DATA EXPLORATORY ON TAXI DATA IN NEW YORK CITY

Christie Natashia
Asia Pacific University of Technology and Innovation

DATA EXPLORATORY ON TAXI DATA IN NEW YORK CITY

Christie Natashia Archie[1], Shubashini Rathina Velu[2]
[1]School of Computing, Asia Pacific University, Kuala Lumpur, Malaysia
[2]School of Computing, Asia Pacific University, Kuala Lumpur, Malaysia
[1]TP041544@mail.apu.edu.my, [2]shubashini@staffemail.apu.edu.my

***Abstract*— With increasing amounts of urban data being collected and made available, new possibilities of data-driven research emerge lead to changes in people living through scientific proof-based decision-making and strategies. In this paper, the researcher concentrates on an urban data set of particular significance which is taxi rides. Taxis are essential instruments, and taxi-related information can provide ground-breaking insight into different facets of city life, from economic activity and human behavior to mobility trends. Taxi data involves geographical components as well as several variables associated with each ride. Throughout this research, data insight is gain through data pre-processing, feature engineering and data exploratory analysis. This transformation of raw data enables researcher, data scientist to have the meaningful insight into the data and understand the mobility pattern of New York City.**

***Index Terms*— Data exploratory, Feature Engineering, Geolocation data, Urban understanding, Taxi dataset**

1. Introduction

Taxi is one of the urban public transport in many busy countries. Unlike other public transports, taxi rides provide accessibility, convenience, yet privacy to passengers [1]. A competitive and reasonable taxi pricing is worth the ride for private car users to switch to a taxi service. Millions of taxi trips data are generated on monthly basis, which this data can be useful to gain the insight of the traffic patterns and obtain a clear view of urban city life [2]. Not only that, by leveraging the given dataset, taxi demand on major events like Christmas and New Year's Eve, can be studied in order to make a better decision making.

The growth of the science of geographic information provides new possibilities for urban understanding and planning [3]. Spatial assembly is an important analytical step for summarizing and perceiving the geographic environment from the demand for taxis [4]. With automated data collection of taxi movement, a city's operation can be extracted from geospatial data in both spatial and temporal point of view. It provides a more accurate depiction of the nature of a region, considering that daily movement and activities found in geospatial data indicate the social-economic properties of urban functions [5]. This paper will be focusing on examining the urban dataset of New York City taxi trips. According to taxi factbook report release by NYC Taxi & Limousine Commission (2018), TLC taxi trips in New York City has a total of over 41 million trips a year between June 2017 and June 2018 [6]. In general, taxi trips comprise spatial elements like longitude and latitude points. Thus, this data enables data scientist to encode geolocation information into an insight of urban traffic movement and activities.

Not only that, other attributes like taxi Id, total fare amount, and number of passengers are also recorded which allows researchers to study the traffic congestion, economics of fare pricing, and optimal fleet size [3]. Not surprisingly, it is challenging to explore those data and results. Commonly used tools and platforms like SAS, Python, Tableau are used in this report. Although analyzing high load of data is hard to perform, scientist first selects a smaller set of data to be analyzed as an alternative.

2. Related Works

Yanga, et al. (2018) proposed a study to investigating the relationship between demand for taxi, land usage, and accessibility in Washington D.C. using GPS and GIS data. These datasets of latitude and longitude values were used to visualize geographic data to discover traffic densities. Further illustration in taxi trips was provided and other data exploration was performed including missing data handling.

Throughout the data analysis by Yange, et al. (2018), the model showed that residential density, employment density, and larger block size have high correlation with taxi demand. Not only that, he also revealed that demand for taxi was lower in the area nearby bus stops and the reverse was true in metro stations exist. Additionally, during winter season the demand for taxis was reduced [7].

The recent study by Zhu, et al. (2017) leveraging three-month taxi trajectory in Beijing, China to explore the urban mobility by analyzing the spatio-temporal pattern on streets, specifically on 5th Ring Road. Before analysis process, data was cleaned, grouped and sorted. Non-correlated attributes were eliminated and merely geolocation records were included for traffic analysis which were pickup and drop off points, date, time, and lastly, the status of trajectory. Data was transformed from coordinate-based taxi trip into a street-based trip by implementing simplified map-matching process techniques. As the result, the transformation of geolocation data into street unit facilitates urban planners and government policy makers to have a better understanding of urban structures, optimizing public resources, and have a greater sense of urban movement in every streets unit [8].

Kong, et al. (2017) proposed a taxi allocation model using three-dimensional attributes namely: time, location and relationship. In their work, the correlation between passengers' getting on and off on particular time and location was deeply explored by leveraging numerous data-driven analysis techniques on taxi trajectory datasets. The aim of Time-Location-Relationship (TLR) taxi matching model was to increase the opportunity of finding vacant taxis for passengers and at the same time reduce transportation resource consumption and air pollution [9].

Kong, et al. (2017) experiments revealed that there was a rise of get-on demand taken place in diplomacy areas for affairs purposes during morning time and immediately leave during lunch time, which contributes a get-on and off demand in restaurants and café districts. It has been highlighted that commercial entertainment areas like shopping malls has higher get-on demand at the peak at 12 PM. Not only that, the taxi demand is lesser during weekend in railway exits area. There was no doubt that data exploration on the pattern of passengers' going up and down in particular time and location can be useful to develop a better model to allocate drivers and make more profitable income to drivers. Not only that, it also reduces the number of resources used during taxi vacant time [9].

To sum up, data is the biggest asset for urban city planning, governments, and institution to solve the real-world traffic problems such as optimizing transportation resources, organizing major events, reduce traffic congestion, sensing the urban mobility and many more [10]. However, urban traffic insight is hard to analyze without data exploration process. Therefore, data preparation and feature engineering are the necessary tasks to transform raw information into meaningful insight of the traffic condition of big cities.

3. Materials and Methods

In this paperwork, data is obtained through Kaggle repository titled "New York Taxi Fare Prediction" which the data composes of taxi ride in New York City from 2009 to 2015. The dataset attributes including key, pickup and drop-off geolocation, date, time, passenger count and finally fare amount will be analyzed. In total, the given dataset composes of nine core attributes. The following is the data description as listed in table 1.

Table 1: NYC taxi dataset description

Features	***Description***
key	The row identification
fare_amount	Total fare calculated by time and distance
pickup_date	Date when the meter is used
pickup_time	Time when the meter is used
pickup_longitude	Longitude at pickup point
pickup_latitude	Latitude at pickup point
dropoff_longitude	Longitude at drop off point
dropoff_latitude	Latitude at drop off point
passenger_count	Number of passengers in the taxi

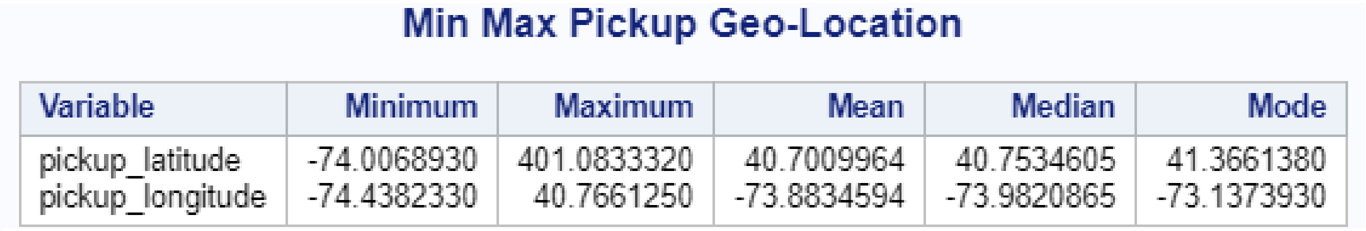
3.1. Data Preperation

It is the best practice to handle missing data and remove outlier before proceeding to data exploration. Missing data happens when data point is not recorded for an observation within a dataset [11]. In this case, technique like mean imputation is used to replace any missing data present in the NYC taxi dataset. Missing data in **fare_amount** attribute and geolocation points are replaced using mean imputation. Additionally, there are outlier data present in this taxi dataset. Outlier is an observation in a random sample from a population that lies an abnormal distance from other values. A mistake during data collection might introduce the outlier in dataset or it is just an indication of variance in the dataset [12].

As taxi dataset is dealing with geolocation both latitude and longitude data. Hence, boxplot or other statistic functions cannot be used to identify outliers in geospatial data. Therefore, in order to get a better view of data, latitude and longitude should plot the coordinates on a map to get a better view. Since the scope of taxi data is focused on yellow taxis, which yellow cabs are the only vehicles which have the right to pick up prearranged passengers and anywhere in New York City. Hence, it is necessary to identify the NYC coordinates. The coordinates and boundary according to NYC Department of City Planning (2020) are listed in table 2 [13].

Table 2: New York City Coordinates

Longitude	**-74.006**
Latitude	**40.7142**
West Longitude	-74.257159
East Longitude	-73.699215
South Latitude	40.495992
North Latitude	40.915568



Min Max Pickup Geo-Location

Variable	Minimum	Maximum	Mean	Median	Mode
pickup_latitude	-74.0068930	401.0833320	40.7009964	40.7534605	41.3661380
pickup_longitude	-74.4382330	40.7661250	-73.8834594	-73.9820865	-73.1373930

Figure 1: Min and max pickup point coordinates value

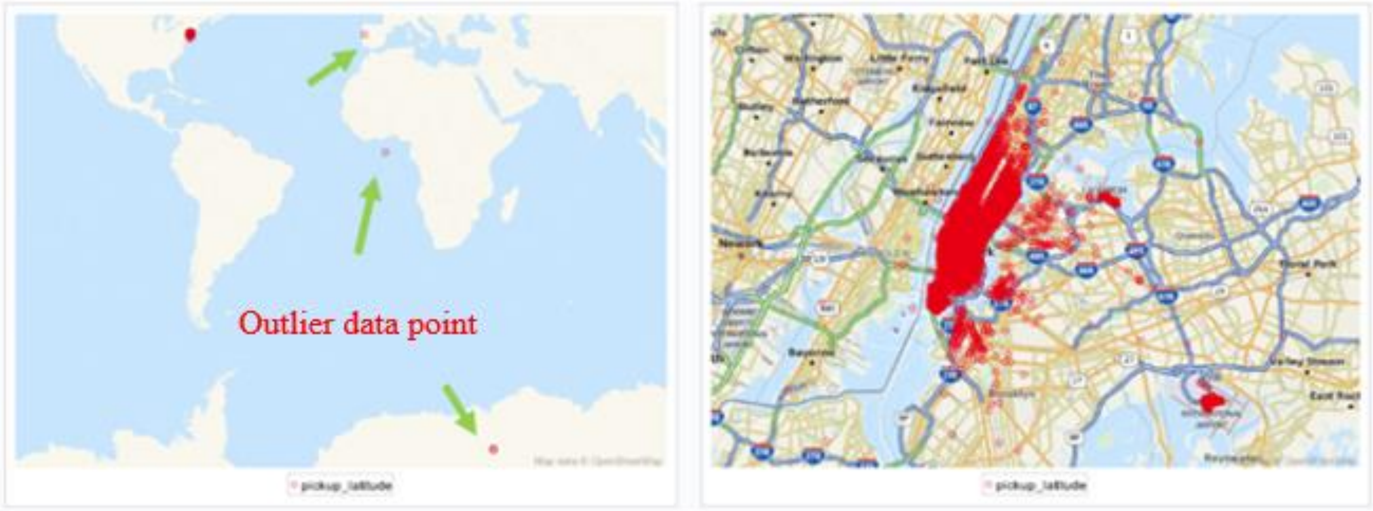
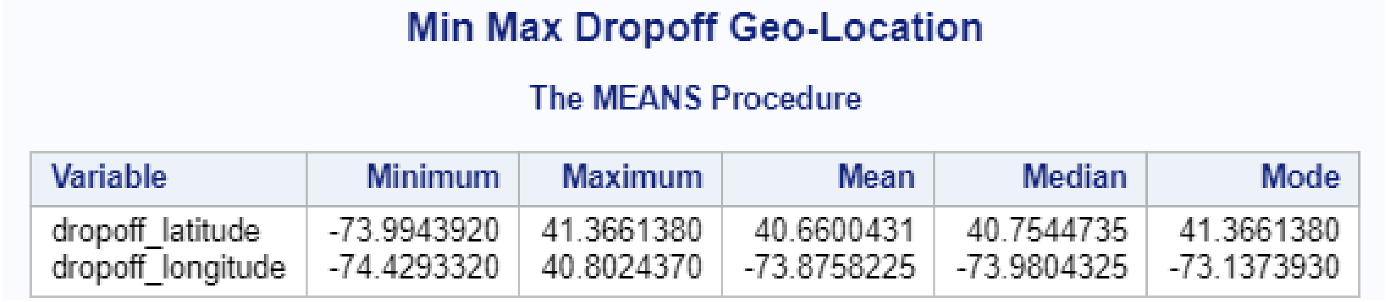


Figure 2: Before outlier treatment (a) and after outlier treatment (b)



Min Max Dropoff Geo-Location

The MEANS Procedure

Variable	Minimum	Maximum	Mean	Median	Mode
dropoff_latitude	-73.9943920	41.3661380	40.6600431	40.7544735	41.3661380
dropoff_longitude	-74.4293320	40.8024370	-73.8758225	-73.9804325	-73.1373930

Figure 3: Min and max drop off point coordinates value

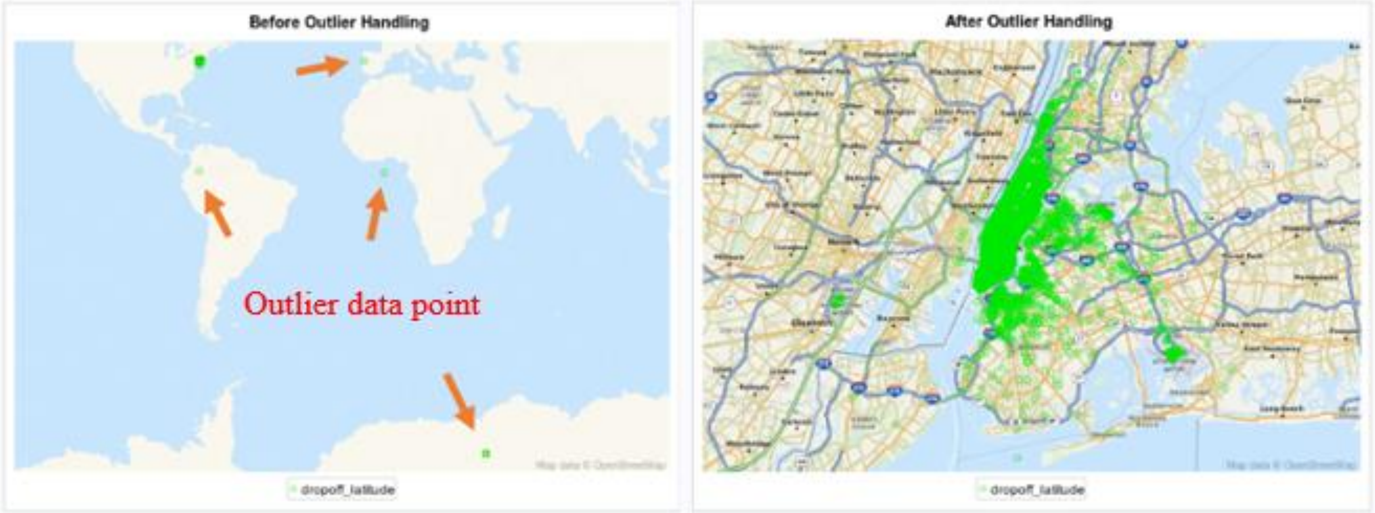


Figure 4: Before outlier treatment (a) and after outlier treatment (b)

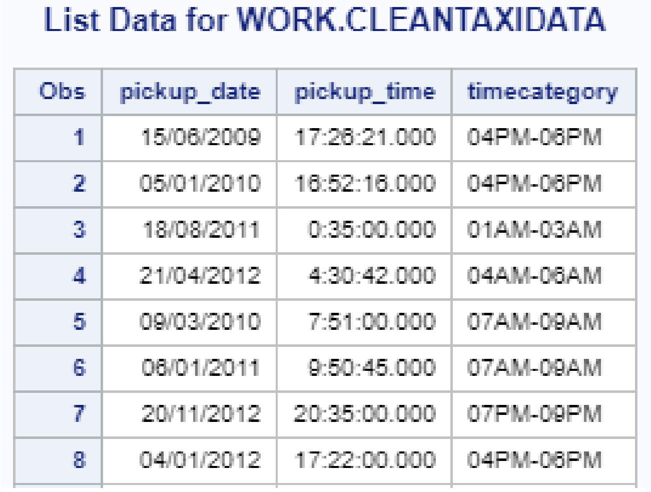
As can be seen in Figure 1, the max point of **pickup_latitude** is **401.7009964** which is an abnormal value. Additionally, there is an outlier in **pickup_longitude**, considering that NYC coordinates lies between **-74.257159** to **-73.699215**. However, the maximum value of **pickup_longitude** value is **40.7661250**. These outliers lead to abnormal pickup points which are illustrated in Figure 2a, where the location pickup points are coordinates to the east side of Africa and Antarctica. Hence, in order to remove the outliers, data is replaced with the mean rate of pickup latitude and longitude. As the result of replacing outlier within the border of NYC coordinates, it shows all the pickup points are focused on New York boundary as illustrated in Figure 2b.

Similar to drop off location cases, there are abnormal values of **dropoff_latitude** and **dropoff_longitude**. According to Table 2, the longitude range lies between -**74.257159** and **-73.699215**. However, according to Figure 3, the maximum rate of **dropoff_latitude** is out of the range of NYC boundary. The same issue also occurs **in dropoff_latitude** value, where the latitude range of NYC is within **40.495992** and **40.915568**. therefore, the minimum rate shows outliers. As the result, with the outliers data, coordinates are plotted on map resulted abnormal drop off points in the east of Africa, South America, and Antarctica as illustrated in Figure 4a. Therefore, in order to remove the outliers, data is replaced with the mean rate of latitude and longitude as shown in Figure 3 which are **40.6600431** and **-73.8758225**. As the result of replacing outlier within the border of NYC coordinates, it shows all the drop off points is coordinated on New York as illustrated in Figure 4b

2.1. Feature Engineering

1. *Time Grouping Operation*

Based on the New York Taxi dataset, the **pickup_time** attribute is in UTC Time format which will be hard to be interpreted into graphs. Therefore, this attribute is binned into eight categories which have 3 hour-interval on each category. As the result, **timecategory** attribute is created as follows:



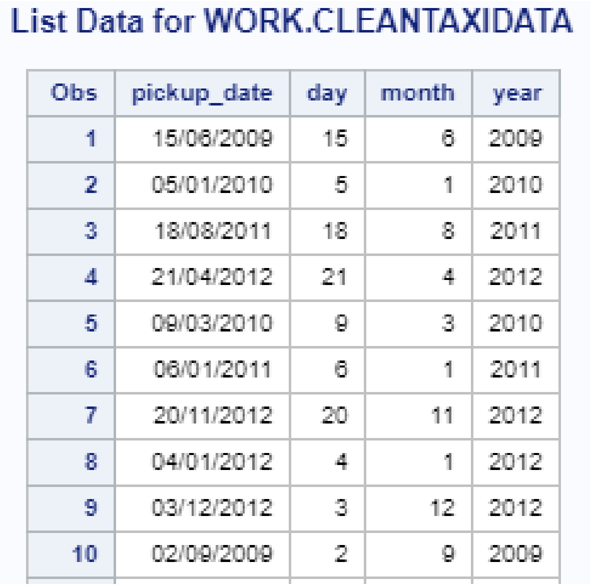
List Data for WORK.CLEANTAXIDATA

Obs	pickup_date	pickup_time	timecategory
1	15/06/2009	17:26:21.000	04PM-06PM
2	05/01/2010	16:52:16.000	04PM-06PM
3	18/08/2011	0:35:00.000	01AM-03AM
4	21/04/2012	4:30:42.000	04AM-06AM
5	09/03/2010	7:51:00.000	07AM-09AM
6	06/01/2011	9:50:45.000	07AM-09AM
7	20/11/2012	20:35:00.000	07PM-09PM
8	04/01/2012	17:22:00.000	04PM-06PM

Figure 5: Feature creation of timecategory attribute

2. *Date Extraction*

Similar with **pickup_time** column, **pickup_date** column is in dd/mm/yyyy date format. Hence, data extraction is performed by extracting the column into separated column of date, month and year by using SAS functions which are **day()**, **month()** and **year()**. As the result, three new attributes are created as follows:



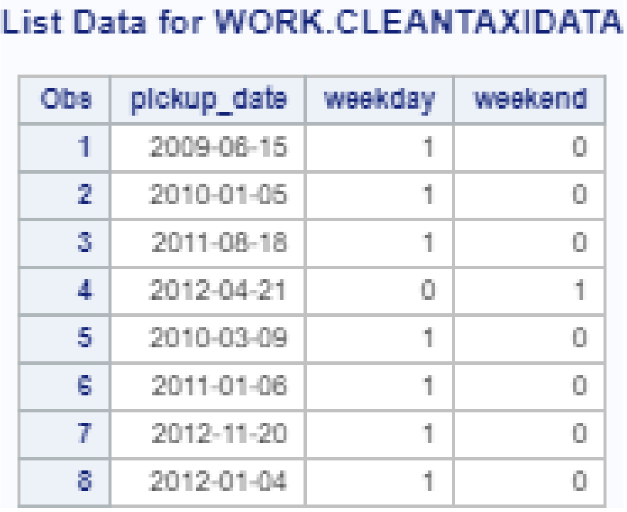
List Data for WORK.CLEANTAXIDATA

Obs	pickup_date	day	month	year
1	15/06/2009	15	6	2009
2	05/01/2010	5	1	2010
3	18/08/2011	18	8	2011
4	21/04/2012	21	4	2012
5	09/03/2010	9	3	2010
6	06/01/2011	6	1	2011
7	20/11/2012	20	11	2012
8	04/01/2012	4	1	2012
9	03/12/2012	3	12	2012
10	02/09/2009	2	9	2009

Figure 6: Feature creation of day, month and year attributes

3. *Weekday/Weekend One-Hot Encoding*

It is important to get an overview of taxi demand during weekdays and weekends. Therefore, **pickup_date** attribute is extracted to determine whether the taxi trips happen on either weekdays or weekends. **Weekday()** SAS function is used to identify either the given data is weekday or weekend. As the result, new columns are created as follows:



List Data for WORK.CLEANTAXIDATA

Obs	pickup_date	weekday	weekend
1	2009-06-15	1	0
2	2010-01-05	1	0
3	2011-08-18	1	0
4	2012-04-21	0	1
5	2010-03-09	1	0
6	2011-01-06	1	0
7	2012-11-20	1	0
8	2012-01-04	1	0

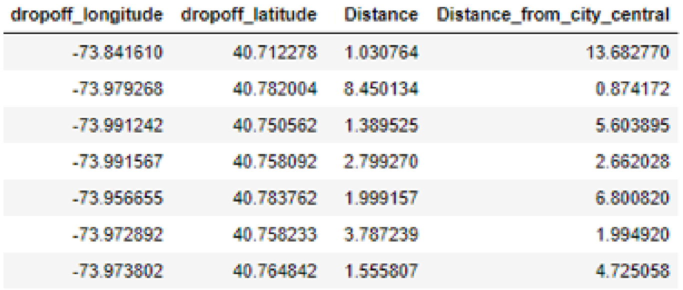
Figure 7: Feature creation of weekday and weekend attributes

4. Taxi Trip Distance

With the given longitude and latitude information, it can be used to measure the geographic distance. To calculate the distance of great circle between two points on the earth Haversine equation is used [14]. Several parameters including **pickup_longitude**, **pickup_latitude**, **dropoff_longitude**, **dropoff_latitude** are used. Following is the Haversine equation [15]:

$$a = \sin^2\left(\frac{\Delta\varphi}{2}\right) + \cos\varphi 1 \cdot \cos\varphi 2 \cdot \sin^2\left(\frac{\Delta\lambda}{2}\right)$$
$$c = 2 \cdot \operatorname{atan2}\left(\sqrt{a}, \sqrt{(1-a)}\right)$$
$$d = R \cdot c \qquad (1)$$

Using python code, Haversine Equation is formulated. There are two feature created which are: (1) **distance** which refers to the distance between pickup and drop-off point and (2) **distance_from_city_central** refers to the distance between pickup point and NY city center point (**-74.006, 40.7142**) which is derived from NY coordinates in table 2. As the result, the new distance attribute is created as follows:



dropoff_longitude	dropoff_latitude	Distance	Distance_from_city_central
-73.841610	40.712278	1.030764	13.682770
-73.979268	40.782004	8.450134	0.874172
-73.991242	40.750562	1.389525	5.603895
-73.991567	40.758092	2.799270	2.662028
-73.956655	40.783762	1.999157	6.800820
-73.972892	40.758233	3.787239	1.994920
-73.973802	40.764842	1.555807	4.725058

Figure 8: Feature creation of distance attribute

2.2. Exploratory Data Analysis (EDA)

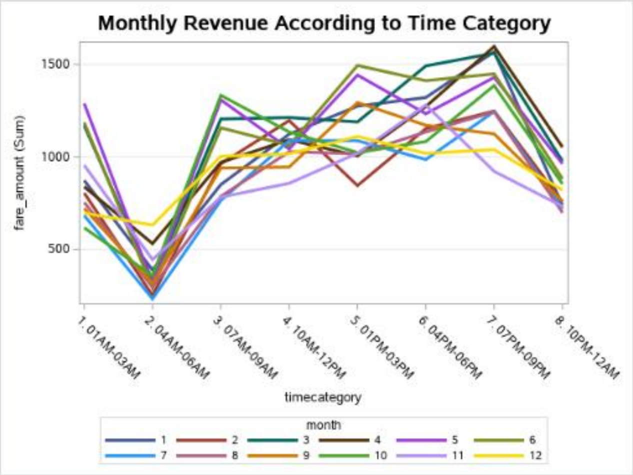


Figure 9: Monthly Revenue According to Time Category

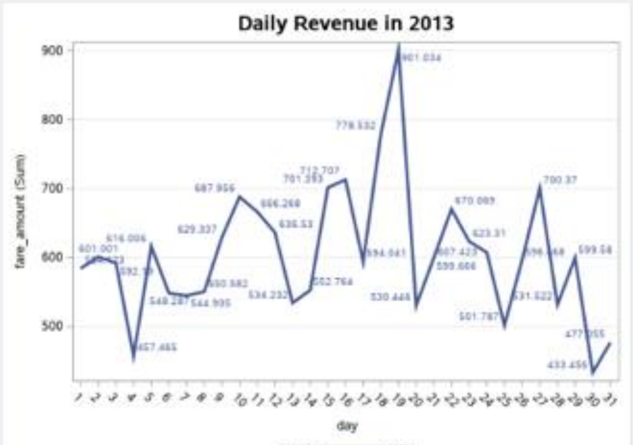


Figure 10: Daily Revenue in 2013

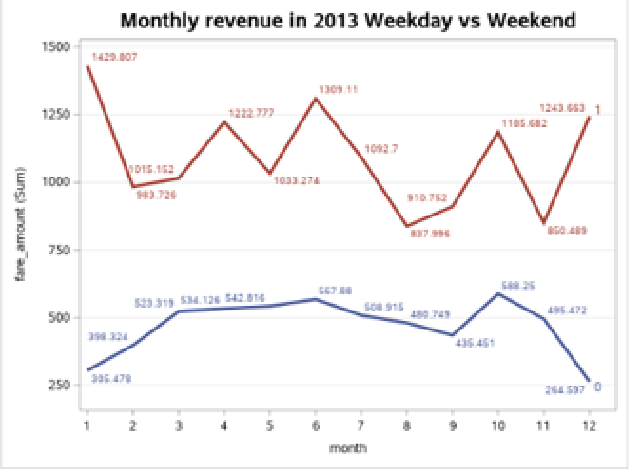


Figure 11: Monthly revenue in 2013 weekday vs weekend

Exploratory Data Analysis (EDA) is defined as the process of performing an investigation on data in the form to identify anomalies, discover trend patterns, and summarize data characteristics by leveraging summary statistics and visual representations. Not only that, EDA helps the researcher to see what the data can explain beyond the formal modelling. This data analysis is essential to gain a better insight [16]. As can be seen from the line graph in Figure 9, the graph shows there is a significant decrease in taxi revenue and demand during **04 AM** to **06 AM** and the demand started to dramatically increase from **07AM** to **09AM** and reach its peak time between **07 PM** to **09 PM**.

Moreover, Figure 10 represents the daily revenue in taxi in 2013 where the trends show that there is a drastic drop every 5 days. This trends leads to the further study that reveals there is significant difference in demand during weekdays and weekends. Figure 11 illustrates is the comparison revenue of NYC Taxi in 2013 where the red line represents weekday revenue and blue like represents weekend revenue. Based on the graph, the taxi demand during weekday is significantly higher compared to the demand during weekends.

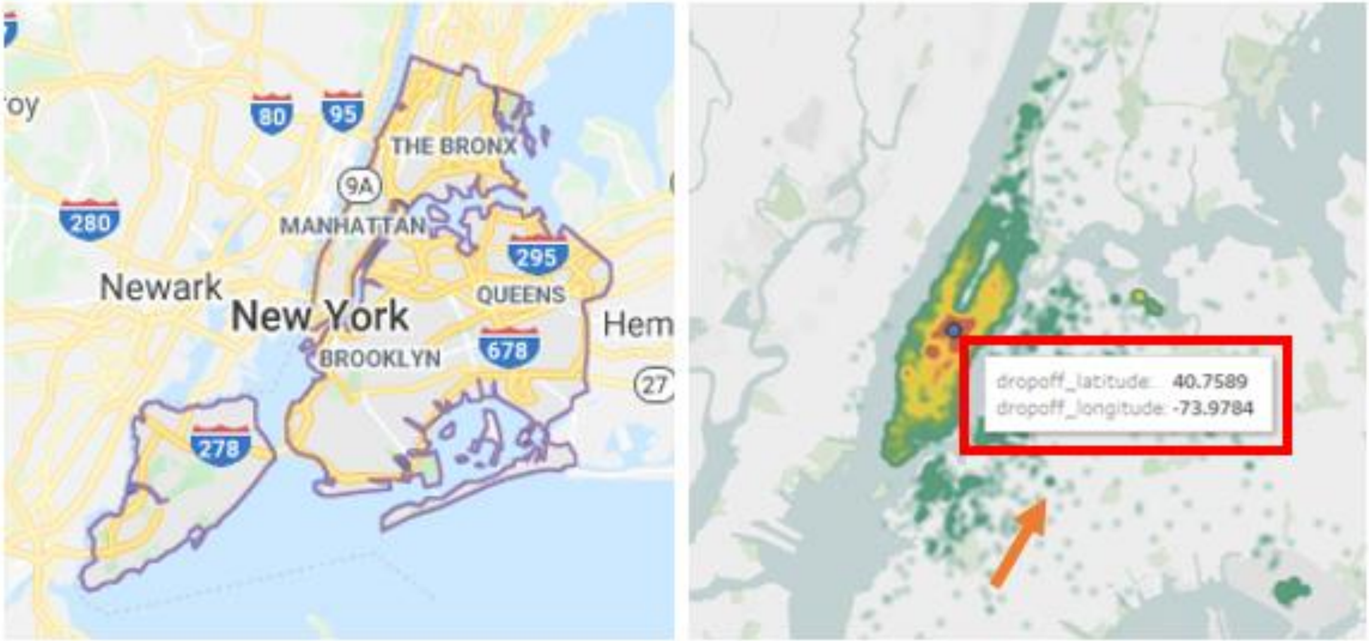


Figure 12: The map of New York City (a) and the midpoint which has highest demand (b)

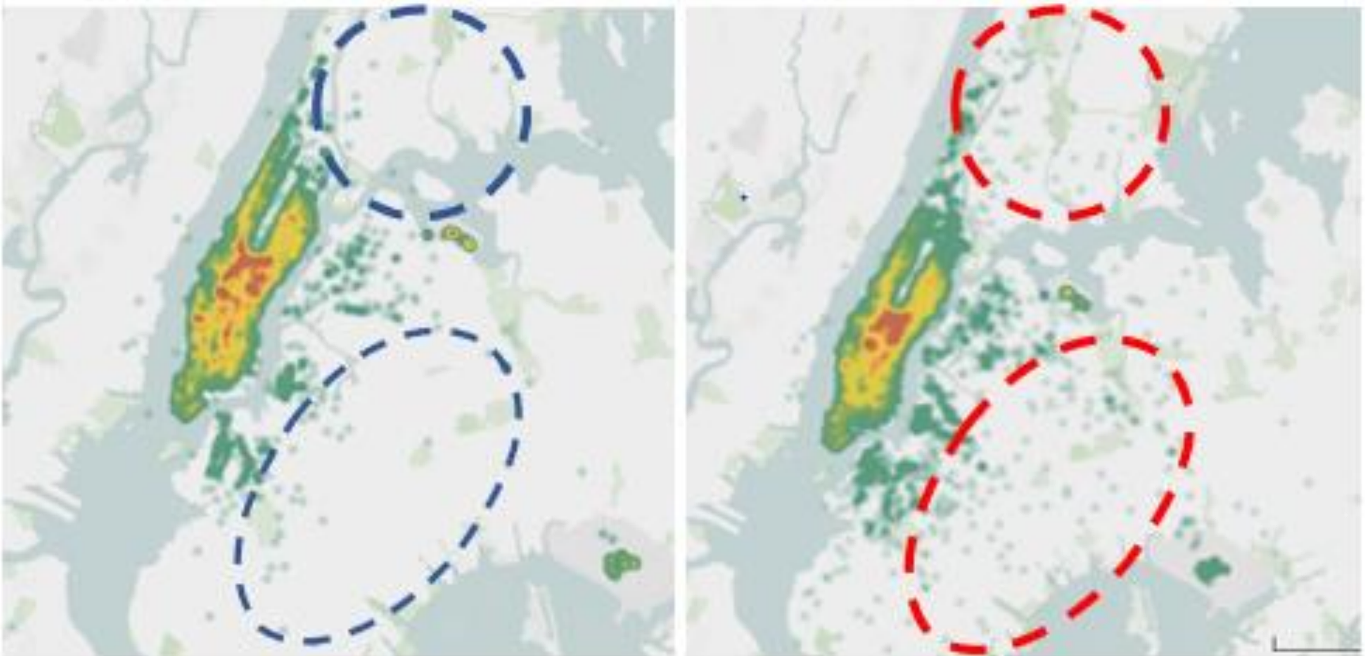


Figure 13: Pickup demand density (a) and drop off demand density (b)

Taxi demand can be clearly visualized by plotting the coordinates into a map using Tableau visualization tool. In Figure 12a, the New York City map covers five boroughs namely The Bronx, Brooklyn, Manhattan, Queens, and Staten Island. The red zone in Figure 12b represents the highest density and also the midpoint which falls under **40.7589° N 73.9784° W** in Manhattan City. With that being said, the demand density indicates that there is a high taxi demand in Manhattan city center.

Moreover, The Figure 13 breaks the data further into pickup and drop off density. The Figure 13a refers to the demand density for **pickup**. While Figure 13b refers to taxi demand density for **drop-off**. The further analysis then reveals that there is **higher demand in Manhattan** for both pickup and dropoff points. On the other hand, there are **drop-off demands** scatter outside Manhattan City (Brooklyn, Queens, and the Bronx) where it is true that there is **less pickup demand** located outside Manhattan.

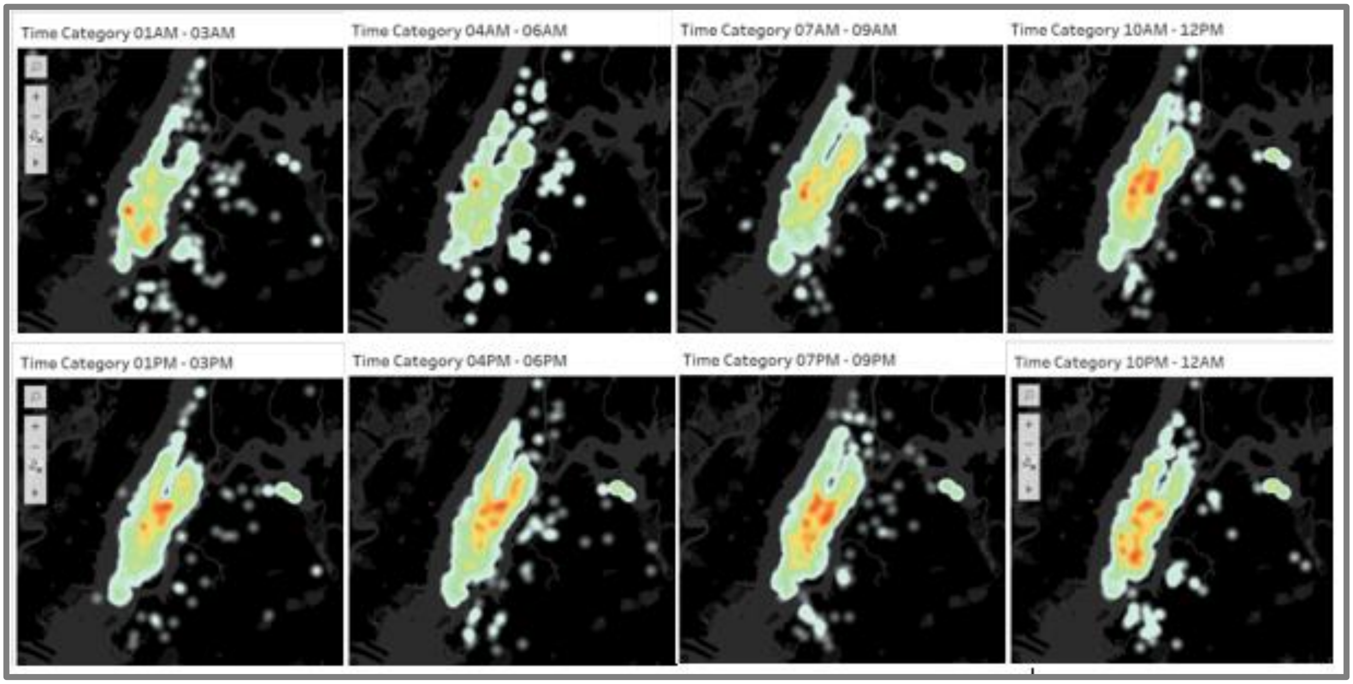

Time Category 01AM - 03AM
Time Category 04AM - 06AM
Time Category 07AM - 09AM
Time Category 10AM - 12PM
Time Category 01PM - 03PM
Time Category 04PM - 06PM
Time Category 07PM - 09PM
Time Category 10PM - 12AM

Figure 14: Pickup demand on timely basis

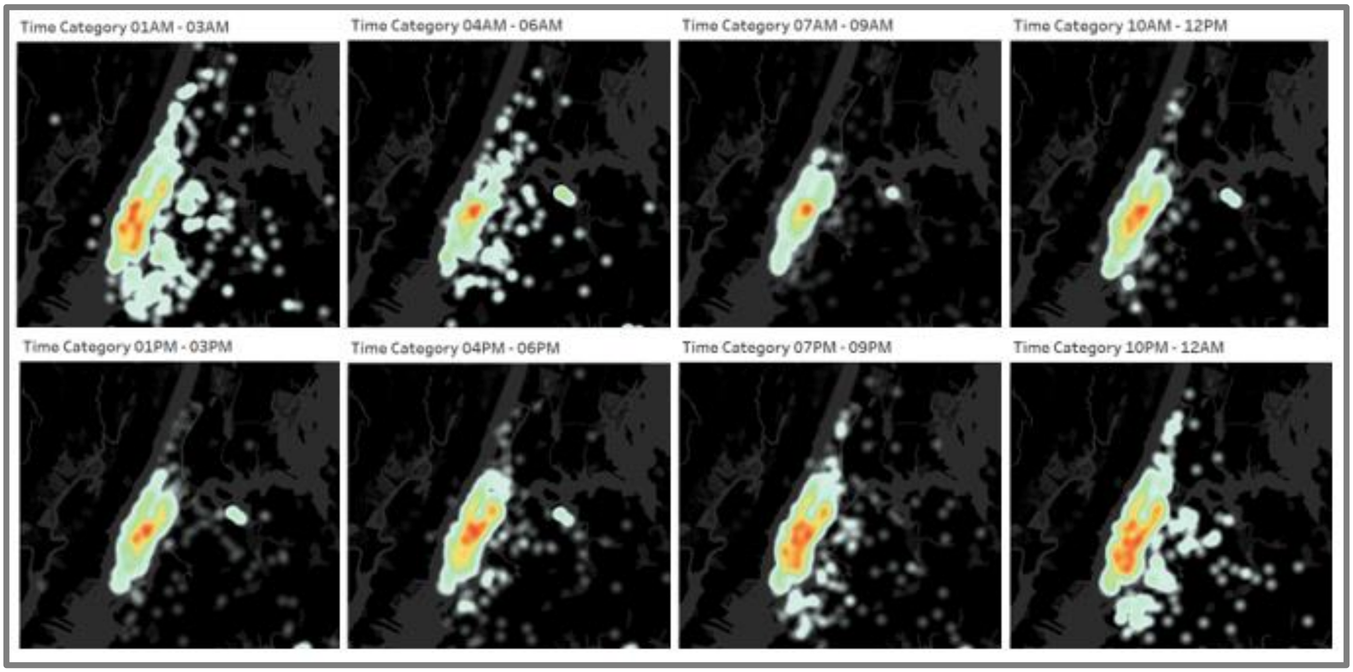

Time Category 01AM - 03AM
Time Category 04AM - 06AM
Time Category 07AM - 09AM
Time Category 10AM - 12PM
Time Category 01PM - 03PM
Time Category 04PM - 06PM
Time Category 07PM - 09PM
Time Category 10PM - 12AM

Figure 15: Drop off demand on timely basis

As can be seen from the density map in Figure 14, it shows the **pickup** taxi demand base on timely basis. The pickup demand is significantly reduced from **04 AM** to **06 AM** where there is only one small point of the red area. The demand started to increase from **07 AM** to **09 AM** and reaches its peak time during **07 PM** to **09 PM** where there are more red spots present during that time. Additionally, there are more trips travels outside Manhattan City during **10 PM** to **03 PM.**

On the other hand, Figure 15 shows the **drop-off** demand on timely basis. Unlike pickup demand, during **07 AM** to **09 AM,** it has the lowest demand. It slowly increases and reached at its peak time at **01 AM** to **03 AM**. Additionally, there are more taxi trips travel in Brooklyn area started from **07 PM** till **04 PM** which it reaches the peak time during **01 AM** to **03 AM.**

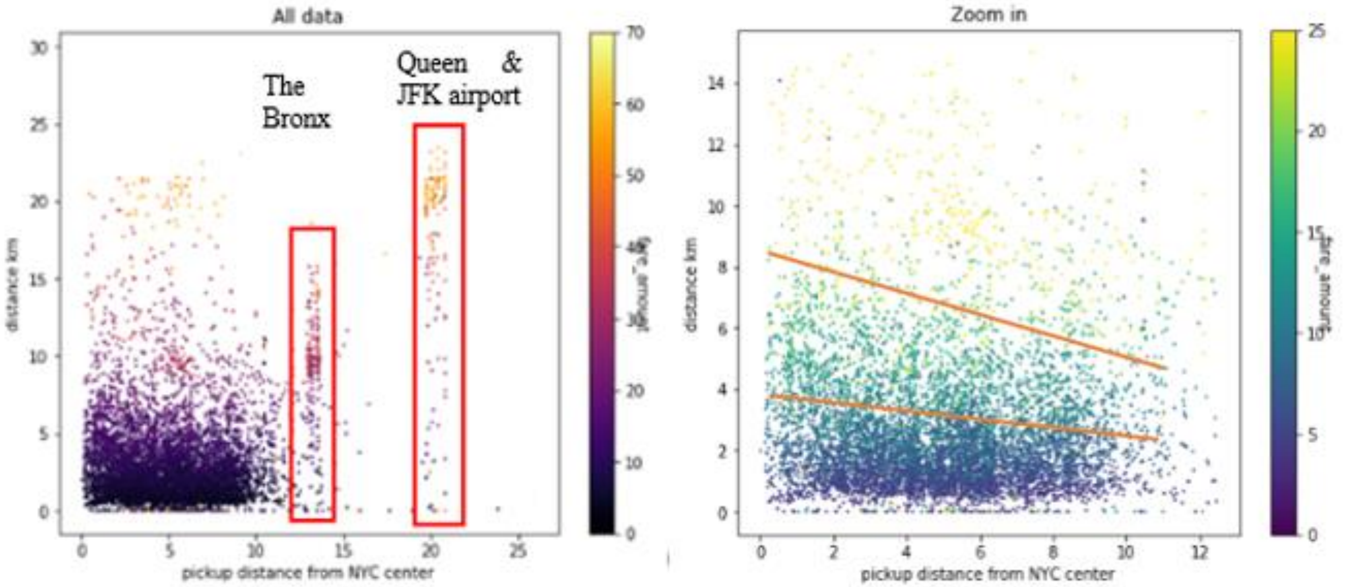
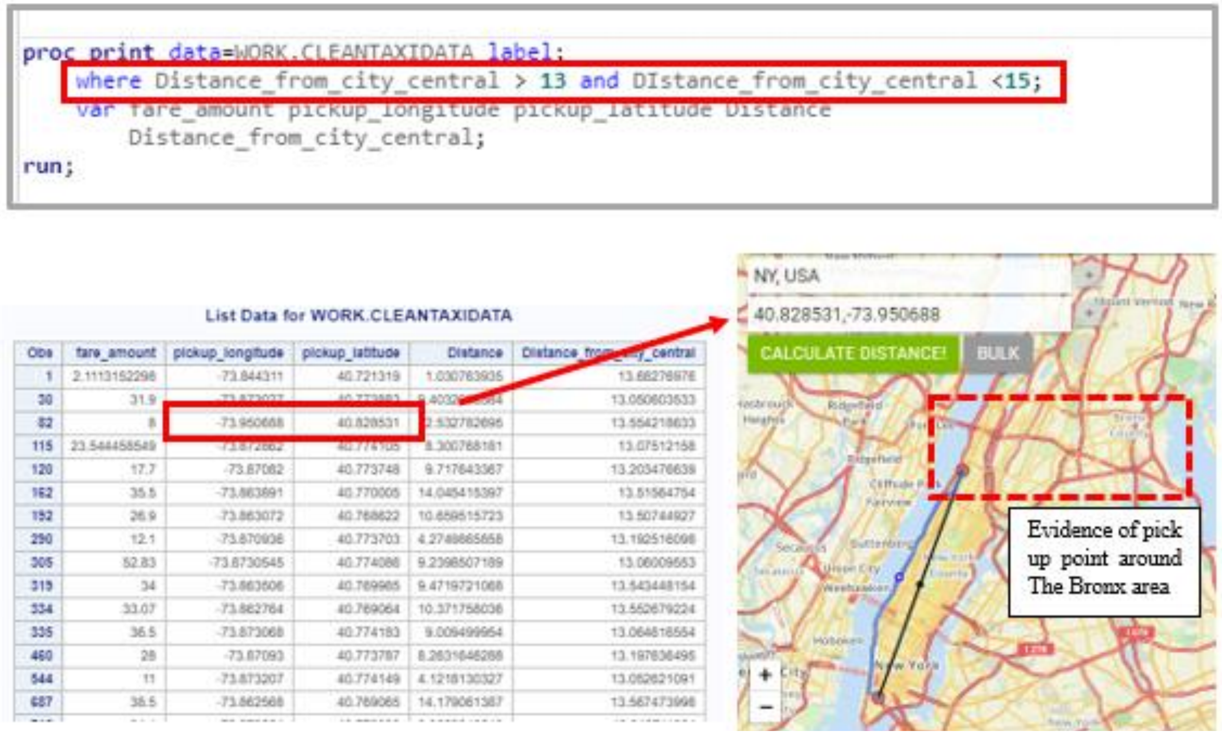


Figure 11: Pickup distance from NYC centre vs fare amount(a) Pickup distance from NYC centre vs fare amount zoom in (b)

The figure above shows the correlation between pickup distance from the NY city center and taxi fare trip. This analysis reveals that the further the pickup distance from the city center, the higher the taxi fare trips. This can be seen as the dark purple dots turn to light purple and orange as the distance from city center is getting further. Moreover, the virtualization above reveals that there is **a straight bar line of light purple dots** that indicates the pickup point is located around 13km away from the city center which is assume to be around **The Bronx City** and it costs around $20 - $30 (refer to Figure 17).

Additionally, the **longer straight bar line of orange dots** indicates the pickup point is located around 20km away from the city center which is assume to be located around either **Queens** or **JFK airport** that cost about $50-$60 (refer to Figure 18). In Figure 16b, the red line indicates as separation of three fare grouping which are dark purple, blue and yellow. the further the pickup location distance from the city center, the fare price tends to be higher which shown on the color that turns lighter to blue and yellow.



```
proc print data=WORK.CLEANTAXIDATA label;
    where Distance_from_city_central > 13 and DIstance_from_city_central <15;
    var fare_amount pickup_longitude pickup_latitude Distance
        Distance_from_city_central;
run;
```

List Data for WORK.CLEANTAXIDATA

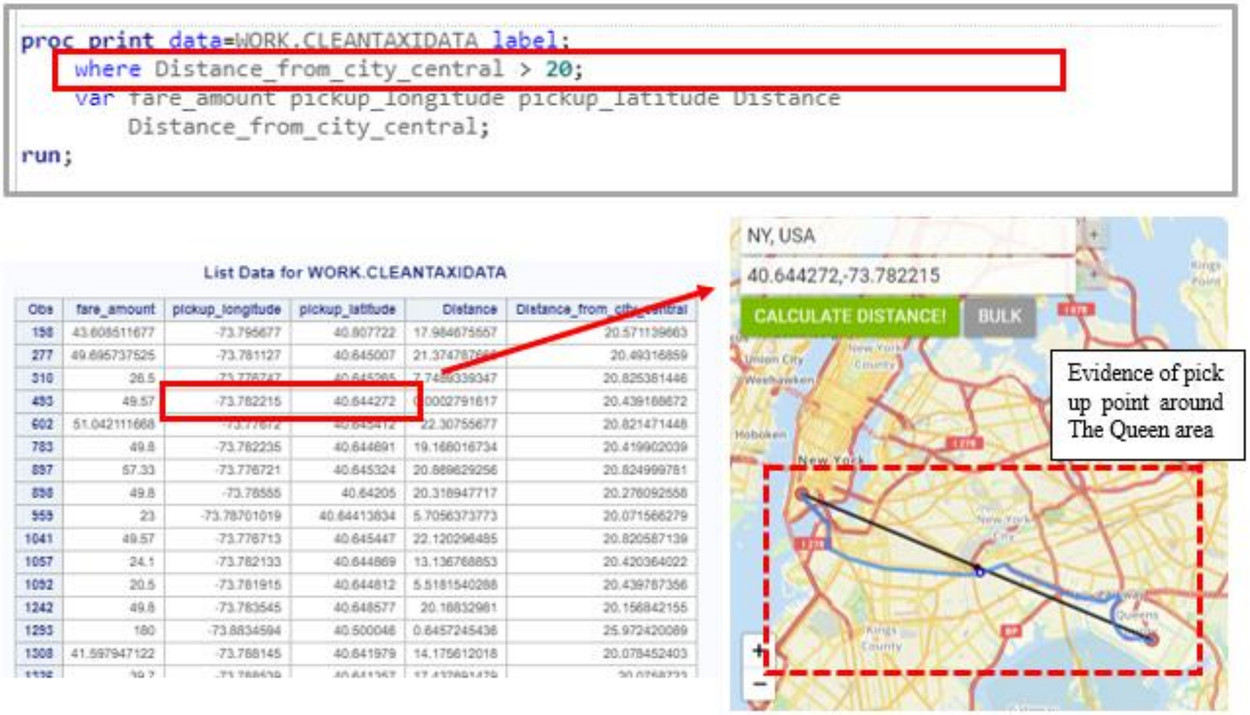
Obs	fare_amount	pickup_longitude	pickup_latitude	Distance	Distance_from_city_central
1	2.1113152298	-73.844311	40.721319	1.030763935	13.68276976
30	31.9	-73.873027	40.773883	[illegible]	13.050603533
82	8	-73.950688	40.828531	2.532782696	13.554218633
115	23.544458549	-73.872862	40.774105	8.300768181	13.07512158
120	17.7	-73.87082	40.773748	9.717643367	13.203476639
162	35.5	-73.863891	40.770005	14.045415397	13.51564754
192	26.9	-73.863072	40.768622	10.659515723	13.50744927
290	12.1	-73.870936	40.773703	4.2749865658	13.192516098
305	52.83	-73.8730545	40.774088	9.2398507189	13.06009563
319	34	-73.863506	40.769965	9.4719721068	13.543448154
334	33.07	-73.862764	40.769064	10.371758036	13.552679224
335	36.5	-73.873068	40.774183	9.009499954	13.064816554
460	28	-73.87093	40.773787	8.2831646288	13.197636495
544	11	-73.873207	40.774149	4.1218130327	13.052621091
687	38.5	-73.862568	40.769065	14.179061367	13.567473998

NY, USA

40.828531,-73.950688

Evidence of pick up point around The Bronx area

Figure 17: Query of Distance from city central between 13km and 15km (a) output query (b)



```
proc print data=WORK.CLEANTAXIDATA label;
    where Distance_from_city_central > 20;
    var fare_amount pickup_longitude pickup_latitude Distance
        Distance_from_city_central;
run;
```

List Data for WORK.CLEANTAXIDATA

Obs	fare_amount	pickup_longitude	pickup_latitude	Distance	Distance_from_city_central
158	43.608511677	-73.795677	40.807722	17.984675557	20.571139663
277	49.695737525	-73.781127	40.645007	21.374787660	20.49316859
310	26.5	-73.776747	40.645285	7.7459339347	20.825361446
493	49.57	-73.782215	40.644272	0.0002791617	20.439188672
602	51.042111668	-73.77672	40.645412	22.30755677	20.821471448
783	49.8	-73.782235	40.644691	19.168016734	20.419902039
897	57.33	-73.776721	40.645324	20.889629256	20.824999781
898	49.8	-73.78555	40.64205	20.318947717	20.276092558
959	23	-73.78701019	40.64413834	5.7056373773	20.071566279
1041	49.57	-73.776713	40.645447	22.120298485	20.820587139
1057	24.1	-73.782133	40.644869	13.136768853	20.420364022
1092	20.5	-73.781915	40.644812	5.5181540288	20.439787356
1242	49.8	-73.783545	40.648577	20.18832981	20.156842155
1293	180	-73.8834594	40.500046	0.6457245438	25.972420089
1308	41.597947122	-73.788145	40.641979	14.175612018	20.078452403
1336	39.7	-73.788539	40.641357	17.437891479	20.0768733

Evidence of pick up point around The Queen area

Figure 18: Query of Distance from city central bigger that 20km (a) output query (b)

In Figure 17a, several taxi trips observations where the distance trips between 13 km and 15 km are listed out using the given SAS query. After the query has been executed, one of the outputs of pickup location is selected and a map calculator is used to prove whether the given latitude and longitude are coordinates around the Bronx. As the result, the map output shows the coordinates are pointed at The Bronx area which is about 13 km away from the city center.

Additionally in Figure 18a, several taxi trips observations where the distance trips greater than 20 km are listed out using the given SAS query. After the query has been executed, one of the outputs of pickup location is selected and a map calculator is used to prove whether the given latitude and longitude are coordinates around either Queens or JFK Airport area. As the result, the output of the map shows the coordinates are pointed at Queens area which is about 20 km away from the city center.

4. Discussion

Throughout the research, this offers new possibilities of urban understanding through the geographic information. Not only that, large geo-data spatial assembly is an important analytical step to summing up and perceiving the geographical environment from taxi demand. The data offers a more accurate definition of a region's existence, taking into account the everyday movement and activities found in geospatial data that indicate the socio-economic properties of urban functions.

The aim of this research is to obtain an insight of taxi demand in New York City. In this research, the data of taxi ride in New York City from 2009 – 2015 is used for data exploration. The data is complex and involves geographical and temporal components as well as several variables associated with each ride. Therefore, the commonly used tools and platforms like SAS, Python, Tableau is used in this report to support the data exploration and pre-processing.

Initial data preparation involves missing data handling and outlier treatment. SAS is used for analytical and data management. Not only that, Python is used to calculate the distance of the taxi by utilizing Haversine equation. There are outlier cases which are: (1) abnormal pickup location and (2) abnormal drop off location. These abnormal coordinates are treated using mean imputation. As the result, data is more presentable where geolocation data is plotted into a map in which the range of data points are within New York City coordinates.

Then, Exploratory Data Analysis (EDA) is performed to identify anomalies, discover trend patterns, summarize data characteristics by leveraging summary statistics and visual representations. A comparison has been made before and after data transformation. Raw data without statistical analysis is hard to be analyzed or even interpreted. However, after data transformation, taxi timestamp is extracted into date, month, year, and pickup time is categorized into bins. With that being said, data can be visualized in either daily, monthly or yearly basis. This enables to gain a better understanding of the taxi revenue and demand. Several figures are presented to gain an overview of what the data can explain beyond the formal modelling. Moreover, the further analysis of geolocation reveals that there is higher demand in Manhattan for both pickup and dropoff points. Besides, there are several drop-off demands scatter in The Bronx, Brooklyn, and Queens which reverse is true in pickup demand on those areas. Not only that, the demand density plot reveal the lowest demand occurs during 04 PM – 06 PM.

5. Conclusions

To sum up, data is the biggest asset for urban city planning. Geographic information helps the government or institution to have a better understanding of urban city in order to solve the real-world problem such as optimizing transportation resources, urban mobility and many more. However, this insight is hardly obtained without data exploration process. Therefore, data preparation and feature engineering are the necessary tasks to organizing major events, reduce traffic congestion, sensing the transform raw information into meaningful insight of the traffic condition of big cities like New York. With increasing amounts of urban data being collected and made available, new possibilities of data-driven research emerge lead to changes in traffic condition through scientific proof-based decision-making and strategies.

References

[1] Jansson, K., Andreasson, I. & Kottenhoff, K., 2016. Public Transport in the Era of ITS: Forms of Public Transport. *Springer Tracts on Transportation and Traffic,* Volume doi:10.1007/978-3-319-25082-3_2 , pp. 29-83.

[2] Çolak, S., Lima, A., C., M. & Gonza´lez, 2016. Understanding congested travel in urban areas. *Nature Communications,* Volume doi:10.1038/ncomms10793 .

[3] Thakuriah, P. (., Tilahun, N. Y. & Zellner, M., 2019. Big Data and Urban Informatics: Innovations and Challenges to Urban Planning and Knowledge Discovery. *Seeing Cities Through Big Data,* Volume doi.org/10.1007/978-3-319-40902-3_2, pp. 11-45.

[4] Tang, J. et al., 2019. Understanding Spatio-Temporal Characteristics of Urban Travel Demand Based on the Combination of GWR and GLM. *Sustainability,* 11(19).

[5] Xu, X., Ren, J., Zhu, L. & Zhang, L., 2019. Intelligent Device System of Urban Transportation. *IEEE 8th Joint International Information Technology and Artificial Intelligence Conference,* Volume doi: 10.1109/ITAIC.2019.8785876, pp. 1729-1732.

[6] New York City Taxi and Limousine Commission, 2018. *TLC Publishes 2018 Factbook,* New York: The New York City Taxi and Limousine Commission.

[7] Yanga, Z. et al., 2018. Analysis of Washington, DC taxi demand using GPS and land-use data. *Journal of Transport Geography,* Volume 66 , pp. 35-44. doi:10.1016/j.jtrangeo.2017.10.021.

[8] Zhu, D., Wang, N., Wu, L. & Liu, Y., 2017. Street as a big geo-data assembly and analysis unit in urban studies: A case study using Beijing taxi data. *Applied Geography,* Volume doi:10.1016/j.apgeog.2017.07.001 , pp. 152-164.

[9] Kong, X. et al., 2017. Time-Location-Relationship Combined Service Recommendation Based on Taxi Trajectory Data. *IEEE Transactions on Industrial Informatics,* 13(3), pp. 1202-1212.

[10] Wei, L. & Yang, S., 2018. Based on Big Data Technology Analysis on the Mode and Countermeasures of Smart City Construction Operation Management. *2018 3rd International Conference on Smart City and Systems Engineering (ICSCSE).* , Volume doi:10.1109/icscse.2018.00043.

[11] Enders, C. K. & Baraldi, A. N., 2018. Missing Data Handling Methods. *The Wiley Handbook of Psychometric Testing,* Volume doi:10.1002/9781118489772.ch6, p. 139–185.

[12] Aggarwal, C. C., 2016. An Introduction to Outlier Analysis. *Outlier Analysis,* Volume doi.org/10.1007/978-3-319-47578-3_1, pp. 1-34.

[13] Department of City Planning, 2020. *New York City Borough Boundary Metadata,* New York: New York City Department of City Planning.

[14] Mahmoud, H. & Akkari, N., 2016. Shortest Path Calculation: A Comparative Study for Location-Based Recommender System. *2016 World Symposium on Computer Applications & Research (WSCAR),* Volume doi: 10.1109/WSCAR.2016.16.

[15] Winarno, E., Hadikurniawati, W. & Rosso, R. N., 2017. Location based service for presence system using haversine method. *2017 International Conference on Innovative and Creative Information,* Volume doi:10.1109/innocit.2017.8319153.

[16] Cox, V., 2017. Exploratory Data Analysis. *Translating Statistics to Make Decisions,* Volume doi.org/10.1007/978-1-4842-2256-0_3, pp. 47-74.