SDAIA
الهيئة السعودية للبيانات
والذكاء الاصطناعي
Saudi Data & AI Authority

TLC Trip Record Data

Yellow Taxi

Outlines

1. Introduction
2. Purpose of the project
3. Chosen year and month
4. EDA
5. Data model
6. Result

Introduction

The New York City Taxi and Limousine Commission (TLC), Over 200,000 TLC licensees complete approximately 1,000,000 trips each day. According to TLC the data it is recorded since 2009 - 2021

preprocess

Purpose of the project

1. Predict the fare amount of the ride.
2. Who effects on the fare amount
3. Visiualize the features

Chosen year and month

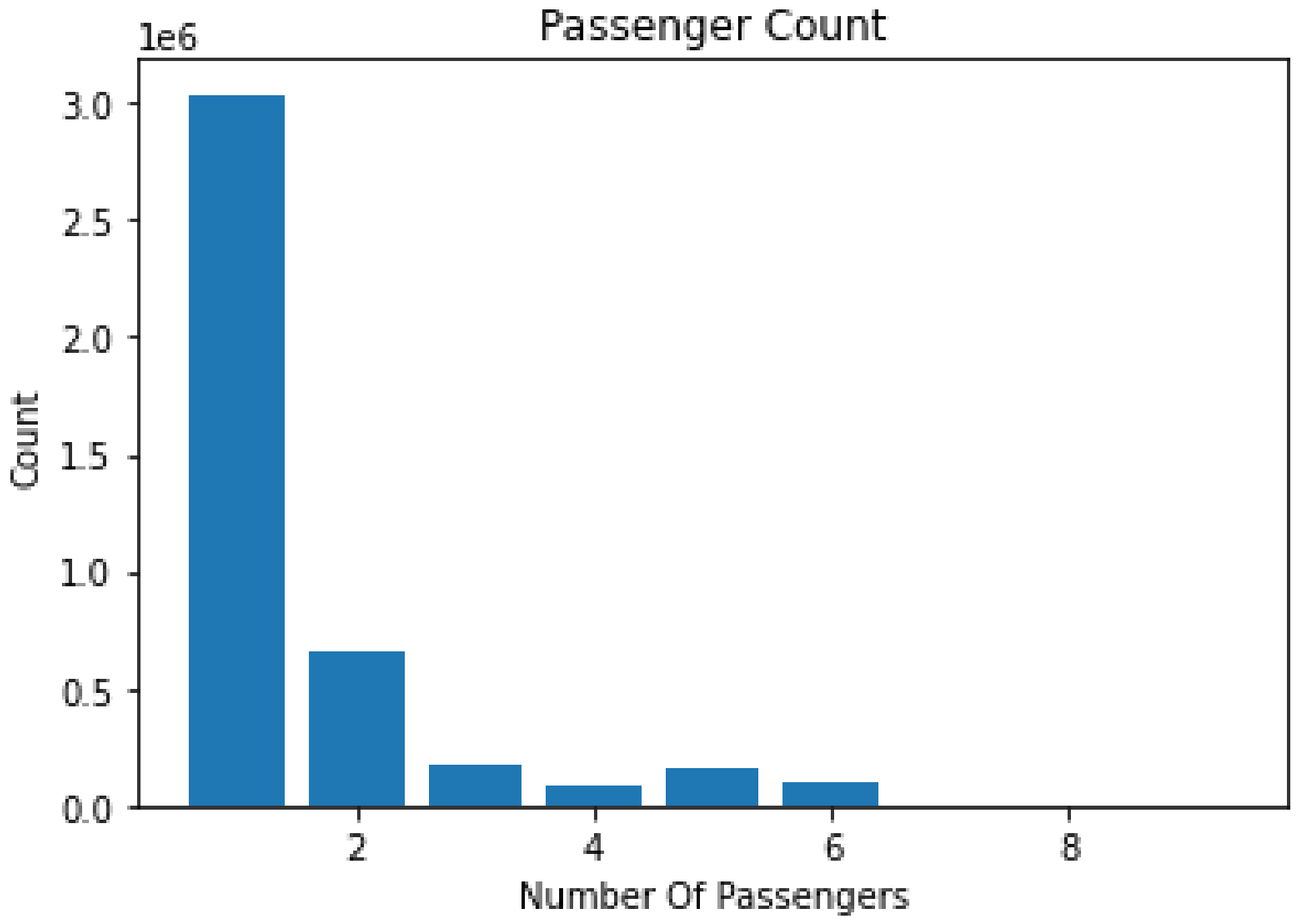
In this project the prediction and visualization will be on the dataset of October 2019

Who effects on the target?

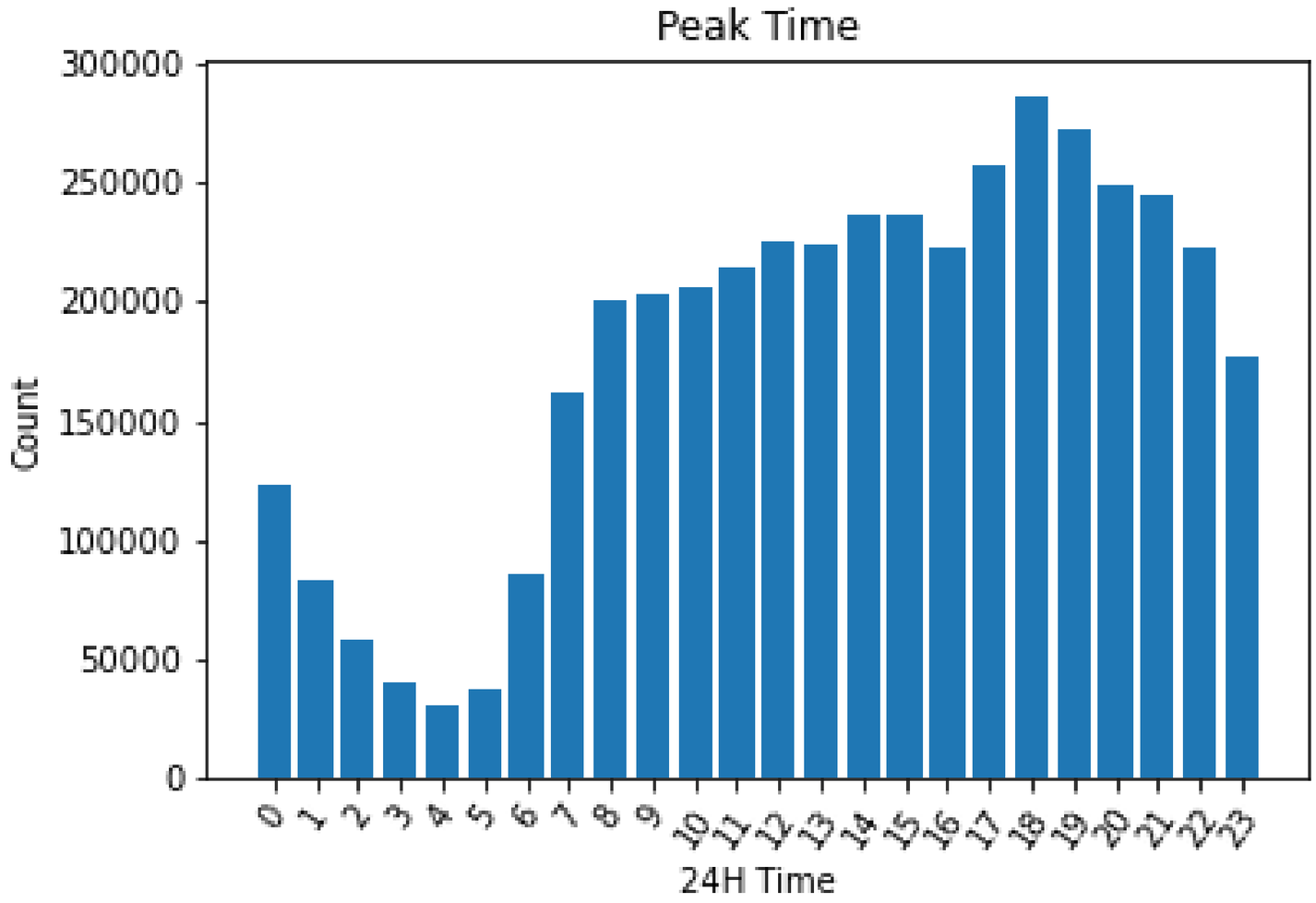
1. Distance
2. Taxi car size
3. Peak hours
4. Peak days
5. Rate code id
6. Duration

Correlation between distance and fare amount

Passengers count



What are the peak hours?



Rate code Id

Rate code in effect at the trip.

1= Standard rate

2=JFK

3=Newark

4=Nassau or Westchester

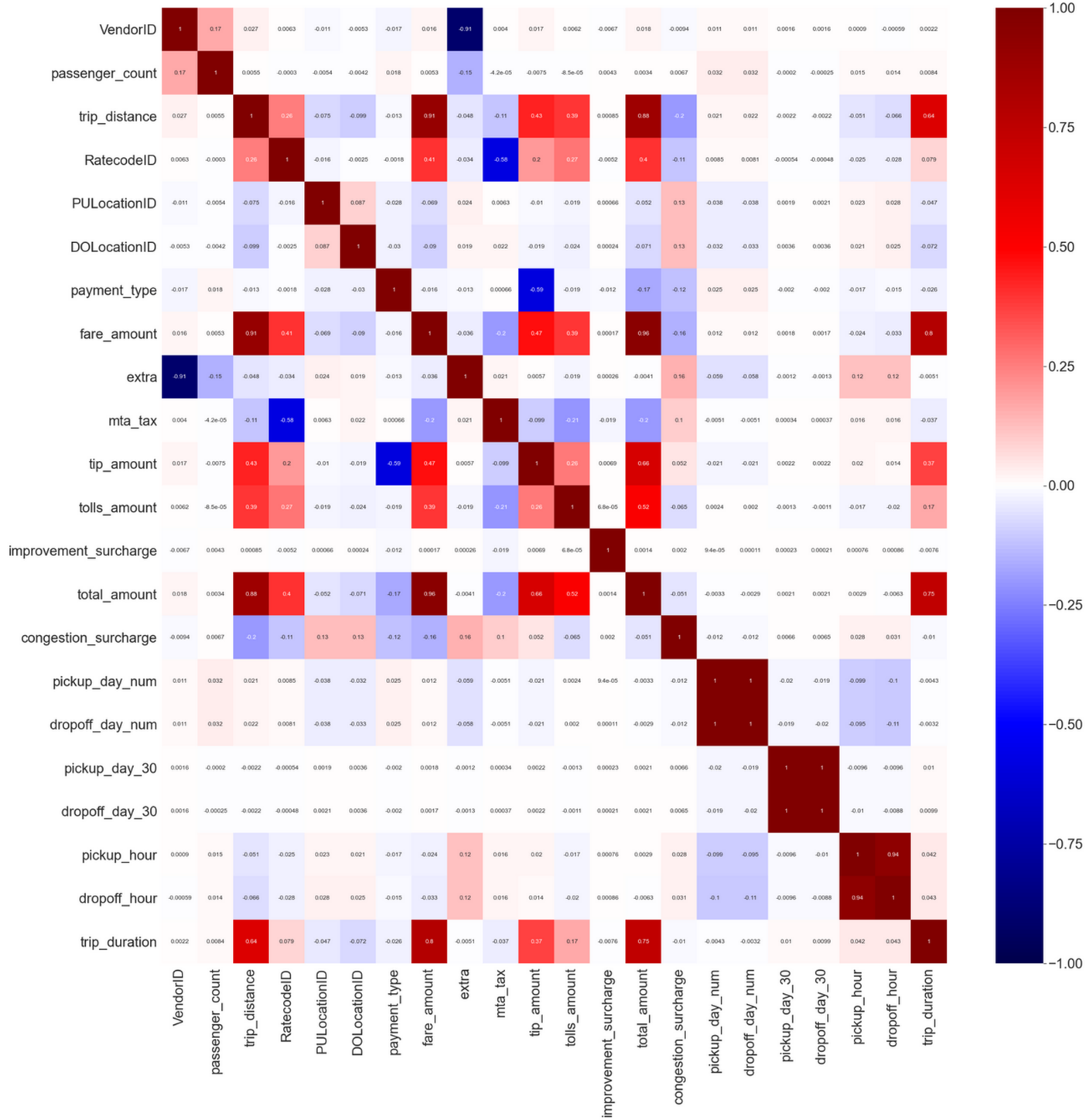
5=Negotiated fare

6=Group ride

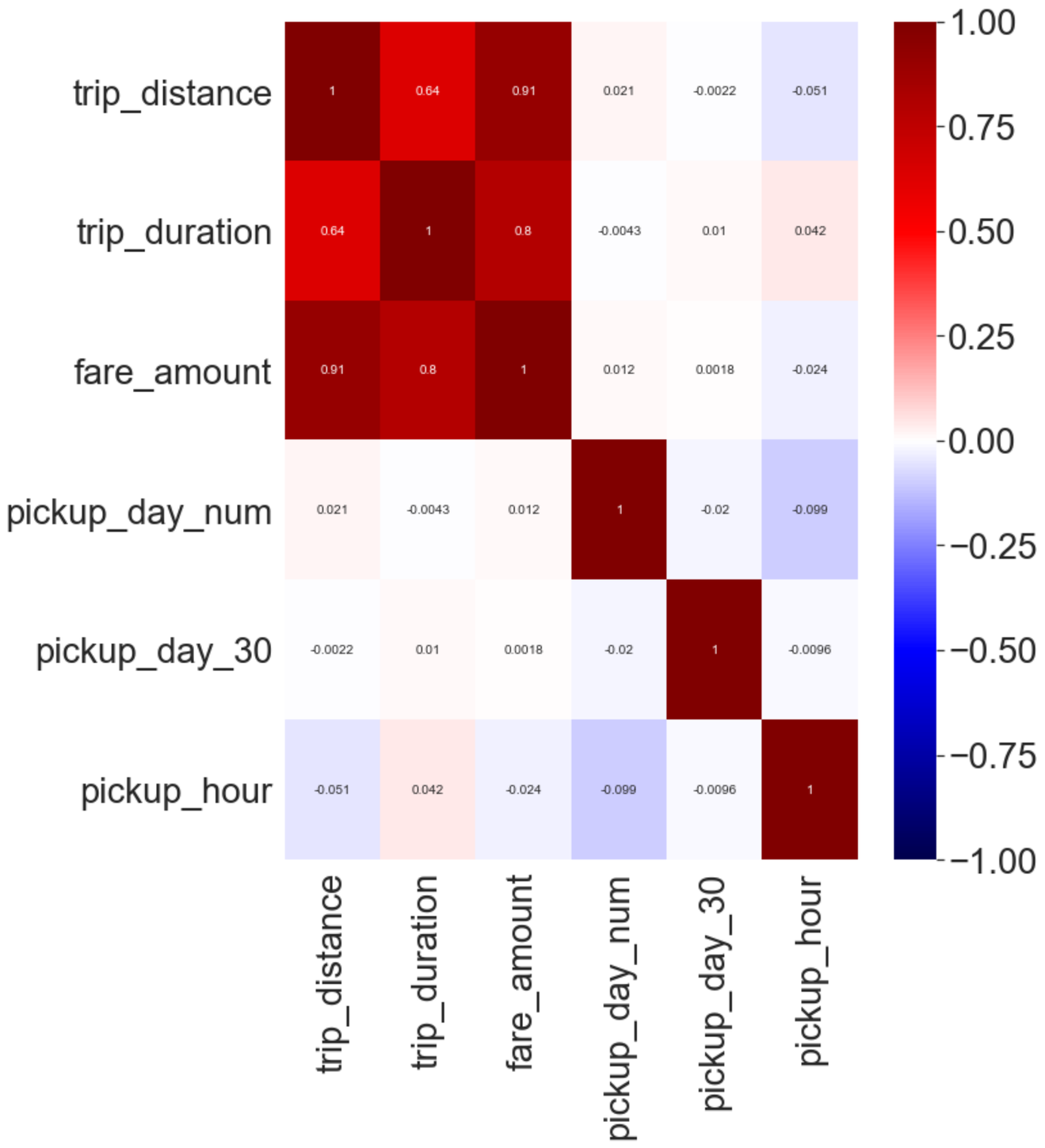
ref: https://www1.nyc.gov/site/tlc/passengers/taxi-fare.page

Data model

All the features in the dataset



Chosen features



Model used

1. Ordinary least square regression
2. Linear regression
3. Polynomial Regression
4. Evaluation regression (RMSE - MAE)

Linear Regression

Ordinary least square regression

OLS Regression Results

Dep. Variable:	fare_amount	**R-squared (uncentered):**	0.970
Model:	OLS	**Adj. R-squared (uncentered):**	0.970
Method:	Least Squares	**F-statistic:**	8.957e+07
Date:	Sun, 05 Dec 2021	**Prob (F-statistic):**	0.00
Time:	09:28:01	**Log-Likelihood:**	-1.2194e+07
No. Observations:	5548803	**AIC:**	2.439e+07
Df Residuals:	5548801	**BIC:**	2.439e+07
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
trip_distance	2.0053	0.001	3596.147	0.000	2.004	2.006
trip_duration	0.4802	0.000	3977.076	0.000	0.480	0.480

Omnibus:	10228504.495	**Durbin-Watson:**	1.913
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	19766216669.659
Skew:	13.829	**Prob(JB):**	0.00
Kurtosis:	294.082	**Cond. No.**	8.79

Polynomial regression

1. Categorical feature to dummy variables
2. Polynomial transformation
3. Interaction term
4. Standard scaling features

Evaluation regression (RMSE - MAE)

RMSE

- 1.4086979799510435

MAE

- 0.41391811577253645

Conclusion

The Followed benchmarks in choosing the best model

Baseline feature set:
~.90789 R^2

Add Category features (RatecodeID,VendorID):
~.95340 R^2

Add polynomial features:
~.95349 R^2

Add Several interaction terms:
~.95352 R^2