction

It can be interpreted as the ratio of the expected frequency (that is to say, the frequency that the rule makes an incorrect prediction) divided by the observed frequency of incorrect predictions.

Apriori algorithm

The Apriori algorithm is the best-known algorithm to mine association rules. It uses a breadth-first search strategy to count the support of item sets and uses a candidate generation function which exploits the downward closure property of support.

Apriori principle: * If an item set is frequent, then all its subsets must also be frequent. That is, any subset of a frequent item set is frequent.

Apriori Advantages:

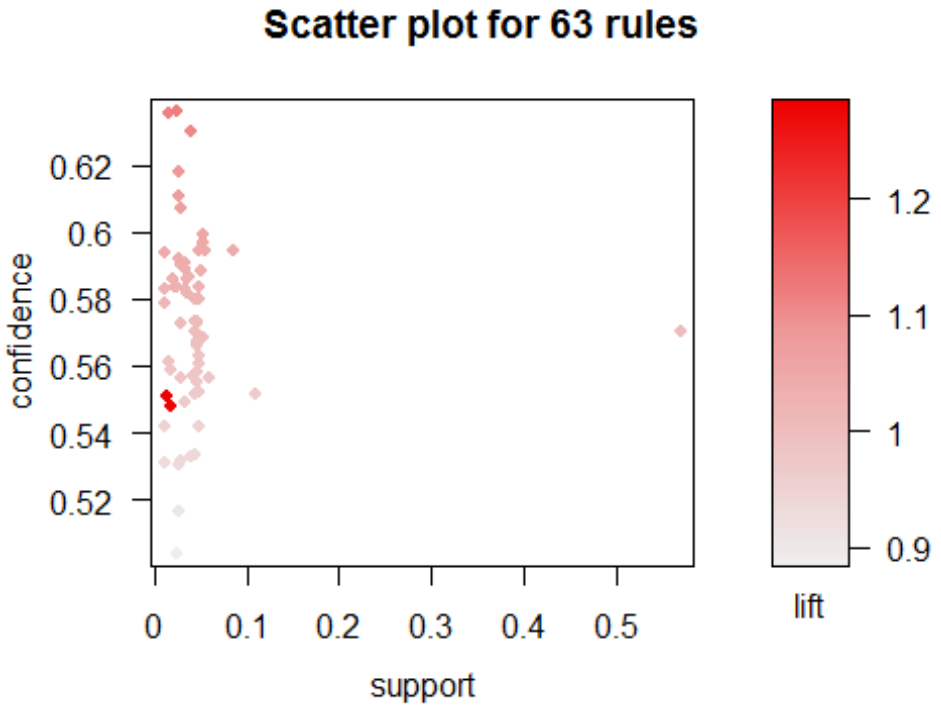
- Uses large item set property.
- Easily parallelized
- Easy to implement.

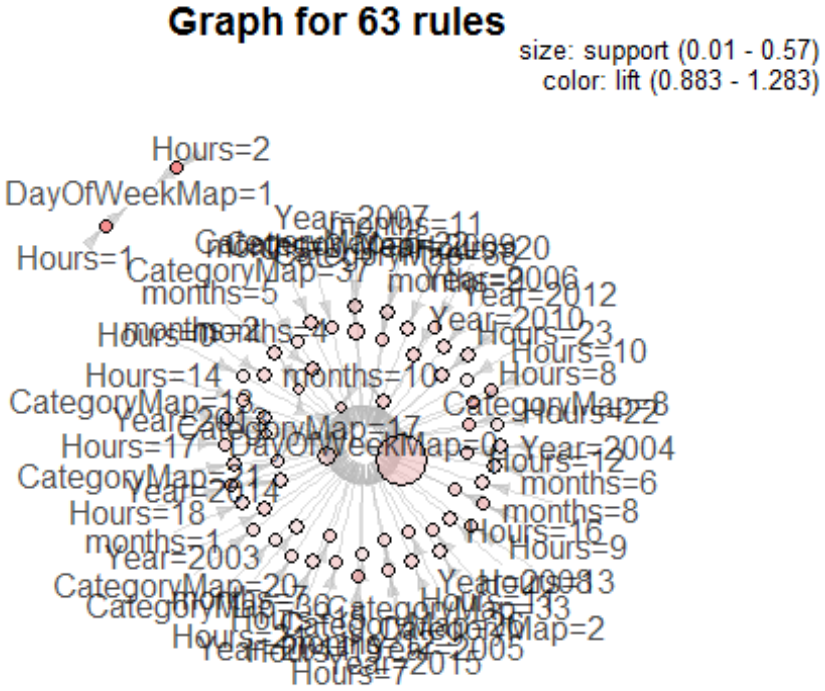
Apriori Disadvantages:

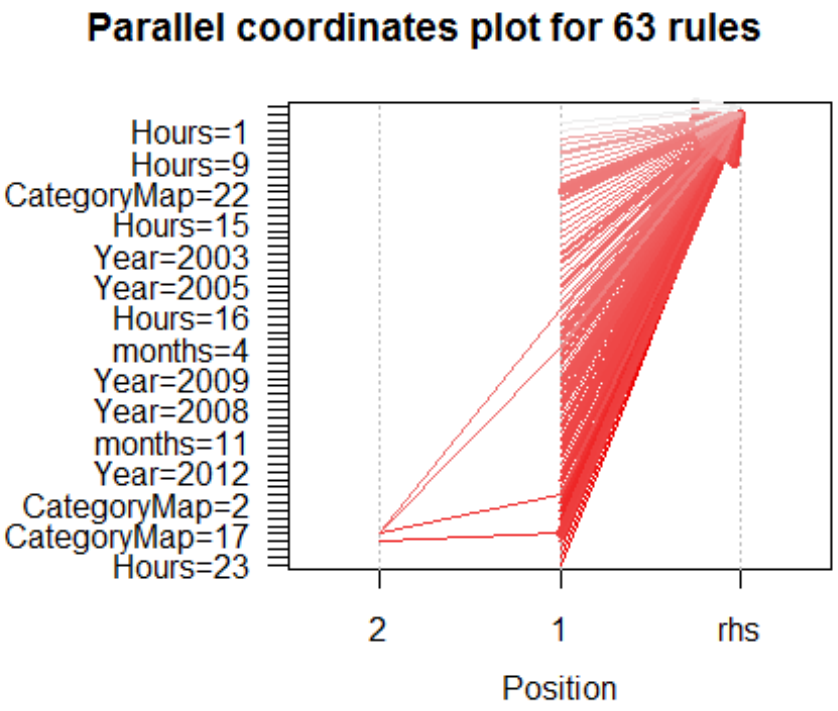
- Assumes transaction database is memory resident.
- Requires many database scans.

Improving Apriori's Efficiency

- Hash-based item set counting: A k-item set whose corresponding hashing bucket count is below the threshold cannot be frequent
- Transaction reduction: A transaction that does not contain any frequent k-item set is useless in subsequent scans
- Partitioning: Any item set that is potentially frequent in DB must be frequent in at least one of the partitions of DB
- Sampling: mining on a subset of given data, lower support threshold + a method to determine the completeness
- Dynamic item set counting: add new candidate item sets only when all of their subsets are estimated to be frequent


Scatter plot for 63 rules
confidence
0.62
0.6
0.58
0.56
0.54
0.52
0
0.1
0.2
0.3
0.4
0.5
support
1.2
1.1
1
0.9
lift


Graph for 63 rules
size: support (0.01 - 0.57)
color: lift (0.883 - 1.283)
Hours=2
DayOfWeekMap=1
Hours=1
Year=2007
CategoryMap=37
months=5
Year=2012
Year=2010
Hours=23
Hours=10
Hours=8
Hours=14
months=10
Hours=17
Year=2004
Hours=12
months=6
CategoryMap=31
Year=2014
Hours=18
months=8
months=1
Hours=16
Hours=9
Year=2003
CategoryMap=20
Hours=7
Year=2015



Time Series Analysis

A time series is a sequence of data points, typically consisting of successive measurements made over a time interval. This is a very common type of data as we frequently measure how something varies over time. usually a time series is an ordered sequence of values of a variable at equally spaced time intervals; but methods exist to deal with irregular sampling.

Examples of time series include:

- Stock Market
- The change of weather
- An electrocardiogram (EKG or ECG), that is, the electrical activity of your heart.
- The popularity of a celebrity or politician
- Economic Forecasting

Univariate (bivariate, multivariate) time series: collection of observations of one (two, several) state variables, that are made in sequential moments in time.

Components of Time Series

The pattern or behavior of the data in a time series can be made up of several components. Theoretically, any time series can be decomposed into:

Trend -

The trend component accounts for the gradual shifting of the time series to relatively higher or lower values over a long period.

Cyclical -

A regular pattern of sequences of values that go above and below the trend line is a cyclical component.

Seasonal -

The seasonal component accounts for regular patterns of variability within certain time periods, such as a year. While seasons could be modeled using cyclical components. Often within-year, within-month, within-week or within-day cycles are treated as "seasonal" behavior.

Irregular -

The irregular components represent erratic, unsystematic, 'residual' fluctuations. That is, noise.

Time Series Analysis

Time series analysis generates a model that accounts an internal structure (such as autocorrelation, trend or seasonal variation) of a set of data points taken over time. This is very frequently used for forecasting. That is, predicting and future event in time based on a sequential historical sample.

Autoregressive Integrated Moving Average [ARIMA]

An autoregressive integrated moving average (ARIMA or ARMA) model combines an autoregressive component with a moving average component in to a single model.

An autoregressive integrated moving average (ARIMA or ARMA) model is a generalization of an autoregressive moving average (ARMA) model. These models are fitted to time series data either to better understand the data or to predict future points in the series (forecasting). They are applied in some cases where data show evidence of non-stationarity, where an initial differencing step (corresponding to the "integrated" part of the model) can be applied to reduce the non-stationarity.

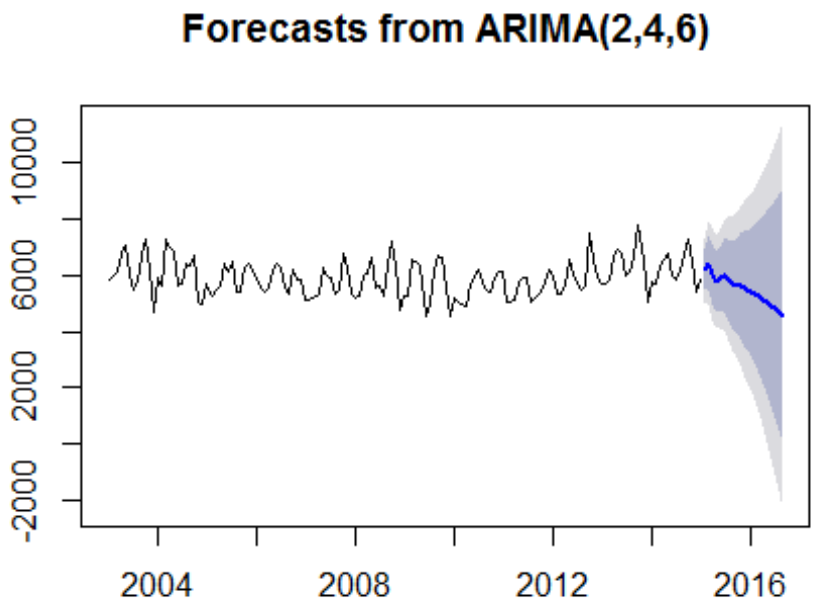
Non-seasonal ARIMA models are generally denoted *ARIMA* (p,d,q) ARIMA (p,d,q) where parameters p, d, *and* q p, d, and q are non-negative integers, pp is the order of the Autoregressive model, (dd) is the degree of differencing, and (qq) is the order of the Moving-average model. The he number of differences (dd) is determined using repeated statistical tests. The values of pp and qq are then chosen by minimizing the AICc after differencing the data dd times.

The ARIMA model uses an iterative three-stage modeling approach:

- Model identification and model selection to make sure that the variables are stationary, identifying seasonality in the dependent series (seasonally differencing it if necessary), and using plots of the autocorrelation and partial autocorrelation functions of the dependent time series to decide which (if any) autoregressive or moving average component should be used in the model.

- Parameter estimation using computation algorithms to arrive at coefficients that best fit the selected ARIMA model. The most common methods use maximum likelihood estimation or non-linear least-squares estimation.
- Model checking by testing whether the estimated model conforms to the specifications of a stationary univariate process. The residuals should be independent of each other and constant in mean and variance over time. If the estimation is inadequate, we must return to step one and attempt to build a better model.

For our project, *the ts() function will convert a numeric vector into an R time series object. Further,* the ARIMA model creating ranges for the possible values for the order parameters p [0:2], d [0:6] and q [0:6]. Thus, we pick the best model, which would be the one with the smallest AIC. Next, we generate the ARIMA model for the best p, d, q order model. Finally, we plot the forecast for the model.



Trend Analysis

Trend Analysis is the practice of collecting information and attempting to spot a pattern, or trend, in the information. Typically, this involves analyzing the variance for a change over time. The null hypothesis: *H*OHO is that there is no trend. Many techniques can be used to identify trends, we'll use an ARMA model again.

Dickey-Fuller Test

The Dickey-Fuller Test is a test for the stationarity of a time series.

The Dickey-Fuller test tests whether a unit root is present in an autoregressive model. simple AR(1) model is

$$y_t=\rho y_{t-1}+u_t \text{yt=ρyt−1+ut}$$

where $y_t$yt is the variable of interest, tt is the time index, ρρ is a coefficient, and $u_t$ut is the error term. A unit root is present if $\rho=1$ρ=1. The model would be non-stationary in this case.
The regression model can be written as

$$\nabla y_t=(\rho-1)y_{t-1}+u_t=\delta y_{t-1}+u_t \nabla\text{yt=(ρ−1)yt−1+ut=δyt−1+ut}$$

where ∇∇ is the first difference operator.

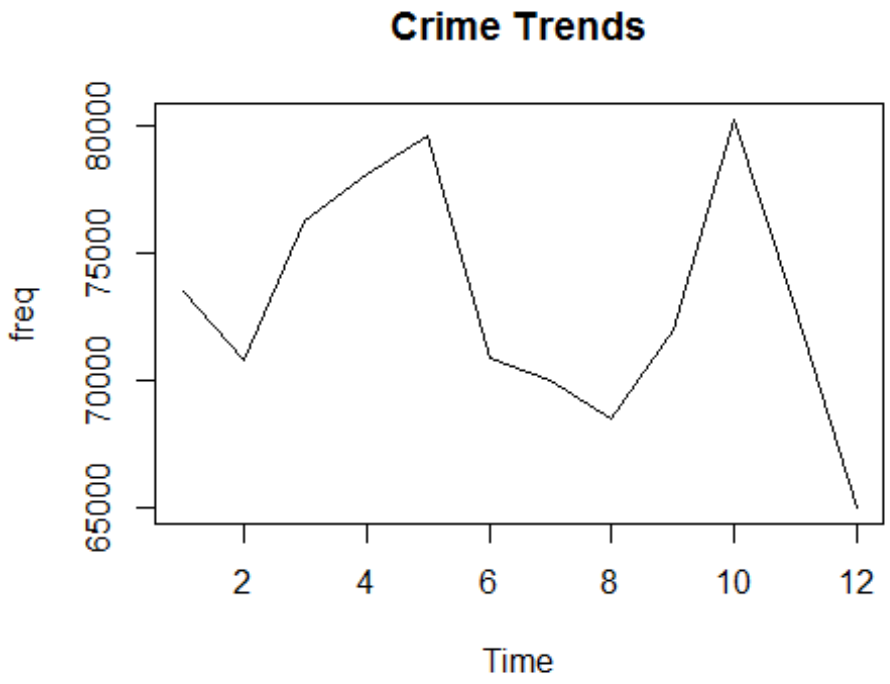
The Dickey-Fuller Test uses a specific distribution simply known as the Dickey-Fuller table to assess whether ∇y_t∇yt is signficant.

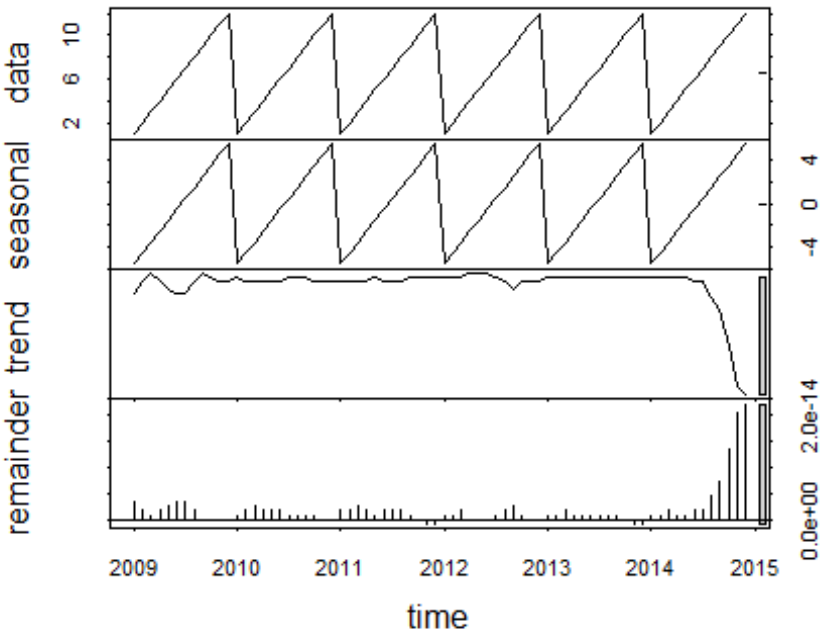
For our project, we first create ‘crimeT’ to contain the month data and the frequency of crime occurrence in that month. Plotting this helps us to view the crime trends. Next, we plot the ‘random walk’ crime trends, using ‘crime_rw’. Once this is done we need to generate the Moving Average Model. For this we require the ‘arima.sim’ function, which is followed by the generation of the Autoregressive model.

This is followed by the following steps:

- Dickey-Fuller [for stationarity]
- Philips-Perron Test [unit root test]
- Seasonal Trend Decomposition

The Seasonal Trend Decomposition uses Loess algorithm to divide up a time series into three components: the trend, seasonality and remainder.


Crime Trends
freq
65000
70000
75000
80000
2
4
6
8
10
12
Time

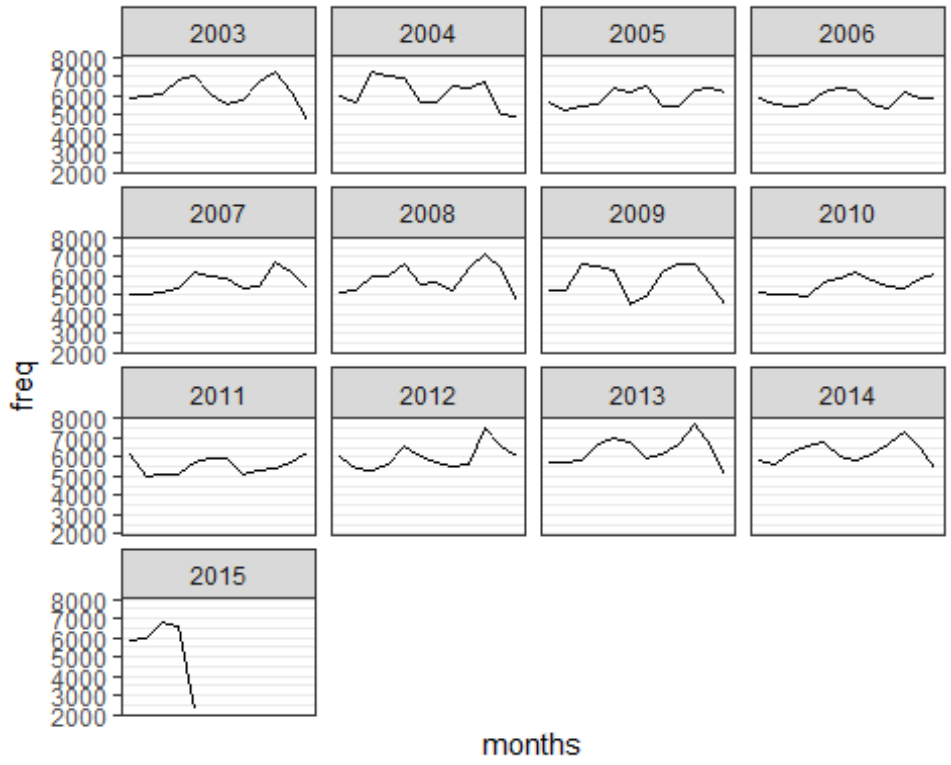

data
seasonal
trend
remainder
2
6
10
-4
0
4
2.0e-14
0.0e+00
2009
2010
2011
2012
2013
2014
2015
time

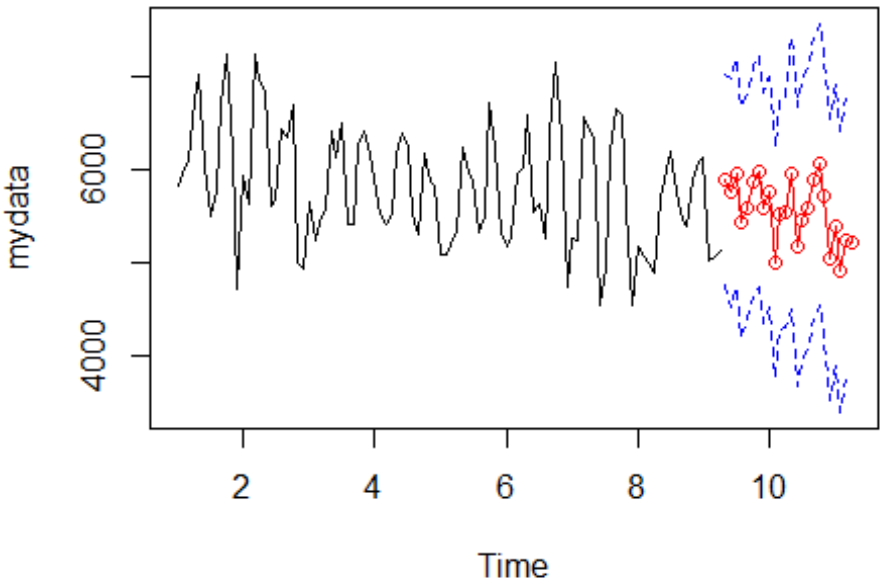
Seasonal Models

When there are patterns that repeat over known, fixed periods of time (i.e. day, week, month, quarter, year, etc.) within the data set it is seasonal variation. One has a model for the periodic fluctuations based on knowledge of the domain.

The seasonal ARIMA model incorporates both non-seasonal and seasonal factors in a multiplicative model. In a seasonal ARIMA model, seasonal parameters predict xt using data values and errors at times with lags that are multiples of S (the span of the seasonality). Before we model for a given data set, one must have an initial guess about the data generation process, that is the span of the seasonality (i.e.day, week, month, quarter, year, etc.)

For our project, we use an ARIMA model to identify seasonality trends by looking for significant seasonal differences. Once, reasonable predictions are done, we can forecast the model. We have considered forecasting 24 months into the future.



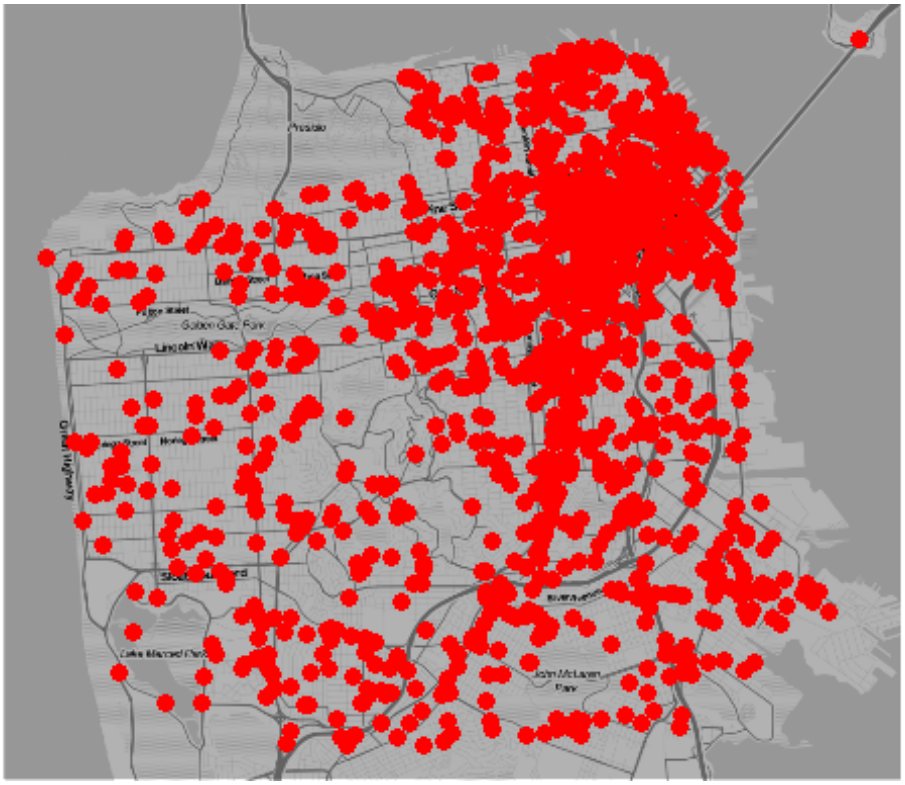


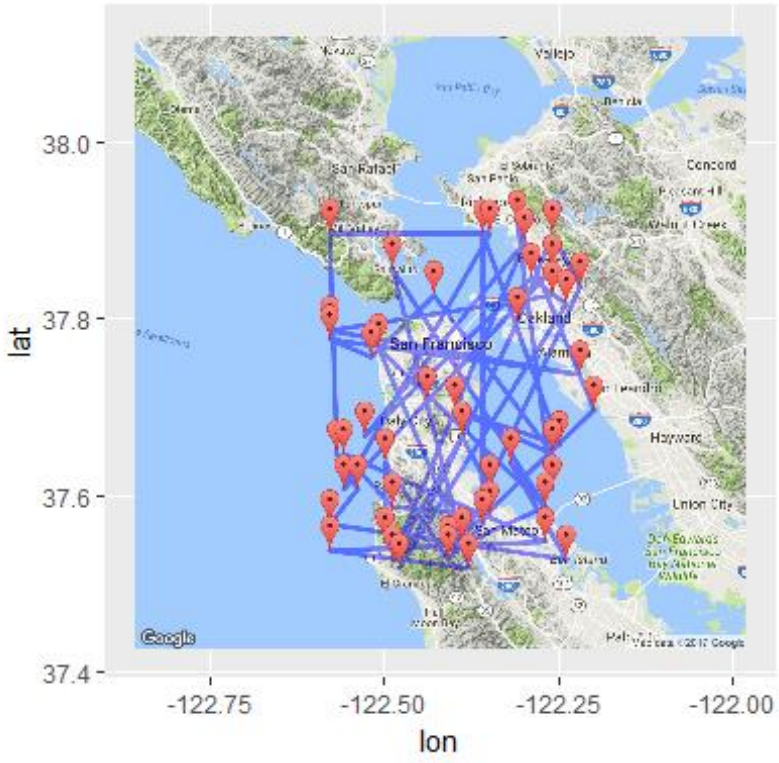
Results

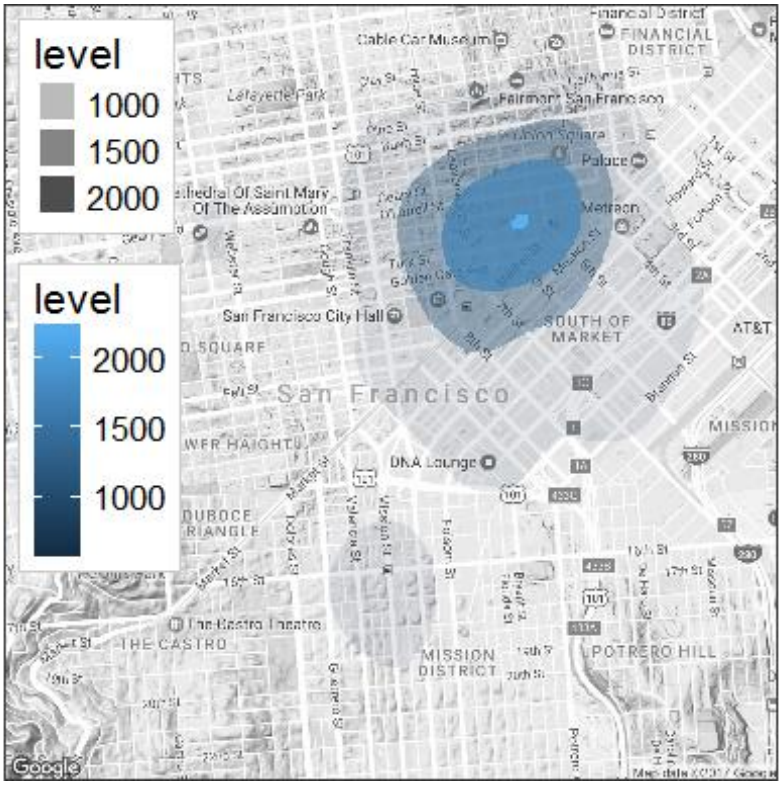
As mentioned above, Our purpose was to analyze the data for prediction model and crime mapping of certain areas in the city. Our method was to access each forecasting approach based on accuracy and ease of use. The importance of data quality is also essential to obtain effective results, therefore we removed the outliers in the beginning of our project for any false output.

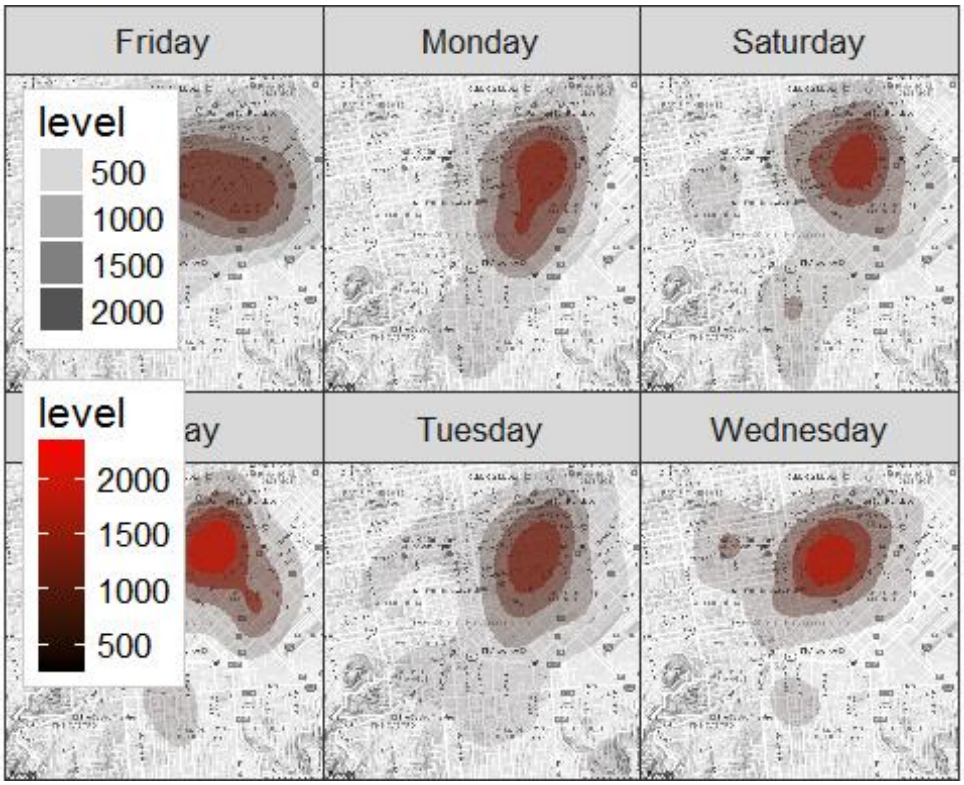
To extract information from the data to predict the future crime pattern was the primary objective of our research. The result of our project can play an important role to decide an effective strategy for future crime. The visualization itself shows the crime pattern over the years for various categories and the resolution for the same. It can show the improvisation in certain aspects of handling few areas where the violent crimes are committed and furthermore these models can be utilized to work on the prevention effort through guiding theory.

Complementing spatial analysis of crime with the knowledge of previous crimes in a given territory, one may identify the reasons and therefore plan accordingly to prevent the crimes thereby reducing the rate. Overall, our project sheds the light on the crime rate in a geographical area and predict the outcome for further years which can be utilized as a part of strategy to reduce the crime rate of the city.




38.0
37.8
37.6
37.4
lat
-122.75
-122.50
-122.25
-122.00
lon
Vallejo
Concord
San Rafael
Oakland
San Francisco
Hayward
Union City
Google


level
1000
1500
2000
level
2000
1500
1000
Cable Car Museum
FINANCIAL DISTRICT
Fairmont San Francisco
Palace
Metreon
San Francisco City Hall
SOUTH OF MARKET
AT&T
San Francisco
DNA Lounge
MISSION
THE CASTRO
MISSION DISTRICT
POTRERO HILL
Google



References

1] San Francisco - Wikipedia

2] Agrawal, R.; Imieliński, T.; Swami, A. (1993). "Mining association rules between sets of items in large databases". Proceedings of the 1993 ACM SIGMOD international conference on Management of data - SIGMOD '93. p. 207.

3] Behara, E. b. M., Krickeberg, K., & Wolfowitz, J. (1973). "Probability and information theory II" Springer-Verlag.

4] Yeung, R. W. 2002. "A first course in information theory Kluwer Academic/Plenum Publishers.

5] Cover, T. M., & Thomas, J. A. (1991). "Elements of Information Theory"" Wiley.

6] Deng,H.; Runger, G.; Tuv, E. (2011). "Bias of importance measures for multi-valued attributes and solutions."" Proceedings of the 21st International Conference on Artificial Neural Networks (ICANN).

7] Emmert-Streib, F., & Dehmer, M. (2009). "Information Theory and Statistical Learning."" Springer-Verlag.

8] Quinlan, J. R. (1987). “Simplifying decision trees”. International Journal of Man-Machine Studies 27 (3): 221. doi:10.1016/S0020-7373(87)80053-6.

9] Shannon, C. E. (1948). “A Mathematical Theory of Communication.”" Bell System Technical Journal, 27(3), 379-423.

10] Theil. (1972). “Statistical Decomposition Analysis.”" Studies in Mathematical and Managerial Economics, 14.

11] Decision Trees

12] R for Time Series Analysis

13] https://www.safaribooksonline.com/library/view/mastering-predictive-analytics/9781783982806/

14] http://www.statoek.wiso.uni-goettingen.de/veranstaltungen/zeitreihen/sommer03/ts_r_intro.pdf

15] http://www.stat.pitt.edu/stoffer/tsa3/R_toot.htm

16] http://www.statoek.wiso.uni-goettingen.de/veranstaltungen/zeitreihen/sommer03/ts_r_intro.pdf

17] https://www.safaribooksonline.com/library/view/mastering-predictive-analytics/9781783982806/

18] http://www.r-bloggers.com/seasonal-trend-decomposition-in-r/

19] http://www.springer.com/us/book/9781441978646#otherversion=9781461427599

20] http://a-little-book-of-r-for-time-series.readthedocs.org/en/latest/src/timeseries.html

21] https://rpubs.com/ryankelly/tsa5

22] Kaggle