UBER DATA EDA

Business context/scenario.

1. Why this data was collected?
   1. Understand Revenue losses because of cancelled trips.
   2. Understand Complains and dissatisfaction of customers.
   3. No. of trips day, month year.
   4. To understands traffic or rush hour pattern.
2. Reading if you want to understand more about how does this business work.
3. Use case:
   1. To find out in which are or region of place we would need more driver.
   2. What time of the day the prices should dynamically change.
   3. Deciding on the minimum distance for search of requested cab.

Business problem:

There has been a rise in cancelled trips and customers are not happy with and file complain regarding the same.

Also, there has been shift in customer preference towards other business competitor Ola, Rapido etc.

Data:

Data Dictionary:

Source:

This is sample of UBER data which consist of 6745 records and 6 columns.

The data was collected to understand trips patterns between city and airport during different time of the day.

This data is collected between 2016-07-13 to 2016-12-07.

1. Request id – Integer or Numeric

* This is unique no. for all cab request that was made.

1. Pickup point – String or categorical

* This is the place from where pickup request was made. There are only values city, airport for this column.

1. Driver id – Integer or Numeric

* This is the unique identifier of the driver.

1. Status – String or categorical

* This represent status of the trip and has 3 values namely Trip completed, Not Available or Cancelled.

1. Request timestamp – Timestamp

* Time at which pickup request was made.

1. Drop timestamp – Timestamp

* Time at which the trip was completed.

2 Numerical columns, 2 Categorical columns and two time stamp or time related columns.

Decide on the tools for your EDA.

Formulate Question that you would like answer?

1. What is the % value or distribution or counts on pickup points?
2. What is the % value or distribution or counts of all categories in Status of trips?
3. What is the time span for different rush hours during day?
4. Do a count on driver\_id , most efficient one?
5. Driver\_id and status with cancelled count?
6. Time frame where trips are mostly getting cancelled.
7. Avg. time, max, min time between pick and drop?
8. Time frame when cars are most available?
9. Request pattern during weekdays and weekends?
10. Pickup point and status combination?

Default:

Analysis points:

Univariate Analysis

* Solving analytical question and representing them with appropriate charts and plots.
* Pick up Numerical columns and understand their distribution. Summary - mean, median ,min, max, 25 ,75 percentile.
* Graphs like histogram or distribution plots , box plots
* Categorical values , find unique values in category understand their proportion.
* Count plot , pie-chart etc.

Bi-variate Analysis

* Solving analytical question and representing them with appropriate charts and plots.
* Combine two univariate values
* Two Numerical values - scatter plot, line chart , box plot ,bubble chart , dot plots, violin plot etc.
* Two categorical values – stacked bar chart, tree map etc.
* One categorical and one Numerical – bar plot, pie-chart etc.
* Multiple numerical values – Heat map , (correlation analysis)
* Date Value and Numerical – line chart, Area chart etc.