WHAT'S YOUR CAR WORTH?

Andy Wang

INTRODUCTION

* **Problem:**
  - **Buying and selling a used car can be tricky**
* **Goal:**
  - **A web application provides a predicted used car value based on its features**

DATA

- **Data from Cars.com**
- **75,253 listings of car information**

Basics

Exterior color	White
Interior color	Black
Drivetrain	Front-wheel Drive
MPG	54–50
Fuel type	Hybrid
Transmission	Automatic CVT
Engine	1.8L I4 16V MPFI DOHC Hybrid
VIN	JTDKARFU7J3055940
Stock #	055940
Mileage	28,246 mi.
Vehicle history	CARFAX Report ↗

Features

Convenience	Adaptive Cruise Control Heated Seats Keyless Start Navigation System
Entertainment	Bluetooth USB Port
Safety	Automatic Emergency Braking Backup Camera Brake Assist Lane Departure Warning Rear Cross Traffic Alert Stability Control

Used

Save

2018 Toyota Prius Two

28,246 mi.

$18,999

Great Deal | $2,863 under

Virtual Appointments

THE PIPELINE

- City zip codes
- Number of pages

List of Hyperparameters

Data Acquisition	Data Cleaning	Data Storage	Model Training	Web Application
Beautiful Soup	Pandas compy.py	SQL SQLAlchemy	scikit-learn tunepy.py	Streamlit

COMPY.PY

Dataframe Information

Data From Web Scraping

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 78868 entries, 0 to 78867
Data columns (total 10 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   Unnamed: 0             78867 non-null  object
 1   price                  78369 non-null  float64
 2   mpg                    71419 non-null  float64
 3   mileage                78862 non-null  float64
 4   drivetrain             78867 non-null  object
 5   fuel_type              78867 non-null  object
 6   transmission           78867 non-null  object
 7   engine                 78867 non-null  object
 8   num_of_entertainment   75149 non-null  float64
 9   num_of_safety          69820 non-null  float64
dtypes: float64(5), object(5)
memory usage: 6.0+ MB
```

Cleaned Data in SQL

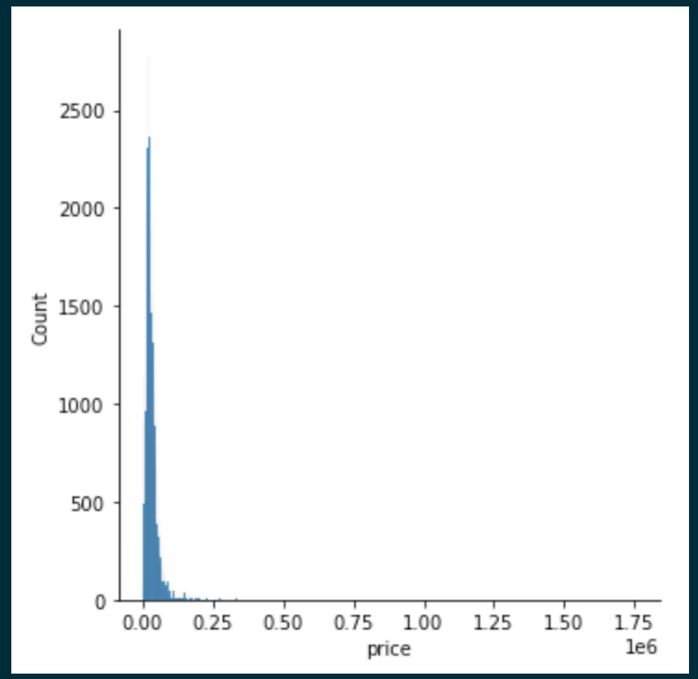
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 75253 entries, 0 to 75252
Data columns (total 13 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   name                   75253 non-null  object
 1   price                  75253 non-null  int64
 2   mpg                    75253 non-null  int64
 3   mileage                75253 non-null  int64
 4   num_of_entertainment   75253 non-null  int64
 5   num_of_safety          75253 non-null  int64
 6   year                   75253 non-null  int64
 7   drivetrain             75253 non-null  object
 8   fuel_type              75253 non-null  object
 9   transmission           75253 non-null  object
 10  engine                 75253 non-null  object
 11  make                   75253 non-null  object
 12  model                  75253 non-null  object
dtypes: int64(6), object(7)
memory usage: 7.5+ MB
```

If significant data are lost during the cleaning stage, we can take action immediately

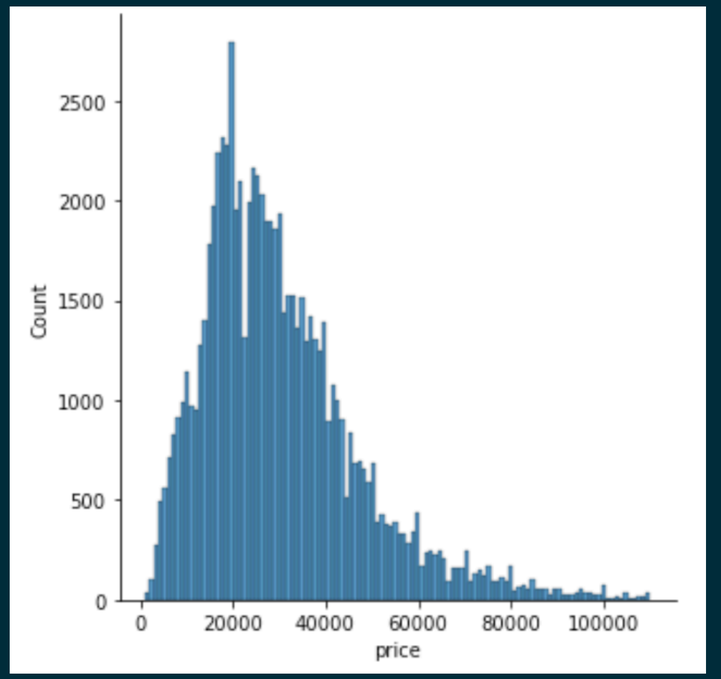
COMPY.PY

Distribution Plots

Price Distribution Before Outlier Removal



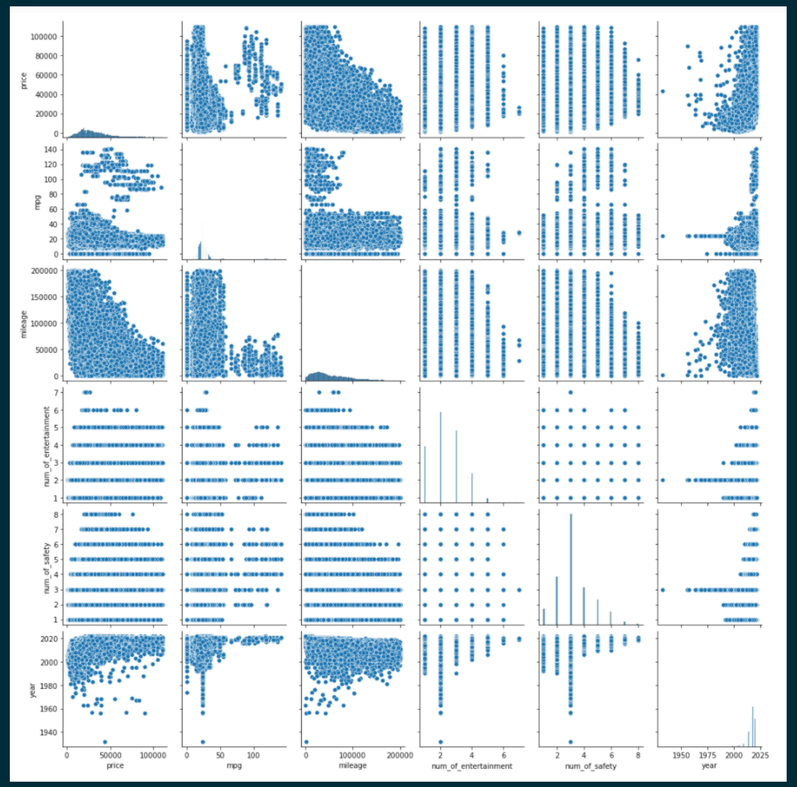
Price Distribution After Outlier Removal



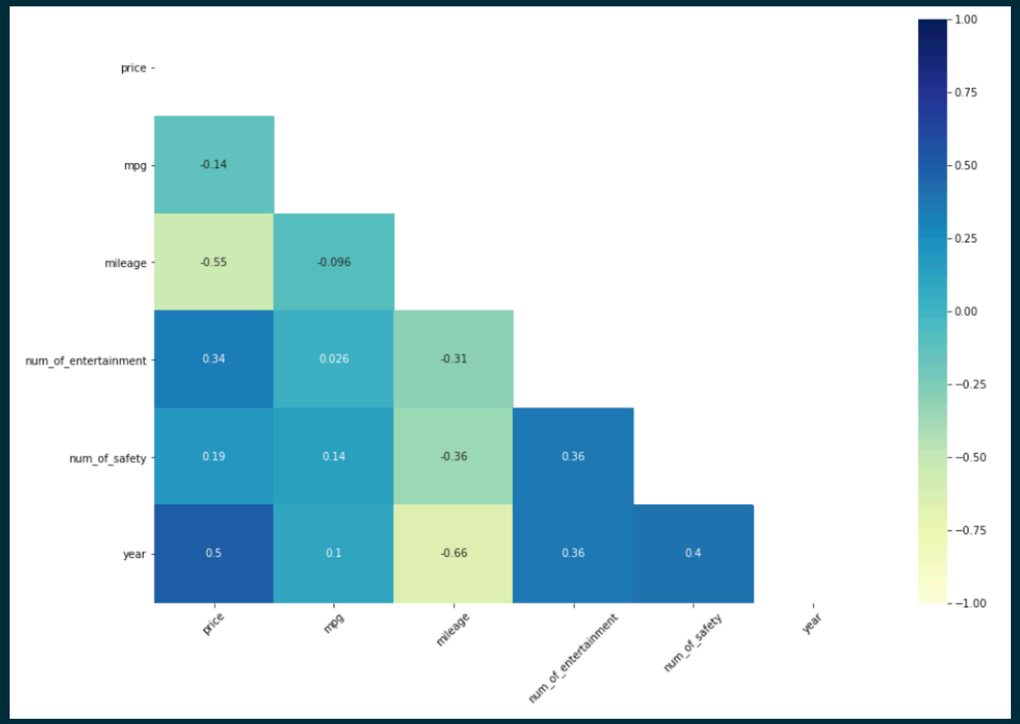
Making sure outliers are removed properly

COMPY.PY

Pair Plot

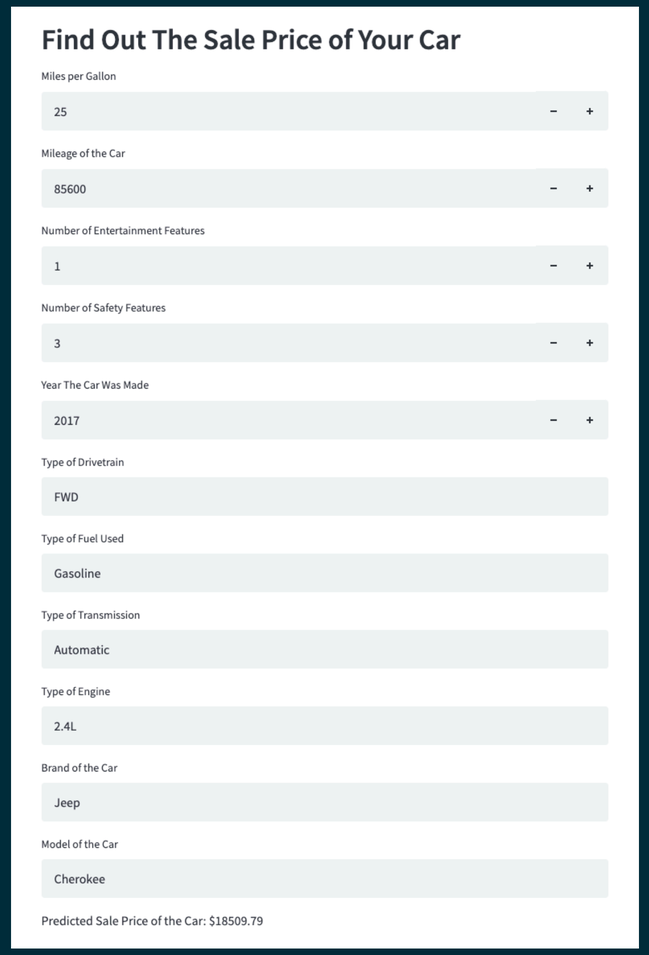


Correlation Plot



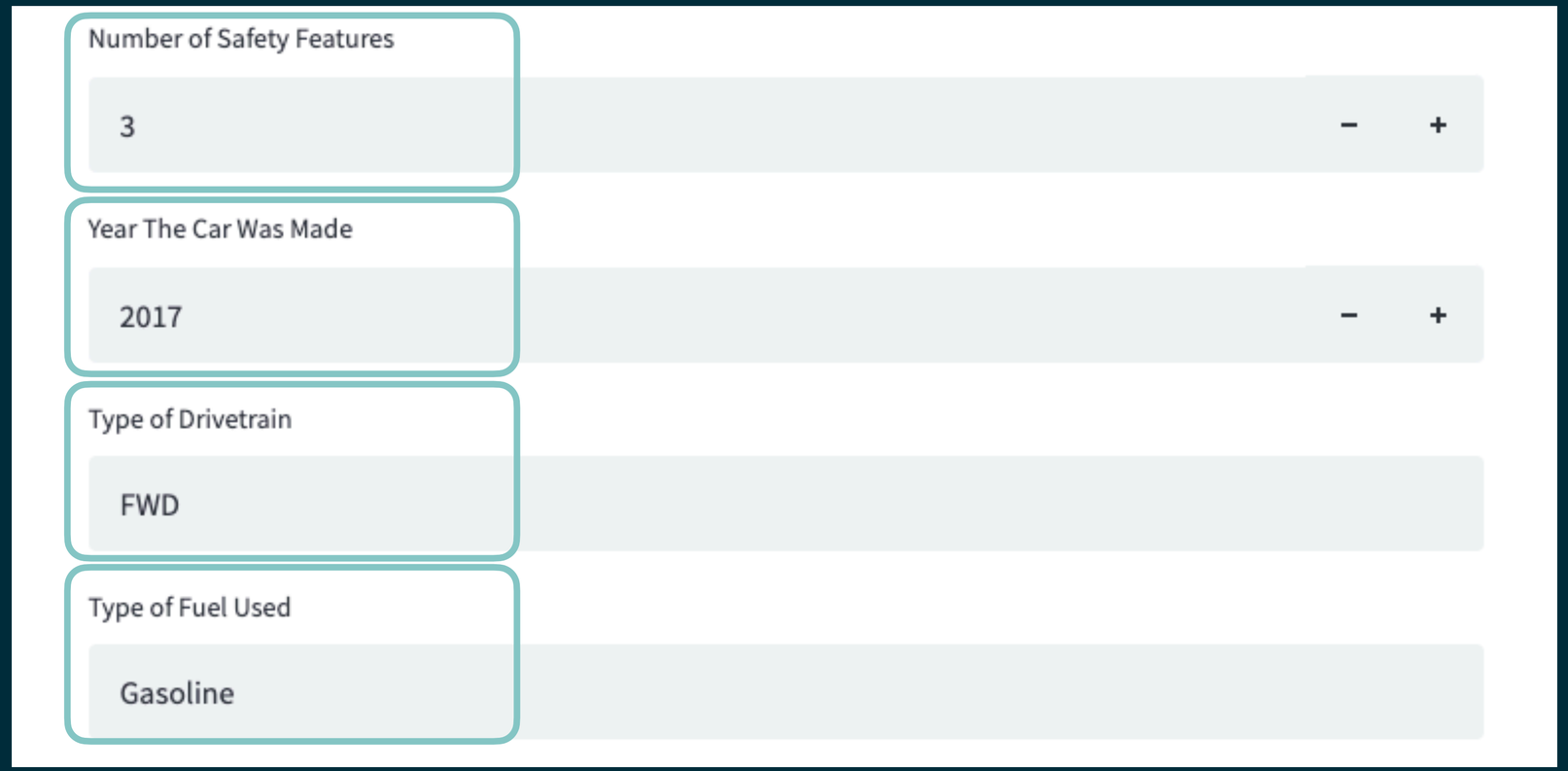
We’re able see if any additional feature engineering can be done with the updated data

THE APPLICATION

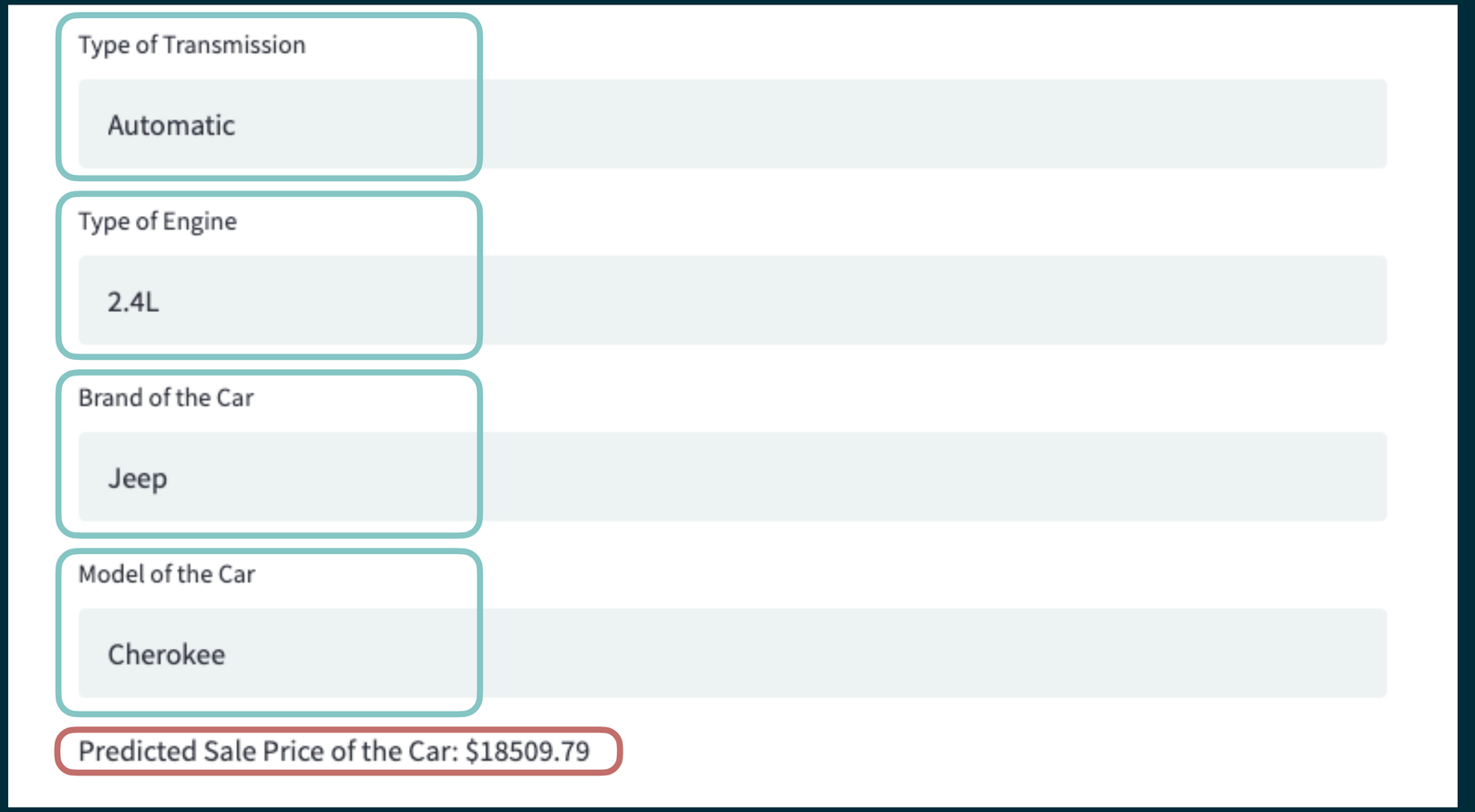


THE APPLICATION

THE APPLICATION



THE APPLICATION



PREDICTION

2019 Toyota 4Runner TRD Off Road

mpg	18
Mileage	26,704
Entertainment	2
Safety	3
Drivetrain	4WD
Fuel	Gasoline
Transmission	Automatic
Engine	4.0L

Actual: $ 44,995

Predicted: $ 45,033.40

PREDICTION

2017 Jeep Cherokee Latitude

mpg	25
Mileage	85,600
Entertainment	1
Safety	3
Drivetrain	FWD
Fuel	Gasoline
Transmission	Automatic
Engine	2.4L

Actual: $ 18,990

Predicted: $ 18,509.79

FUTURE WORK

- Change some application features to drop box with options
- Automate compy.py with Cron
- Deploy application with Heroku

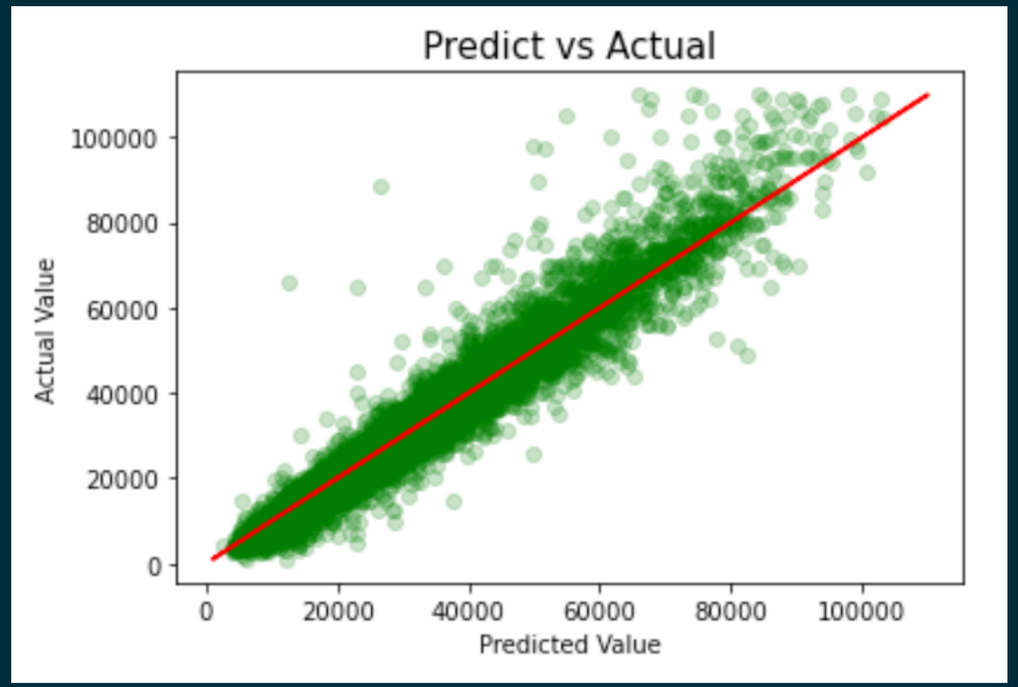
THANK YOU!

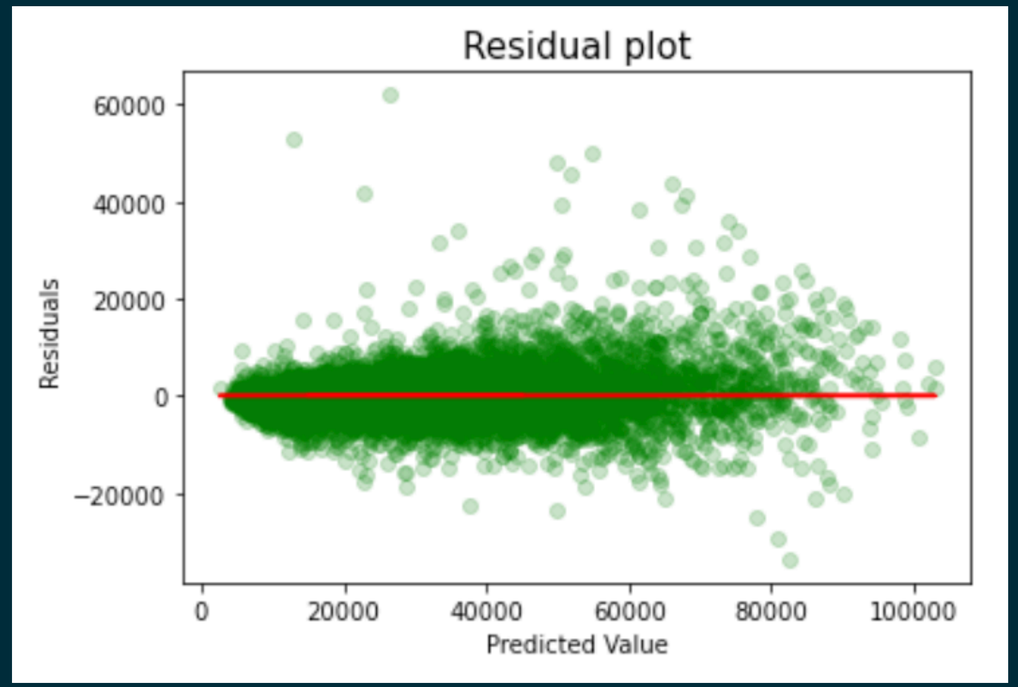
APPENDIX

Final Model:
Random Forest Regressor

Train R2	0.99
Test R2	0.94
MAE	2658.98

APPENDIX


Predict vs Actual
Actual Value
Predicted Value
0
20000
40000
60000
80000
100000


Residual plot
Residuals
Predicted Value
60000
40000
20000
0
−20000
0
20000
40000
60000
80000
100000