PROJECT 4 – CAR PRICING PREDICTOR

COLIN ROBERTS

MITCHELL HATCHETT

JAYLEN WHITTAKER

CONNERY HINSON

PAUL ANDERSON

PROJECT SUMMARY

The "Car Price Analysis and Prediction" project involves delving into a dataset encompassing various attributes of used cars, ranging from price and make to fuel type (electric, hybrid, gasoline), color, and horsepower. Through data analysis, we aim to uncover the key factors influencing car prices. Moreover, predictive modeling will enable us to estimate the price of cars based on their attributes, empowering private sellers and dealers to make informed pricing decisions. The data could then be used by an automobile seller as a guide on how to target their customer base and maximize sales.

PROCESS

- Find and clean the data, using pandas, create a few visualizations to ensure the data is clean and usable. Create csv files for export and use in the next step
- Use a PostGreSQL database to create and join tables, the goal being to have all the pertinent data contained in one or two tables. Perform basic queries to ensure data integrity
- Create visualizations in Tableau
- Create machine learning model, probably in Colab Jupyter Notebook to handle the predictive analysis that is the ultimate goal of this study.
- The interactive piece would allow a seller to input information about their car, and get an estimate of what it would sell for on the market.

DATA CLEANUP

- We used several csv files from various sources.
- Used pandas to clean up the files, dropping some irrelevant columns, adding an index in some cases.
- Code for the cleanup found in the repository

SQL DATABASE

After the data were cleaned, we created a SQL database for easy queries and to check the quality of the data. Three tables were eventually used. This one took into account items like horsepower and number of owners

	index [PK] integer	year_made integer	fuel_type character varying (30)	seats integer	mileage integer	ownership integer	transmission character varying (20)	fuel_economy numeric	engine_cc numeric	horsepower numeric	torque_nm numeric	price numeric	make character varying (30)	model character varying (30)
1	0	17	Petrol	5	34796	1	Automatic	7.81	2996.0	2996.0	333.0	86062.5	Mercedes-Benz	S-Class
2	1	21	Petrol	5	19023	1	Automatic	17.4	999.0	999.0	9863.0	12136.5	Nissan	Magnite
3	2	18	Diesel	5	14912	1	Automatic	20.68	1995.0	1995.0	188.0	32062.5	BMW	X1
4	3	19	Petrol	5	11419	1	Manual	16.5	1353.0	1353.0	13808.0	18306.0	Kia	Seltos
5	4	19	Petrol	5	27899	1	Automatic	14.67	1798.0	1798.0	17746.0	32400.0	Skoda	Superb
6	5	17	Petrol	5	26097	1	Manual	18.7	1199.0	1199.0	887.0	7357.5	Honda	Jazz
7	6	19	Petrol	5	22828	1	Manual	18.9	1197.0	1197.0	8186.0	6912.0	Hyundai	Grand
8	7	18	Petrol	5	47224	1	Manual	15.8	1591.0	1591.0	1213.0	12555.000000000002	Hyundai	Creta
9	8	15	Diesel	5	42253	2	Automatic	13.5	2987.0	2987.0	25479.0	56700.0	Mercedes-Benz	S-Class
10	9	19	Petrol	5	17884	1	Manual	17.0	1198.0	1198.0	1085.0	10827.0	Tata	Nexon
11	10	20	Petrol	5	24854	1	Automatic	17.4	1497.0	1497.0	1176.0	14782.499999999998	Honda	City
12	11	22	Petrol	5	10800	1	Manual	16.42	1498.0	1498.0	10455.0	12136.5	Renault	Duster
13	12	20	Diesel	5	74564	1	Automatic	18.88	1995.0	1995.0	184.0	10057.5	BMW	3
14	13	22	Petrol	5	10563	1	Automatic	18.15	998.0	998.0	11835.0	14782.499999999998	Hyundai	Venue

SECOND TABLE

This table added additional parameters like color and state

	index [PK] integer	price integer	make character varying (30)	model character varying (30)	year integer	title_state character varying (40)	mileage numeric	color character varying (50)	vin character varying (30)	state character varying (30)
1	0	6300	toyota	cruiser	2008	clean vehicle	274117	black	jtezu11f88k007763	new jersey
2	1	2899	ford	se	2011	clean vehicle	190552	silver	2fmdk3gc4bbb02217	tennessee
3	2	5350	dodge	mpv	2018	clean vehicle	39590	silver	3c4pdcgg5jt346413	georgia
4	3	25000	ford	door	2014	clean vehicle	64146	blue	1ftfw1et4efc23745	virginia
5	4	27700	chevrolet	1500	2018	clean vehicle	6654	red	3gcpcrec2jg473991	florida
6	5	5700	dodge	mpv	2018	clean vehicle	45561	white	2c4rdgeg9jr237989	texas
7	6	7300	chevrolet	pk	2010	clean vehicle	149050	black	1gcsksea1az121133	georgia
8	7	13350	gmc	door	2017	clean vehicle	23525	gray	1gks2gkc3hr326762	california
9	8	14600	chevrolet	malibu	2018	clean vehicle	9371	silver	1g1zd5st5jf191860	florida
10	9	5250	ford	mpv	2017	clean vehicle	63418	black	2fmpk3j92hbc12542	texas
11	10	10400	dodge	coupe	2009	clean vehicle	107856	orange	2b3lj54t49h509675	georgia
12	11	12920	gmc	mpv	2017	clean vehicle	39650	white	1gks2bkc6hr136280	california
13	12	31900	chevrolet	1500	2018	clean vehicle	22909	black	3gcukrec0jg176059	tennessee
14	13	5430	chrysler	wagon	2017	clean vehicle	138650	gray	2c4rc1cg5hr616095	texas

THE SQL DATABASE MADE IT EASY TO SPOT OUTLIERS AND POSSIBLE MISTAKES IN THE DATA, AND WE USED THE CLEAN TABLES FOR OUR MACHINE LEARNING MODEL, VISUALIZATIONS AND THE INTERACTIVE PARTS OF OUR PROJECTS AS WELL.

```sql
select make, price, model
FROM the_used_cars
order by price desc
```

What is a Ford Figo and why does it cost $94.5 million? Obviously an error in the dataset.

	make character varying (30)	price numeric	model character varying (30)
1	Ford	94500000.0	Figo
2	Ford	94500000.0	Figo
3	BMW	133650.0	X7
4	BMW	133650.0	X7
5	BMW	132975.0	X7
6	Mercedes-Benz	132975.0	S-Class
7	Mercedes-Benz	132975.0	S-Class
8	BMW	132975.0	X7
9	Mercedes-Benz	130950.0	S-Class
10	Mercedes-Benz	130950.0	S-Class
11	Mercedes-Benz	125550.0	S-Class
12	Mercedes-Benz	125550.0	S-Class
13	Audi	117450.0	Q7
14	Audi	117450.0	Q7

USED CAR PRICE PREDICTOR USING FLASK

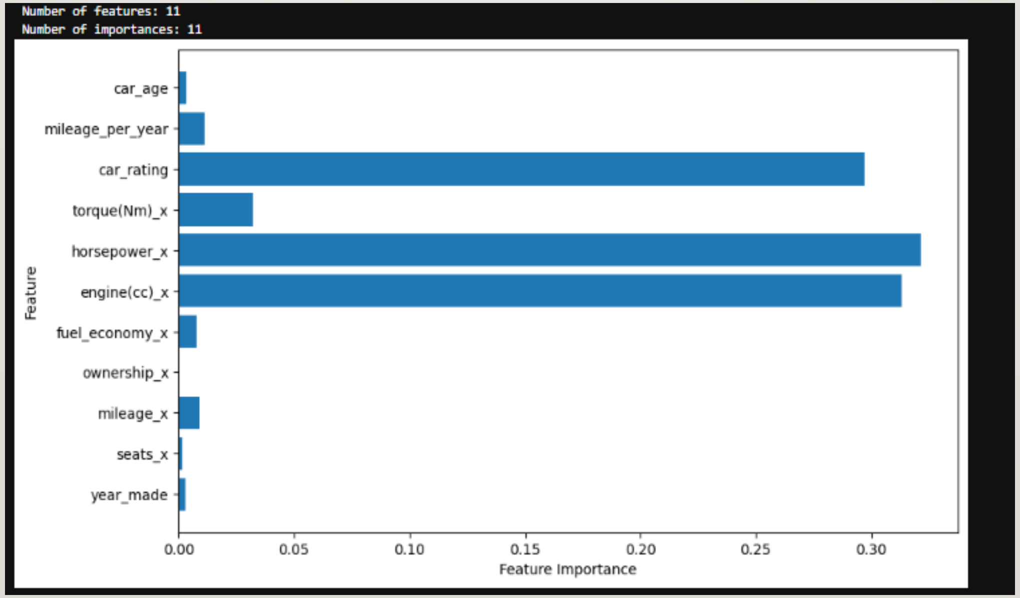
- Several columns are converted to numeric types to ensure proper analysis. The price column is converted from "lakhs" to USD using a conversion rate of 1 lakh = 1350 USD. The original price column is then dropped. Unnecessary columns, such as 'Unnamed: 0', are dropped. The 'ownership' column is renamed from a misspelled version to ensure consistency in the dataset. cleaning and preparing the dataset for further analysis, ensuring that all relevant data is in the correct format for subsequent modeling and evaluation. The we identify missing values in both datasets and drops rows with missing data to clean the datasets further. datasets are merged based on common columns such as index, Make, and Model. We used various machine learning tools such as LinearRegression and RandomForestRegressor to plot future prices in next five years.

USED CAR PRICE PREDICTOR USING FLASK

- Using Pandas Numpy and SKLearn libraries
  All the factors are normalized using Min-Max scaling to a scale of 0 to 1. This step ensures that each factor contributes proportionally to the final score, making the different criteria comparable by 3.
  Then sum up the normalized values of the positive factors. Those being
  Fuel efficiency, torque, horse power and year made.
  Subtracting the normalized values of the negative factors.
  Engine size, mileage, price and previous owners.
  The resulting score is then normalized to a 0-100 scale to make it easier to interpret.

USED CAR PRICE PREDICTOR USING FLASK

- Then identified the rows that were most important for the price prediction model. These columns being the car rating, horsepower and engine.



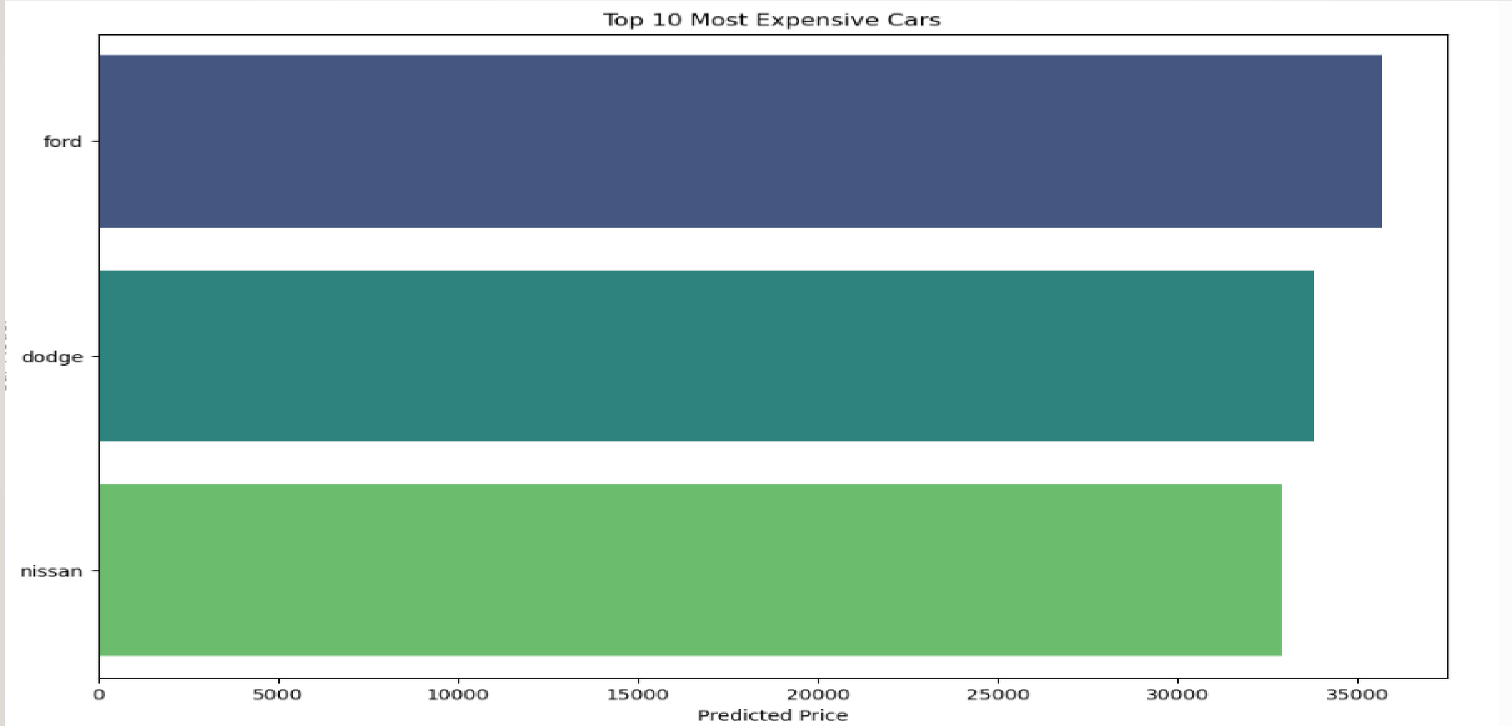
PY APP WITH FLASK FOR THE USED CAR ESTIMATOR

Using flask and python upon initialization, the app loads two key datasets. web application is designed to estimate the value of used cars and provide a detailed rating based on various criteria These dropdowns are dynamically populated based on the unique options extracted from the datasets, ensuring that users only see relevant and available choices. Upon form submission, the app attempts to find an exact match in the dataset to calculate an average price based on similar listings. If no exact match is found, it broadens the search or defaults to an overall average price of the make and models average . The app also retrieves a rating for the selected car make and model, displaying it alongside the estimated price. The application is designed with basic error handling and is deployed on Heroku through github.

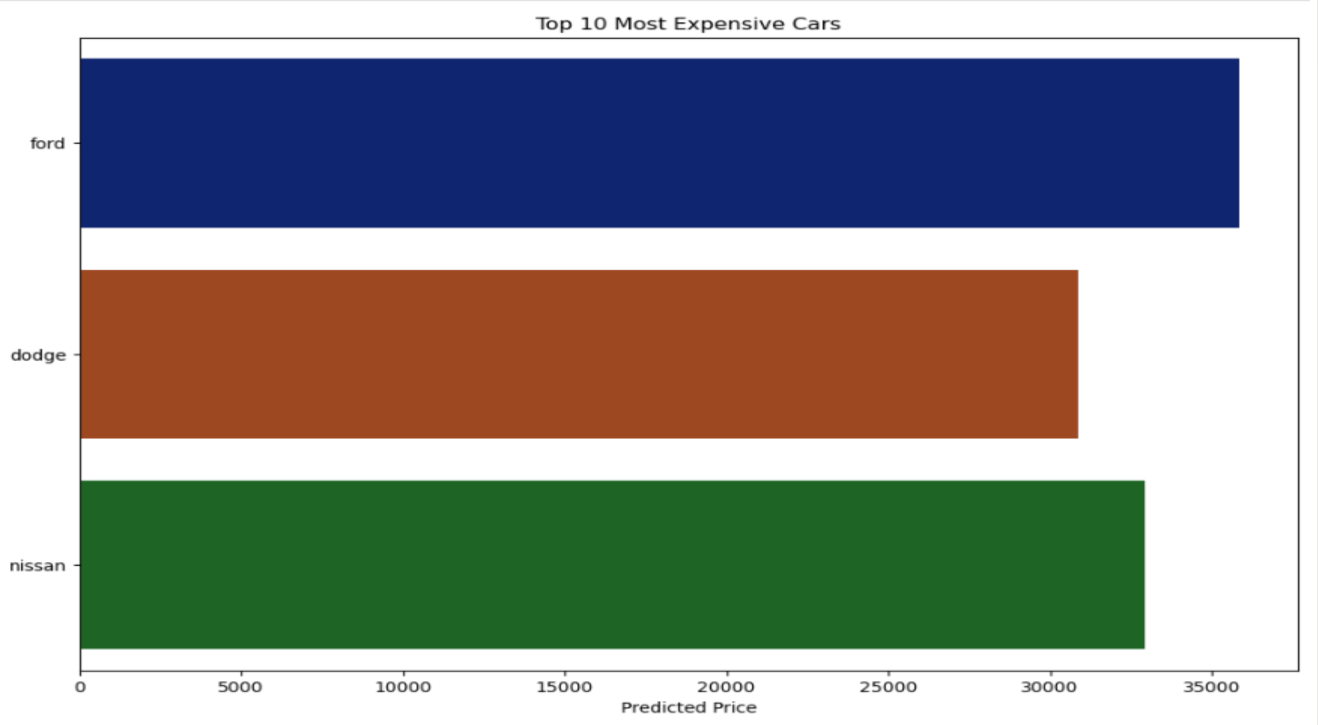
THE MACHINE LEARNING MODEL

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as pltf
import seaborn as sns
from sklearn.model_selection import train_test_split, GridSearchCV
from sklearn.preprocessing import StandardScaler
from sklearn.ensemble import GradientBoostingRegressor
from sklearn.metrics import mean_squared_error
```

WE USED OUR CLEANED DATA IN PYTHON/PANDAS TO CONSTRUCT A MODEL TO PREDICT THE PRICE OF USED CARS, AND COMPARED OUR PREDICTIONS TO THE DATA



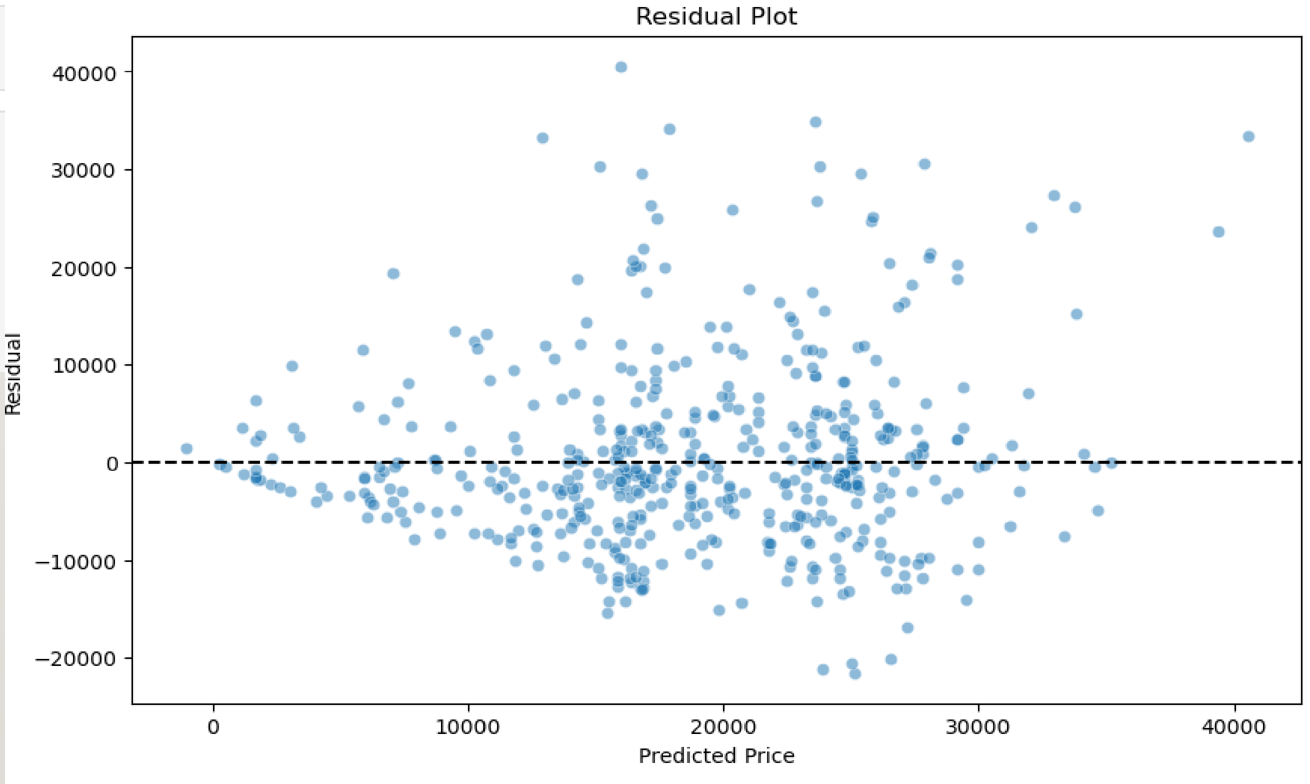
REFINING THE MODEL



A RESIDUAL PLOT OF THE MODEL

```
residuals = y_test - y_pred
```

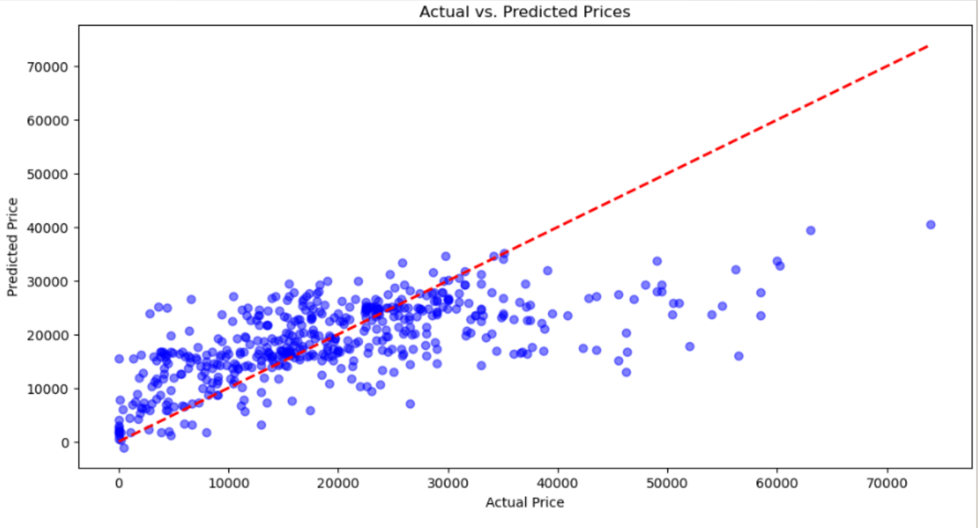
```
plt.figure(figsize=(10, 6))
sns.scatterplot(x=y_pred, y=residuals, alpha=0.5)
plt.axhline(0, color='k', linestyle='--')
plt.title('Residual Plot')
plt.xlabel('Predicted Price')
plt.ylabel('Residual')
plt.show()
```



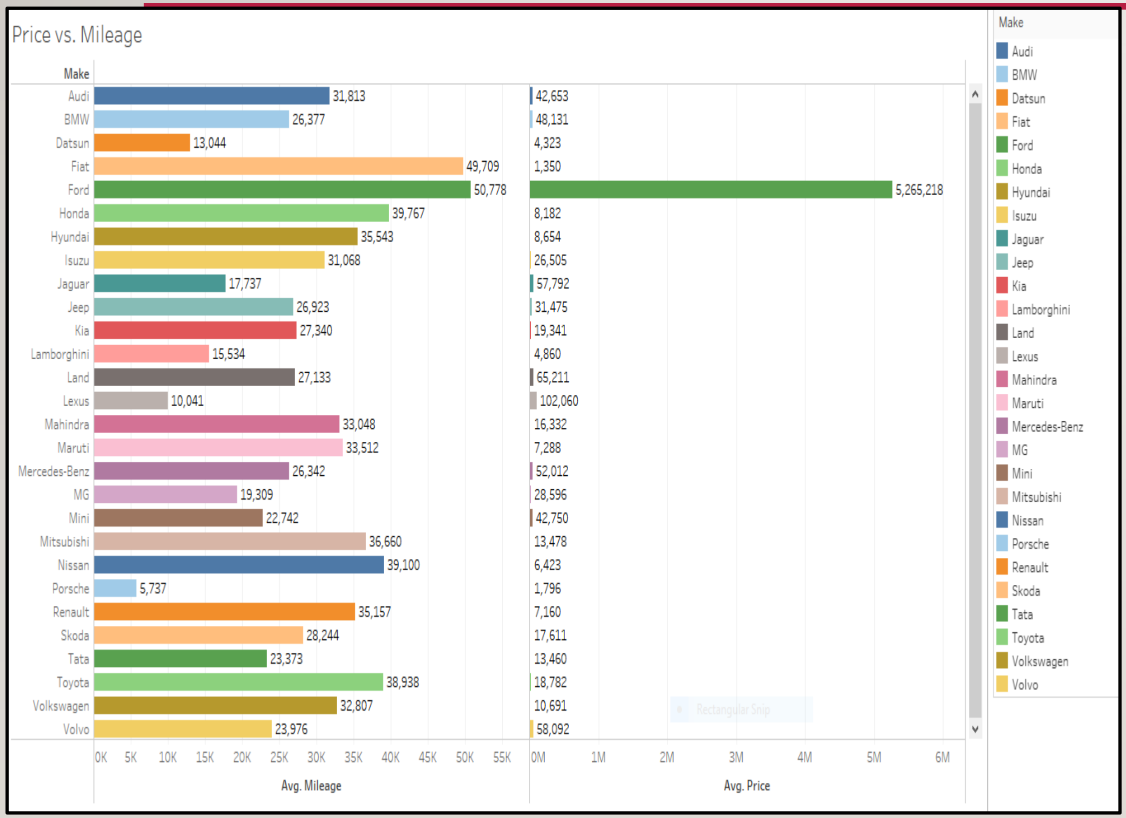
A PLOT OF ACTUAL VS. PREDICTED PRICES

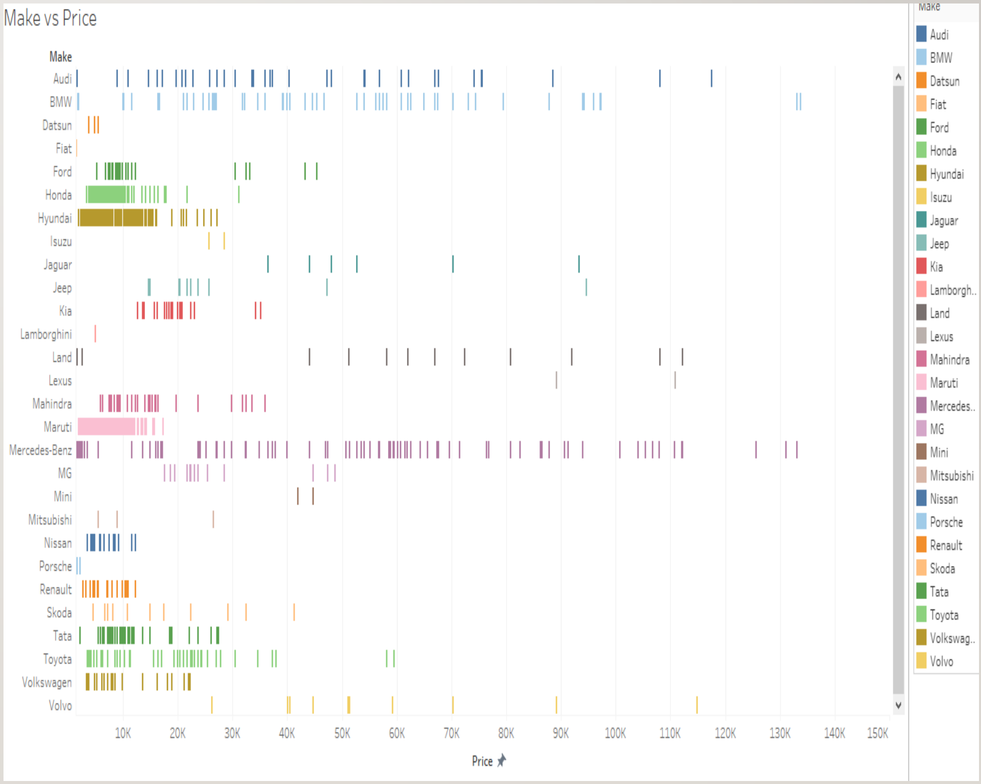
```
y_test_pred = model.predict(X_test)
```

```
plt.figure(figsize=(12, 6))
plt.scatter(y_test, y_test_pred, alpha=0.5, color='blue')
plt.plot([y_test.min(), y_test.max()], [y_test.min(), y_test.max()], 'k--', color='r', lw=2)
plt.title('Actual vs. Predicted Prices')
plt.xlabel('Actual Price')
plt.ylabel('Predicted Price')
plt.show()
```

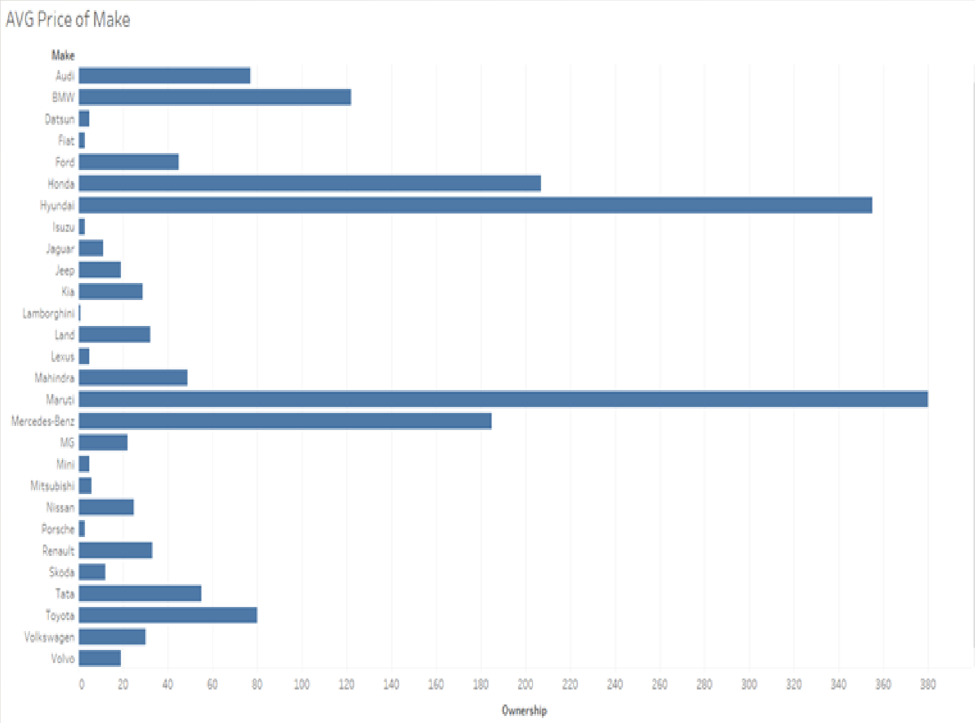


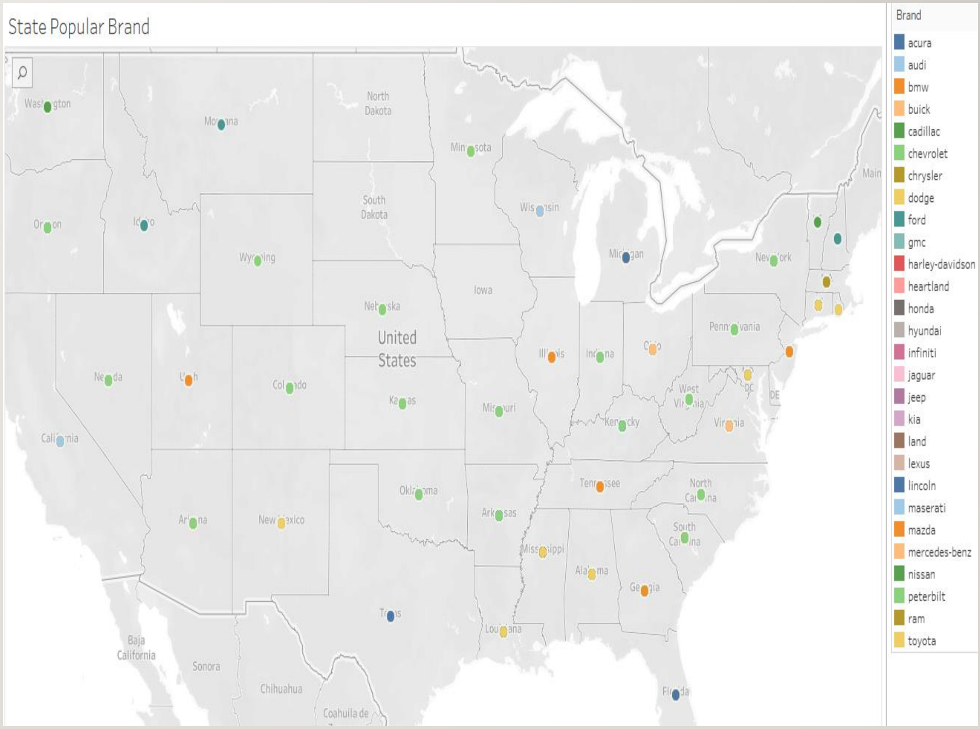
NEXT, WE USED THE DATA TO CREATE VISUALS IN TABLEAU. HERE ARE SOME EXAMPLES. THE REST CAN BE FOUND AT THE LINK TO TABLEAU PUBLIC IN THE REPOSITORY





VISUALIZATIONS





VISUALIZATIONS

Fuel Type

	Fuel Type		
Make	CNG	Diesel	Petrol
Audi		Audi	Audi
BMW		BMW	BMW
Datsun			Datsun
Fiat		Fiat	
Ford		Ford	Ford
Honda		Honda	Honda
Hyundai	Hyundai	Hyundai	Hyundai
Isuzu		Isuzu	
Jaguar		Jaguar	Jaguar
Jeep		Jeep	Jeep
Kia		Kia	Kia
Lamborghini			Lamborghini
Land		Land	Land
Lexus			Lexus
Mahindra		Mahindra	Mahindra
Maruti	Maruti	Maruti	Maruti
Mercedes-Benz		Mercedes-Benz	Mercedes-Benz
MG		MG	MG
Mini			Mini
Mitsubishi			Mitsubishi
Nissan		Nissan	Nissan
Porsche			Porsche
Renault		Renault	Renault
Skoda		Skoda	Skoda
Tata	Tata	Tata	Tata
Toyota		Toyota	Toyota
Volkswagen		Volkswagen	Volkswagen
Volvo		Volvo	Volvo

RESULTS AND FEATURES

- Using a combination of datasets and APIs from sources on the internet, we were able to construct a machine learning model that accurately predicted the price of a used car based on various criteria, such as price, make, model, and year.
- We were able to construct visualizations to illustrate the model
- We were also able to incorporate an interactive link, where users could estimate the value of a car they were trying to sell, based again, on similar criteria that they input

QUESTIONS