Project Name: Taxi Service Price Prediction

1. **Merge two datasets:**

   It was done by dividing the column (time_stamp) by 1000 in the taxi dataset to convert the time into minutes and making the last four numbers in the column (time_stamp) zero in the weather dataset, then removing the duplicated values using the two columns (location , time_stamp) in the weather dataset, then change the name of a column(location) in weather dataset to (destination),then merge the two datasets using merge function using columns (destination ,time_stamp)

2. **Preprocessing:**
   - Remove all duplicated values from final dataset after the merge
   - Using encoding(labelEncoder) on columns(name ,cab_type ,destination ,source ,product_id)
   - Fill in all null values in each column using the mean of the column

3. **Feature Selection:**

   Using **p_value(**Backward Elimination),using statsmodels.api library
   - Run model with all features and choice the feature has highest p_value >.05
   - Remove wind has P_value=.0962 then run again
   - Remove pressure has p_value=.858 and run again
   - Remove rain has p_value=.774 and run again
   - Remove de has p_value=.657 and run again
   - Remove time_stamp has p_value=.435 and run again
   - Remove temp has p_value=.078 and run again
   - Finally when run this steps the error from models not change that means they not affect in prediction
   - When remove any one from remaining features the error increase (very high)
   - Finally, remove ['id', 'wind', 'time_stamp', 'rain', 'pressure', 'temp'] and use ['distance', 'cab_type', 'destination', 'source', 'surge_multiplier', 'product_id', 'name', 'clouds', 'humidity'] as features

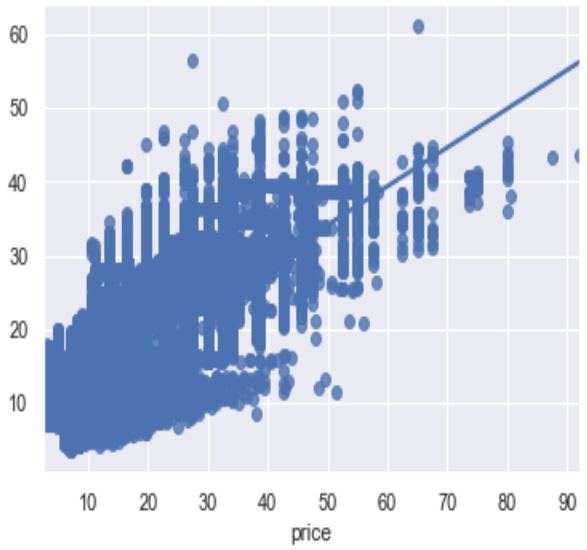
4. **Training and Testing:** split the data 80% for training and 20% for testing

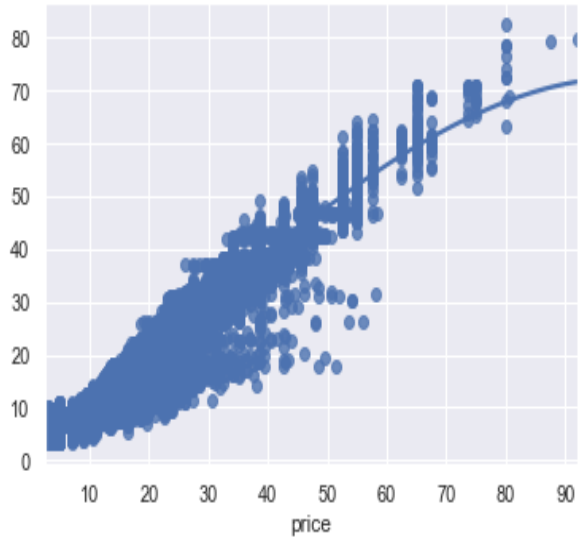
5. **Models:**
   - **Multilinear regression :** mse_error_traning:37.9911460
     Mse_error_testing:38.055132774
     Training_time: 0.0827794075012207
   - **Polynomial regression :** mse_train_poly: 3.030822000620637
     mse_test_poly: 3.008244316059448
     Training_time: 18.54770064353943
   - Finally, using the polynomial model because, has most accuracy.

6. **Further techniques that were used to improve the results :**

- P_value.

7. **screenshots of the resultants :**





8. **conclusion**

when we saw the data for the first time, we thought that the features like ['wind','time_stamp','rain','pressure','temp'] will affect in our model. But when we tried to calculate the correlation and P_Value , They disproved our opinions. And we know later that feature like ['distance', 'cab_type', 'destination', 'source', 'surge_multiplier', 'product_id', 'name', 'clouds', 'humidity'] would affect the model.