UBER FARE ANALYSIS

CHAPTER 1: INTRODUCTION

CHAPTER 2: DATA COLLECTION AND SAMPLING PROCEDURE

CHAPTER 3: DESCRIPTIVE STATISTICS

CHAPTER 4: ESTIMATION

‘

ABSTRACT

This project is about one of the world’s largest taxi company, Uber Inc. Uber is an international company located in 69 countries and around 900 cities around the world. The main objective of the project is to perform statistical analysis on Uber’s data to better understand the trends in its pricing. This project comprises Descriptive statistics and Inferential Statistics and provides useful insights which can be used to improve the business model of Uber.

CHAPTER 1

INTRODUCTION

1.1 Application

Uber delivers services to lakhs of people daily all over the world. This results in a lot of data being generated which needs to be managed properly. Statistical analysis of the data can help realize various trends and produce new business ideas for the benefit of the company. Since Uber uses dynamic pricing, further analysis would also support much accurate predictions of the fare.

1.2 Objective

The main objectives of our project are:

1. To Understand the dataset and perform any clean-up if required
2. To Analyze descriptive statistics
3. To Perform Regression and Analysis
4. To Analyze Inferential statistics
5. To perform Hypothesis testing

1.3 Hypothesis

Our hypothesis for the project is that the fare estimated for a trip and the distance traveled are linearly related. This hypothesis will further be tested in the coming sections.

CHAPTER 2

DATA COLLECTION

2.1 Source for the data

The dataset used for the project is the Uber Fares Dataset taken from Kaggle. It consists of 7 features. The dataset contains the following fields:

key - a unique identifier for each trip

fare_amount - the cost of each trip in usd

pickup_datetime - date and time when the meter was engaged

passenger_count - the number of passengers in the vehicle (driver entered value)

pickup_longitude - the longitude where the meter was engaged

pickup_latitude - the latitude where the meter was engaged

dropoff_longitude - the longitude where the meter was disengaged

dropoff_latitude - the latitude where the meter was disengaged

2.2 Random variables

The Continuous Random variables considered are Distance and Price(fare_amount)

2.3 Population and Sample

The entire dataset is taken as the population. Samples are a subset of the dataset and are chosen at random for statistical analysis.

CHAPTER 3

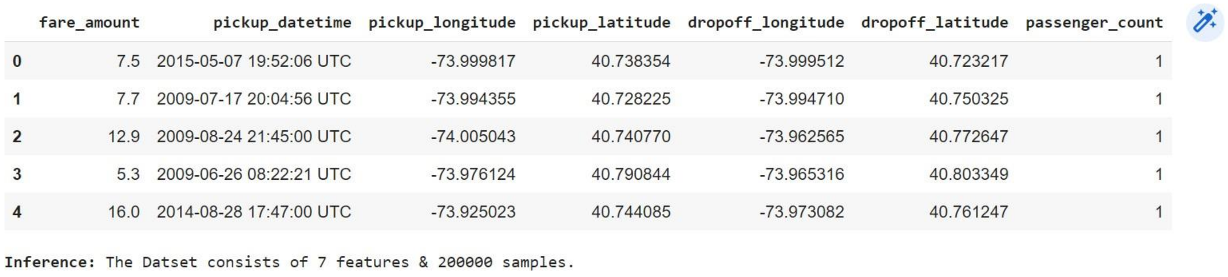
DESCRIPTIVE STATISTICS

This section collates the various visualizations as part of the exploratory data analysis performed. Descriptive statistics is a means of describing features of a data set by generating summaries about data samples.

Before performing descriptive statistics, Data preprocessing is done to drop unnecessary columns and obtain new columns in the dataset

DATA PREPROCESSING:

After the dataset is obtained, 2 unnecessary rows aOur dataset contains 7 features and 200k samples.

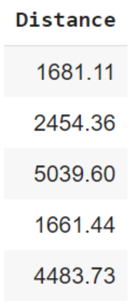


	fare_amount	pickup_datetime	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude	passenger_count
0	7.5	2015-05-07 19:52:06 UTC	-73.999817	40.738354	-73.999512	40.723217	1
1	7.7	2009-07-17 20:04:56 UTC	-73.994355	40.728225	-73.994710	40.750325	1
2	12.9	2009-08-24 21:45:00 UTC	-74.005043	40.740770	-73.962565	40.772647	1
3	5.3	2009-06-26 08:22:21 UTC	-73.976124	40.790844	-73.965316	40.803349	1
4	16.0	2014-08-28 17:47:00 UTC	-73.925023	40.744085	-73.973082	40.761247	1

```
Inference: The Datset consists of 7 features & 200000 samples.
```

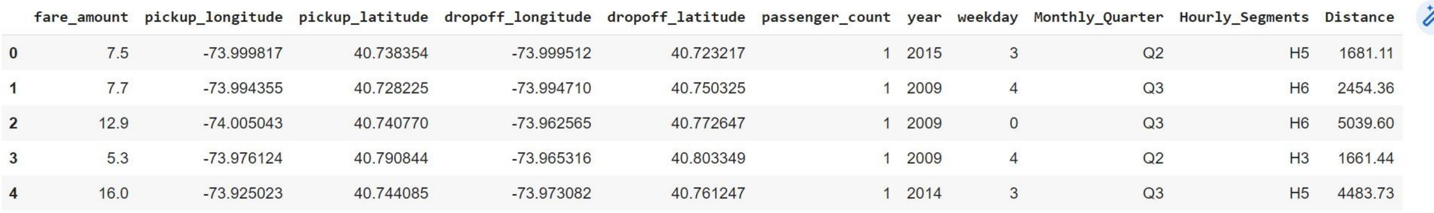
Then null values in the dataset are checked and dropping those particular rows results in 199987 samples.

A new column called ‘Distance’ is created which contains the distance values that are obtained from 4 existing columns namely ‘pickup_latitude, pickup_longitude, dropoff_latitude, dropoff_longitude’



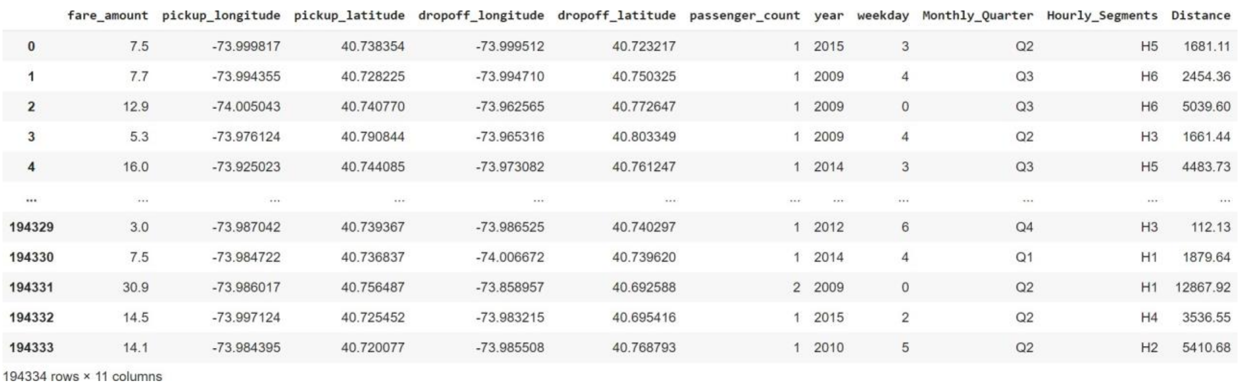
Distance
1681.11
2454.36
5039.60
1661.44
4483.73

We created other columns called ‘year’, ‘Monthly Quarter’ and ‘Hourly segments’. These are obtained from a single existing data column called ‘pickup_datetime’.
Here, the Monthly quarter represents 4 different categories namely Q1, Q2, Q3, Q4 where each of them represents a quarter of the year. For example, Q1 consists of the first 4 months and so on. Similarly, Hourly segments represent 6 different categories H1, H2, H3, H4, H5, H6 where each of them represents a section of the day. a 24 hour day is divided into 6 groups and a set of 4 consecutive hours is assigned each hourly segment.



	fare_amount	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude	passenger_count	year	weekday	Monthly_Quarter	Hourly_Segments	Distance
0	7.5	-73.999817	40.738354	-73.999512	40.723217	1	2015	3	Q2	H5	1681.11
1	7.7	-73.994355	40.728225	-73.994710	40.750325	1	2009	4	Q3	H6	2454.36
2	12.9	-74.005043	40.740770	-73.962565	40.772647	1	2009	0	Q3	H6	5039.60
3	5.3	-73.976124	40.790844	-73.965316	40.803349	1	2009	4	Q2	H3	1661.44
4	16.0	-73.925023	40.744085	-73.973082	40.761247	1	2014	3	Q3	H5	4483.73

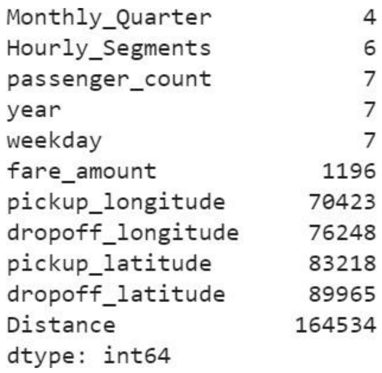
After performing further data cleaning, the number of samples reduced to 194334



	fare_amount	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude	passenger_count	year	weekday	Monthly_Quarter	Hourly_Segments	Distance
0	7.5	-73.999817	40.738354	-73.999512	40.723217	1	2015	3	Q2	H5	1681.11
1	7.7	-73.994355	40.728225	-73.994710	40.750325	1	2009	4	Q3	H6	2454.36
2	12.9	-74.005043	40.740770	-73.962565	40.772647	1	2009	0	Q3	H6	5039.60
3	5.3	-73.976124	40.790844	-73.965316	40.803349	1	2009	4	Q2	H3	1661.44
4	16.0	-73.925023	40.744085	-73.973082	40.761247	1	2014	3	Q3	H5	4483.73
...	...	...	...	...	...	...	...	...	...	...	...
194329	3.0	-73.987042	40.739367	-73.986525	40.740297	1	2012	6	Q4	H3	112.13
194330	7.5	-73.984722	40.736837	-74.006672	40.739620	1	2014	4	Q1	H1	1879.64
194331	30.9	-73.986017	40.756487	-73.858957	40.692588	2	2009	0	Q2	H1	12867.92
194332	14.5	-73.997124	40.725452	-73.983215	40.695416	1	2015	2	Q2	H4	3536.55
194333	14.1	-73.984395	40.720077	-73.985508	40.768793	1	2010	5	Q2	H2	5410.68

194334 rows × 11 columns

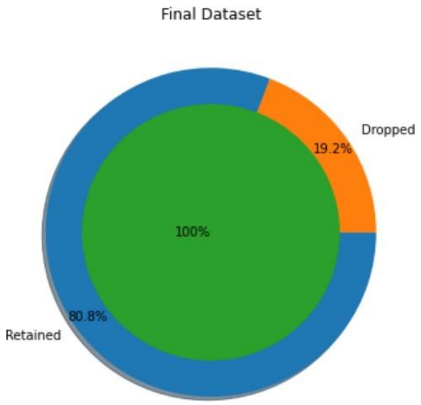
Following is the number of unique values in each features



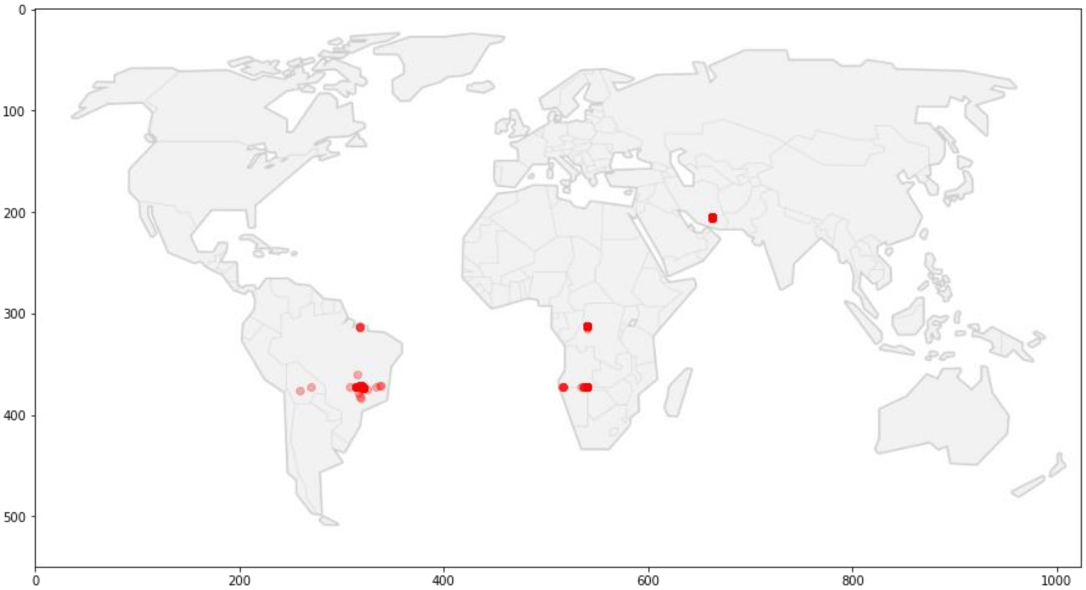
```
Monthly_Quarter           4
Hourly_Segments           6
passenger_count           7
year                      7
weekday                   7
fare_amount            1196
pickup_longitude      70423
dropoff_longitude     76248
pickup_latitude       83218
dropoff_latitude      89965
Distance             164534
dtype: int64
```

Our dataset contains 5 numerical and 5 categorical features.

There were 5653 duplicate entries found in the dataset which were then dropped. Before removal of outliers, the dataset had 194334 samples. After removal of outliers, the dataset now has 16129 samples. After the clean-up process, 38358 samples were dropped, which is 19.18% of the data.

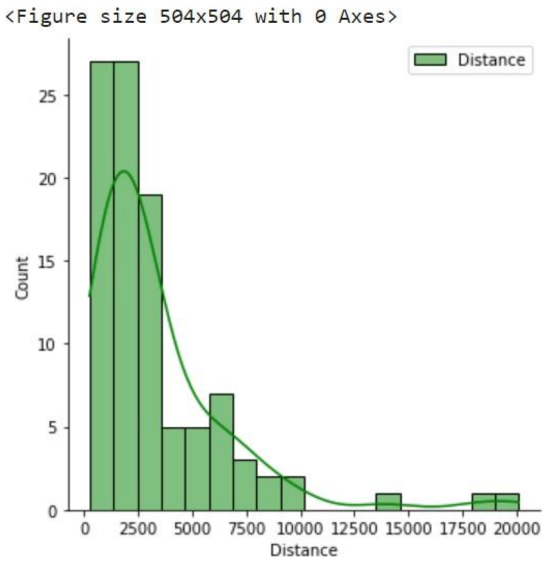


3.1 Plots



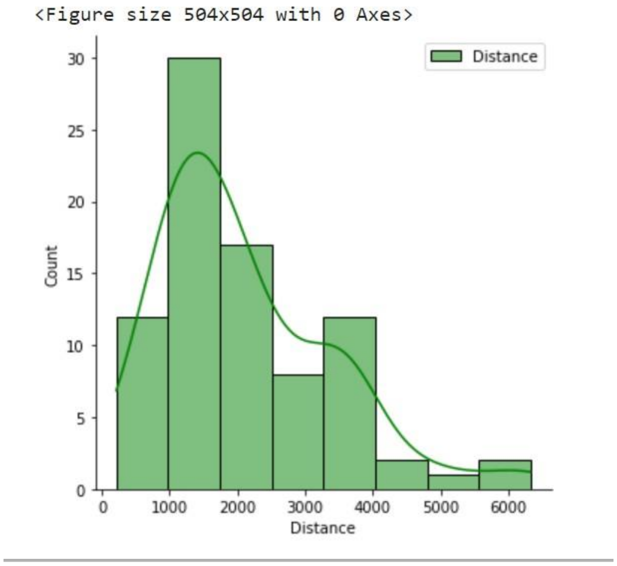
The data is distributed over South America, Africa, and The Middle East with most data in South America and the least data in The Middle East.

A histogram is an approximate representation of the distribution of numerical data. The following histogram has been plotted to analyze the distribution of the Distance.



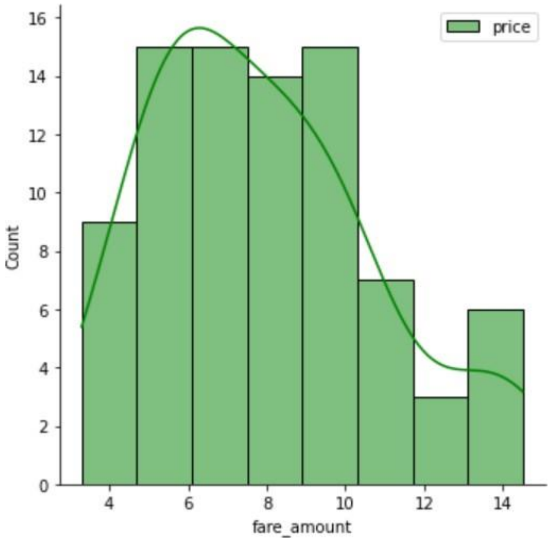
```
<Figure size 504x504 with 0 Axes>
```

This depicts that data is highly skewed to the left, So, after removing the outliers, the following histogram is obtained.



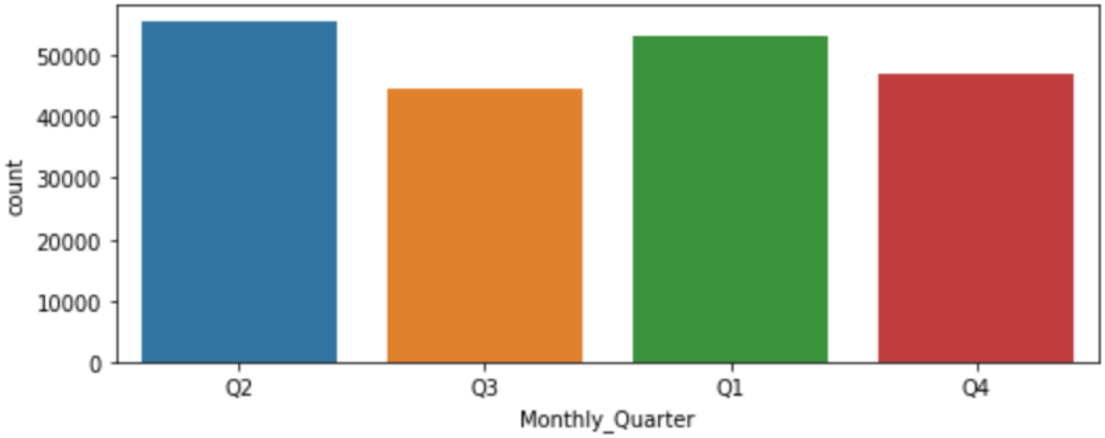
```
<Figure size 504x504 with 0 Axes>
```

Similarly, After removing the outliers, the following histogram has been plotted to analyze the distribution of the price.

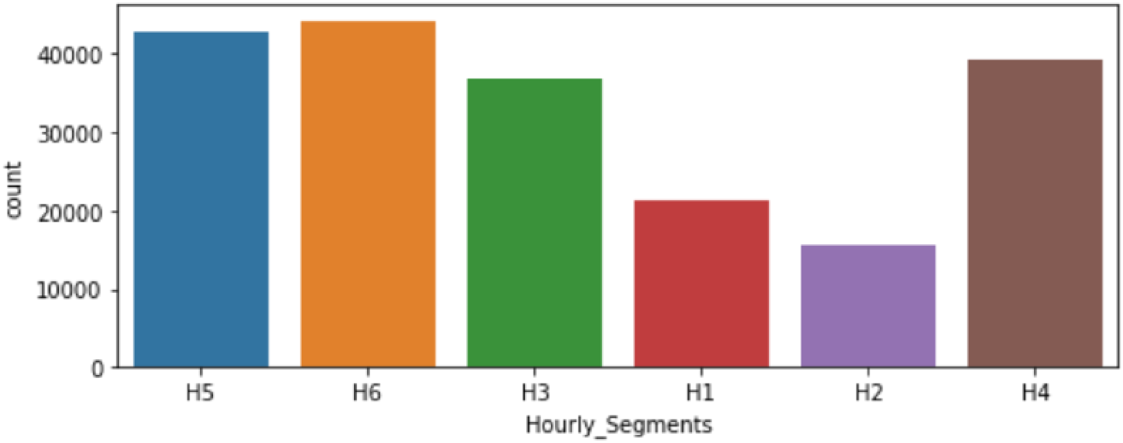


We then visualized various categorical features such as number of trips per each quarter, hourly segment, passenger count, year, and weekday.

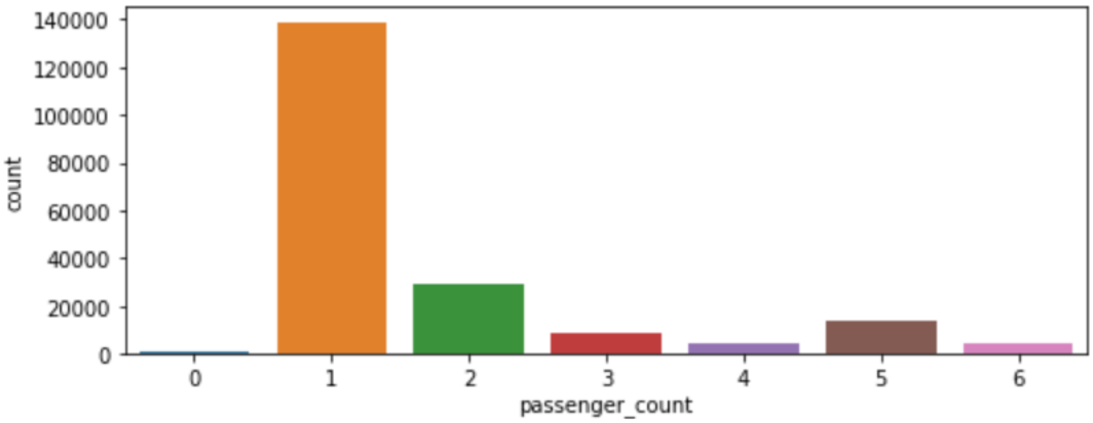
A countplot is used to represent the occurrence(counts) of the observation present in the categorical variable. It uses the concept of a bar chart for the visual depiction.
The following countplot represents the number of trips per each quarter of the year where every 4 months make up a quarter.



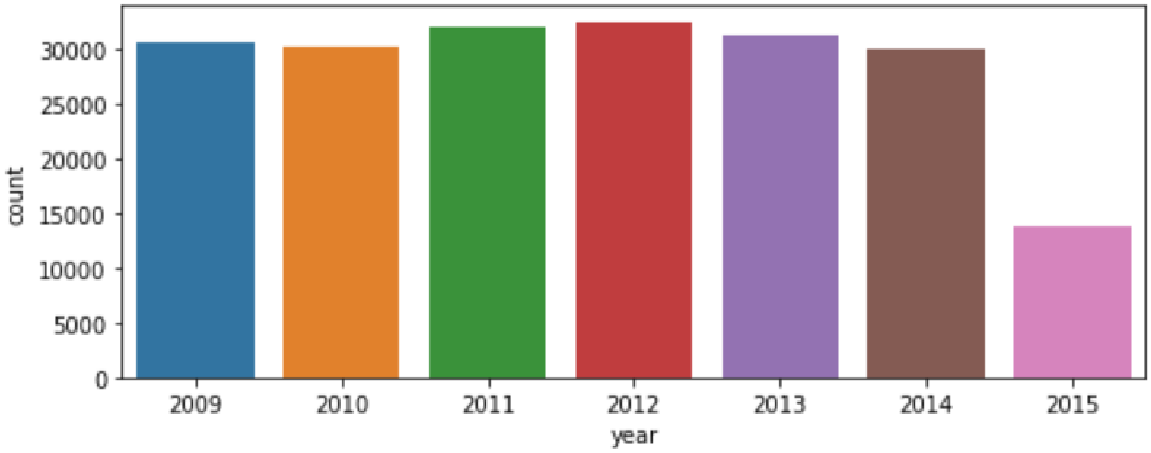
Considering the day is divided into 6 segments with 4 hours in each segment, the following countplot represents the number of trips in each hourly segment.



The next countplot depicts the count of trips based on the number of passengers. It can be inferred that most of the trips occurred had only one passenger.

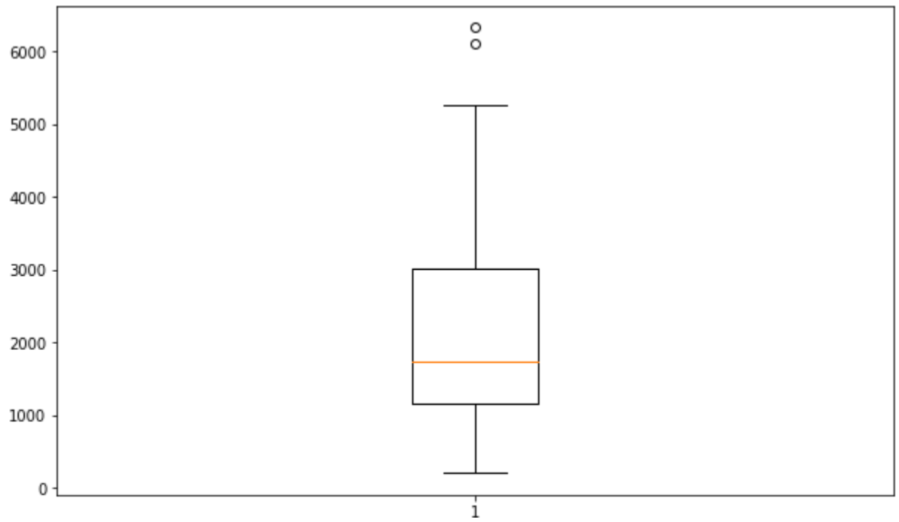


The final countplot depicts the number of trips per year. It can be observed that the number of trips fell by approximately 50% in the year 2015 when compared to the previous six years.

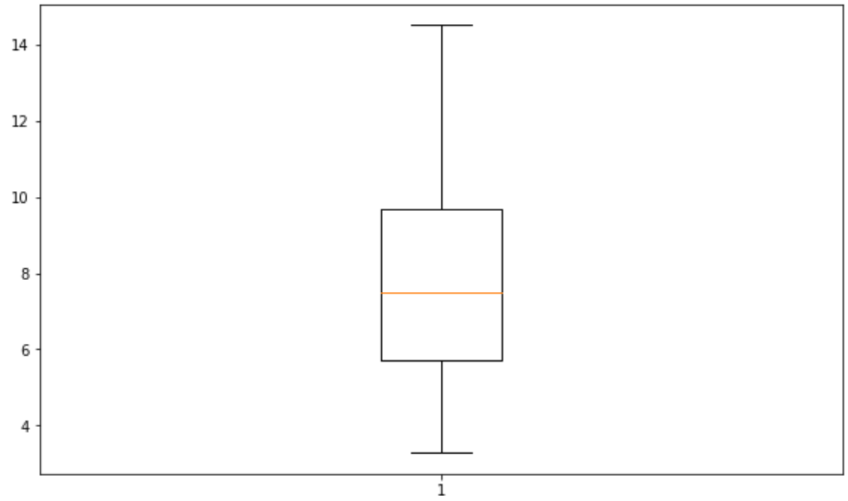


A box whisker plot, also known as a box plot, displays the five-number summary of a set of data. The five number summary is the minimum, first quartile, median, third quartile, and maximum.

The box plot below represents the five-number summary for ‘Distance’

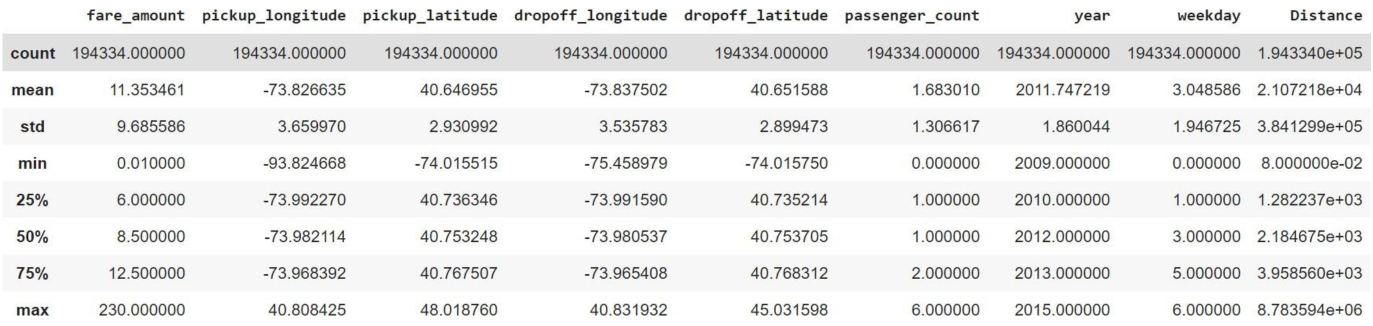


The next box plot represents the five-number summary for ‘fare amount’



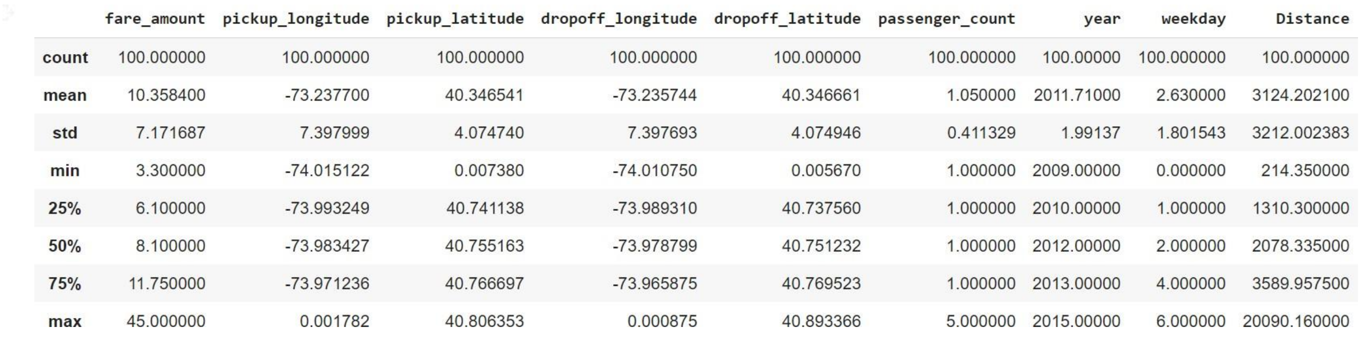
3.2 Basic statistical analysis

Basic statistics of Population is as follows:



	fare_amount	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude	passenger_count	year	weekday	Distance
count	194334.000000	194334.000000	194334.000000	194334.000000	194334.000000	194334.000000	194334.000000	194334.000000	1.943340e+05
mean	11.353461	-73.826635	40.646955	-73.837502	40.651588	1.683010	2011.747219	3.048586	2.107218e+04
std	9.685586	3.659970	2.930992	3.535783	2.899473	1.306617	1.860044	1.946725	3.841299e+05
min	0.010000	-93.824668	-74.015515	-75.458979	-74.015750	0.000000	2009.000000	0.000000	8.000000e-02
25%	6.000000	-73.992270	40.736346	-73.991590	40.735214	1.000000	2010.000000	1.000000	1.282237e+03
50%	8.500000	-73.982114	40.753248	-73.980537	40.753705	1.000000	2012.000000	3.000000	2.184675e+03
75%	12.500000	-73.968392	40.767507	-73.965408	40.768312	2.000000	2013.000000	5.000000	3.958560e+03
max	230.000000	40.808425	48.018760	40.831932	45.031598	6.000000	2015.000000	6.000000	8.783594e+06

Basic statistics of sample of 100 records is as follows:



	fare_amount	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude	passenger_count	year	weekday	Distance
count	100.000000	100.000000	100.000000	100.000000	100.000000	100.000000	100.00000	100.000000	100.000000
mean	10.358400	-73.237700	40.346541	-73.235744	40.346661	1.050000	2011.71000	2.630000	3124.202100
std	7.171687	7.397999	4.074740	7.397693	4.074946	0.411329	1.99137	1.801543	3212.002383
min	3.300000	-74.015122	0.007380	-74.010750	0.005670	1.000000	2009.00000	0.000000	214.350000
25%	6.100000	-73.993249	40.741138	-73.989310	40.737560	1.000000	2010.00000	1.000000	1310.300000
50%	8.100000	-73.983427	40.755163	-73.978799	40.751232	1.000000	2012.00000	2.000000	2078.335000
75%	11.750000	-73.971236	40.766697	-73.965875	40.769523	1.000000	2013.00000	4.000000	3589.957500
max	45.000000	0.001782	40.806353	0.000875	40.893366	5.000000	2015.00000	6.000000	20090.160000

.

CHAPTER 4

ESTIMATION

INTERVAL ESTIMATION:

A confidence interval, in statistics, refers to the probability that a population parameter will fall between a set of values for a certain proportion of times. Analysts often use confidence intervals than contain either 95% or 99% of expected observations.
A confidence interval is a range of values, bounded above and below the statistic's mean, that likely would contain an unknown population parameter. Confidence level refers to the percentage of probability, or certainty, that the confidence interval would contain the true population parameter when you draw a random sample many times.

Confidence interval obtained for price is
[7.834327344761008, 9.553672655238989]

Confidence interval obtained for distance is
[1963.0949451236718, 2538.649854876328]

CHAPTER 5

REGRESSION AND ANALYSIS

5.1 Linear Regression

Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.

For our analysis, ‘Distance’ is taken as the independent variable and ‘fare_amount’ is the variable dependent on ‘Distance’. Linear regression attempts to model the relationship between two variables by fitting a linear equation to the observed data. The code for linear regression analysis is as follows:

CODE:

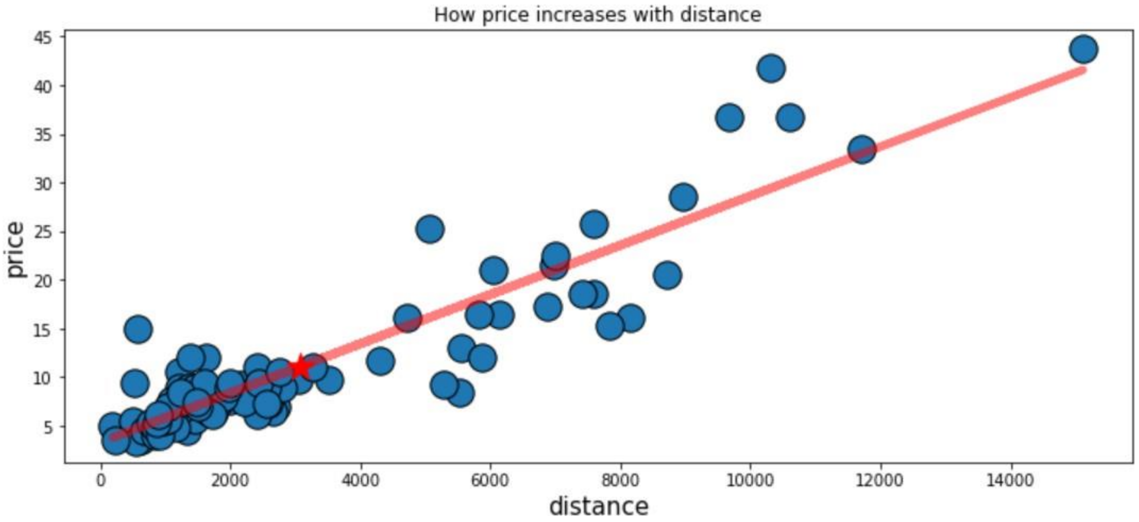
```
import numpy as nmp
import matplotlib.pyplot as mtplt
def estimate_coeff(p, q):
# Here, we will estimate the total number of points or observation
    n1 = nmp.size(p)
# Now, we will calculate the mean of a and b vector
    m_p = nmp.mean(p)
    m_q = nmp.mean(q)
# here, we will calculate the cross deviation and deviation about a
    SS_pq = nmp.sum(q * p) - n1 * m_q * m_p
    SS_pp = nmp.sum(p * p) - n1 * m_p * m_p
# here, we will calculate the regression coefficients
    b_1 = SS_pq / SS_pp
    b_0 = m_q - b_1 * m_p
```

```
     return (b_0, b_1)

def plot_regression_line(p, q, b):
# Now, we will plot the actual points or observation as scatter plot
    mtplt.scatter(p, q, color = "m",
            marker = "o", s = 30)
# here, we will calculate the predicted response vector
    q_pred = b[0] + b[1] * p
# here, we will plot the regression line
    mtplt.plot(p, q_pred, color = "g")
# here, we will put the labels
    mtplt.xlabel('p')
    mtplt.ylabel('q')
# here, we will define the function to show plot
    mtplt.show()

p=df_sample['Distance']
q=df_sample['fare_amount']
# now, we will estimate the coefficients
b = estimate_coeff(p, q)
print("Estimated coefficients are :\nb_0 = {} \  \nb_1 = {}".format(b[0], b[1]))
```

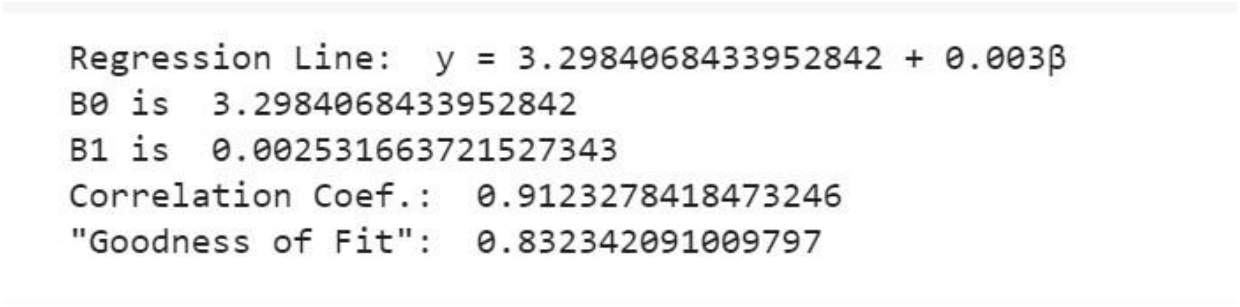
Below is a plot of the linear regression line. The goodness-of-fit test compares the observed values to the expected values. The goodness of fit obtained for the analysis is 0.8323.



.

5.2 Estimates of regression coefficients

The regression line obtained is Y = 3.2984068433952842 + 0.003β, and the correlation coefficient obtained is 0.9123
Estimated coefficients are :
b_0 = 3.2984068433952842
b_1 = 0.003



```
Regression Line:  y = 3.2984068433952842 + 0.003β
B0 is  3.2984068433952842
B1 is  0.002531663721527343
Correlation Coef.:  0.9123278418473246
"Goodness of Fit":  0.832342091009797
```

5.3 Analysis of Variance

Ordinary Least Squares regression (OLS) is a common technique for estimating coefficients of linear regression equations which describe the relationship between one or more independent quantitative variables and a dependent variable.

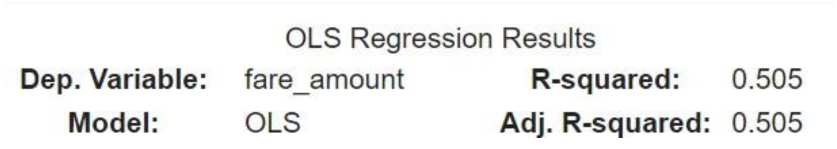
Mean Absolute Error is a model evaluation metric used with regression models. The mean absolute error of a model with respect to a test set is the mean of the absolute values of the individual prediction errors on all instances in the test set.

Mean Squared Error (MSE) is used to check how close estimates or forecasts are to actual values. Lower the MSE, the closer is forecast to actual. This is used as a model evaluation measure for regression models and the lower value indicates a better fit.

R-squared error represents the fraction of variance of the actual value of the response variable captured by the regression model rather than the MSE which captures the residual error.

Weighted Least Squares Regression (WLS) regression is an extension of the ordinary least squares (OLS) regression that weights each observation unequally.

In our analysis, R squared error is 0.505



OLS Regression Results

Dep. Variable:	fare_amount	**R-squared:**	0.505
Model:	OLS	**Adj. R-squared:**	0.505

Weighted least square is 0.5

Mean absolute - 1.5724613930646771
Mean squared error - 7.6193696009534815
R2 score- 0.541622637022696

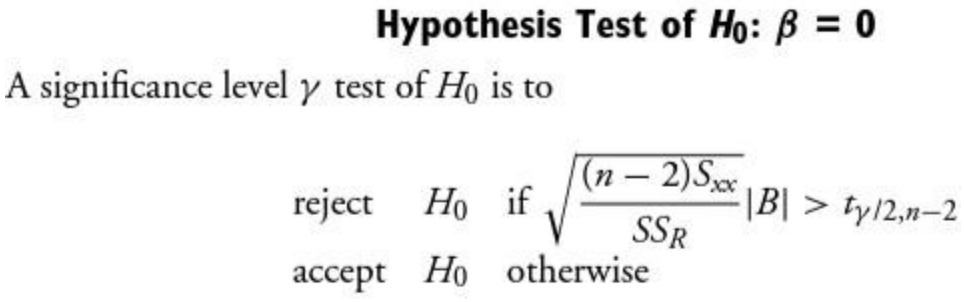
CHAPTER 6

INFERENCE STATISTICS

Hypothesis testing is a form of statistical inference that uses data from a sample to draw conclusions about a population parameter or a population probability distribution. First, a tentative assumption is made about the parameter or distribution. This assumption is called the null hypothesis and is denoted by H0. An alternative hypothesis (denoted Ha), which is the opposite of what is stated in the null hypothesis, is then defined. The hypothesis-testing procedure involves using sample data to determine whether or not H0 can be rejected. If H0 is rejected, the statistical conclusion is that the alternative hypothesis Ha is true.

The null hypothesis Ho is that Distance and fare_amount are not linearly related.
The alternate hypothesis Ha is that Distance and fare_amount are linearly related.

To test this, we consider the confidence interval of β.



Hypothesis Test of H_0: $\beta = 0$

A significance level γ test of H_0 is to

$$\begin{array}{lll} \text{reject} & H_0 & \text{if } \sqrt{\dfrac{(n-2)S_{xx}}{SS_R}}\,|B| > t_{\gamma/2,n-2} \\ \text{accept} & H_0 & \text{otherwise} \end{array}$$

The code for the significance level test of Ho (with Alpha value as 0.05 and n-2 as 98) is as follows

```
import numpy as nmp
import matplotlib.pyplot as mtplt
```

```
p = hk_sample_dist['Distance']
q = hk_sample_dist['fare_amount']
n1 = nmp.size(p)
m_p = nmp.mean(p)
m_q = nmp.mean(q)

SS_pp = nmp.sum(p * p) - n1 * m_p * m_p
SS_qq = nmp.sum(q * q) - n1 * m_q * m_q
SS_qq
SS_pq = nmp.sum(q * p) - n1 * m_q * m_p
SSR_N=(SS_pp*SS_qq)-(SS_pq*SS_pq)
SSR=SSR_N/SS_pp
a=(98*SS_pp)/SSR
b=math.sqrt(a)
K=b*0.002
K

c=b*1.984
[0.002-c, 0.002+c]
```

We get the LHS as **9.036524485986964** and the RHS as **1.984**. As mentioned above, since LHS > RHS, we reject the null hypothesis that β=0. This implies that the variables ‘Distance’ and ‘fare_price’ are linearly related, thus rejecting the original null hypothesis and accepting the alternative hypothesis.

The confidence interval for β is obtained to be
[0.0015199611955963022, 0.0024800388044036977]

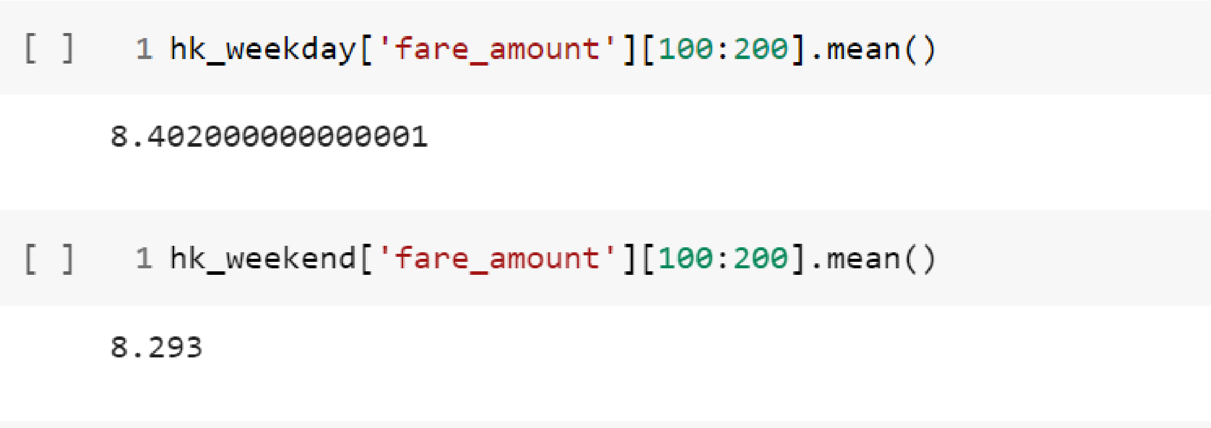
$$\left(B - \sqrt{\frac{SS_R}{(n-2)S_{xx}}}\,t_{\alpha/2,n-2},\ B + \sqrt{\frac{SS_R}{(n-2)S_{xx}}}\,t_{\alpha/2,n-2}\right)$$

Another hypothesis is that the fare amount on weekdays is greater than the fare amount on weekends which needs to be tested.

Null hypothesis (Ho) :
The average fare amount on weekdays = the average fare amount on
weekends.

Alternate Hypothesis (Ha) :
The average fare amount on weekdays > the average fare amount on weekends.

After testing, we get the average fare amount on weekdays as 8.402 and on weekends as 8.293. Since 8.402 > 8.293, we reject the null hypothesis and accept the a;ternate hypothesis.



```
hk_weekday['fare_amount'][100:200].mean()
```

```
8.402000000000001
```

```
hk_weekend['fare_amount'][100:200].mean()
```

```
8.293
```

CHAPTER 7

CONCLUSION

In this project, we have successfully performed different statistical analysis on the dataset. We started off by trying to better understand the data and then performed a clean-up by removing all the duplicate entries. Next, we proceeded to perform descriptive statistics with the help of histograms, boxplots, countplots, etc. We then successfully estimated the confidence intervals for the parameters ‘Distance’ and ‘fare_amount’. Regression analysis of the data was performed to obtain the correlation coefficients, which were finally used to test the hypothesis.