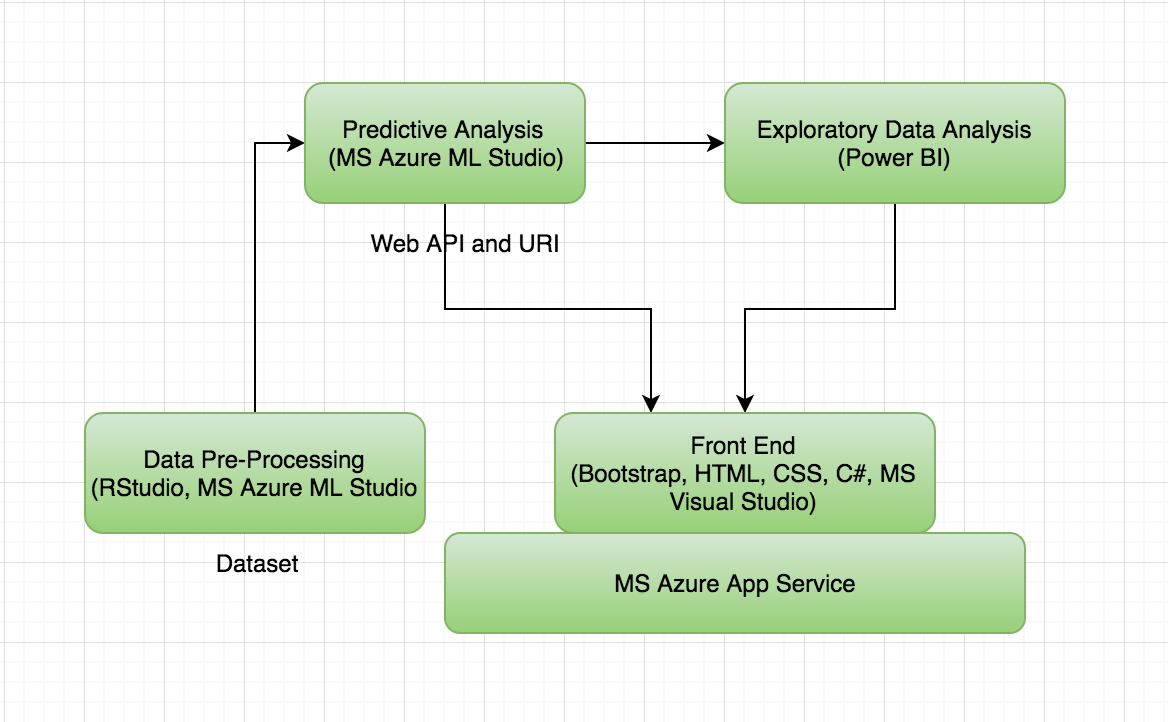
**Project Proposal:**

The main objective of our project is to predict a crime incident that has the highest probability of occurrence at a given street, at a particular date and at a point of time.

The end users such as the Law Enforcement Agencies will be benefitted by the analysis and visualization of crime occurrence patterns.

We have considered the Boston Crime Dataset to perform exploratory data analysis and build a predictive model to achieve our resolve. The predictive model was then deployed as a web service using MS Azure App Service so that it is available to the end user as a web application.

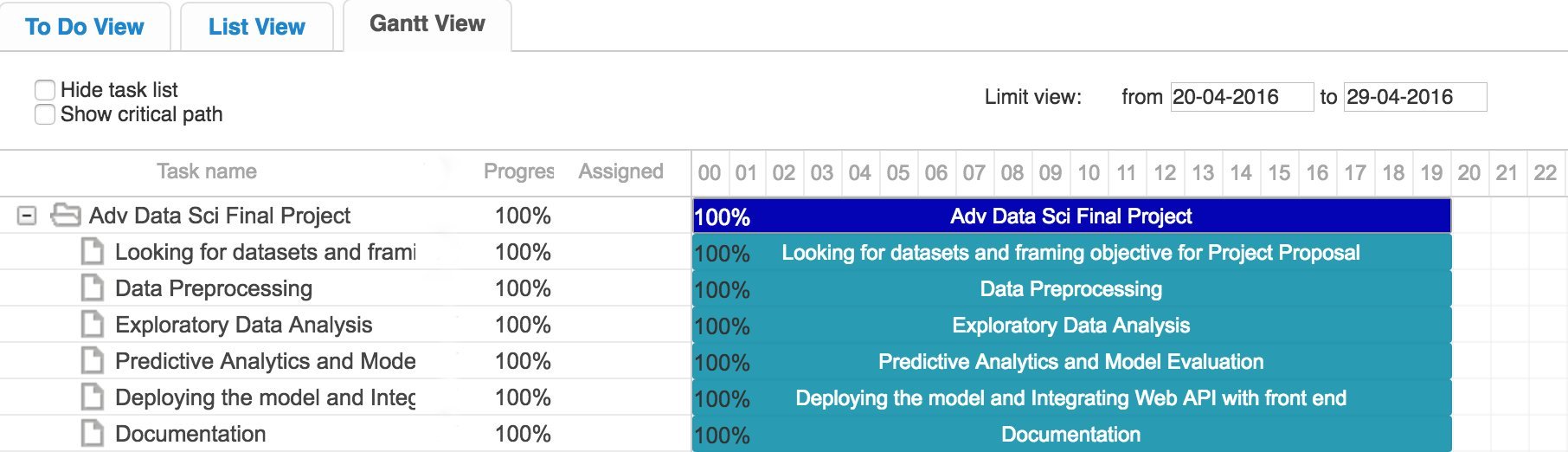
**Architecture:**



The above figure explains the components of the architecture:

1. Data Pre-processing was done on MS Azure ML Studio as well as RStudio.
2. The Predictive analysis was performed on MS Azure ML Studio where the web service was deployed by generating the web API.
3. The Exploratory data analysis was carried out on Power BI.
4. The final predictive model and the Power BI visualization was integrated to the front-end using Bootstrap, HTML, CSS and C# on a MS Visual Studio IDE.

**Project Plan:**

****

The above Gantt View illustrates the road map to the completion of our project plan.

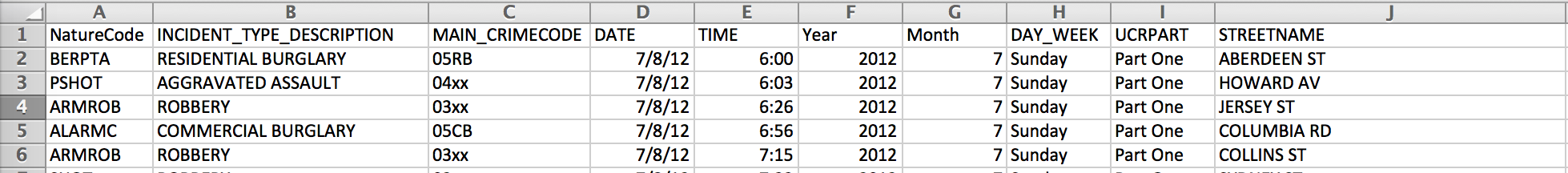
**Dataset and Data pre-processing:**

The Boston crime dataset comprised records of crime incidents from 2012 to 2015 for different streets in the city of Boston. The original dataset had the following variables.

1. Nature code - Nature of crime code.
2. Incident Type (24 categories: Vandalism, Larceny, Simple Assault, Drug Charges, Fraud, Property Lost, Residential Burglary, Towed, Robbery, Trespass, Commercial Burglary, Forgery, Weapons Charge, Landlord dispute, Harassment, Violation of liquor laws, Crimes Against Children, Drug Violation, Bomb, License Violation, Offenses Against Child/Family, Sex Offender Registration, Criminal Harassment, Prostitution)
3. MAIN\_CRIMECODE - is the Boston Regional Intelligence Center classification of the crime code
4. UCRPART - FBI’s Uniform Crime Reporting (UCR) Program divides offenses into groups to measure the level and scope of crime and collect arrest data.

5. STREETNAME – Street names of Boston where the crime occurred.

1. DAY
2. TIME
3. Year
4. Month
5. DAY\_WEEK



Our objective is to predict the highest probability of the crimes that can happen on a particular street, given a particular day and time. If we consider the dataset above, the variables NatureCode, Main\_CrimeCode and UCRPART have a one-to-one relationship with the INCIDENT TYPE outcome variable.

* MAIN\_CRIMECODE: is the Boston Regional Intelligence Center classification of the crime code
* UCRPART: FBI’s Uniform Crime Reporting (UCR) Program divides offenses into groups to measure the level and scope of crime and collect arrest data.
* NatureCode: Nature of crime code.

This defeats the purpose of our objective since the predictions will be highly accurate with an over-fitted model.

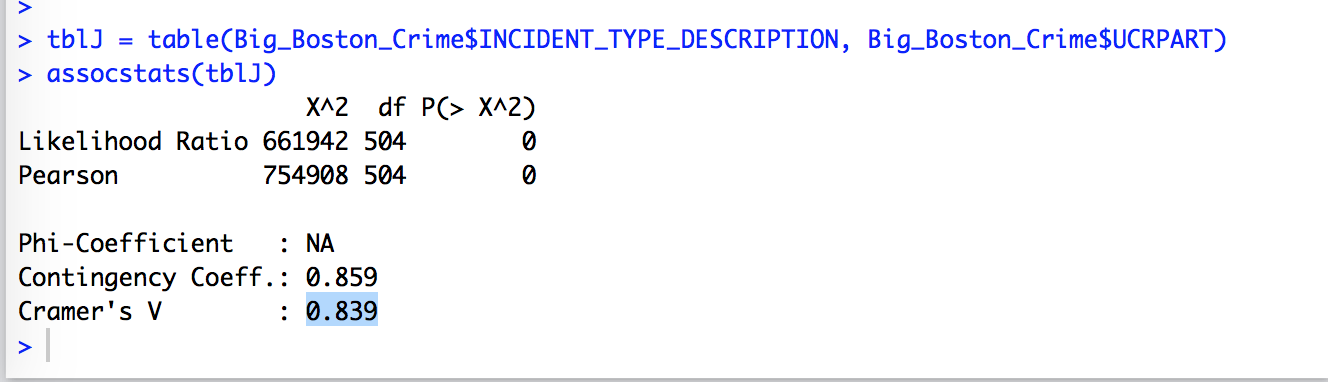
Moreover we performed a Cramer’s V test of association between categorical variables.

**[I] Association**

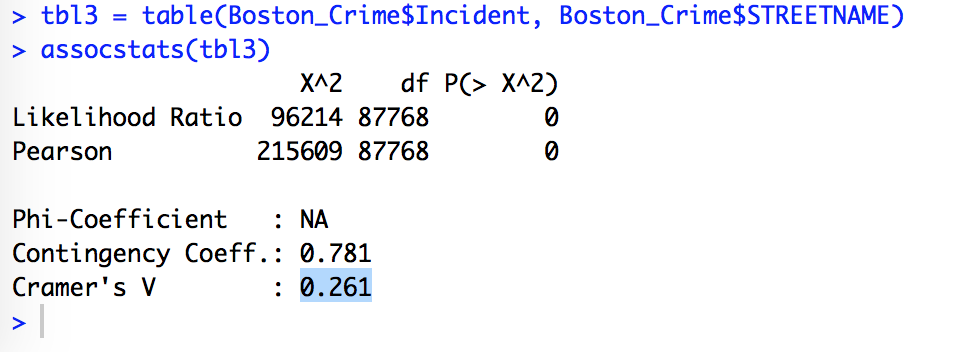
**Cramer’s V association measure for categorical variables**

*Cramer’s V ranges from 0 (no relationship) to +1 (perfect relationship)*

We considered effect of predictor variable UCRPART on INCIDENT TYPE outcome variable to find a high *Cramer’s V value of 0.839.* Therefore we exclude UCRPART from our dataset along with other predictor variables like NatureCode and MAIN\_CRIMECODE.

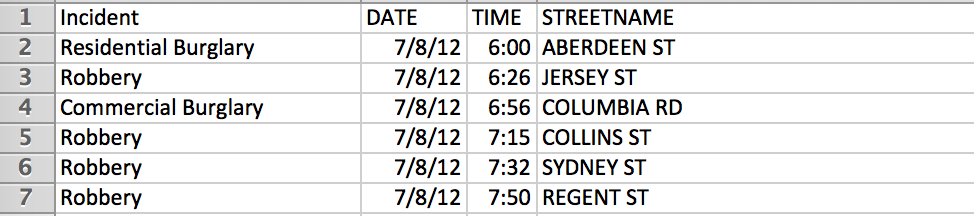


On the other hand we performed a Cramer’s V test of association between INCIDENT TYPE (outcome) and STREETNAME (predictor) to find *a Cramer’s V value of 0.261.* So based on our objective we keep the STREETNAME predictor variable to build our model.



**Our final dataset looks like the following with the variables:**

1. Incident – Outcome variable
2. Date
3. Time
4. Streetname



**Data processing:**

The dataset that consists of roughly a hundred and forty thousand records had many missing, null, zero values that needed to be gotten rid of. We used RStudio and MS Azure Machine Learning Studio for processing the dataset.

**[II] Classification**

**Building the Predictive Model:**

After preprocessing the data, removing duplicate categories and unwanted variables, the csv file was imported to MS Azure ML Studio to build the predictive model.

The following steps were taken for building the predictive model:

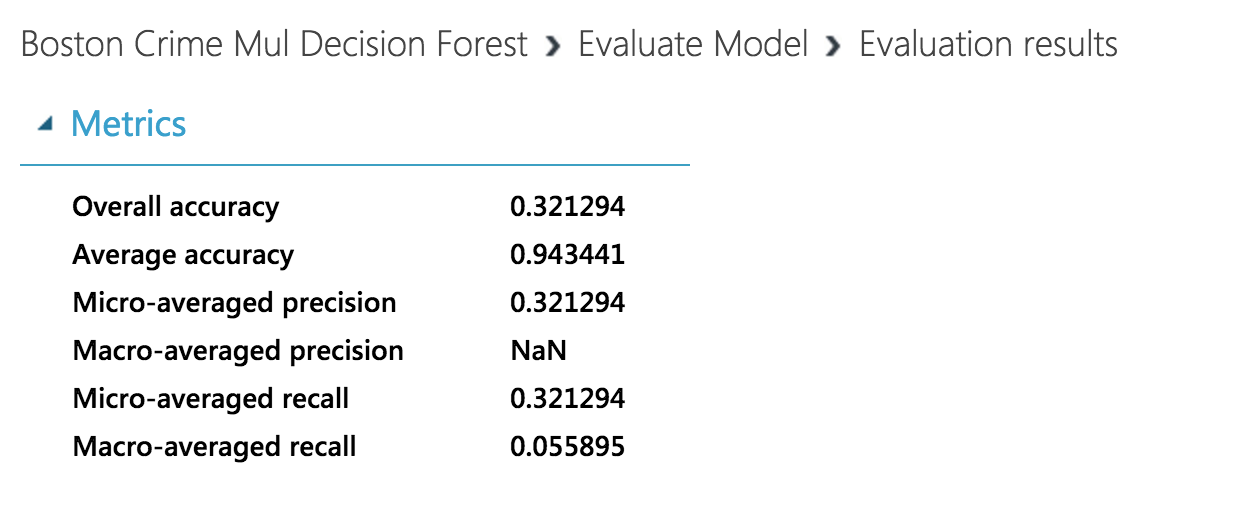
1. The missing values were gotten rid of using the clean missing data module.
2. The outcome variable “incident” type was set to categorical type using the Edit Metadata module.
3. The Machine learning modules used in building the predictive model were classification algorithms since the outcome variable considered is “incident” that has 24 different categories. So our algorithm will have 24 classifiers.

* Multiclass Decision Forest
* Multiclass Decision Jungle
* Multiclass Logistic Regression

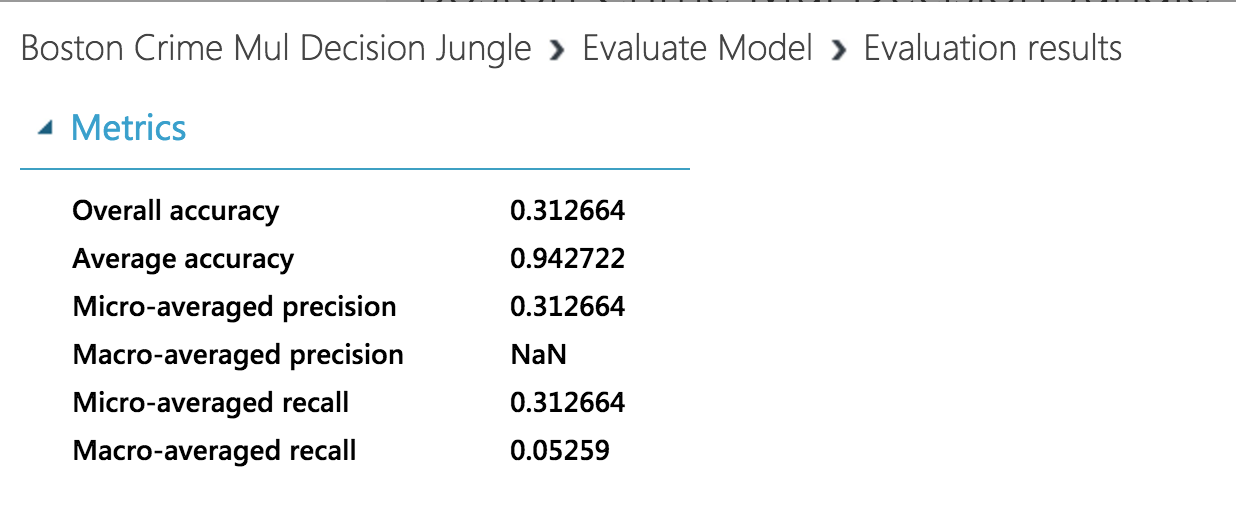
**Model Evaluation:**

The following snapshot displays the over all accuracy measure for all the 3 classification models:

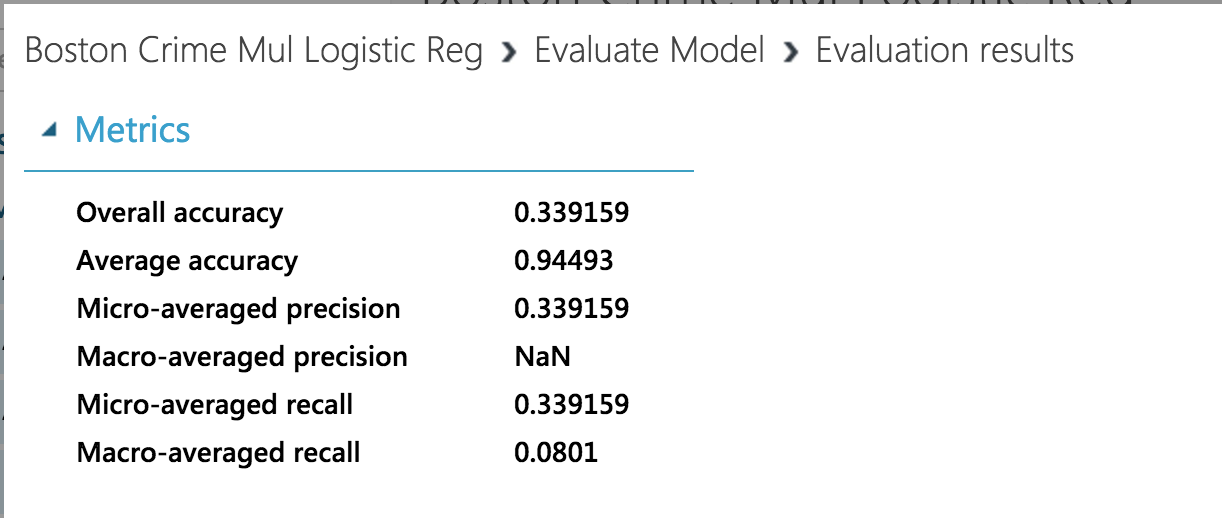
Multiclass Decision Forest:



Multiclass Decision Jungle:

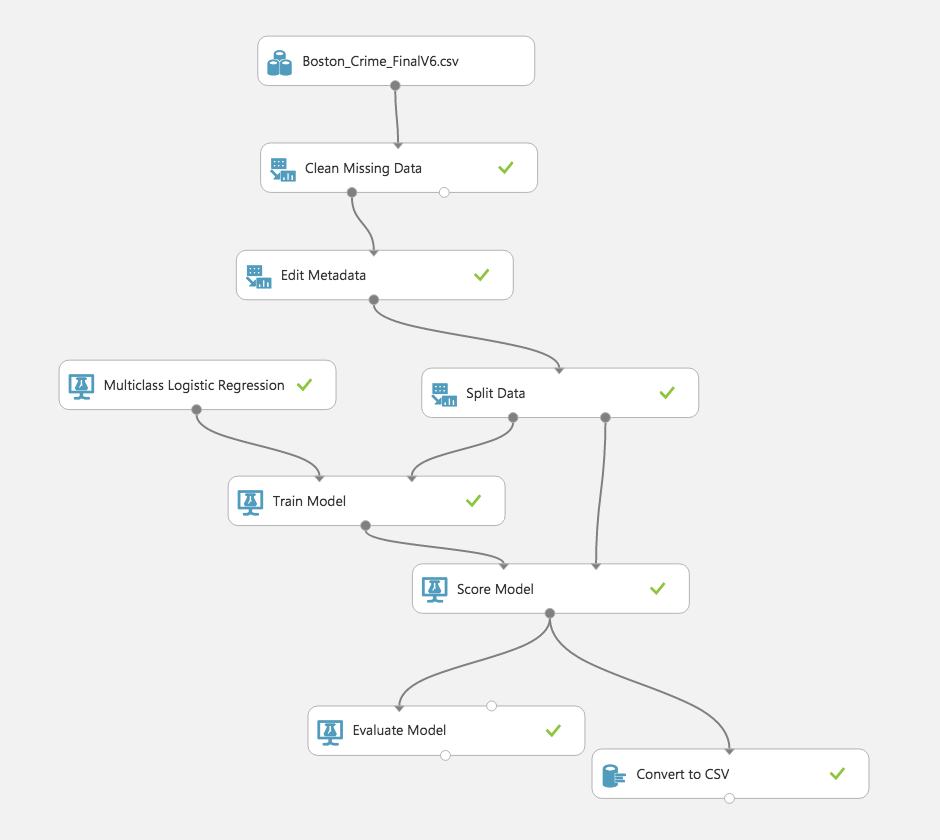


Multiclass Logistic Regression:



Hence our predictive model was built using Multiclass Logistic Regression Classification Model, evaluated based on the accuracy of ~0.34 as deduced from the snapshots above.

**The following is the image of the model we used for the Boston Crime Dataset.**



1. Furthermore the Multiclass Logistic Regression Classification Model was trained and scored using 60:40, 70:30 and 80:20 splits.

60:40 split overall accuracy = 0.33

70:30 split overall accuracy = 0.34

80:20 split overall accuracy = 0.10

So, finally the data was split into a 70:30 partition where 70% of the data was used for the training model and 30% was used for the scoring the.

1. The web service was deployed and the API key was obtained.
2. The predictive model was deployed and the **web API key** was used in R script to confirm the predicted values.

**Here is the R script and the output:**

install.packages("RCurl")

install.packages("rjson")

library("RCurl")

library("rjson")

# Accept SSL certificates issued by public Certificate Authorities

options(RCurlOptions = list(cainfo = system.file("CurlSSL", "cacert.pem", package = "RCurl")))

h = basicTextGatherer()

hdr = basicHeaderGatherer()

req = list(

Inputs = list(

"input1" = list(

"ColumnNames" = list("DATE", "TIME", "STREETNAME"),

"Values" = list( list( "4/30/2016", "21:00", "TREMONT ST" ), list( "5/2/2016", "8:00", "WESTLAND AVE" ) )

) ),

GlobalParameters = setNames(fromJSON('{}'), character(0))

)

body = enc2utf8(toJSON(req))

api\_key = "tq8Bi7E93dxhU2FIg26yCys3+TbdBkZWRho/0J0Mnre6JlAQUaAXk0D1xwwfyCeeMH2GUBKKkMUIjW3U79MsPw==" # Replace this with the API key for the web service

authz\_hdr = paste('Bearer', api\_key, sep=' ')

h$reset()

curlPerform(url = "https://ussouthcentral.services.azureml.net/workspaces/7a42d134b6c64c51b80b0f36259de4c0/services/0856063ef4d5485281a494ad9e05efe2/execute?api-version=2.0&details=true",

httpheader=c('Content-Type' = "application/json", 'Authorization' = authz\_hdr),

postfields=body,

writefunction = h$update,

headerfunction = hdr$update,

verbose = TRUE

)

headers = hdr$value()

httpStatus = headers["status"]

if (httpStatus >= 400)

{

print(paste("The request failed with status code:", httpStatus, sep=" "))

# Print the headers - they include the requert ID and the timestamp, which are useful for debugging the failure

print(headers)

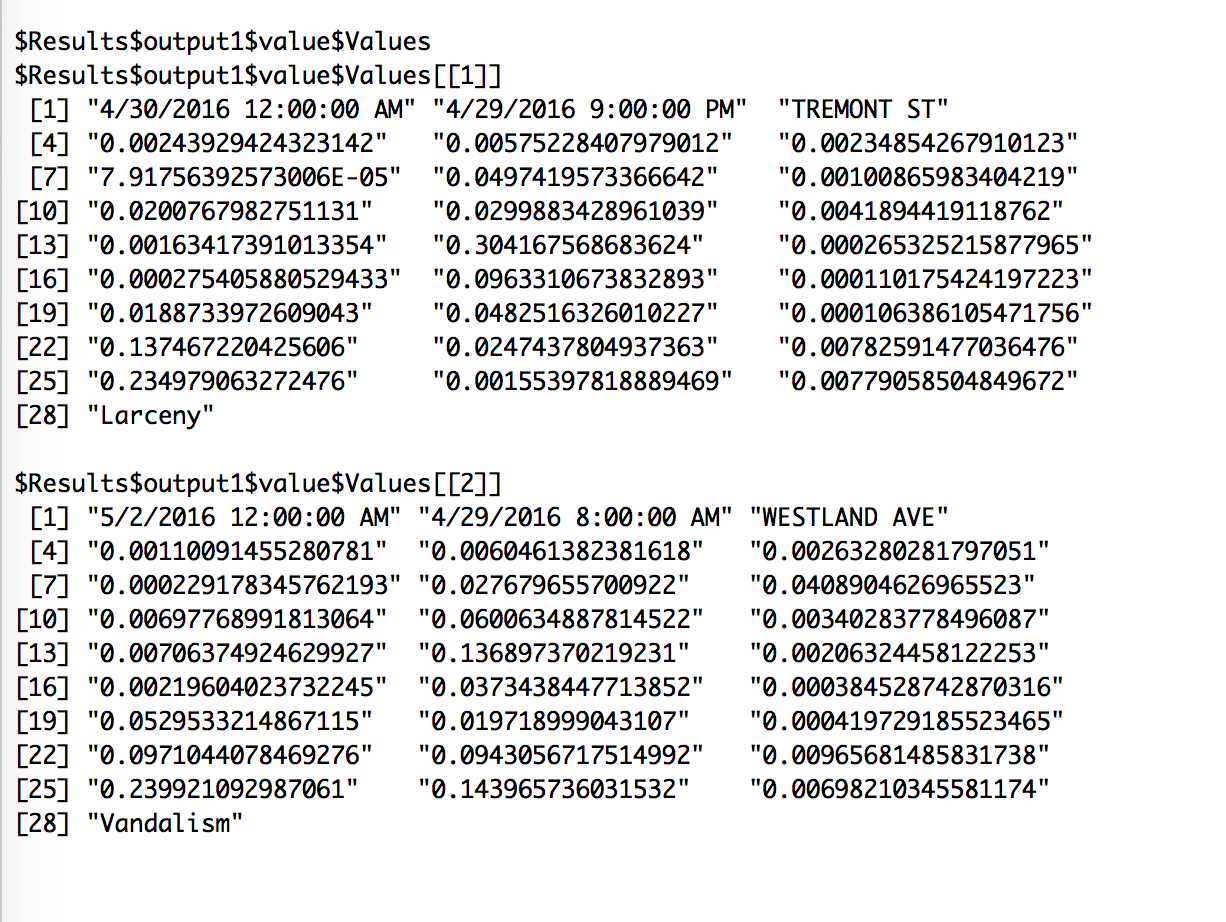
}

print("Result:")

result = h$value()

print(fromJSON(result))

**The output is as follows:**



**Input1:**

DATE = 4/30/2016

TIME = 21:00

STREETNAME = TREMONT ST

**Output:** Shows the probabilities of all the types of the “incidents” that can take place on TREMONT ST with the highest probability of “Larceny”.

**Input2:**

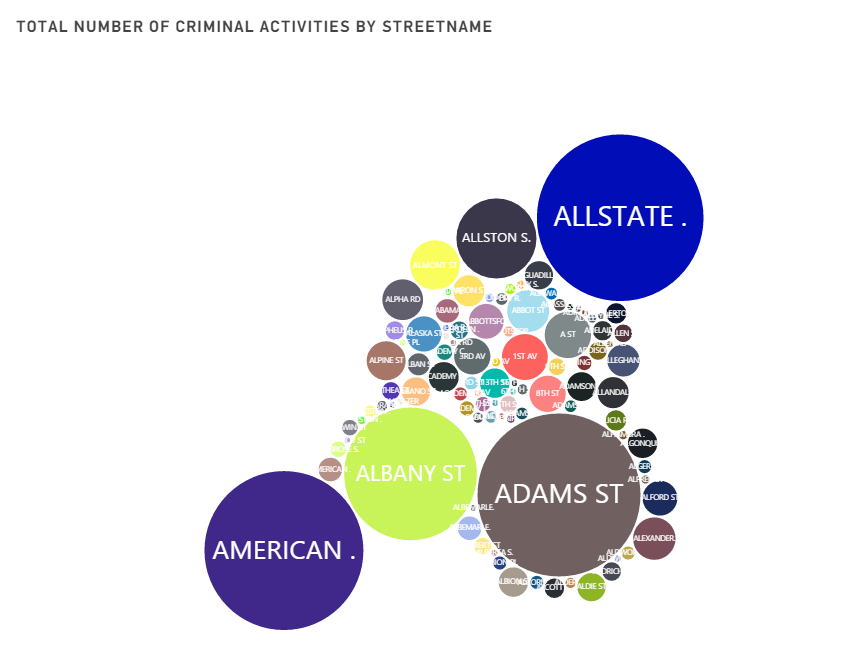
DATE = 5/2/2016

TIME = 8:00

STREETNAME = WESTLAND AVE

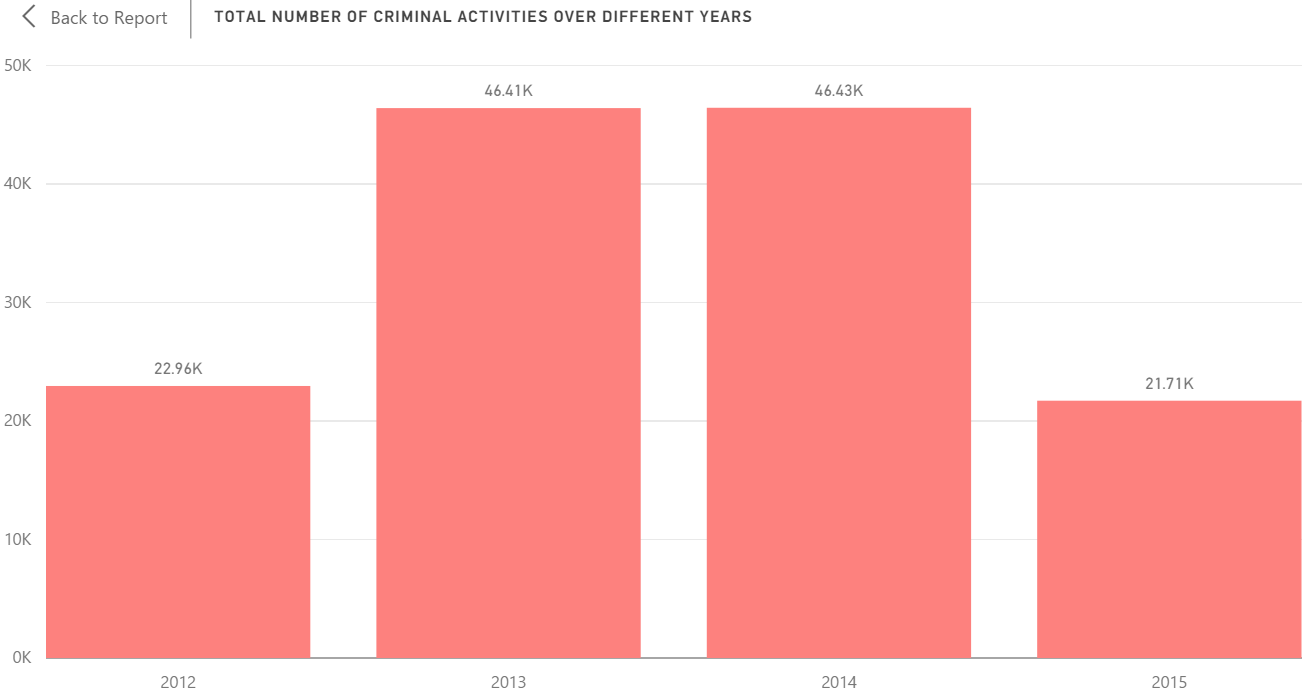
**Output:** Shows the probabilities of all the types of the “incidents” that can take place on WESTLAND AVE with the highest probability of “Vandalism”.

**[III] Exploratory Data Analysis**



(Figure: 1)

The above bubble chart shows the count of total number of different criminal activities occurring in Boston. The size of the bubble depends on the count of occurrence of the crime. From the above graph it is observed that Allstate, American St and Adam st are the most affected areas with respect to criminal activities.

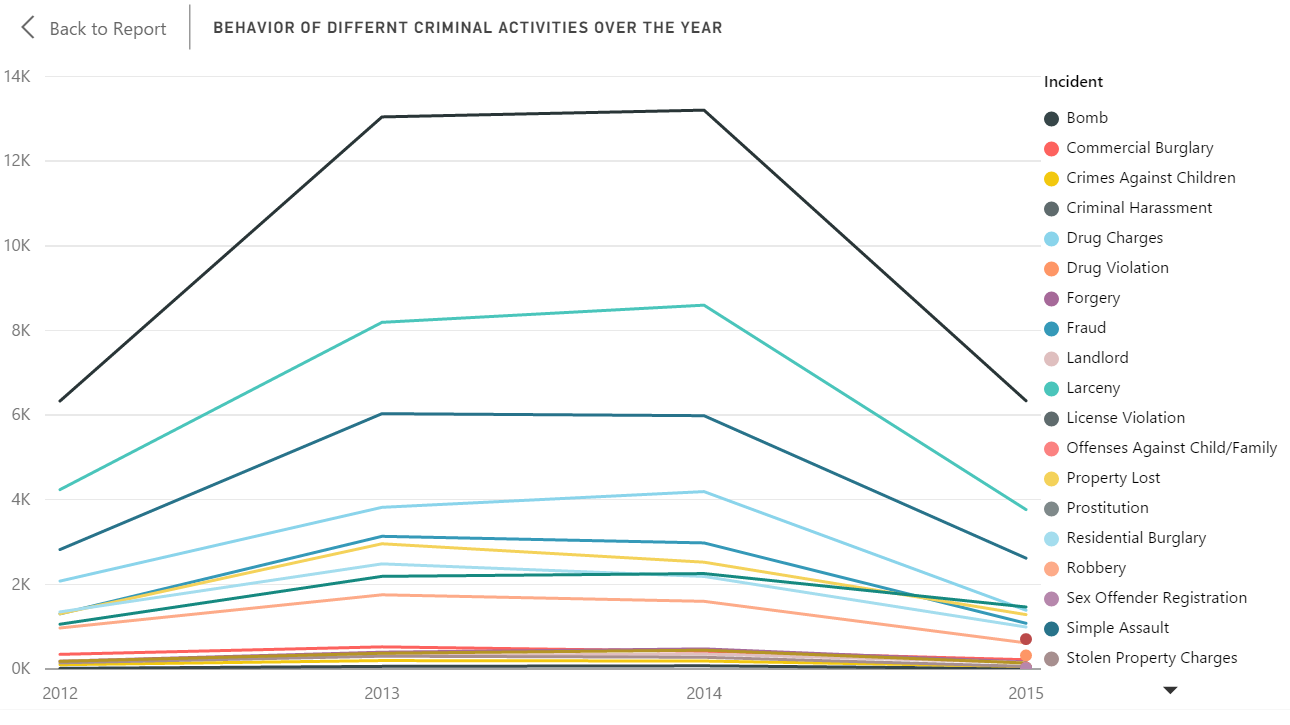


(Figure: 2)

This chart represents the total number of occurrence of different crimes over 4 year starting from 2012 to 2015. From the above graph it can be concluded that in 2013 and 2014 maximum criminal activities were recorded as compared to 2012 and 2015.This can be further drilled down to see the data of different quarters for a particular year and it can be further drilled to different moths and days. PFB the drilled down charts

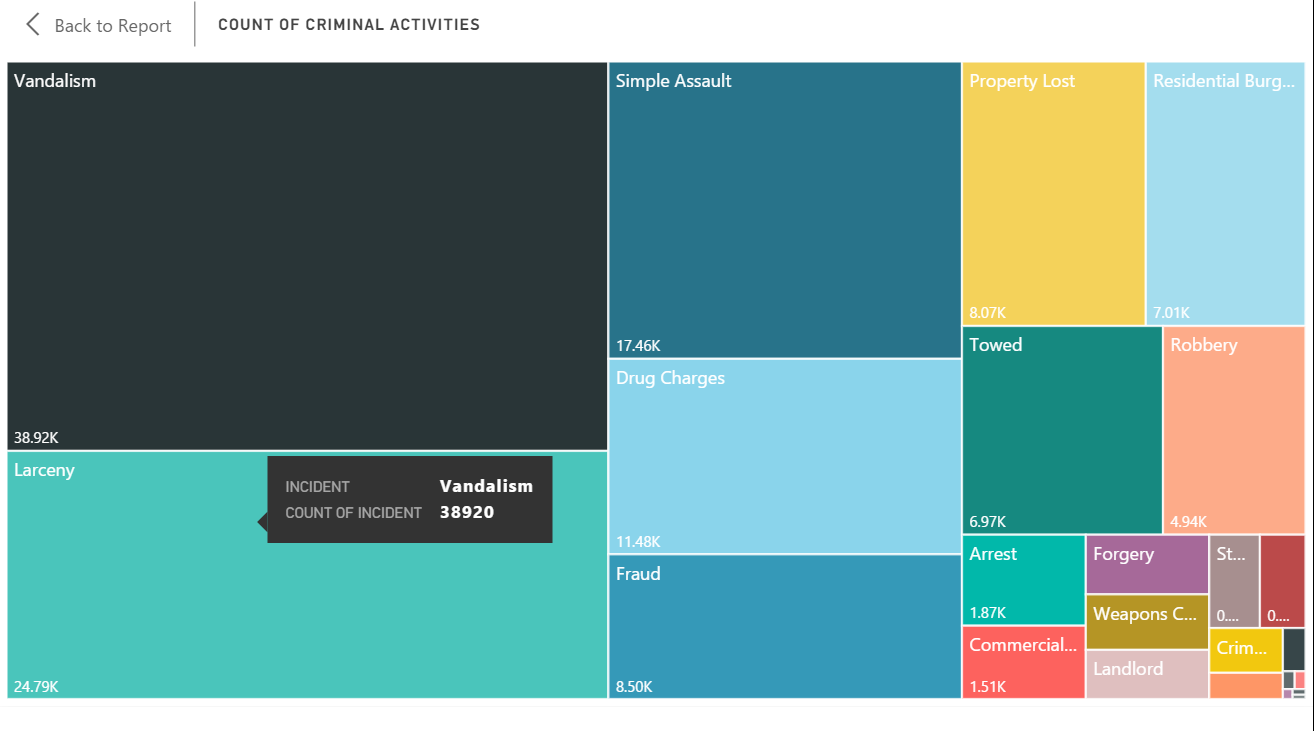
|  |  |  |
| --- | --- | --- |
| Crimes over Quarters | Crimes Over Months | Crime over Days |
|  |  |  |

(Figure: 3)



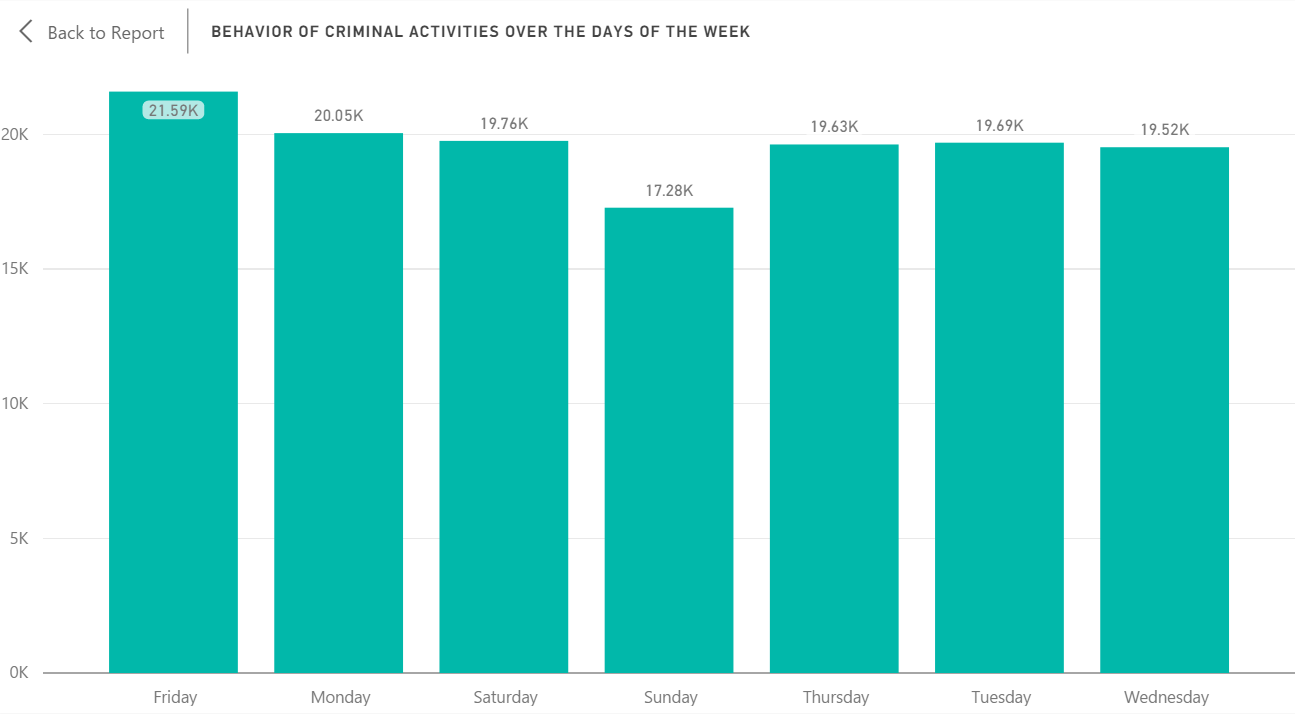
(Figure: 4)

The above graph represents the trend of different criminal activities over different years. From the above graph it is seen that there is an increase in the criminal analysis from the year 2012 to 2013 and its quite steady from 2102 to 2013 and there is drop in the criminal activities from 2014 to 2015.



(Figure: 5)

The above chart classifies different categories of crimes depending on their occurrences over 4 years. From the chart it’s observed that Vandalism is the most prevalent crime over these 4 years.



(Figure: 6)

The above chart provides behavior of criminal activities depending on the days of the week. From the chart it’s observed that highest number of activities occur on Fridays.



(Figure: 7)

The above graph shows the behavior of criminal activities over different months. From the graph it’s seen that the crime rate is quite high in the month of July.

**[IV] Front End Integration:**

Technologies used: C#, Bootstrap, CSS, HTML

Tools used: Visual Studio, Power BI and MS Azure Web App

<http://bostoncrimeanalysis.azurewebsites.net/>