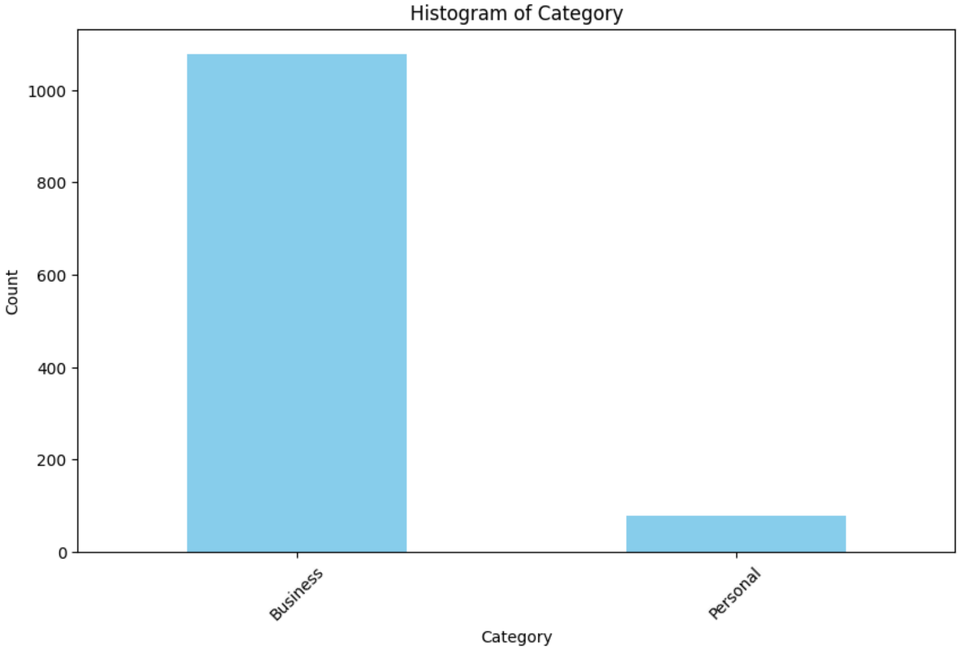
EDA Of Uber Drives

Name: Ali Sufyan
ID: 23037749

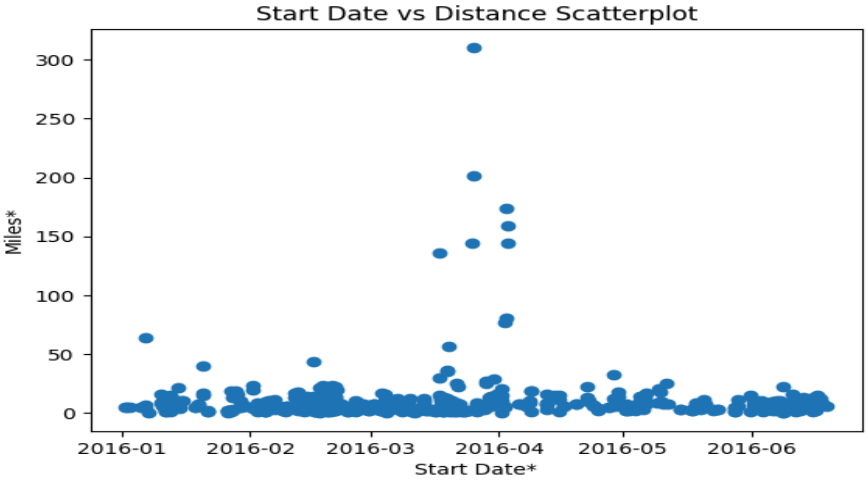
Abstract:
This short report was done as per instructions provided in Assignment 2. The report is an exploratory data analysis on a kaggle dataset. In this report I have explored the dataset and performed visualizations. The statistics of the variable miles has also been explored briefly. And a scatter plot was made to see the relationship between two variables: the start and distance(Miles).

Github: https://github.com/Alisufyangondal/Assignment_2
Dataset: https://www.kaggle.com/datasets/zusmani/uberdrives/data

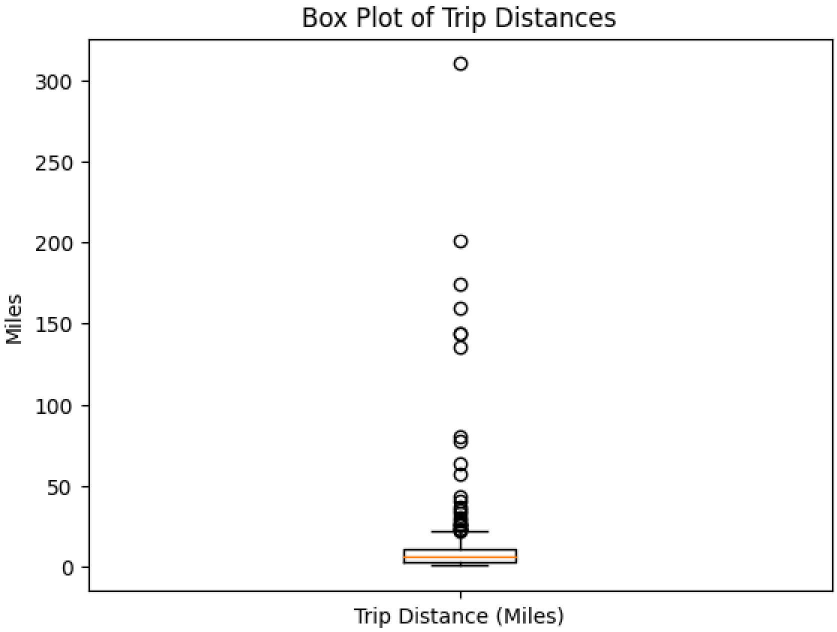
The dataset was downloaded from Kaggle. The data set consists of 1155 rows. Details are as follows. The dataset contains columns consisting of start date, end date, category, start, stop miles and purpose. After printing the head we came to know this. At the end of the dataset there was a total which we removed. So for EDA the first thing we did was make a bar chart of categories of the trips made.



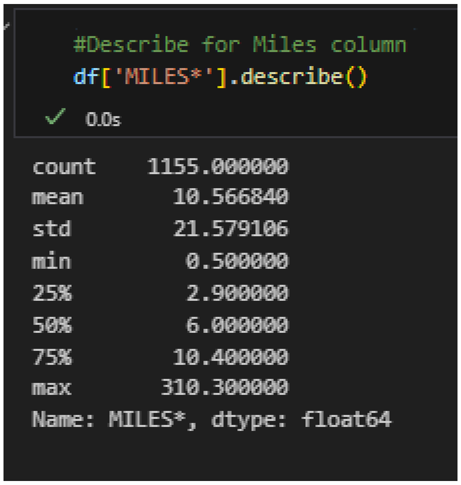
As we can clearly see from the bar chart that the business trips were made way more than the personal trips. So here the bar chart has made us realize very easily that business trips were made too much during the time of data recorded in our dataset.



The scatter plot basically tells us about the relationship between the two quantitative variables. Here in our dataset I have checked for the relationship between start and miles from our dataset. There appears to be a weak positive correlation, meaning that as the start date gets more recent, the distance traveled generally tends to increase slightly. However, the points are fairly scattered, suggesting there's considerable variation in distances traveled regardless of the start date. There might be a few outliers with either unusually short or long distances traveled for their start dates.



This box plot indicates that the trips in our dataset tend to be on the shorter side, with most trips falling within a range of 1.5 to 25 miles. There are a few outliers too. Box plot helps us realize the spread of the data easily. The data in our miles column is not distributed properly as it can be seen in the box plot easily.



```
#Describe for Miles column
df['MILES*'].describe()
```

```
count    1155.000000
mean       10.566840
std        21.579106
min         0.500000
25%         2.900000
50%         6.000000
75%        10.400000
max       310.300000
Name: MILES*, dtype: float64
```

The descriptive statistics reveal that there are 1155 recorded trips. Mean of the distance traveled is approximately 10.57 miles. Standard deviation of about 21.58 which tells us considerable variability in the distances covered. The minimum distance traveled is recorded at 0.5 miles, while the maximum distance reaches 310.3 miles that tell us that wide range of trip lengths. The 25th percentile indicates that 25% of the trips are shorter than 2.9 miles. The median (50th percentile) suggests that half of the trips are shorter than 6 miles. And the 75th percentile shows that 75% of the trips are shorter than 10.4 miles. These statistics collectively offer insights into the distribution and variability of trip distances in the dataset.