311 Service Requests Analysis

Anonymous Author(s)

ABSTRACT

311 services data from 2010 to till date was collected for data analysis. With increase in different platforms for registering the complaints and requesting for information and services in 311 services. The data set contained 27 million rows consisting of features including created date,closed date and resolution description. The main goal of this project was to understand the data set and its trend and find the features which resolve the complaints faster than the other. The data was cleaned and analyzed using Big data techniques used to understand the data. Data cleaning was the first and foremost step in data analysis. Visualizing the data in graphs and charts augmented the understanding of the data. The regression algorithms were applied to find the importance of the features in the data set which will aid in improving the 311 services to the people.

After data cleaning, the implementation of data models for predicting the resolution of the requests was done. It considered the features of created time, closed time and place from where the request was raised. As the features in the data set were of string data type, the features cannot be correlated. We neglected the features using the threshold value. Implementation of various regression algorithms like Linear Regression algorithm and k-nearest algorithm helped in predicting the importance of the features. It predicted the potential of the features for improved quality service to the people of the city. For classifying the type of complaint, the project used algorithms like logistic regression and random forest classifier for classification. These models were implemented and the algorithms predicted the output in higher precision. The legal and ethical consideration of the handling 311 services helped to learn the ability of handling public data. During the research, lessons such as data analysis and effective visualization of data for finding the essential features which affect the prediction of the output were learned.

KEYWORDS

Logistic Regression; k-nearest algorithm; data processing; data cleaning; pyspark

1 MOTIVATION

A 311 Service request refers to a non-emergency number citizens can use within different cities to report any issues they are facing, place complaints, or request information about specific city services.[16]The 311 call centers came into the public sphere in the 2010s. They have enabled direct communication between the governments and citizens regarding non-emergency information. The predictive model from the above Big Data analytics will help provide timely responses to the requests in the future and help to make decisions to improve the quality of 311 services for fulfilling the needs of people.

The motivation behind choosing this topic for the group project is to analyze the number and type of issues brought up for each request. Through further analysis of the service requests based on the region of their origin, better understanding of the factors that are influencing residents of a particular area to create service requests can be gained. The information gained from these analysis can then be used as a measure of the service demands based on certain attributes and living conditions of a particular area.

Through further analysis of the service requests based on the region of their origin, a better understanding of the factors that are influencing residents of a particular area to create service requests can be understood.The information gained from the analysis helps to improve the service demands based on significant attributes and living conditions of a particular area. These predictions would help for highlighting the complaints types. A large number of requests are being collected and allow governments to take preemptive action in response.

Our main goal is to explore the 311 service requests from New York using big data analytical techniques based on the information gathered and the data visualizations.

The main steps are to create a prediction and prevention model are as follows

(1) Gathering the 311 Service Request data Set
(2) Performing data prepossessing functions on the data set
(3) Analyzing the data set
(4) Recognizing Patterns within the data set
(5) Analyze predictions results
(6) Data Analysis Visualization

The data set which we perform data analysis is 311 service requests made between 2010 to Present. The data set is provided by the New York City OpenData website.[6].Every row entry in the data is an individual 311 service request. Every entry has a unique numeric key. Each entry has a created and closed date which signifies when the request is created and closed. Each 311 entry also contains fields for the agency that responded to the call and their corresponding acronym. Each entry has a complaint type and descriptor, which determines the purpose of the call ranging from a noise complaint, maintenance request, highway conditions to a blocked road.

The location type determines whether the incident occurred on the street, in a residential area, or some other specific place. Many of the other fields provide further information to narrow down the exact location where the incident occurred. The location information includes zip code, address, city, borough, street name, cross streets, intersection streets, nearby landmarks, community board, borough-block-lot number, state-wide x-y coordinates, and latitude-longitude coordinates. The address type field determines the kind of address that was provided which is typically "address", or "block", but occasionally null. If a city facility was involved in this request, it is also included in its associated field. The due date field provides information about the service request is supposed to respond by, but some values are null. The data set also has a resolution action taken data field, which provides the information about the date when the recent resolution action is provided.

There is an open data [17] channel field that determines how requests are submitted. Most of them are either online, phone, other or unknown. The remaining fields are incident-specific. If the

incident happened in a park, some fields signify the park's name and borough. If the incident took place on a bridge or a highway, there are fields for the bridge's highway name, direction, road ramp, and segment. Lastly, if the incident happened in a taxi, there are fields for the type of taxi, the taxi company, and the taxi pickup location.

1.1 Sample Data Set and Applications

Table 1: 311 Services Requests Sample Data set from 2010 to till date. Unique Key, Created Date, Agency, Complaint type is shown

Unique Key	Created Date	Agency	Complaint Type
18302847	1/23/2019	DOT	Sidewalk Condition
26378416	3/06/2020	DOT	Highway Condition
32948361	6/30/2021	DPR	Maintenance or Facility
42827394	1/09/2021	TLC	New Tree Request

The table provided is a sample data set that shows some of the features from the 311 service requests. The original data set contains 41 features and 26.7 M rows. The data set provides information as to when a service request was created along with its complaint type and address.

First of all, we perform Exploratory Data Analysis data which focuses on exploring data and selecting specific features in the data set. These features are used in handling outliers which will then allow us to effectively predict results. For instance, we can find the location where more Service Requests are being raised and from this, gain information about the factors responsible for the specific request being raised. With this, preventative measures can be taken to minimize the number of requests being submitted. Data Analysis visualizations can be done in the form of bar graphs and line graphs which could provide us with further information such as the time taken to resolve certain incidents.

Along with this we could plot graphs by categorizing the service request based on the Open data channel type, which is classified into mobile, online, etc. After completely analysing the data and gaining possible insights .[13], the next step would be to create a prediction model based on the features we selected.

As we can classify complaints based on the type, it can be considered as supervised learning. It is a multivariate data type as we have more than two categories. We can apply various classification techniques such as Multinomial Logistic Regression which can classify multiple classes, k- nearest neighbour algorithm which can classify the data into partitions based on rules. Logistic Regression uses the sigmoid function to classify between two classes.

The main goal of this data analysis is to gain insight regarding the complaints types and location of the crime. With this information it is possible to reduce the causes of complaints being raised and provide better services for the people.

The report is organized in the following order, as follows, section 2 describes the goal of the project. Section 3 describes the research work related to the project. Further, section 4 describes the design and implementation of the project and the data model and architecture of the project. Section 5 describes the analysis of the 311 services data. The lessons learned from the project are explained in the section. Section 6 describes the legal and ethical considerations of the project. The status and the future work of the project are discussed in section 7. The report is concluded in the final section 8. These models predict the expected result in higher precision.

2 PROJECT GOALS

The first and foremost goal is to handle large amount of data. As the pandas data frame cannot handle 27 million of data, we used Spark data frame which does the data processing in efficient way. Instead of taking a subset from a data set which will make the data cleaning easy. We tried using spark data frame considering most of the rows which will make the prediction as accurate as possible.

The main goal of the project is to find the trend and pattern in the data set. It also involves understanding the different types of complaints and the duration taken for fixing the raised requests. As the data set is updated every day it has real-world values, it needs to be cleaned for analyzing the data and finding insights from them for future prediction.As the data is getting updated every day, it contains real-time data, which needs data cleaning for further analysis. For instance, the date value in closed date, resolution date are missing, and all of the date values are in string format that needs date format conversion.The data has some outliers, which might affect the prediction of the expected value. The exploratory analysis involves data visualization, which depicts the distribution of data and provides essential information. The project also includes data modeling, which includes separating the data set into training and test data to avoid bias and variance and fit the unseen data in the model. Regression predicting algorithms such as Random forest classifier and gradient boost classifier is built for predicting the importance of the features.

We examined the related work of request services for a better understanding of a similar data set. These related work helped to move in the right direction for improving the model. We aimed to understand the legal and ethical considerations as the data considered for processing needs to be handled with caution as it contains important information about people.

3 RELATED WORK

Some see 311 services as a move towards "smart cities", cities where private and public services are seamlessly integrated, technology facilitates their efficient fulfillment's, and technology enables them to be forward thinking and moving.

Taewoo Nam and Theresa A.Pardo's paper[14] shows how 311 services aid in fulfilment of the vision of smart cities. Their study showed how 311 services forced a change in their organizational structure and increased efficiency in Philadelphia. They found that many departments were able to free up resources after the establishment of these services. This is due to the fact that departments can put 911 calls on top priority and can handle 311 requests on a per urgency and as requested basis. This paper also touched on how 311 data allowed for the collection of data by officials to inform future decisions about infrastructure and service hot zones. This paper makes it clear that there are a lot of new valuable insights that lay hidden in 311 data.

Some have taken up the task of discovering these insights. Baek-Young Choi et al. found in their paper[8] that certain cities had higher requests for specific types of services. For example Los Angeles had higher requests for graffiti removal while Kansas City had higher requests pertaining to water related issues. These insights can allow for the respective cities to allocate more of their budgets to combat these specific issues more effectively. Their paper also showed that requests of a particular type had higher volumes during certain times of the year. For example New York City saw higher requests for hot water and heating during the winter than any other time of the year. These insights also help services effectively plan for higher call volumes of a specific type so need can be met ahead of time.

Hagen et al. showed some deeper infrastructural insights in their paper[10]. They utilized k-means clustering to find similar 311 calls in Miami. Informed by their clusters, they discovered that their request types may have a socio-economic characteristic. They found that the cluster with the highest proportion of relatively poor and foreign Hispanics usually lived in areas where there were higher volumes of Community Code Enforcement requests were made. They also found that areas with a higher volume of calls about roads and infrastructure were made in mostly black unemployed communities. Analysis of this kind may shed light on some sort of systemic problems that exist.

The paper Structure of 311 service requests as a signature of urban location by Wang et al. [15]. describes the socioeconomic features which will be a major factor for finding the future trends in urban planning[7]. It classifies the services based on the categories for making decisions and planning the urban areas. After clustering of data, the model finds the average income of the cities based on patterns of distinctive socioeconomic features.

4 DESIGN AND IMPLEMENTATION

4.1 Data Model and Architecture

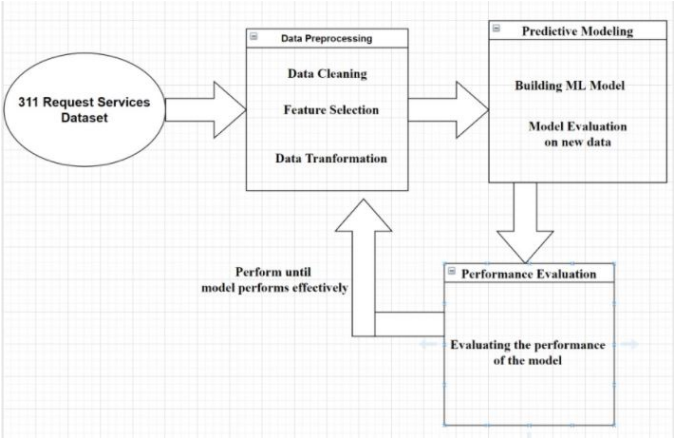


Figure 1: An Overview of the Architecture Model used in the paper. It consists of data cleaning and data processing

The above data model depicts the life cycle for our data analytics. As the first and foremost step, we are collecting the data set and converting it into a data frame for performing data pre-processing which will enrich the value of the data. We perform data visualization of the data sets using bar charts, pie charts, and histograms which aids in gaining insights into the data. We can build predictive models for finding the trends in the data and predict the output for new data. These data are fed to machine learning algorithms to do performance evaluations. Repeat the cycle starting from data pre-processing until the model produces the desired result.

We visualized the data sets into bar-charts, pie charts and bar-graph which aids in gaining insights about the data. We build predictive models for finding the trends in the data and predict the output for new data which are fed to machine learning algorithms and do performance evaluation. Repeat the cycle starting from data pre-processing until the model produces the desired result.

4.2 Function of PySpark

PySpark is an Apache Spark interface that utilizes Python to create Spark Applications by using built in Python APIs. It also provides a Spark python shell which can be used to analyze our data sets in a distributed environment. The PySpark service also provides us with multiple features that can be used to analyze data such as **Spark SQL** which can function as a distributed query engine as well as be used to abstract Data Frames, **MLlib** Libraries that allow us to integrate advanced APIs to create and modify machine learning pipelines as well as use existing Machine Learning algorithms such as K-means clustering and logistic regression, **Streaming** which provides access to analytical applications and finally the **Spark Core** which is the execution engine for the entire platform.[4]

The data analysis software PySpark was used to analyze the provided data set. Our data set has around 26.7 million rows which meant that we needed an analysis tool that could handle data sets at a large scale and PySpark seems to have been the right option as it enables us to store our data in memory and process it in real time. By storing it in memory we are able to reduce access time and it is not necessary to retrieve the data from disk memory.

4.3 Data Pre-Processing

Data Preprocessing is a Data mining technique that transforms the data into meaningful and efficient data. There are various steps in data pre-processing such as data cleaning, data transformation, and data reduction. As the data is large, the pandas would lack the potential for operating data. So, we are using a PySpark data frame that has an inbuilt API for faster performance. As the first and foremost step we have cleaned our 311 services data set by filling in the missing values. In the 311 services data set, we have columns such as **Created Date**, **Closed Date**, **Due Date** and **Resolution Action Updated Date** which have some null values. So, we are replacing the **Created Date** null values with the corresponding **Resolution Action Updated Date** non-null values and the corresponding **Closed Date** values since some form of action has been taken on the specific service requests. Similarly we have filled the missing values of **Due Date**, and **Closed Date** respectively. We have also filtered outliers which will help in improving the performance of the data.

Feature Selection is the process where we select the features which will contribute to the outcome of our prediction. The following are the benefits of feature selection,

(1) Greater Precision
(2) Reduces Over-fitting
(3) Reduces training time

We are setting the threshold for the total number of null values in the feature as fifty percent. As the number of null values is higher it would not add significance to the prediction model. If the number of null values in the feature is greater than the threshold, we are removing the below threshold feature as it would not add significance to the prediction model. So we are selecting the columns which have null and None values and storing the total count of each feature in the data frame and converting them into the pandas data frame to find the number of features that are higher than the threshold and removing the features from the data frame. Initially, there were 41 features in the dataset after removing the feature based on the threshold value. The features get reduced to 31 more significant features.

By examining the schema of the spark data frame which we created from the 311 services data, the datatype of dates is in string type. For converting the string to date format we use the date format inbuilt function from pyspark SQL function which converts the string to date format. We are converting the string to UNIX timestamp and casting them to timestamp type for converting them from "MM/dd/yyyy hh:mm:ss a" into the "yyyy-MM-dd" date format. As we need dates in the proper date format for categorizing based on month and year for future data analysis. So, we are converting the string data type into conventional date format and time format and adding these two features to the dataset.

4.4 Exploratory Data Analysis

EDA is the process of exploring and investigating data sets to discover some form of patterns or inferences with the help of any form of visualization tools and techniques. While analysing the data distribution, we visualized the bar-graph of agency and its corresponding agency count using the Matplotlib library which creates interactive data visualizations using Python. It is evident from the below bar-graph that some of the agencies such as EDC and DOF frequencies are insignificant. Therefore, we are neglecting the negligible frequencies of the agent and plotting the bar-graph which have frequencies greater than 15000 occurrences. The below table shows the agency and its corresponding frequencies which are in considerable amounts.

As it can be seen from the bar-graph some of the agencies are barely even visible on the graph which implies that they have a very small almost negligible number of service requests being submitted. Therefore, e decided to neglect the agencies that have very low frequencies.

After cleaning the data set of low frequency Agencies, we are left with the following values represented in a table.

After creating this table, we re-visualized the bar-graph excluding the negligible frequency values.

From this visualization we can see that out of all the agencies some of the highest number of 311 service requests were submitted from the NYPD and HPD.

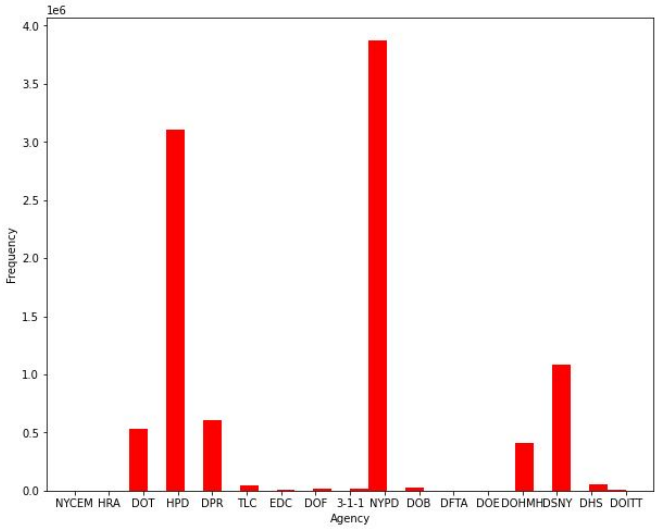


Figure 2: Agency Count of the data set through which the complaints were raised in 311

Table 2: Frequency distribution of 311 requests in Agencies based on the different complaint types

ID	Agency	Count
2	DOT	535888
3	HPD	3105680
4	DPR	602273
5	TLC	43218
7	DOF	15367
8	3-1-1	15860
9	NYPD	3876425
10	DOB	27626
13	DOHMH	413562
14	DSNY	1088089
15	DHS	53744

Further exploration was done by creating a PySpark data frame and grouping the service requests by "Park Borough". We then converted the PySpark frame into a panda data frame for visualization. The following table was created.

For this visualization we decided to create a pie graph which will show the percentage of the number of service requests being submitted from areas under "Park Borough".

It can be inferred from the Pie chart that the majority of the requests are raised from Brooklyn with 30.4 percent followed by Queens with 21.5 percent and Staten Island. Further analysis on Park Borough is essential for determining the reason for each and

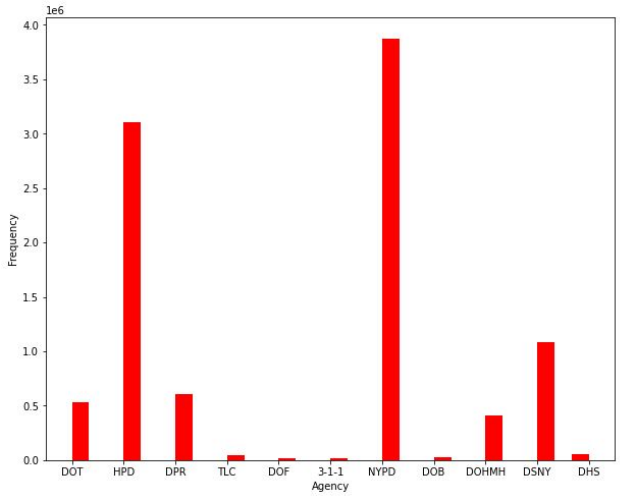


Figure 3: Agency Count

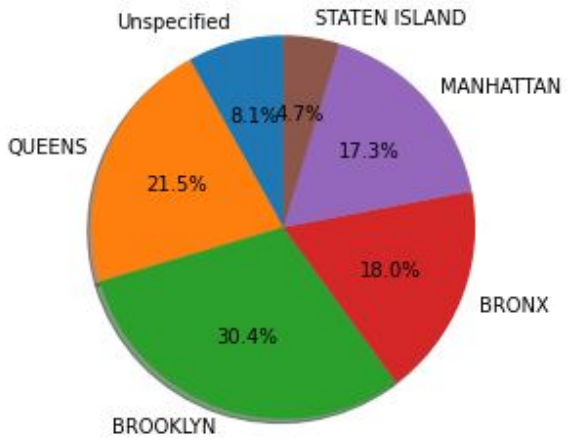


Figure 4: Distribution of 311 services requests based on Park borough and its is visualized as pie chart

Table 3: Distribution of 311 services requests based on Park borough

ID	Park Borough	Count
0	Unspecified	793413
1	Queens	2101492
2	Brooklyn	2977987
3	Bronx	1761696
4	Manhattan	1690670
5	Staten Island	461832

Table 4: Top five frequency count of complaint types registered in the 311 services

ID	COMPLAINT tYPE	Count
36	Noise - Residential	1498724
126	HEATING	861184
55	Blocked Driveway	798396
33	Illegal Parking	571085
19	GENERAL CONSTRUCTION	469519

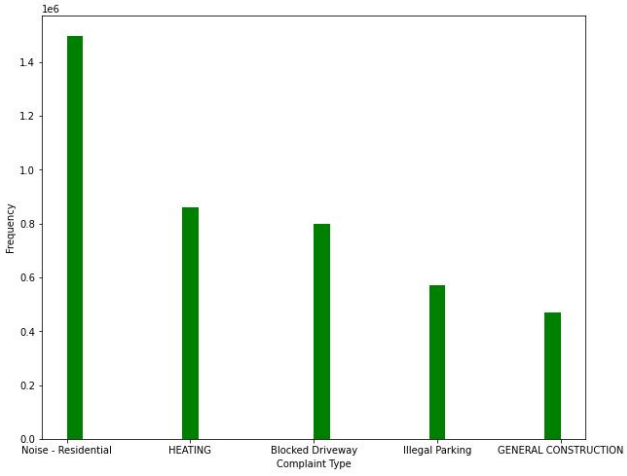


Figure 5: Complaint Count in Bar Graph

every Park Borough for making a future decision for improving the services. For interpreting the different types of complaints, we created a data frame which contains the total number of complaint types. As the number of complaints are large, we are plotting the top 5 most raised complaints and it is evident from the bar plot that non-residential complaints are higher than the other complaints.

The next data visualization performed was on the frequency of different complaint types within the data set. The table of these values can be seen below.

Similarly, the data can be visualized in the complaint type bar graph which has been inserted below.

It can be inferred from the bar graph that the complaint type with the highest frequency is the **Noise - Residential** and the lowest frequency is the **GENERAL CONSTRUCTION** complaint type. This means that noise pollution is a major cause for service requests being submitted.

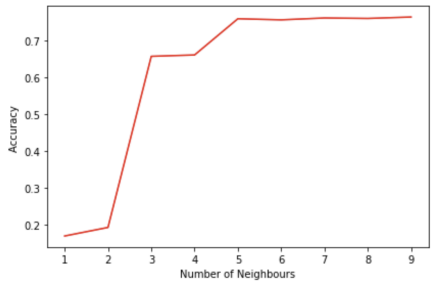
The data is split into train and test data of 7:3 ratio. Test data is used for finding the accurate prediction after the data is trained by the model. It tests the model with a unknown data for checking its correctness. Logistic Regression is the first model which is applied to the data set. The date difference from the created date and closed are calculated and they are categorized based on the

```
In [243]: from sklearn.linear_model import LogisticRegression

          clf = LogisticRegression(random_state=0).fit(X_train,Y_train)
          clf.score(X_test, Y_test)

          C:\Users\Vignesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:63: DataConversionWarning: A column-vector y was passe
          d when a 1d array was expected. Please change the shape of y to (n_samples, ), for example using ravel().
            return f(*args, **kwargs)
          C:\Users\Vignesh\anaconda3\lib\site-packages\sklearn\linear_model\_logistic.py:763: ConvergenceWarning: lbfgs failed to converg
          e (status=1):
          STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

          Increase the number of iterations (max_iter) or scale the data as shown in:
              https://scikit-learn.org/stable/modules/preprocessing.html
          Please also refer to the documentation for alternative solver options:
              https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression
            n_iter_i = _check_optimize_result(
Out[243]: 0.6048
```

Figure 6: Code Snippet of Logistic Regression applied to the agency and corresponding to its resolution date

agencies. Logistic Regression is applied when the data can be categorized. As the agency data can be categorized, the multivariate regression model can be applied. We have achieved a R2 score of 0.6048 which is a average model for predicting the agency based on the date difference. Further, feature extraction and feature selection and bagging techniques can be used to improve the r2 score.

The next model which is applied to the data frame is k nearest neighbour algorithm. It is supervised algorithm for finding the predicted value based on the labeled input value. k-nearest neighbour classifier is applied to the data set for categorizing the data set. The total number of neighbours and its accuracy of the prediction is calculated.




```
The K value with the highest accuracy is 9 with an accuracy of 0.7643666666666666
```

Figure 7: Graph that shows the result of k-nearest Algorithm and its accuracy in prediction

From the above graph it can be found that the k value is 9 and has the highest accuracy of 0.76436666. The above model is a average model which can predict the output most of the time. As the data set is in large amount, the accuracy of the output is of 0.76. The accuracy can be improved by feature extraction, bagging technique and hyper parameter tuning. The model accuracy can increased.

5 ANALYSIS

We have cleaned and transformed the data for finding trends in the data by building a machine learning model. As the data distribution is random, the model will perform efficiently on new data without bias and variance.The lesson learned from the project is the ability to handle large amounts of data. We learned the process of data cleaning, data transformation using Pyspark. Pyspark is more effective in handling a large amount of data than pandas. We learned how to analyse the data set of millions of data, how to handle using pyspark and ability to do data management and data munging using pyspark. We learned how to write code in pyspark for retrieval of data, manipulation of data and data visualization of enormous amounts of data. We used Matplotlib for data visualization which provided a good visualization of data and made us understand the pattern in the data. We neglected the features which do not add value to the final prediction. We learned to neglect the least essential features by setting up a threshold for removing them. We learned that we cannot correlate the features as most of them are not correlated to each other. We also learned how to create new features to the column using SQL functions from spark data frame and did a conversion of date from string format to date format function and transformed the date into conventional date format. In addition to that, we gained insights about the agency, complaint types, and Park Borough data distribution by performing data visualizations. We learned to find the trend in request in 311 services and made us use Logistic Regression and k-nearest algorithm for finding the importance of the agency feature in the dataset.We learned the legal and ethical consideration of data as it is a public data that needs to be handled without any disclosure.

6 LEGAL AND ETHICAL CONSIDERATIONS

The disclosure of national public data can help citizens better understand the state of the country. This is the right of citizens. Each state also has laws on the disclosure of public data.[9] Public data can be used and reused by the public without restrictions, free of charge and not restricted by intellectual property rights or licenses.The data set used in our project is the public 311 data set released by the New York government, so the copyright issue is not involved. The public data set has no copyright protection and is free.

The 311 call requires the address information of the caller to solve the reported problem. The caller's phone number is usually required to provide feedback on whether the problem is resolved. Although the 311 call can be anonymous, the name of the caller is generally required. These all involve the personal privacy of the caller. So according to the requirements of NYC Local Law 47 of 2005 [3], the publicly released 311 data set has already removed the part of the caller's personal privacy such as their name and phone number or physical address.[1]

The government is also considering legislation to require 3111 callers to provide their names and phone numbers, but the 311 hotline reports are usually minor matters. The requested information may be used to track callers, such as people who use the 311 hotline to "threat" government officials, and DoITT may record who calls frequently, causing information abuse.[12] In our project, combining the location information of the caller provided in the data set and the reported events, we can analyze the number of times a certain event occurred in which area, and which area has the most events, etc. However, New York is a multi-ethnic metropolis, and their living conditions often tend to live in clusters. Information about the settlements of various races and their communities in the city can be easily obtained.[5] If combined with the incidents analyzed above, we may see which incidents occurred in which ethnic communities are more likely to occur. These results may be used by people with ulterior motives to conduct personal attacks or promote racial discrimination. But in fact, the specific circumstances

and causes of various problems still require police or government officials to go to the scene for actual inspection and analysis. We cannot jump to conclusions or over-interpret the data results. The data analysis results of this project can only help the government to better manage the city, and the conclusion analysis should be rational and objective.

Ethics and privacy issues are extremely important in big data analysis. We must ensure that the results conform to public ethics [11] and are not affected by personal emotions, and the consequences of actions do not harm others. Any operation done in the analysis process has a reasonable ethical explanation. Big data may sometimes provide too much information and damage the privacy of individuals or organizations. Although different cultures have different views on privacy, many governments have introduced different degrees of privacy protection related laws. In data analysis, data should not be used to monitor and track others. And we should not want to obtain users' personal data without people's knowledge, being forced or deceived, as some social media have done.[2]

7 STATUS & FUTURE WORK

The initial data set has been pre-processed by performing data cleaning and data transformation so that it can be used to create a strong prediction model. Data exploration and data visualization were the next steps completed, which explains the data distribution of features of the data set. Further analysis would be plotting a geo-plot of New York that displays the locations based on their request frequency. Analysis can be done to understand the peak time service requests are raised. This analysis could also help in calculating the average response time to the requests based on when they are raised.

After data analysis the next step would be to do create predictive model which predicts new data and outputs the resolution time taken for a particular service request. We should work on building a model which accepts the real-time data and considers it as input and predicts the expected value. We will divide the data into training data, test data and cross validation data to enhance the quality of the predictions as well as to avoid over-fitting and under-fitting. If the prediction is not accurate we need to repeat data pre-processing and feature engineering to improve the accuracy of the output.

8 CONCLUSIONS

The analysis of all rows of the dataset helped us to gain insights about the features that will affect the timely action on the requests. We did data cleaning, data manipulation and data modeling for finding valuable information from the data. We learned how to use libraries such as pandas,pyspark,sckitlearn and matplotlib for prediction The predictive model from the above Big Data analytics will help provide timely responses to the requests in the future and help to make decisions to improve the quality of 311 services for fulfilling the needs of people. We can conclude that feature agency has more influence on the model than the other features. Regression algorithm in large data set like 311 is difficult and it has wide range of complaints and it needs to be categorized under some common categories.Agency plays a significance role in affecting the model prediction.Fulfilling the request of people plays a significant role in improving the people quality life. From the above insights we can make decision making from improving the services.

REFERENCES

[1] [n.d.]. https://www1.nyc.gov/assets/doitt/downloads/pdf/311-privacy-policy.pdf
[2] [n.d.]. https://link.springer.com/chapter/10.1007/978-3-030-45002-1_14
[3] [n.d.]. About 311 reporting. https://www1.nyc.gov/site/311reporting/index.page
[4] [n.d.]. PySpark documentation¶. http://spark.apache.org/docs/latest/api/python/
[5] 2021. New York City ethnic enclaves. https://en.wikipedia.org/wiki/New_York_City_ethnic_enclaves
[6] DoITT 311. 2021. 311 Service Requests from 2010 to Present: NYC Open Data. https://data.cityofnewyork.us/Social-Services/311-Service-Requests-from-2010-to-Present/erm2-nwe9
[7] Luís Bettencourt, José Lobo, Deborah Strumsky, and Geoffrey West. 2010. Urban Scaling and Its Deviations: Revealing Structure of Wealth, Innovation and Crime across Cities. *PloS one* 5 (11 2010), e13541. https://doi.org/10.1371/journal.pone.0013541
[8] Baek-Young Choi, Mohammed K. Al-Mansoori, Rafida Zaman, and Ahmed A. Albishri. 2018. Understanding what residents ask cities. *Proceedings of the Workshop Program of the 19th International Conference on Distributed Computing and Networking* (2018). https://doi.org/10.1145/3170521.3170531
[9] Pam Greenberg. [n.d.]. State Open Data Laws and Policies. https://www.ncsl.org/research/telecommunications-and-information-technology/state-open-data-laws-and-policies.aspx
[10] Loni Hagen, Hye Seon Yi, Siana Pietri, and Thomas E. Keller. 2019. Processes, Potential Benefits, and Limitations of Big Data Analytics: A Case Analysis of 311 Data from City of Miami. *Proceedings of the 20th Annual International Conference on Digital Government Research* (2019). https://doi.org/10.1145/3325112.3325212
[11] Edmund Iii and Falicia Elenberg. 2020. Ethical Challenges Posed by Big Data. *Innovations in clinical neuroscience* 17 (10 2020), 24–30.
[12] Jennbsp;Chung, Jakenbsp;Offenhartz, Jaclynnbsp;Jeffrey-Wilensky, Elizabethnbsp;Kim, and Betsynbsp;Ladyzhets. 2008. Making sure 311 isn't abused. https://gothamist.com/news/making-sure-311-isnt-abused
[13] Steve Lavalle, Eric Lesser, Rebecca Shockley, Michael Hopkins, and Nina Kruschwitz. 2011. Big Data, Analytics and the Path From Insights to Value. *MIT Sloan Management Review* 52 (12 2011), 21–32.
[14] Taewoo Nam and Theresa A. Pardo. 2012. Transforming city government. *Proceedings of the 6th International Conference on Theory and Practice of Electronic Governance - ICEGOV 12* (2012). https://doi.org/10.1145/2463728.2463787
[15] Lingjing Wang, Cheng Qian, Philipp Kats, Constantine Kontokosta, and Stanislav Sobolevsky. 2017. Structure of 311 service requests as a signature of urban location. *PLoS ONE* 12, 10 (Oct. 2017), e0186314. https://doi.org/10.1371/journal.pone.0186314 arXiv:1611.06660 [physics.soc-ph]
[16] Colin Wood. 2021. What Is 311? https://www.govtech.com/dc/articles/what-is-311.html
[17] Anneke Zuiderwijk, Marijn Janssen, and Chris Davis. 2014. Innovation with open data: Essential elements of open data ecosystems. *Inf. Polity* 19 (2014), 17–33.