Springboard Data Analytics Course

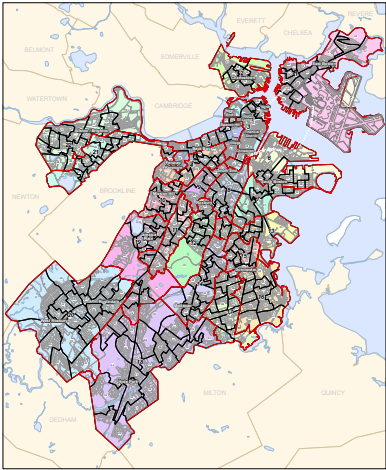
Final Report for Capstone II

Comparing Ontime and Overdue Boston 311 Calls

November 2019

John L. Parsons

**Map of the Boston Wards and Precincts in Massachusetts**



Contents

[**Introduction** 3](#_Toc24317145)

[**1.1 Problem Statement** 3](#_Toc24317146)

[**1.2 Dataset Summary** 5](#_Toc24317147)

[**Data Wrangling** 7](#_Toc24317148)

[**2.1 Missing Values** 7](#_Toc24317149)

[**2.2 Transformation and Binning** 8](#_Toc24317150)

[**2.2 Time Analysis** 13](#_Toc24317151)

[**2.3 Geo Pandas** 15](#_Toc24317152)

[**2.4 Correlation Plots** 15](#_Toc24317153)

[**Results** 17](#_Toc24317154)

[**4.1 Completion Time Analysis** 17](#_Toc24317155)

[**4.2 Completion Time Analysis for Boston and Chicago Departments** 19](#_Toc24317156)

[**4.3 Model Development** 21](#_Toc24317157)

[**4.4 Baseline Model** 22](#_Toc24317158)

[**4.5 Model Optimization** 23](#_Toc24317159)

[**4.6 Champion Model** 24](#_Toc24317160)

[**4.7 Geo Pandas** 27](#_Toc24317161)

[**Discussion** 30](#_Toc24317162)

[**Conclusion** 38](#_Toc24317163)

[**References** 39](#_Toc24317164)

[**Appendix** 40](#_Toc24317165)

# **Introduction**

## **1.1 Problem Statement**

This Final Report will be the first of several audits that will be conducted from an independent agency to show the residents a more accountable and transparent government.

This audit is a three-year analysis for all 311 Boston Calls made from October 1, 2017 to October 31, 2019 to determine how many calls were completed ONTIME compared to OVERDUE Calls for this three-year period. 311 Boston is a Government service that connects Boston Residents with highly trained Constituent Service Representatives for all non-emergency City Services and the goal of this study is to determine how effective this agency is in completing non-emergency issues ONTIME and within budget. The four milestones for the Internal Audit are;

1. Analyze the time (days) it takes to complete all 311 service calls for the three-year and the 2019 time period. The close\_open\_diff variable was created to measure the length of time to complete all calls for Boston and will be compared to San Francisco, Chicago and New York City in 2019. This time analysis will determine if Boston’s completion time rates for 311 calls is similar to other cities or they are significantly longer.
2. Develop a champion machine learning model that will accurately predict which 311 calls made from Boston Residents have a higher probability for not being completed on time. The higher risk calls will be forwarded to the special task force that will be qualified to handle problem cases to minimize the time it takes to complete these incidents.
3. Map specific locations using the latitude and longitude coordinates with Geo Pandas for select overdue 311 Calls in the Boston Area to determine if location may be a factor in length of time for completing these cases.
4. Submit a Final Report that will list possible areas that can be improved on to cut cost and time it takes the city to service 311 calls and be more efficient to the public.

This Final Report will contain the condensed version of Milestone Reports 1A and 1B, the results from the internal audit and the recommendations to reduce overdue completion time rates. The complete data wrangling and modeling results from these two milestone reports can be seen in the following GitHub repository link: <https://github.com/JLorenzP/Capstone-II-Project>.

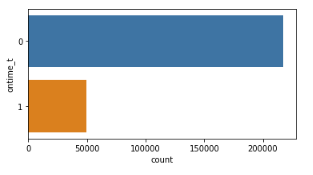
## **1.2 Dataset Summary**

The 311 Boston Open Source Data file can be found from the following website: <https://data.boston.gov/dataset/311-service-requests/resource/2968e2c0-d479-49ba-a884-4ef523ada3c0>. The 311 Service Requests was created by the City Constituent Relationship Management System (CRM) in 2011 and is updated daily. The 311 Boston open source data currently contains a total of 1,629,245 data entries and the data file contain 30 attributes. This study will only review the data from October 1, 2017 until the October 31, 2019 and this data file contains 267,312 entries of data. The Data Quality Report can be seen in Figure 1A, 2A and 3A and shows a list of the original variables and the ones that have been transformed to complete the analysis for this audit. All attributes that end with a “\_t”,” \_t1” or” \_t2” have been transformed and will be discussed in the data wrangling section of this report. The following is a brief description of the data set and the description can be seen in Figure 2.

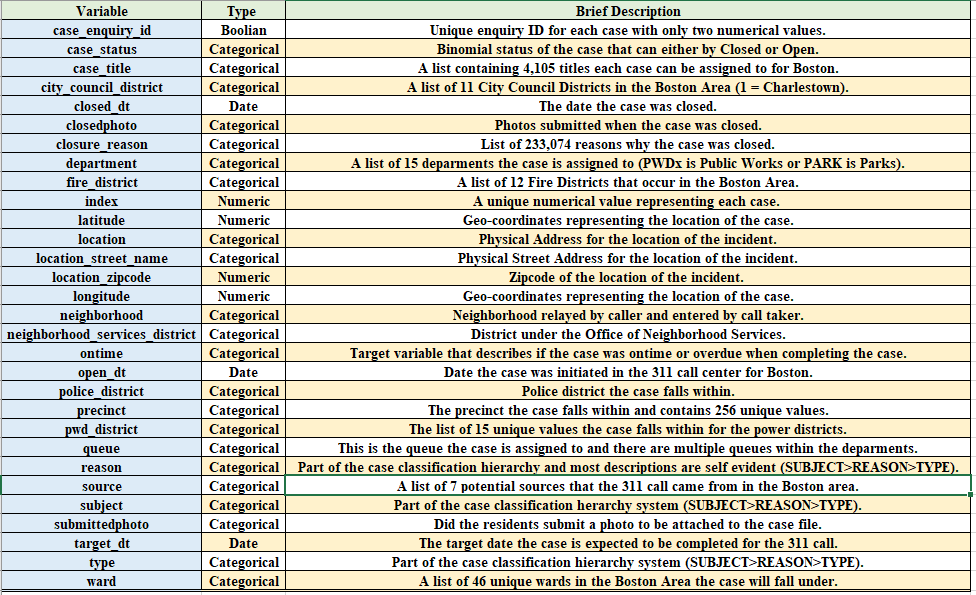
The ontime variable is a categorical attribute and will be the target variable for this study. This is a binary target variable and has been transformed to ontime\_t to convert the text to a binary value. The ontime\_t value will be represented with a level of 0 and the overdue value will be represented with a value of 1. This is an imbalanced dataset and level 0 contains 217,605 rows of data and level 1 only contains 49,707 rows of data. The remaining 29 attributes contains a total of 21 categorical variables ranging from a binary to a multifactorial variable, four numerical attributes, three dates and one Boolean attribute for the data set.

The 29 input variables can be divided into location, time, description and unique identifiers. The location input variables had many different levels of descriptions such as the physical location of the event or the district the event took place.

**Figure 1: Bar plot of the Target ontime\_t variable.**



**Figure 2: Boston 311 Description of the 30 Attributes.**



The specific location was recorded as the latitude, longitude, location, location\_street\_name, and location\_zipcode for the 311 events and will not be used for model development. The district location was divided up into a total of five districts found in the Boston Area. These five districts are; city\_council\_distict, fire\_district, neighborhood\_services\_distict, police\_district and pwd\_district. The last type of description for the 311 calls was placed into a case hierarchy system. This hierarchy system was subject (denotes which department the case is assigned to), the reason of the incident and then the type of incident that occurred. The time attributes for the Boston 311 dataset were given as the open\_dt, target\_dt and the closed\_dt of the event for the dataset. The open\_dt is the time the 311 case was opened and then the Boston Government gave a specific time the case was expected to be completed and the closed\_dt is the time the case actually took to complete. The unique identifiers for the dataset was the index, case\_enquiry\_id and case status and will not be used in model development. The dataset also contained photos of the event if submitted by Boston residents and government officials. These input variables were submittedphoto and closedphoto. The complete description of the attributes can be seen in Figure 2 and Figures 1A, 2A and 3A in the Appendix.

# **Data Wrangling**

## **2.1 Missing Values**

The Boston 311 dataset had a total of 16 attributes with missing values and these are; ontime, fire\_district, city\_council\_district, police\_district, pwd\_district, neighborhood\_services\_district, neighborhood, location\_zipcode, ward, precinct, case\_title, closedphoto, location\_street\_name, submittedphoto, target\_dt and closed\_dt. The list of the missing values and percent missing can be seen in Figure 4A in the Appendix. The missing values that had counts less than 35 were removed from the data set and these are; ontime, city\_council\_district, neighborhood\_services\_distict, ward and case\_title. The date and time variables target\_dt and closed\_dt had a total of 33,600 and 31,983 missing variables and were included in the dataset for descriptive, bivariate and model development phase. The missing values from the target\_dt and closed\_dt variables were removed for the time analysis section of this report to determine the average number of days it took to complete the case.

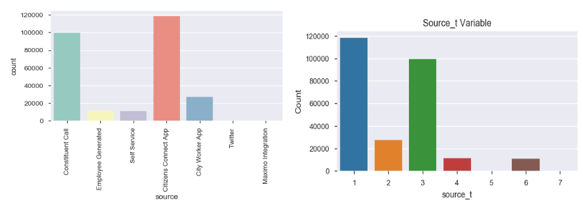
There are five input variables that had missing values and were not removed from the dataset. These five input variables with the number of missing values in parenthesis are; fire\_district (615), police\_district (125), pwd\_district (130), neighborhood (325) and precinct (194). These attributes with missing values had the missing value replaced with 0 and then were binned into a specific value called 33. The 33 bin will be used for most of the attributes selected for model development.

## **2.2 Transformation and Binning**

The input categorical variables in the dataset had descriptive names to explain the category like “Traffic Division” for subject or 10A for the pwd\_district that needed to be replaced with an integer for analysis. The input variables had a lot of missing values represented with ‘ ‘ or NaN in the dataset that needed to be removed (discussed earlier) or binned into a distinct value. The categorical variables also had a lot of sparse categories that will lower model performance and will be binned into a group to help increase model performance. The complete transformation and binning process can be seen in Milestone Report 1B and a brief description of this process will be discussed below.

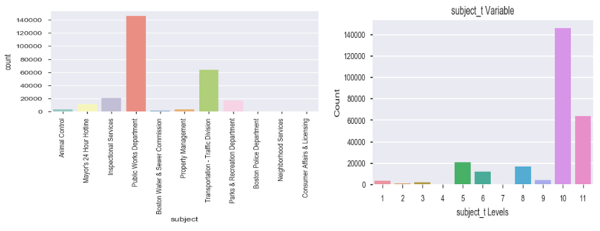
The source variable, subject and department attributes did not have any missing values for this data set. The source variable had a total seven levels and all seven levels can be seen in Figure 3 below. The levels were arranged in alphabetical order and then assigned a number from 1 to 7 and placed in a new attribute called source\_t. Levels 5 (Maximo Integration) and 7 (Twitter) were very sparse and another variable was created called source\_t2 and level 5 and 7 were binned into level six (Figure 3).

**Figure 3:** **Count plot for the source and transformed source\_t variables.**



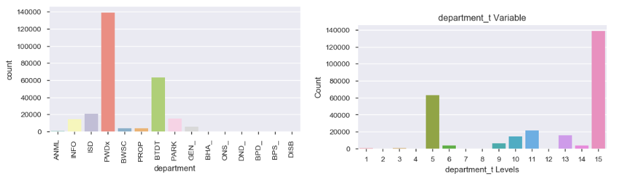
The subject attribute had a total of 11 levels for the data set. The subjects were arranged in alphabetical order and then replaced with a value of 1 thru 11 and placed in the new variable called subject\_t. The levels 1 to 4 were very sparse and then another attribute called subject\_t2 was created and all four of these levels were binned into level 7 to increase model performance. The results can be seen in Figure 4 for the transformation and binning of the subject variable.

**Figure 4: Count plot for the subject and transformed subject variables.**



The department attribute had a total of 15 levels and was arranged in alphabetical order and replaced a number for a name (1 for “ANML” up to 15 for “PWDx”). The count plot for the original and replaced values can be seen in Figure 5. The department\_t2 attribute was created to bin the sparse levels in the dataset. Levels 1, 2, 3, 4 and 7 were all binned into level 8 for this variable. An explanation of these values can be seen in Figure 5 and 5A.

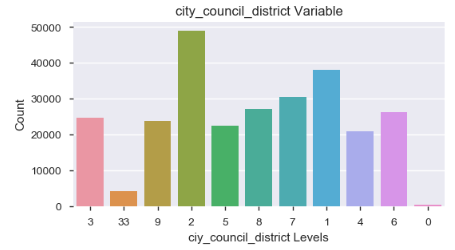
**Figure 5: Count plot for the department and transformed department\_t variables.**



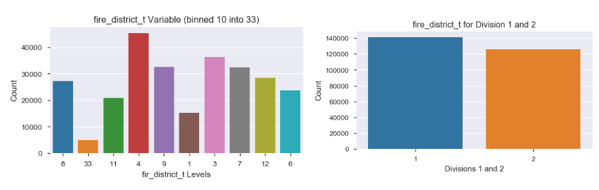
The city\_council\_district, fire\_district, pwd\_district, neighborhood\_services\_district and police\_district attributes had missing values listed as “ “ or nan and these were replaced the number 33 and will be used as the missing number bin for this study. The city\_council\_district had a total of 10 levels and missing values listed as ‘ ‘ in the data column. The results from the replacement method and binning of empty values to 33 can be seen in Figure 6. The level 0 for the dataset is also very sparse and was binned into the level 33 and this can be found in the new attribute called city\_council\_district\_t.

The fire\_district had a total of 10 levels and the results from the replacement method and binning of empty values to 33 can be seen in Figure 7. The level 10 for the dataset is also very sparse and was binned into the level 33 and this can be found in the new attribute called fire\_district\_t. The city of Boston has two Fire District Divisions containing nine Districts and these were placed in new attribute called fire\_district\_t2. Districts 1, 3, 4, 6 and 11 were placed in Division 1 and 7, 8, 9, 10, 12 and 33 were placed in Division 2 (Figure 7). District 10 is not originally from the two Divisions and 33 is the empty values for data set. The names of the districts can be seen in Figure 6A in the Appendix.

**Figure 6. Count plot for the city\_council\_district variable.**



**Figure 7: Count plot for the fire\_district\_t and transformed fire\_distirct\_t2 variables.**



The pwd\_district had a total of 13 unique levels in the dataset and had a mixed set of names which mostly used numbers to represent the district but also contain letters A or B for 1 and 10 pwd\_districts. The pwd\_district\_t was created and the letters from these pwd\_districts were replaced with a unique number starting with 20 for 1A to 24 for 10B.

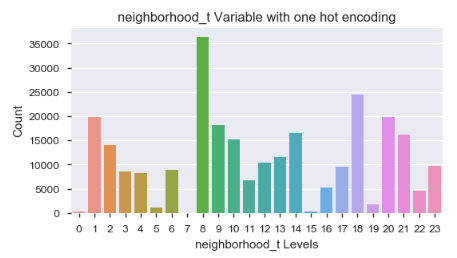
The neighborhood\_services\_district attribute contains the name of all 17 neighborhoods (including missing values) relayed by the caller and entered into the data base. All values in this variable were integers and will be used for the analysis.

The police district is the location where the case falls within for the Boston residents. There is a total of 12 districts and are labeled as a letter attached to the number. The missing variables were binned into level 33 and then all letters were removed and placed in a new attribute called police\_district\_t.

The neighborhood, ward and reason attributes had between 24 and 44 levels and One Hot Encoding was used to create unique identifiers for these levels to be used for the model development phase of this report. The neighborhood attribute contains the name of all 24 neighborhoods (including missing values) relayed by the caller and entered into the data base. This attribute also contained missing values that were replaced with the name “UNKNOWN”. The attribute neighborhood\_t was created and then One Hot Encoding was used to assign specific values to the neighborhoods. The results can be seen in Figure 8 and the names of all the neighborhoods can be seen in Figure 7A in the Appendix. The number of calls for levels 0, 5, 7, 15 and 19 are sparse compared to the other neighborhood levels and all will be binned into level 19 in the new attribute called neigborhood\_t2 for model development.

**Figure 8: Count plot for the neighborhood and transformed**

**neighborhood\_t variables.**



The ward attribute had a total of 44 levels and is the election ward a case will fall into for this variable. The ward category had an integer designated to the ward or a letter and integer designated each ward. A new attribute called ward\_t was created and One Hot Encoding was used to create a unique integer for each ward.

The reason variable was a text describing what the problem was and there is a total of 46 different reasons from the raw dataset but only 26 listed for the last three years. This attribute did not have any missing values and another attribute called reason\_t was created and One Hot Encoding was used to create unique numbers for this variable. The list of all reasons can be seen

in Figure 8A and 9A (Appendix).

The last two attributes are the type and precinct for this study. The type of complaint had a total of 134 unique categories ranging from “Animal Lost” to “Walk-In Service Inquiry”. There were no missing values for this input variable. A new attribute type\_t was created and One Hot Encoding was used to generate unique numbers for analysis. The precinct variable had a total 255 levels and is the Election Precinct a case falls within for the 311 calls. There were missing values for this variable and these were replaced with the number 33 and precinct 0502A

was replaced with 0502001. A new attribute called precinct\_t was created and One Hot Encoding was used to generate unique numbers for this variable.

The last variable of interest was the submittedphoto category and this had a hyperlink attached to a picture which could not be uploaded or it was empty with no link to a photo. A new attribute called submittedphoto\_t was created and all missing rows of data was given a level of 0 (no photos) and all rows of data with a submitted photo was given a level of 1.

## **2.2 Time Analysis**

The three attributes that were listed as time and date the call was made is classified as an object in python. These three attributes are open\_dt, target\_dt and closed\_dt for the Boston 311 calls. The target\_dt had a total of 33,600 missing values and the closed\_dt had a total for 31,933 missing values. The raw data set was converted to the datetime function and the dataframe was filtered on all rows with an open\_dt greater or equal to October 1, 2017 for analysis. The two “time” attributes with missing values were removed and placed in another file called time. There was a total of 207,386 cases and only 28,168 of these were classified as overdue when all missing values were removed. The time dataset was separated into the on\_time and the over\_time data files and contained either all level 0 (ontime) or level 1 (overdue) attributes for the time analysis section.

Three large metropolitan cities were chosen to compare the mean number of days for the 2019-year to complete a project and determine how Boston’s completion compares to these cities. The three cities are Chicago, New York City and San Francisco. All three files came from the Open Source Data and was originally filtered on date from the website. The open and closed dates were converted to a datetime stamp in python and all opening dates that started in the year 2019 were selected for the study. A new attribute called close\_open\_diff for Boston and closed\_created\_diff (the names are slightly different) was created for the other three cities and is referred to as completion time for the entire year. All missing values and values that were negative were removed from the dataset for the closed\_created\_diff categories. The description of the completion time, bar graphs and student T Tests were used to compare Boston’s completion time to the other three cities.

The completion time was drilled down to the Public Works and Transportation Department for the Boston and Chicago datasets for the 2019 year. The PWDx contains a total of 121,891 cases and the BTDT contains a total of 54, 857 cases out of a total of 217,605 calls for a three-year period (please note the missing time values were dropped). These two departments represent 81.2% of all 311 Boston calls and is standard department between cities. The PWDx and BTDT is the Public Works and Transportation Department for Boston. The DWM and CDOT are the Public Works and Transportation Department for Chicago. The description of the completion time, bar graphs and student T Tests were used to compare Boston’s to Chicago completion time for the Public Works and the Transportation Department

## **2.3 Geo Pandas**

The attributes used for the Geo Pandas was the latitude and longitude for location and the target variable ontime\_t. The latitude and longitude attributes did not have any missing values and every call was assigned a specific coordinate. A base map had to be downloaded from the

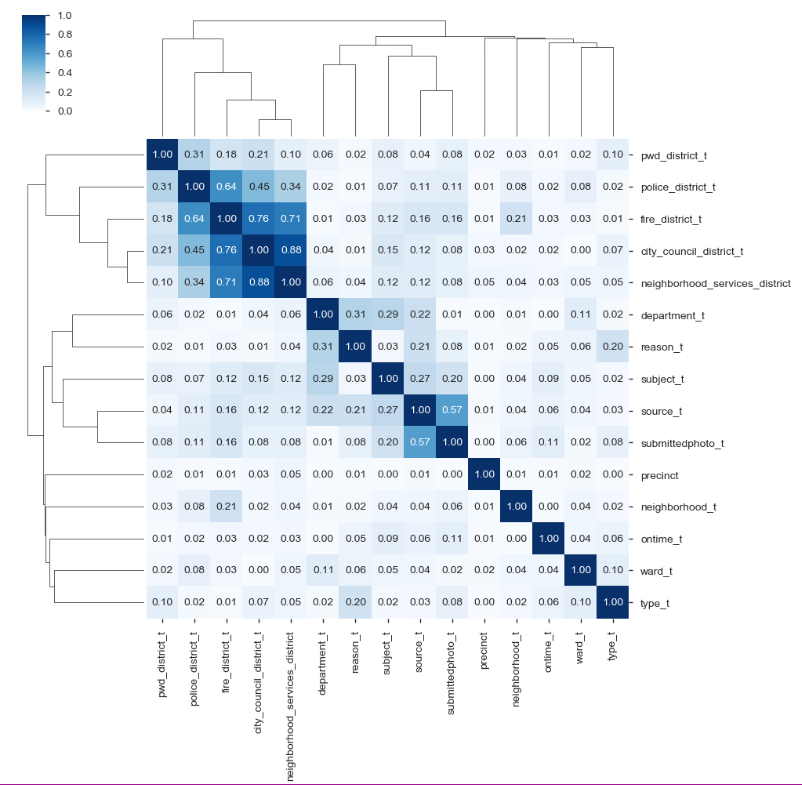
Boston Open Data web site and the two base maps that will be used is the Fire District map and Ward Map. The data was separated into ontime and overdue cases and the coordinates were plotted for all data points in this file.

## **2.4 Correlation Plots**

The pdf file for Boston 311 discusses how several cases could be placed under different attributes such as districts. The correlation cluster map in Python was created to determine which attributes were closely related to each other and will not be used in the final model. The Cluster map can be seen in Figure 9 and the five districts that are grouped together have a higher correlation than the other attributes except for source\_t and submittedphoto\_t (57%). The city\_council\_district\_t and neighborhood\_district\_t had a correlation value of 88% and the next highest value was between fire\_district\_t and city\_council\_district\_t with a value of 76%. The last variable with a high correlation value is fire\_district\_t and neighborhood\_district\_t with a value of 71%. These will not be used together in the final champion model to reduce multicollinearity from the dataset.

The correlation values were generated between the target and the input variables in Python and the LogWorth values in JMP SAS to help in the attribute selection process for the champion model. The target variable ontime\_t was compared to the input variables that were

**Figure 9: Seaborn Cluster Map for 15 variables.**



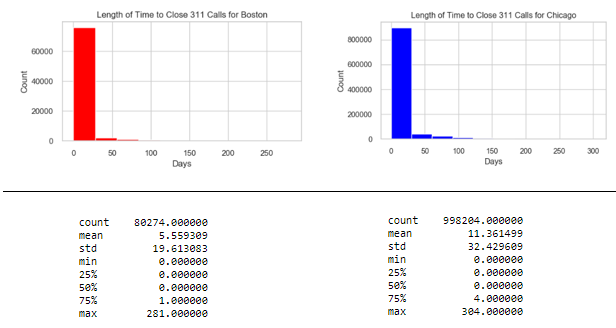
transformed in the data set and the number of days (time differential) it took to close the case between all three time-date input variables was only included in the first correlation matrix. The open\_dt, closed\_dt and target\_dt time-date stamps was only included in the LogWorth values for JMP SAS. The time-date input values will only be used as the metric to be compared between the three metropolitan cities for time to completion for 311 cases and the Geo Pandas. The results can be seen in Figure 10A in the Appendix. These results show the top six correlated values (excluding time differential attributes) between the input and target variables are; submittedphoto\_t, subject\_t2, source\_t, type\_t, reason\_t and fire\_district\_t2.

# **Results**

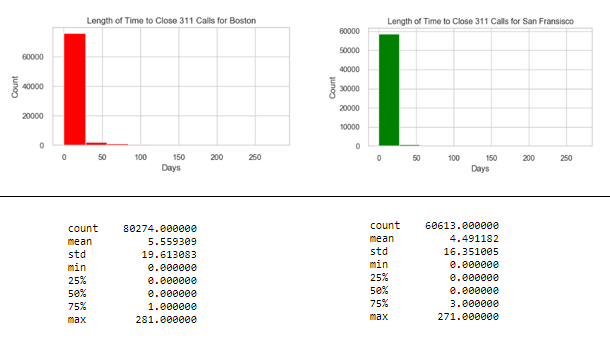
## **4.1 Completion Time Analysis**

The histogram and descriptive statistics can be seen in Figure 10, 11 and 12. The mean number of days to complete the case for Boston was 5.6 compared to 11.4 for Chicago, 4.5 days for San Francisco residents and 8.0 days for New York City Residents. The minimum time to complete the case was 0 days for all four cities and the maximum time to complete a case was 304 days for Chicago. Chicago had the greatest dispersion around the mean with the standard deviation of 32.4 days and San Francisco had the lowest standard deviation value of 16.4. Boston had the second lowest standard deviation with a value of 19.6. The maximum values for the 2019 year was between 271 days for San Francisco and 304 days for Chicago. The students T test for independent samples was used to compare the means for all three cities and the cv (critical values) and p-values were all significant with a p-value of less than 0.0 (Figure 13).

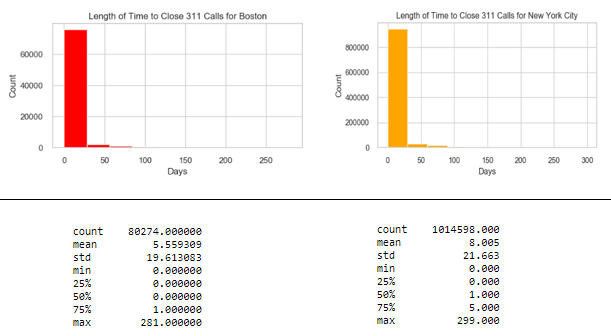
**Figure 10: Histogram and Descriptive Statistics of 311 calls for Boston and Chicago.**



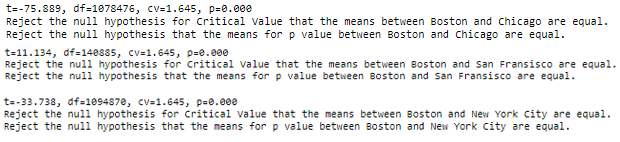
**Figure 11: Histogram and Descriptive Statistics of 311 calls for Boston and San Francisco.**



**Figure 12: Histogram and Descriptive Statistics for 311 calls for Boston and New York.**



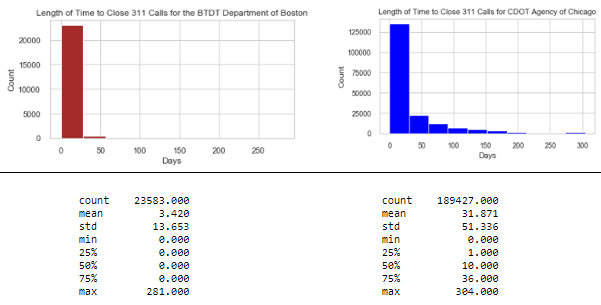
**Figure 13: T Test for Independent Samples Comparing the mean Completion Time for 311 Calls.**



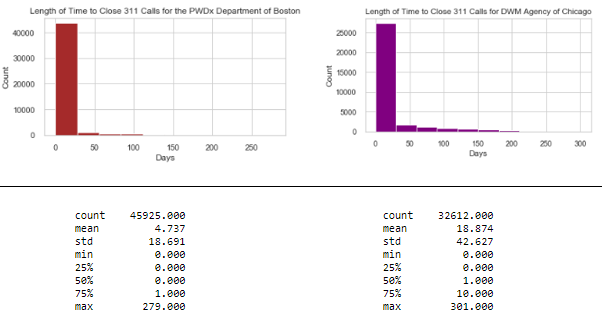
## **4.2 Completion Time Analysis for Boston and Chicago Departments**

The results for these comparisons between Boston and Chicago for the Public Works and Transportation Department can be seen in Figures 14, 15 and 16. Boston PWDx mean number of days to complete a call was 3.4 compared to 31.9 for Chicago and the maximum number of days to complete a project was 281 compared to 304 days. The mean number of days to complete the 311 calls for the PWDx was 4.77 compared to 18.8 for the Chicago DPM and the longest time to complete the case ranged from 279 to 301 days. The Students T Test for independent samples compared the means for the PWDx and DWM and both were significant for the cv and p-value of mean days to complete the project. The results show that Boston does a good job in completing projects for these two departments when compared to Chicago.

**Figure 14: Histogram and Descriptive Statistics for 311 calls for Boston and Chicago for between the BDTD and CDOT.**



**Figure 15: Histogram and Descriptive Statistics for 311 calls for Boston and Chicago for between the PWDx and DWM.**



**Figure 16: T Test for Independent Samples Comparing the Completion Time for 311 Calls**

**Between Boston and Chicago for the Water Departments.**



## **4.3 Model Development**

The Random Forest Model Classifier will be used to create the base line model and the f1 score and classification report results will be the metric used to determine model performance. The model development section will have three phases for the evolution of the champion model. The first and second stage will contain a group of seven models to determine the best f1 score for the ontime\_t variable and variable importance from the Random Forest Model. These seven models are the Random Forest, KNN Nearest Neighbor, Extra Tree Classifier, Support Vector Machines, Ada Boost and Gradient Boosting and multilayer perceptron. The second iteration will contain the same seven models but will only have select attributes from based on previous feature importance values, correlation matrix and logworth values. The third iteration will contain the Ada Boost, Extra Tree Classifier, Random Forest Model and Gradient Boosting for the hyperparameter testing to find the best model. The final two models will be selected with the best f1 scores and tuned parameters to generate the champion model for detecting ontime\_t cases in the Boston 311 Dataset and will use the balanced approach for the unbalanced data set to improve model performance.

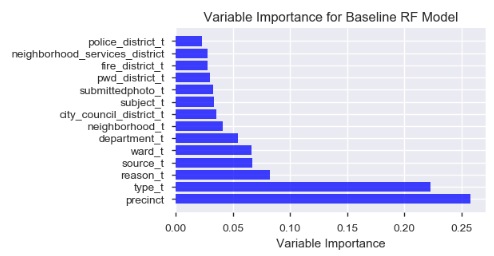
The base line model and first stage of model optimization will be using 14 attributes from the dataset. These 14 input attributes and target variable are; source\_t, city\_council\_district\_t, fire\_district\_t, subject\_t, pwd\_district\_t, department\_t, neighborhood\_services\_district, police\_district\_t, neighborhood\_t, ward\_t, reason\_t, type\_t, precinct\_t, submittedphoto\_t and ontime\_t. Only seven input variables were used for the second stage and hyperparameter optimization models and these are (including the target variable); ontime\_t, precinct\_t, type\_t, reason\_t, ward\_t, source\_t, neighborhood\_t and department\_t. The entire dataset was used for the baseline model and the champion models. The first and second stage only used 20% of the data set and 26,960 samples were used as the training set and 14,517 were used as the valuation set (65:35 Split). The third optimization stage used a total of 25% of the data and 33,699 samples were used for the training set and 18,147 samples were used for validation set. A five kfold cross validation was used to tune the models and the f1 score will be the accuracy measure for the three phases of model development and the champion model.

## **4.4 Baseline Model**

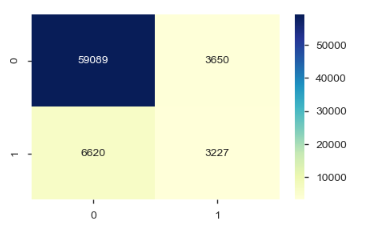
The Random Forest Model results can be seen in Figures 17, 18 and 19.

The top seven attributes were precinct, type\_t, reason\_t, source\_t, ward\_t, department\_t and neighborhood\_t. The Confusion matrix and Classification report shows the model does a good job in classifying the True Negative values (TN) with an accuracy score of 85.9%, but has a hard time classifying the True Positive Value (TP) and misclassified 6,620 of the False Negatives (FN). The f1 score for the model was 38.6%.

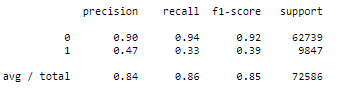
**Figure 17: Base line Variable importance for the Random Forest Model**



**Figure 18: Confusion Matrix for the Baseline Random Forest Model.**



**Figure 19: Classification Report for the Baseline Random Forest Model.**



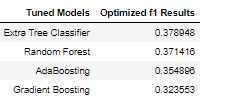
## **4.5 Model Optimization**

The first stage for model optimization contained all 14 attributes and the f1 scores for the seven machine learning models ranged from 9.7% for the Extra Trees and 33% for the SVC Classifier. The top seven variable importance attributes are precinct, type\_t, reason\_t, ward\_t, source\_t, neighborhood\_t and department. The variable importance values ranged from 4.6% up to 26% and were used for the second stage of model development. The Random Forest in the second model had the highest f1 score with 34.9% and the SVC was slightly better from the previous run with a score of 33.7% for this optimization run. The precinct\_t is still the most important variable and department\_t is the least important factor for the Random Forest Model.

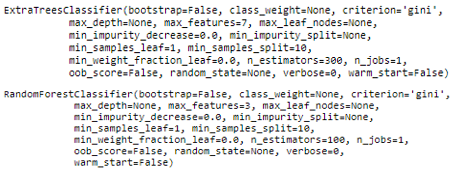
The hyperparameter optimization of the four models was the final stage to boost model performance and select the best model. The results from these runs shows an increase from

8.6 % up to 37.9% for the Extra Trees and 34.9% up to 37.1% for the Random Forest. The Extra Trees and Random Forest Models were the top models and the results can be seen in Figure 20. These two models will be used for the champion model and will include the best model parameters listed in Figure 21. The class\_weights will be for the balanced approach to improve the performance of the model and find the champion model.

**Figure 20: Hyperparameter f1 Results.**



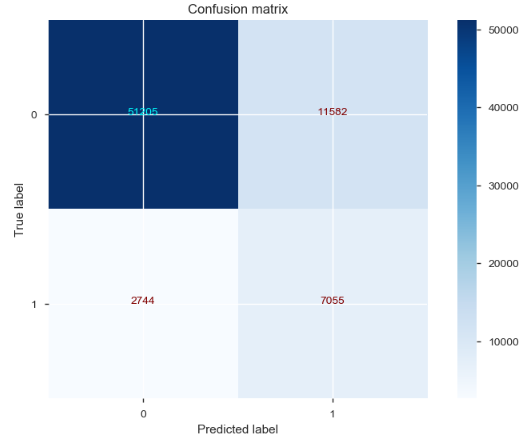
**Figure 21: Extra Trees and Random Forest Best Model Parameters**



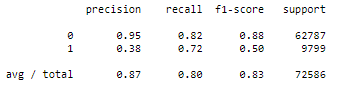
## **4.6 Champion Model**

The Extra Tree and Random Forest Model used the best parameters from the previous run and the balanced approach for the class\_weight. The model that had the best performance was the Random Forest model for the last model discussed and the results from the initial Confusion Matrix and Classification Report can be seen in Figures 22 and 23. The Random Forest had an accuracy rate of 80.2% and the f1 score was 49.7%. The Random Forest baseline model had a f1 score of 38.6%. The optimized model increased f1 performance and did a better job in classifying the TP values for the data set. The model did not do well in classifying the FP values for the data set and is an area that needs to be improved on to increase the f1 scores. The Extra Tree Model had an f1 score of 52.3% and the accuracy rate was 83% and they both did a good job and were a better model than the base line model. The results from the Extra Trees Classifier can be seen in Figures 24 and 25.

**Figure 22: Confusion Matrix for Random Forest Model.**



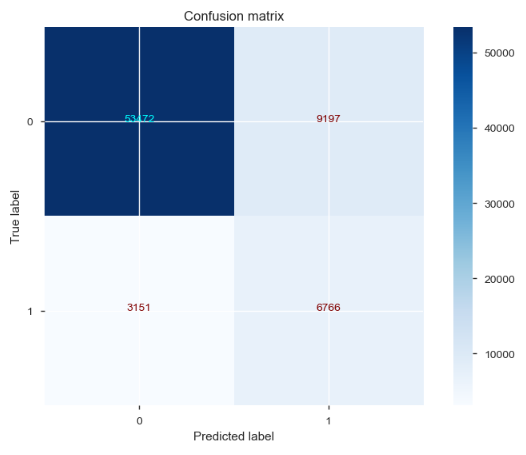
**Figure 23: Classification Report for the Random Forest.**



The correlation matrix had the date-time variables with the highest values compared to the other input variables. The day, week and month variable attributes were created from the open\_dt time variable and tested with the previous Random Forest Model and the same parameters. The only two changes for this model was the addition of the week input variable and the binned neighborhood\_t2 variable. The f1 results increased to 54.5% and the Accuracy Score was 84%. This model did a better job in selecting only 8,505 FP compared to 11,582 FP values, but had 275 more FN values than the previous model. The results can be seen in Figures 26 and 27.

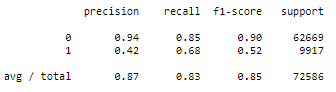
**Figure 24: Confusion Matrix for the Extra Trees**

**Classifier Model.**



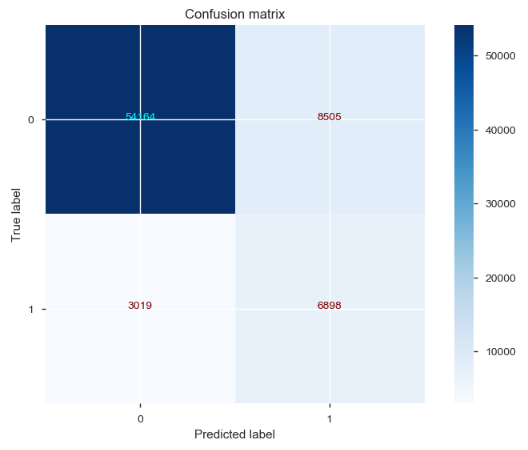
**Figure 25: Classification Report for the Extra**

**Trees Classifier Model.**



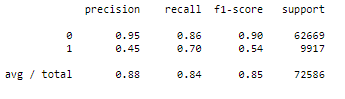
**Figure 26: Confusion Matrix for Random Forest with the**

**addition of the week and neighborhood\_t2 input variable.**



**Figure 27: Classification Report for Random Forest Model with the**

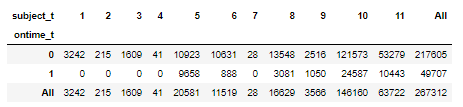
**Addition of the week and neighborhood\_t2 input variable.**



## **4.7 Geo Pandas**

The ISP (Inspectional Services listed as # 5) for subject and the DISP (Disabilities) for department were two areas that were evaluated in the Geo Pandas maps. The goal is to see if certain locations based on the longitude and latitude from the Fire District or Wards were not evenly distributed in the maps. The Inspectional Services had 9,658 overtime cases out of a total of 20,581 cases and was almost half of the number of cases that did not get finished ontime (Figure 28).

**Figure 28: PD Crosstabs for subject\_t and ontime\_t**

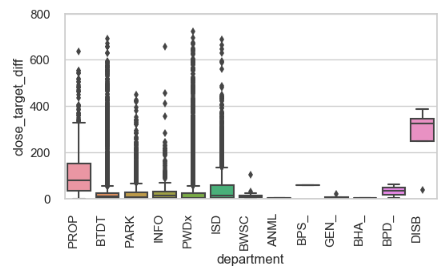


The box plot in Figure 29 shows the mean number of days calculated between the target date and closing date for overdue cases only and shows DISB has the greatest mean number of days compared to the other 12 attributes. The first six attributes ranging from PROP to ISD had more outliers and the PROP mean as the second highest mean compared to the other attributes in the graph. The GEO Panda maps were plotted for the ISP-Subject and DISB-Department and the results can be seen in Figures 30 and 31.

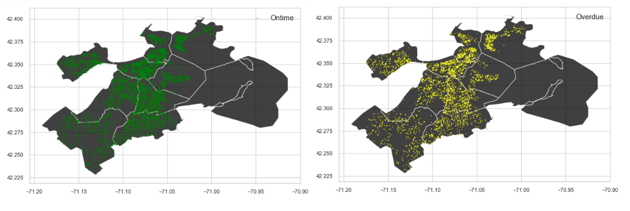
The overall spread of the data points for the ontime and overdue data points for the ISD-Subject does not show specific hotspots for either group of data points for the Fire District Map used in the analysis. The empty Fire District on the upper right side is called Fire District 1\_3 and did not contain any samples from this study of 311 calls.

**Figure 29: Box Plot for all overtime Department Variables for the mean number**

**of days calculated between the target date and the closing date for 311 calls.**



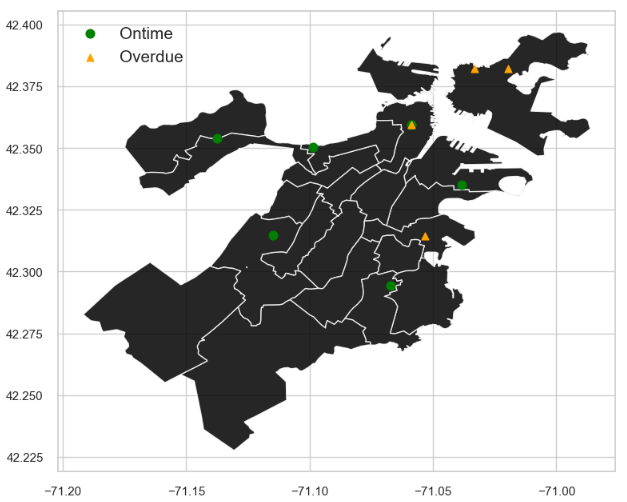
**Figure 30: Geo Pandas for Department filtered for ISD Cases only (Inspectional Services).**



The DISB-Department variable shows good separation between the ontime and overdue calls (Figure 31). When the missing times were removed from the data set only six ontime and four overdue variables were left and seven still have not been finished at this time. A total of three overdue cases were only found in Wards 1 and 13 and number 3 was shared with one ontime case. Wards 1 is East Boston and Ward 13 is North Dorchester. The Wards that only contained ontime cases are 5, 6, 17, 19 and one on the line for 21 and 22. This is another area that needs further investigation because this variable has the greatest number of mean days compared to the other variables and there is clear separation between three of the four overdue cases compared to ontime cases for the data set.

**Figure 31: Geo Pandas Wards for Department filtered for**

**DISB only (Disabilities – General).**



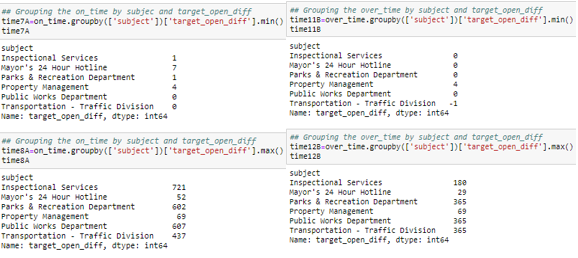
# **Discussion**

Completion time for closing all 311 calls is the metric used to compare Boston to three other metropolitan cities. The total average completion time in days which is measured by the total number of days it takes to close a case after it opens for the 2019 year. All cases that were not closed were removed from this analysis. Boston’s mean completion time to close all cases for the 2019 year was 5.6 days. The mean number of days to close a case ranged from 4.5 days for San Francisco up to 11.4 days for Chicago. The maximum number of days to close a case ranged from 271 for San Francisco up to 304 days for Chicago. The student T Test was used to compare the means for independent samples and all the means were significant when compared to the Boston mean completion time. The 311 data set was drilled down to the Public Works Department and the Department of Transportation because these two areas contained 70% of the overdue cases and this metric was similar when compared to Chicago. Boston completed the BTDT (Transportation) cases in 3.4 days compared to 31.9 days for Chicago’s CDOT Department. Boston only took 4.7 days for the PWDx (Public Works) compared to 18.9 for the Chicago DWM Department. Boston is doing a good job in completing the average number of 311 calls for residents when randomly compared to three large metropolitan cities in the United States.

The City of Boston has a unique attribute called target\_dt and this is the date the 311

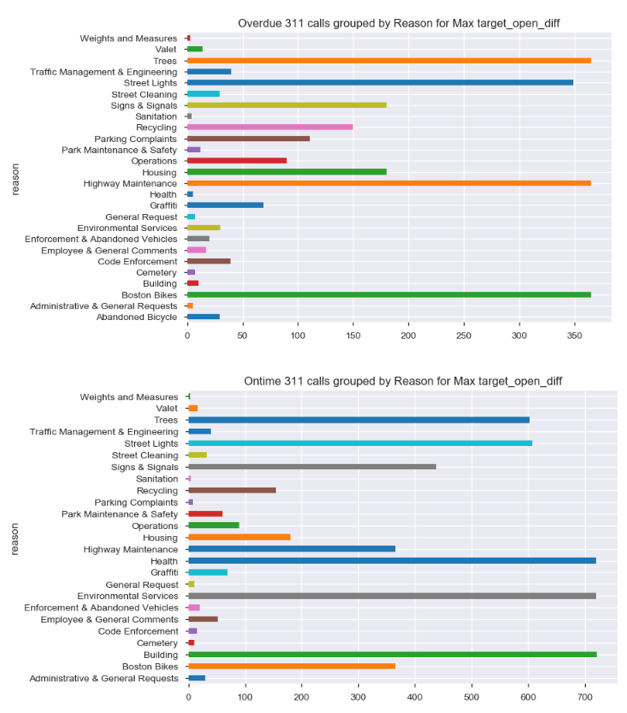
project should be completed as designated by the Constituent Service Representatives. This attribute was not used by the other three cities which only had the open and close date recorded in the data base for 311 calls. This is a great time variable to add for all 311 calls and gives the residents an estimation date of completion. The problem with the target\_dt is there is no documentation which states that if a call has these specifications then it will take “x” amount of days to complete. The Boston time data was separated into the ontime file (ontime\_t = 0) and the overtime file (ontime\_t = 1) and then several groupby functions were set up to determine the length of time it should take to close a case and get a baseline estimate for Inspectional Services or Public Works Department. The total number of days was calculated from the target\_dt and the open\_dt and was called the target\_open\_diff. In theory, the ontime mean number of days for the minimum and maximum number should be the range or estimated completion time for this category and should not be greater than the overdue maximum number of days. The groupby functions and formulas can be seen in Figure 32 below. The minimum number of days to for the target completion time from the opening date seems reasonable and is at 0 or less than or equal to 7 for both the ontime and overdue cases. There is also another negative day for the Transportation Department and is a common value seen with the other cities databases. The maximum number for the target completion time estimated to complete a case varies from each department and the ontime variable. The ontime maximum time varies from 52 days for the Mayor’s 24-hour hotline up to 721 days for the Inspectional Services. Now compare this to the overdue cases and the maximum number for the target completion time of days varies from 29 days for the Mayor’s hotline up to only 365 days for a total of three departments. The targeted number of days to complete a case is only 365 days for the Public Works Department and is considered overdue and yet the target\_dt maximum time is 607 days and it is considered ontime.

**Figure 32: Groupby functions for the min and max values of the ontime and overdue cases.**



The groupby function and reason for the call was plotted for the ontime and the overdue cases. These results did a better job in separating the mean days for the target date based on the ontime and overdue category. Figure 33 shows the maximum number of days for the reason by target variable and most of the ontime estimated completion times are less than the overdue target completion times. The raw data can be seen in Figure 13A in the appendix and will list the number of days by category. The Highway Maintenance and Boston Bikes appear to have an estimated target completion date of 365 days and the Operations is 90 days based on these

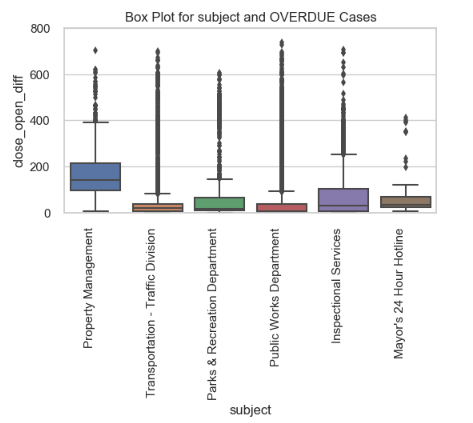
**Figure 33: Barplot groupby function by reason and max target\_open\_diff for the target variable.**



graphs. However, there are a few attributes from the ontime category that have higher target completion times than the overdue categories. These are Code Enforcement which has listed as 39 days for ontime and only 15 days for overdue. The second variable is Parking Complaints and this has 111 days for ontime and only 8 for overdue.

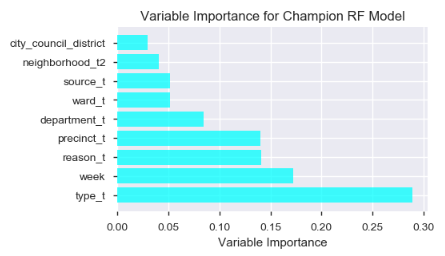
The recommendation for this audit is to look more in depth for the estimated target completion dates and establish a standard operating procedure for all cases that are created to be transparent to the public. There is a lot of variation in the number of days for the targeted completion time which may be due to the wrong completion dates being entered or there are too many categories and need to be condensed. The box plot in Figure 34 shows the close\_open\_diff for the OVERDUE cases which have a lot of outliers for all six categories listed. The audit will look at the estimated target dates assigned to each category to determine if all categories are correctly labeled.

**Figure 34: Box plot for the overdue cases grouped by subject.**



The best predictive model selected for this study is the Random Forest Model with a balanced approach that contains a total of eight attributes from the database. These eight attributes are listed in Figure 35 below. The Random Forest Model did a great job in selecting the TN (54,164) and good job in selecting the TP (6,898) for the study. The model struggled with the FP rate (8,505) and did a much better job in reducing the FN (3,019). The f1 score was 54.5% and the accuracy rate was 84%. This model will be used to predict which cases

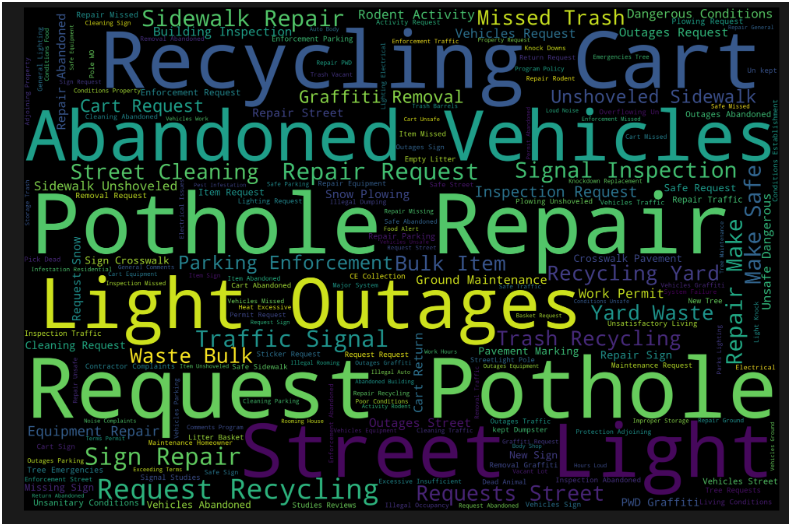
**Figure 35: Variable Importance for the Champion RF Model.**



will have a higher probability for being overdue and then a special trained Constituent Service Representatives will handle the call and try to mitigate the problems before they become overdue. The model does misclassify the FP and the extra customer service will be a welcomed asset for the Boston residents as the f1 score is improved on during the next two audits.

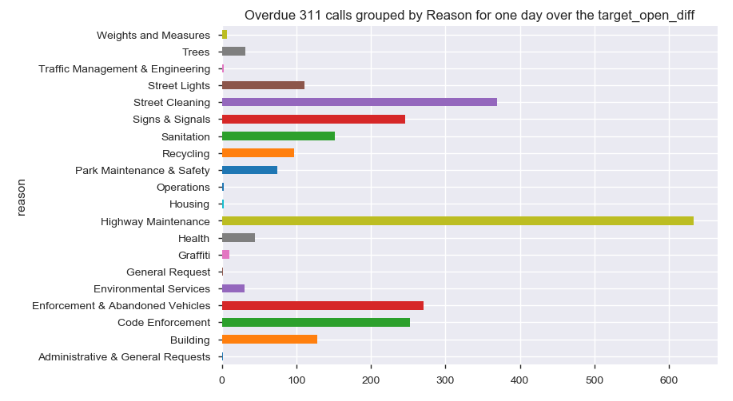
The Geo Panda mapping function did not separate the ontime and overdue cases when there was to many datapoints and most of the maps did not show good separation between these two values. The one map that did show nice separation between these to categories was the department filtered by “DISP” or Disabilities. There were only ten cases and four of these were overdue and they were spread out on the Ward district. The overdue cases were found in Wards 1, 3 and 13. All 10 cases came from the Transportation Department (8), Public Works (1) and the Mayor’s Hotline (1). The four overdue cases were all from the Transportation Department and three were for New Sign, Crosswalk or Pavement (2), Sign Repair (1) and Parking Enforcement (1). The Transportation Department had the second highest number of overdue cases and is another area that needs to be reviewed to streamline the process and minimize the time it takes to complete these cases. A Word Plot was generated for the type category which contained a total of 175 distinct types. The standard operation for text analysis was completed which removed stop words and white space to generate the plot. The word plot can be seen in figure 36 and it does a good job in showing the most complained types for the three-year period of only overdue calls and “Pothole Repair”, “Abandoned Vehicles”, “Light Outages” and “Recycling Cart” are some of the most requested overdue cases for the Boston 311 Calls.

**Figure 36: Word Plot for Overdue cases for Type with 311 Boston Calls.**

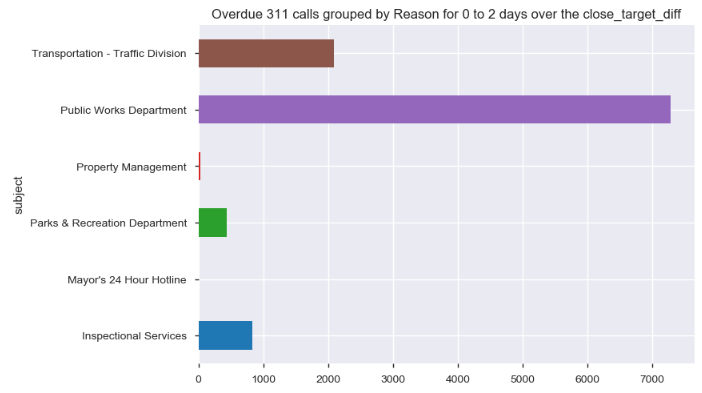


The final recommendation for the Boston 311 calls is to increase the target date by one or even two days to reduce the number of overdue calls to the community. Figures 37 and 38 show a graph for the number of overdue calls that are only one or two days late for the data set. The Highway maintenance calls contains 633 calls and the street cleaning calls contains 369 calls for

**Figure 37: Total count for one day over the target date for all overdue calls.**



**Figure 38: Total count for two days over the target date for all overdue calls.**



being one day late and this can reduce the number of overdue calls by 3.6%. The number of overdue calls that were two days late for the Public Works Department is 7,283 and if all of these calls were reduced this could reduce the number of overdue calls by 26% (based on the 28,168

overdue calls). The public can decide if they want to pay more money in taxes to create more jobs and reduce the case load or increase the time it takes to compete the calls to reduce the number of overdue calls.

# **Conclusion**

The 2017 to 2019 Internal Audit Report for Boston 311 calls show the Governmental Agency is doing a good job in handling the overall 311 calls and the calls for the PWDx and BDTD departments. The 2019 completion rate calls from Boston were compared to Chicago, San Francisco and New York City and Boston did a good job in closing these cases. The Random Forest Model with eight attributes and the balanced design was the best model with a f1 score of 54.5% and an accuracy score of 84%.

The Geo Pandas was used to map categories to determine if specific areas have higher overdue calls. The ISD variable only had 10 datapoints and shows interesting results with three overdue cases only found in Wards 1 and 13 and number 3 was shared with one ontime case. Wards 1 is East Boston and Ward 13 is North Dorchester. All the DISB cases will be analyzed beginning from the 2011 year to determine if this is a trend for this division or just a coincidence.

The target date needs to be investigated and written up so the entire community knows how long each specific 311 call will take to complete. There seems to be variation between different departments and this may be a reason why the model performance is lower than expected. The last suggestion for this audit is to increase the taxes to hire more personnel or to increase the target dates by one or two days to reduce the overdue case load.

.

# **References**

Albon, C. (2018). Machine Learning with Python Cookbook, Practical solutions

preprocessing to deep learning. O’Reilly Media Inc in Boston, First Addition.

Analyze Boston. (ND). Welcome to Boston Maps, Open Data, Geospatial Datasets. Retrieved

from: <http://bostonopendata-boston.opendata.arcgis.com/>

Analyze Boston. (ND). 311 Service Requests. Retrieved

from: <https://data.boston.gov/dataset/311-service-requests/resource/2968e2c0-d479-49ba-a884-4ef523ada3c0>

Chicago Data Portal. (ND). 311 Service Requests. Retrieved from:

<https://data.cityofchicago.org/Service-Requests/311-Service-Requests/v6vf-nfxy>

DataSF. (ND). 311 Cases. Retrieved from:

<https://data.sfgov.org/City-Infrastructure/311-Cases/vw6y-z8j6/data>

Hastie, T., Tibshirani, R. and Friedman, J. (2017). The Elements of Statistical Learning: Data

Mining, Inference and Prediction. Springer Series in Statistics, 2nd Edition.

Hendrickson, J. (2019). Property Taxes Need Accountability and Transparency.

Retrieved from: <https://caffeinatedthoughts.com/2019/04/property-taxes-need-accountability-and-transparency/>

OPENGOV. (ND). Analyze Boston. 311 Service Requests. Retrieved from:

<https://data.boston.gov/dataset/311-service-requests>

NYC Data. (ND). NYC OpenData. Open Data for all New Yorkers Retrieved from:

<https://opendata.cityofnewyork.us/>

Patch.com. (ND). 2019 Residential Property Tax Rates for the 351 MA Communities

Retrieved From: <https://patch.com/massachusetts/boston/ma-residential-property-tax-rates-each-community>

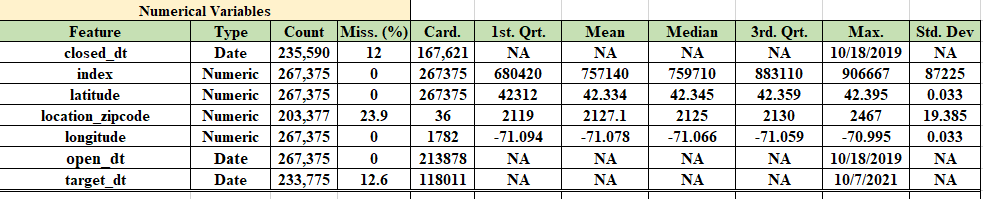
Pichhi, A. (2019). Which homeowners around the U.S. pay the highest property taxes?

Retrieved from: <https://www.cbsnews.com/news/property-tax-which-homeowners-around-the-u-s-pay-the-highest/>

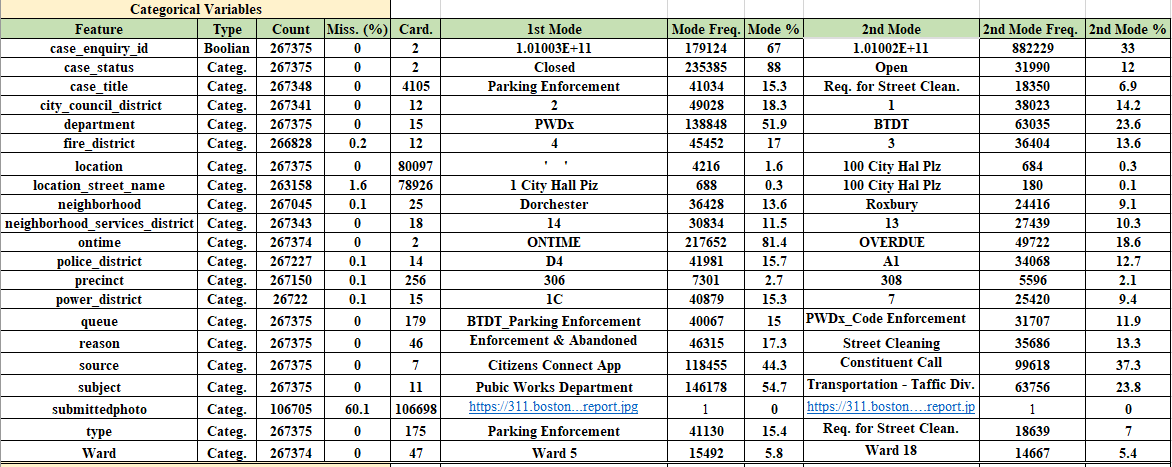
# 

# **Appendix**

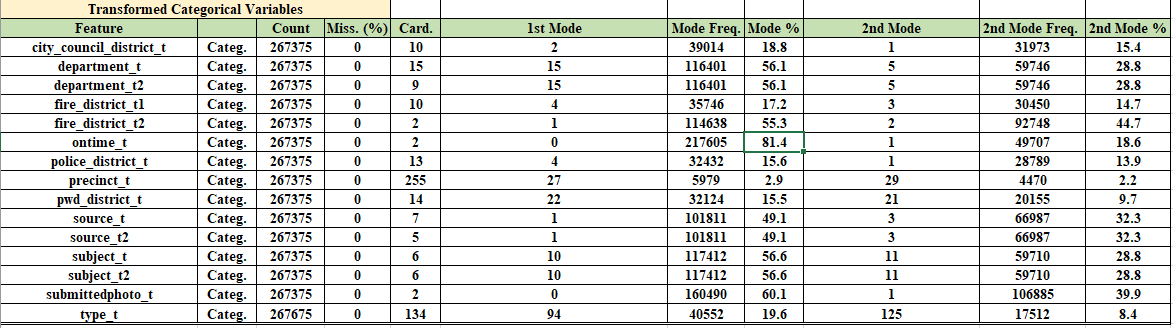
**Figure: 1A Data Quality Report for Numerical Attributes.**



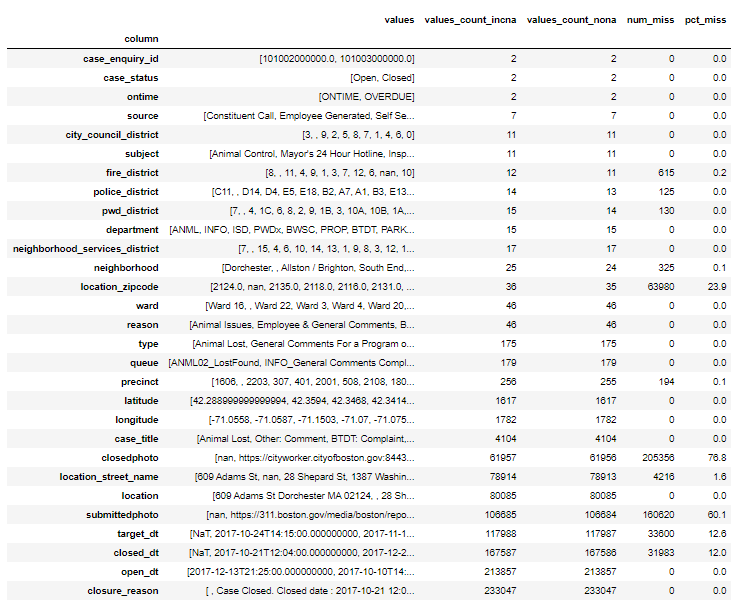
**Figure 2A: Data Quality Report for Categorical Attributes.**



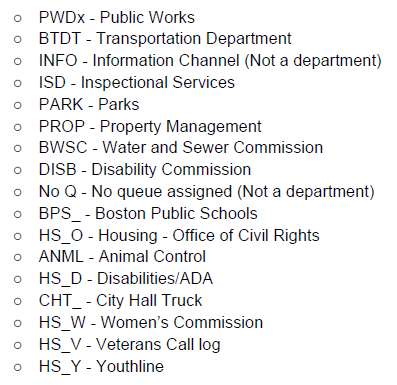
**Figure 3A: Data Quality Report for Transformed Categorical Attributes.**



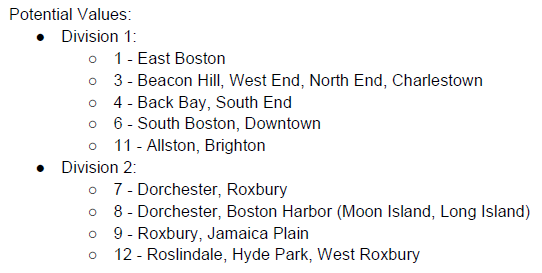
**Figure 4A: Unique Counts and Missing Value Report.**



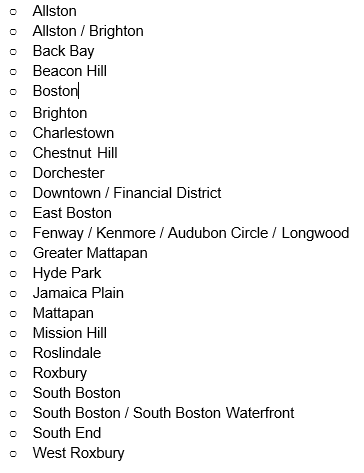
**Figure 5A: Description of the Department Names.**



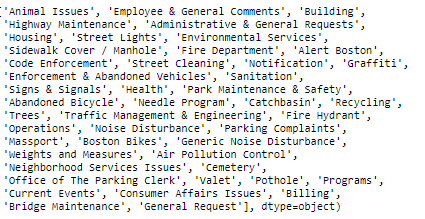
**Figure 6A: Description of the two Boston Divisions.**



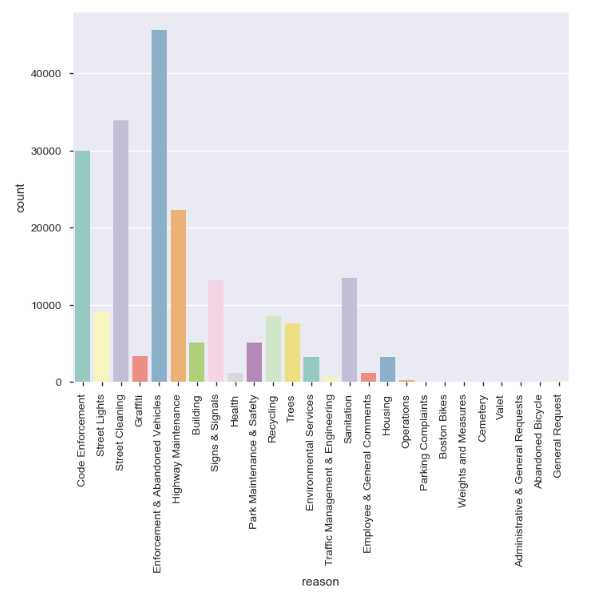
**Figure 7A: Description of the Boston Neighborhoods.**



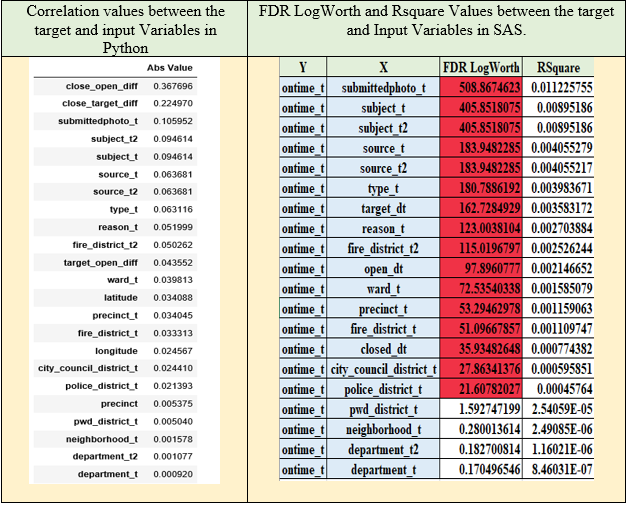
**Figure 8A: List of all 46 Reasons assigned to the 311 Calls.**



**Figure 9A: Count plot for the 26 different levels listed as a reason.**

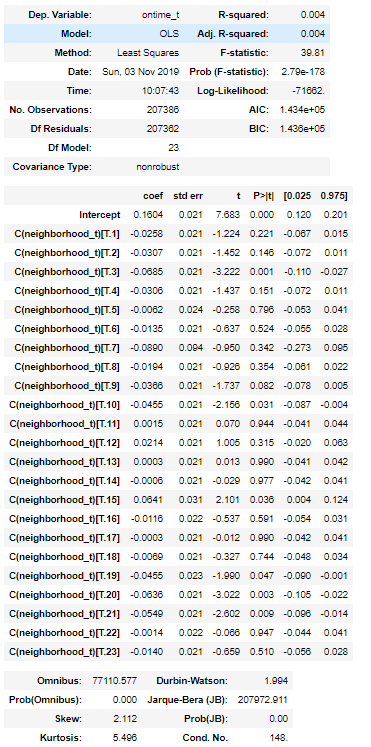


**Figure 10A: Correlation values, Logworth and Rsquare Values.**



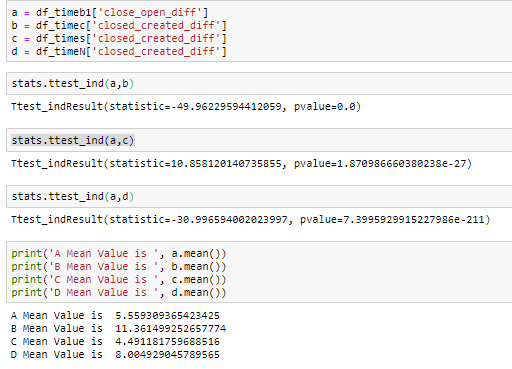
**Figure 11A: OLS Regression Model for**

**neighborhood\_t.**



**Figure 12A: Basic Test Results for Completion Time (a=Boston, b =Chicago,**

**c=San Francisco and d = New York City).**



**Figure 13A: Groupby output for the ontime and overdue cases by the reason category.**

