TLC Project –Executive Summary

PACE Stage : Plan/Analyse

OVERVIEW

Develop regression model that helps estimate taxi fares before the ride based on the data TLC has gathered.

PROJECT STATUS

Initial Data exploration

- ✓ Import data
- ✓ Inspect data
- ✓ Data value nonnull counts
- ✓ Identify Relevant and irrelevant Columns

NEXT STEPS

Data exploration and cleaning

- ❑ Data Cleaning
- ❑ Data Structuring
- ❑ Exploratory Data Analysis

KEY INSIGHTS

- There are no null values in the data frame
- There are some variables that are still in object form e.g. tpep_pickup_datetime , tpep_dropoff_datetime which have to be converted to Datetime format.
- There is one column "Unnamed“
- Trip distance cannot be 0, its abnormal value.
- Total amount cannot be negative

- Key Variables: Fare_amount
  trip_distance

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 22699 entries, 0 to 22698
Data columns (total 18 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   Unnamed: 0             22699 non-null  int64
 1   VendorID               22699 non-null  int64
 2   tpep_pickup_datetime   22699 non-null  object
 3   tpep_dropoff_datetime  22699 non-null  object
 4   passenger_count        22699 non-null  int64
 5   trip_distance          22699 non-null  float64
 6   RatecodeID             22699 non-null  int64
 7   store_and_fwd_flag     22699 non-null  object
 8   PULocationID           22699 non-null  int64
 9   DOLocationID           22699 non-null  int64
 10  payment_type           22699 non-null  int64
 11  fare_amount            22699 non-null  float64
 12  extra                  22699 non-null  float64
 13  mta_tax                22699 non-null  float64
 14  tip_amount             22699 non-null  float64
 15  tolls_amount           22699 non-null  float64
 16  improvement_surcharge  22699 non-null  float64
 17  total_amount           22699 non-null  float64
dtypes: float64(8), int64(7), object(3)
```

Data info: 22699 observations(rows)
18 (columns)