ENTITY ACADEMY PRESENTATION

DISCOVERING THE MOST INFLUENTIAL FEATURES
TO PREDICT THE PRICE OF USED CARS

NAVIGATING THE GREAT DEMAND

TIA HARRIS 2022

ABOUT

PASSIONATELY OBSESSED WITH THE FUSION OF ART AND TECHNOLOGY

ABOUT THE PRESENTER

DESIGNER | DEVELOPER | ANALYST

THROUGH THE TIME

AN “EXPLORATORY ANALYSIS” ABOUT THE AUTHOR

TIA HARRIS

2011

Graduated with degree in Graphic Design from The Art Institute of Pittsburg (Online). Shortly pursued a Flight attendant career soon after in addition to freelance graphic design projects.

2015

Continued work as a freelance designer, started web career at Go Daddy managing and repairing WordPress websites

2017

Switched from GoDaddy to Bluehost, to work as a web designer for client WordPress websites

2018

Went back to school for BA in web development. Started working at a Digital Marketing Agency to handle local SEO development for clients.

2020

Switched degree from Web Development, to Computer Science focus. Started a job with DHL Express and a Web Content Administrator on the UX/UI team.

2021

Enrolled in the entity data science program.

2022

Achieved Data Science certificate from entity academy. Started a job at Indeed as an Account Manager

INTRO

PREDICTING USED CAR PRICES GIVEN ITS SPECIFICATION

PROJECT INTRODUCTION

INTRODUCTION

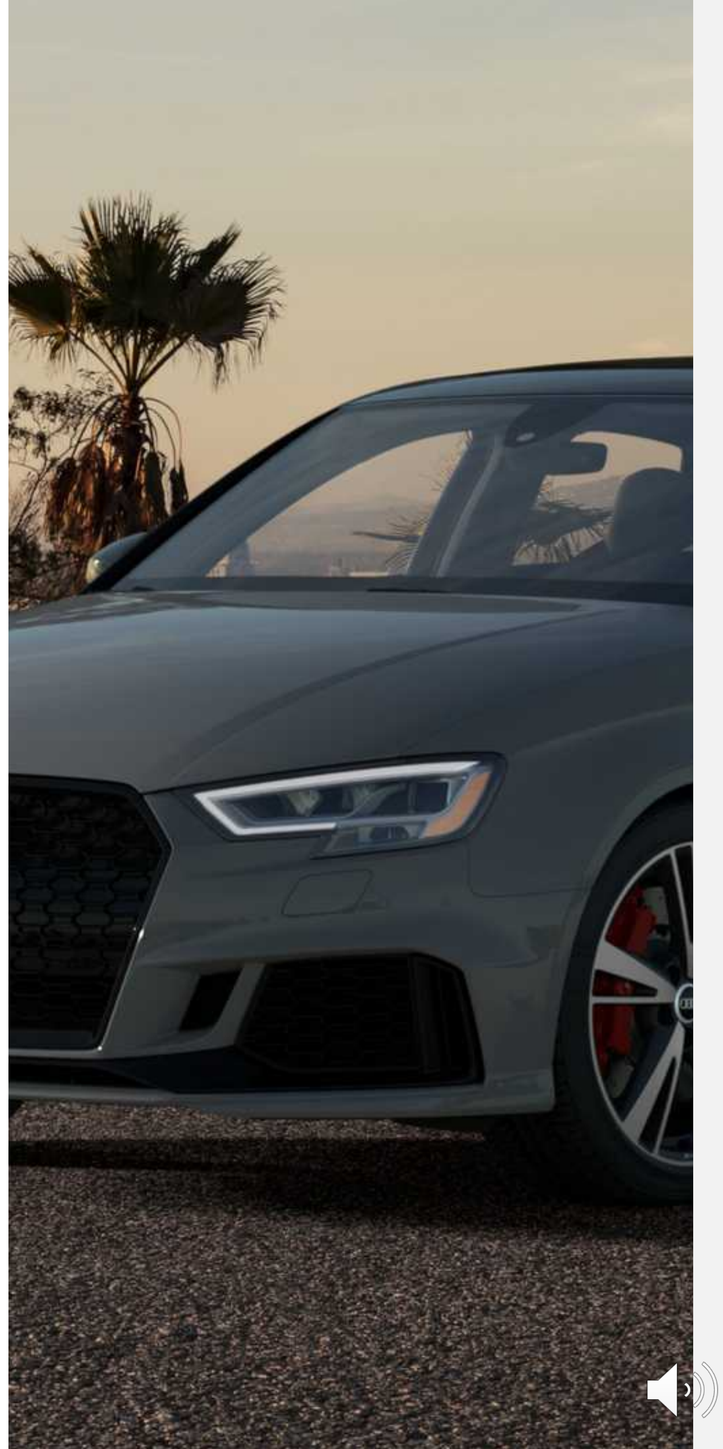
APPROXIMATELY 40.9 MILLION

USED CARDS WERE SOLD IN 2021 IN THE US

Earlier this year, Atlanta based global automotive business unit, Cox Automotive reported that the number of used vehicles sold in 2021 in the US, hit an all-time record.

The COVID-19 pandemic kept millions of Americans off the road due to the shelter – in place order. This drastically slowed down the car shopping of consumers. This would later create a demand for new cars

A microchip shortage limited manufacture's ability to build enough new cars to keep up with this new demand. This shifted would be “new car buyers” into the used car market. Thus, creating a second demand for used cars.



THE USED CAR MARKET PROBLEM

LOW SUPPLY HIGH DEMAND

As buyers flooded the used car market with “new-car money” to spend, the prices for used cars also soared. This did not drive consumers away, however as the demand for buying cars remained.

With so much demand on the current availability of used cars, and with automakers building fewer new cars, there is a bit of uncertainty in the future of the used car market. Fewer new cars to sale, means less trade-ins arriving on used car lots. Which could could then be reflected by a shortage in the pre-owned vehicle market.

CAR SHORTAGES LEAD TO INDUSTRY CHANGES

FEWER + MORE EXPENSIVE OPTIONS

Buying a used car in the U.S has never been more expensive. The average price of a used car is now $30,000, which is $10,00 above average mockup, according to data collected by Kelly Blue Book and CoPilot.

With fewer cars available, dealerships and private sellers have raised priced on even the most rebarbative pre - owned vehicles, in order to continue to maximize their profit from sales. Despite the inflation, buyers are still actively shopping the market. These conditions have created the perfect “seller’s market” for both used and new vehicles.

The entire industry’s prices have gone up, and the demand for used cars has made it difficult for sellers to keep the prices competitive, however there are key factors at play that can help determine optimal pricing for car manufactures and resellers alike.

WHAT WILL WE IDENTIFY?

OBJECTIVE

Identify and understand the physical factors affecting the prices of used cars, to remain competitive in the market. This analysis can be used to determine prices of new cars or accurately set prices of used cars based on the market.

THE SIGNIFICANCE OF THE STUDY

BUSINESS GOALS

- HELP DECISION MAKERS AND CAR MANUFACTUERS DETERMINE CAR PRICES FOR BUSINESS GROWTH AND MARKET MANAGEMENT
- HELP SELLERS UNDERSTAND THE PRICING OF USED CARS IN THE CURRENT MARKET
- LEARN WHICH FACTORS ARE MOST SIGNIFICANT AT DETERMINING THE PRICE OF A CAR
- WHAT FEATURES ARE CONSUMERS WILLING TO PAY FOR, EVEN AT AN INFLATED PRICE?

METHOD 01

MULTIPLE LINEAR REGRESSION

THERE ARE MULTIPLE INDEPENDENT VARIABLES WE ARE CHECKING TO SEE IF THEY INFLUENCE THE DEPENDENT VARIABLE (PRICE)

METHOD 02

PYTHON

PERSONAL PEFERENCE, AND I LIKE TESTING FOR ASSUMPTIONS AND BUILFING VISUALIZATIONS IN PYTHON VERSUS R.

METHOD 03

PACKAGES

THE PACKAGES AND LIBRARIES NECESSARY FOR THIS ANALYSIS VARY FROM PANDAS, SKLEARN, SEABORN, NUMPY,AND MATPLOTLIB,

EXTRAS

OTHER

UTILIZED JUPITER NOTEBOOKS FOR THE ANALYSIS
DID SOME DATA EXPLORATION AND VISUALS IN TABLEAU
MINOR IMAGE EDITING IN ADOBE PHOTOSHOP
PRESENTATION WAS CRAFTED WITH MS POWERPOINT

ABOUT

EXPLORATORY ANALYSIS

ABOUT THE DATA

CAR SALES DATA — TABLEAU

Abc Car-Sales-Data.csv Brand	# Car-Sales-Data.csv Price	Abc Car-Sales-Data.csv Body	# Car-Sales-Data.csv Mileage	# Car-Sales-Data.csv Engine V	Abc Car-Sales-Data.csv Engine Type	Abc Car-Sales-Data.csv Registration	# Car-Sales-Data.csv Year	Abc Car-Sales-Data.csv Model
BMW	4,200.00	sedan	277	2.00000	Petrol	yes	1991	320
Mercedes-Benz	7,900.00	van	427	2.90000	Diesel	yes	1999	Sprinter 212
Mercedes-Benz	13,300.00	sedan	358	5.00000	Gas	yes	2003	S 500
Audi	23,000.00	crossover	240	4.20000	Petrol	yes	2007	Q7
Toyota	18,300.00	crossover	120	2.00000	Petrol	yes	2011	Rav 4
Mercedes-Benz	199,999.00	crossover	0	5.50000	Petrol	yes	2016	GLS 63
BMW	6,100.00	sedan	438	2.00000	Gas	yes	1997	320
Audi	14,200.00	vagon	200	2.70000	Diesel	yes	2006	A6
Renault	10,799.00	vagon	193	1.50000	Diesel	yes	2012	Megane
Volkswagen	1,400.00	other	212	1.80000	Gas	no	1999	Golf IV
Renault	11,950.00	vagon	177	1.50000	Diesel	yes	2011	Megane
Renault	2,500.00	sedan	260	1.79000	Petrol	yes	1994	19
Audi	9,500.00	vagon	165	2.70000	Gas	yes	2003	A6 Allroad
Volkswagen	10,500.00	sedan	100	1.80000	Petrol	yes	2008	Passat B6
Toyota	16,000.00	crossover	250	4.70000	Gas	yes	2001	Land Cruiser ...
Renault	8,600.00	hatch	84	1.50000	Diesel	yes	2012	Clio
BMW	2,990.00	other	203	2.00000	Petrol	no	2001	318

THE STORY DATA TELLS

LONG STORY SHORT.

WHAT CAN THE RAW DATA TELL US?

WHAT CAN WE SPOT?

We can observe a bit from this data, before we even do anything to it. We know that a BMW is more expensive than a Toyota. We also know that the more mileage a car has, the cheaper that car will be. And we also can conclude that the older the car, the more mileage we can assume it has.

DESCRIPTIVE STATISTICS

```
cars.describe(include = 'all')
```

	Brand	Price	Body	Mileage	EngineV	Engine Type	Registration	Year	Model
count	4345	4173.000000	4345	4345.000000	4195.000000	4345	4345	4345.000000	4345
unique	7	NaN	6	NaN	NaN	4	2	NaN	312
top	Volkswagen	NaN	sedan	NaN	NaN	Diesel	yes	NaN	E-Class
freq	936	NaN	1649	NaN	NaN	2019	3947	NaN	199
mean	NaN	19418.746935	NaN	161.237284	2.790734	NaN	NaN	2006.550058	NaN
std	NaN	25584.242620	NaN	105.705797	5.066437	NaN	NaN	6.719097	NaN
min	NaN	600.000000	NaN	0.000000	0.600000	NaN	NaN	1969.000000	NaN
25%	NaN	6999.000000	NaN	86.000000	1.800000	NaN	NaN	2003.000000	NaN
50%	NaN	11500.000000	NaN	155.000000	2.200000	NaN	NaN	2008.000000	NaN
75%	NaN	21700.000000	NaN	230.000000	3.000000	NaN	NaN	2012.000000	NaN
max	NaN	300000.000000	NaN	980.000000	99.990000	NaN	NaN	2016.000000	NaN

DETECTING AND DEALING WITH MISSING VALUES

DROPPING DATA

Using the **.dropna function** in pandas, we can remove the entire record that contains missing data in any column. This is also known as Listwise deletion, and this method can be dangerous, as deleting to many rows can bias the data significantly or render the data useless for analysis.

SINCE NOT MUCH OF THE DATA SEEMED TO BE MISSING (less than 5%), I DROPPED THE RECORDS FROM THE DATASET.

```
Cars.isnull().sum()
```

```
Brand             0
Price           172
Body              0
Mileage           0
EngineV         150
Engine Type       0
Registration      0
Year              0
dtype: int64
```

- Dropped the Model column

```
Cars = cars.drop(['Model'], axis =1)
```

- Used the isnull functions to determine how many missing values where in the dataset.
- There were 172 rows missing data from the price column and 150 rows missing from the engine column. These were also dropped from the dataset.

DESCRIPTIVE STATISTICS

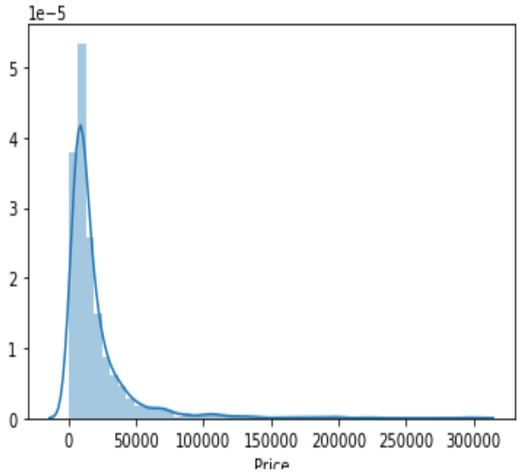
```
cars_data.describe(include='all')
```

	Brand	Price	Body	Mileage	EngineV	Engine Type	Registration	Year
count	4025	4025.000000	4025	4025.000000	4025.000000	4025	4025	4025.000000
unique	7	NaN	6	NaN	NaN	4	2	NaN
top	Volkswagen	NaN	sedan	NaN	NaN	Diesel	yes	NaN
freq	880	NaN	1534	NaN	NaN	1861	3654	NaN
mean	NaN	19552.308065	NaN	163.572174	2.764586	NaN	NaN	2006.379627
std	NaN	25815.734988	NaN	103.394703	4.935941	NaN	NaN	6.695595
min	NaN	600.000000	NaN	0.000000	0.600000	NaN	NaN	1969.000000
25%	NaN	6999.000000	NaN	90.000000	1.800000	NaN	NaN	2003.000000
50%	NaN	11500.000000	NaN	158.000000	2.200000	NaN	NaN	2007.000000
75%	NaN	21900.000000	NaN	230.000000	3.000000	NaN	NaN	2012.000000
max	NaN	300000.000000	NaN	980.000000	99.990000	NaN	NaN	2016.000000

TESTING ASSUMPTIONS

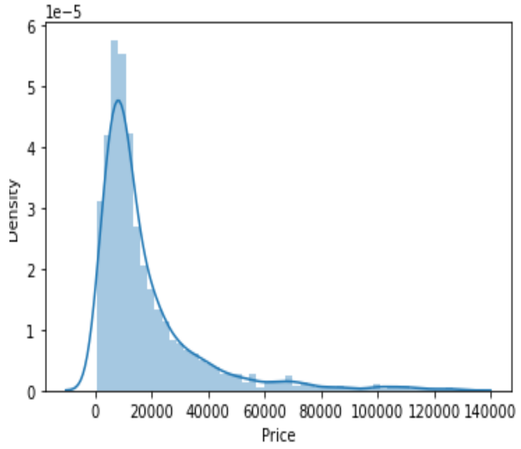
TESTING FOR LINEARITY AND NORMALITY

NOW THAT THE DATA HAS BEEN CLEANED OF MISSING VALUES, WE CAN GET A FEEL FOR THE DISTRIBUTION OF THE DATA. A PICTURE IS TRULY WORTH A THOUSAND WORDS, SO LETS VISIUALIZE THE DATA AND USE DATA VISUALIZATION TO TRY TO GRASP A NARRATIVE OF THE DATA



HISTOGRAM

Used seaborn to apply a best fit line to the histogram. The data looks skewed, which was due to outliers.



NO OUTLIERS

Since the data was not normally distributed, I calculated the outlier data using the quantile method in pandas and a custom function to output the outliers.

```
number of outliers: 355
max outlier value: 300000.0
min outlier value:44600.0
min price value:600.0
max price value:300000.0

5        199999.0
37        67500.0
41        63000.0
62       133000.0
64        50000.0
           ...
4318     300000.0
4322     100000.0
4327      80999.0
4331      45000.0
4340     125000.0
Name: Price, Length: 355, dtype: float64
```

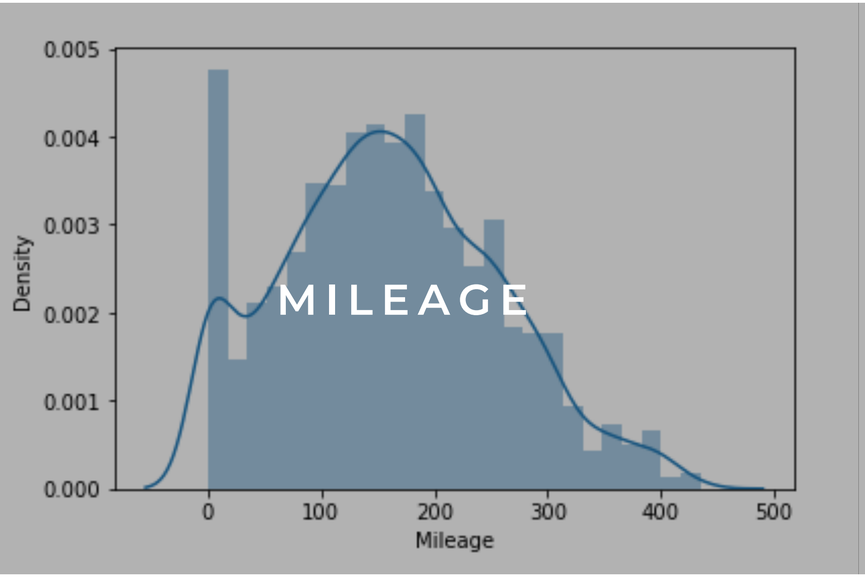
RESULTS

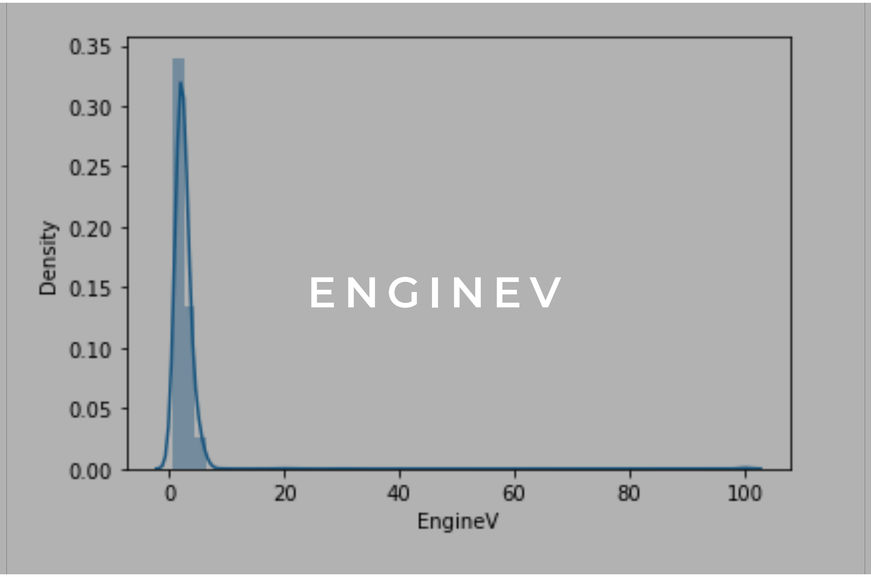
According to the function, there were 355 outliers present in the data. Dropping the outliers would result in data loss, so by rule, only 1% of the data was dropped and all data within the 99th percentile was kept.

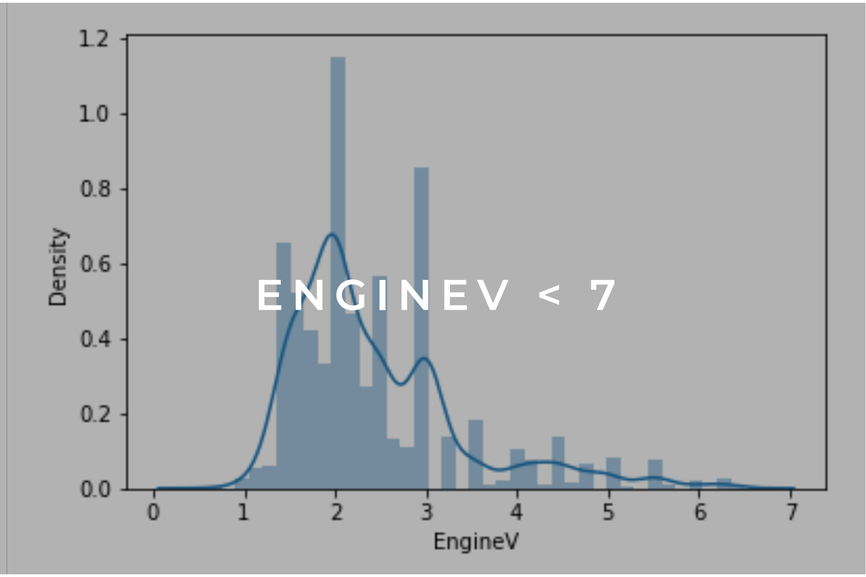
CONTINUE TESTING ASSUMPTIONS

REMOVE OUTLIERS

THE QUANTILE METHOD AND CUSTOM FUNCTION WAS ALSO RUN ON THE INDEPENDENT VARIABLES TO REMOVE THE OUTLIERS







The data for mileage looks normally distributed after dropping the outliers.

This data appeared to be heavily skewed. Dropping outliers did not appear to be enough.

Determined that any value over 7 can be considered an outlier and dropped those values from being used in the analysis.

DESCRIPTIVE STATISTICS

```
carsDataCleaned.describe(include='all')
```

	Brand	Price	Body	Mileage	EngineV	Engine Type	Registration	Year
count	3868	3868.000000	3868	3868.000000	3868.000000	3868	3868	3868.000000
unique	7	NaN	6	NaN	NaN	4	2	NaN
top	Volkswagen	NaN	sedan	NaN	NaN	Diesel	yes	NaN
freq	848	NaN	1468	NaN	NaN	1807	3506	NaN
mean	NaN	18198.929708	NaN	160.542399	2.451487	NaN	NaN	2006.710186
std	NaN	19085.415722	NaN	95.620925	0.951474	NaN	NaN	6.103116
min	NaN	800.000000	NaN	0.000000	0.600000	NaN	NaN	1988.000000
25%	NaN	7200.000000	NaN	91.000000	1.800000	NaN	NaN	2003.000000
50%	NaN	11700.000000	NaN	157.000000	2.200000	NaN	NaN	2008.000000
75%	NaN	21700.000000	NaN	225.000000	3.000000	NaN	NaN	2012.000000
max	NaN	129222.000000	NaN	435.000000	6.500000	NaN	NaN	2016.000000

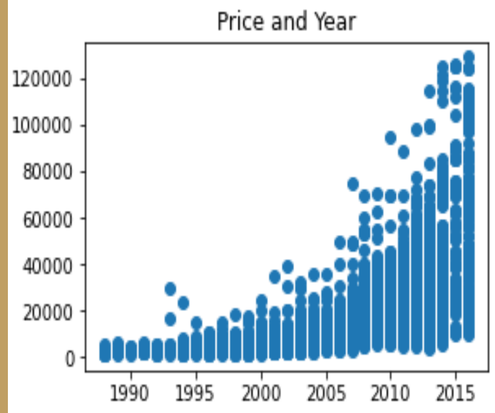
TESTING ASSUMPTIONS

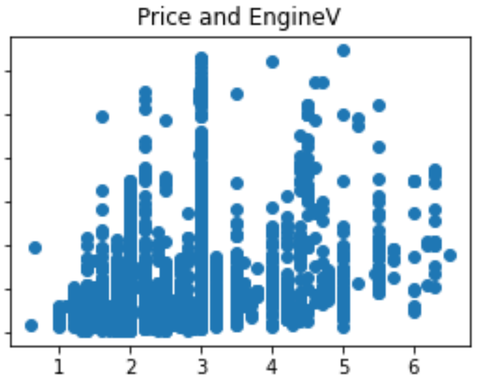
LINEAR RELATIONSHIPS

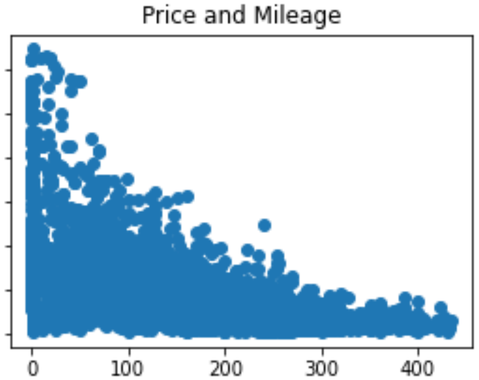
CHECKING RELATIONSHIPS BETWEEN THE INDEPENDENT AND DEPENDENT VARIABLES

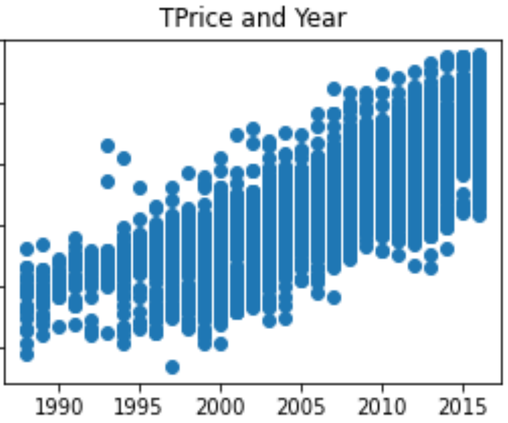
SCATTER PLOTS

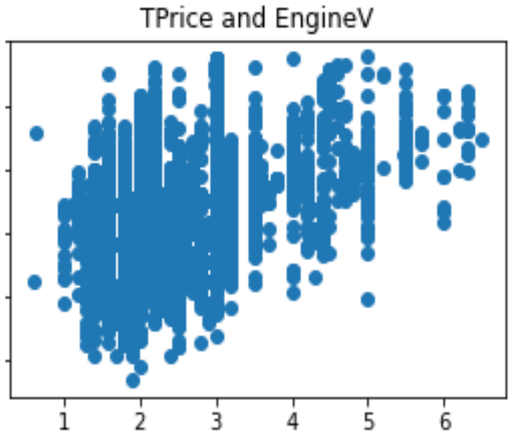
Create a scatter plot, using the scatter method from matplotlib to print graphs that show the relationships and examine for linearity.

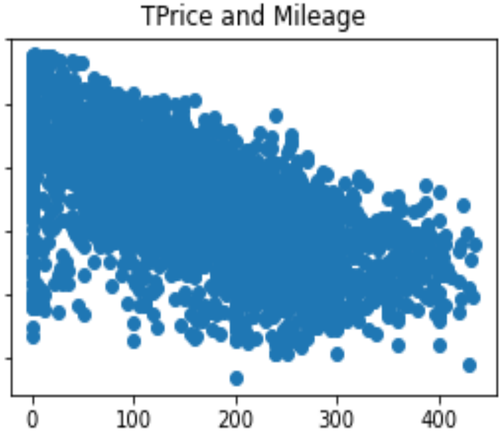












LOG TRANSFORMATION

After running a log transformation on the price variable, a linear relationship can now be found in all of the plots.

TESTING FOR MULTICOLLINEARITY

TESTING FOR INDEPENDENCE, WHERE ONE VARIABLE DOESN'T TELL ANYTHING ABOUT ANY OTHER DATA POINT

THERE APPEARED TO BE A HIGH CORRELATION BETWEEN THE YEAR AND THE MILEAGE, DROP THE YEAR.

```
carsData_cleaned.corr()
```

	Mileage	EngineV	Year	TPrice
Mileage	1.000000	0.005690	-0.715357	-0.614691
EngineV	0.005690	1.000000	-0.019872	0.412223
Year	-0.715357	-0.019872	1.000000	0.746827
TPrice	-0.614691	0.412223	0.746827	1.000000

WRAPPING UP DATA WRANGLING

DUMMY VARIABLES AND REARRANGING COLUMNS

AFTER DROPPING THE COLUMNS, REMOVING THE OUTLIERS, DELETING NULL DATA ROWS, CREATING DUMMY VARIABLES AND REARRANGING THE COLUMNS, OUR DATA IS CLEANED AND READY TO PROCEED WITH MODELING WITH LINEAR REGRESSION

	TPrice	Mileage	EngineV	Brand_BMW	Brand_Mercedes-Benz	Brand_Mitsubishi	Brand_Renault	Brand_Toyota	Brand_Volkswagen	Body_hatch	Body_other	Body_sed
0	8.342840	277	2.00	1	0	0	0	0	0	0	0	
1	8.974618	427	2.90	0	1	0	0	0	0	0	0	
2	9.495519	358	5.00	0	1	0	0	0	0	0	0	
3	10.043249	240	4.20	0	0	0	0	0	0	0	0	
4	9.814656	120	2.00	0	0	0	0	1	0	0	0	
5	9.560997	200	2.70	0	0	0	0	0	0	0	0	
6	9.287209	193	1.50	0	0	0	1	0	0	0	0	
7	7.244228	212	1.80	0	0	0	0	0	1	0	1	
8	9.388487	177	1.50	0	0	0	1	0	0	0	0	
9	7.824046	260	1.79	0	0	0	1	0	0	0	0	

LINEAR REGRESSION MODEL

PREDICTING

USED CAR PRICES

Exploring which variables are significant in predicting the price of a car using multiple linear regression

CREATE

THE BASE MODEL

```
x =carsData_dum[['Mileage','EngineV', 'Brand_BMW',
       'Brand_Mercedes-Benz', 'Brand_Mitsubishi', 'Brand_Renault',
       'Brand_Toyota', 'Brand_Volkswagen', 'Body_hatch', 'Body_other',
       'Body_sedan', 'Body_vagon', 'Body_van', 'Engine Type_Gas',
       'Engine Type_Other', 'Engine Type_Petrol', 'Registration_yes']]
```

```
y = carsData_dum['TPrice']
```

Since our regression model can't take a full dataframe, assign each set of the variables to x and y

TRAIN

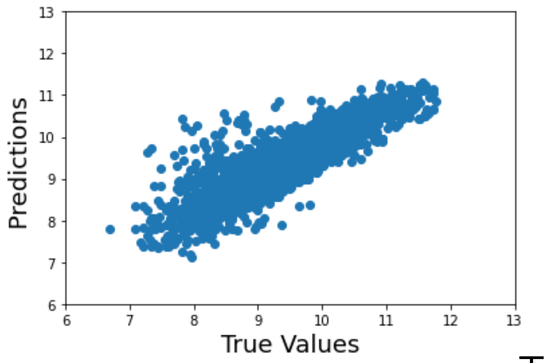
THE REGRESSION MODEL

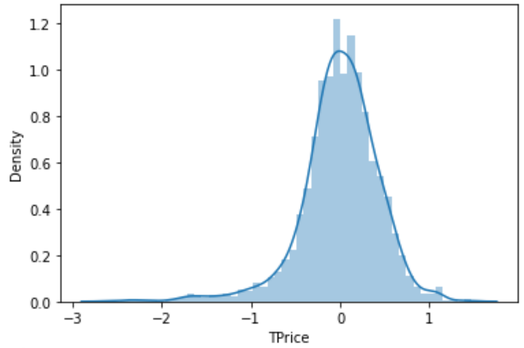
```
x_train, x_test, y_train, y_test = train_test_split(x,y, test_size = .4, random_state=101)
```

```
print(x_train.shape, y_train.shape)
print(x_test.shape, y_test.shape)
```

```
(2320, 17) (2320,)
(1548, 17) (1548,)
```

From the sklearn, package run the train and test method using the model created from the independent variables and dependent variable





TEST

FOR HOMOSCEDASTICITY

After the base model is created, we can calculate the residuals. This is the true values minus the predicted values the model found.

PLOT

THE RESIDUALS

Subtracting the predicted values from the true values produces a new item called a residual. Residuals can be graphed on a scatter plot or run with statistical tests.

FINAL ANALYSIS

GATHERING THE ANALYSIS RESULTS

INTERPRET THE REGRESSION SCORE, COEF AND INTERCEPT

75%

Our model is accurate 75% of the time, when using the predictor variable meets to predict the response variable.

```
lm.score(x_train, y_train)
```

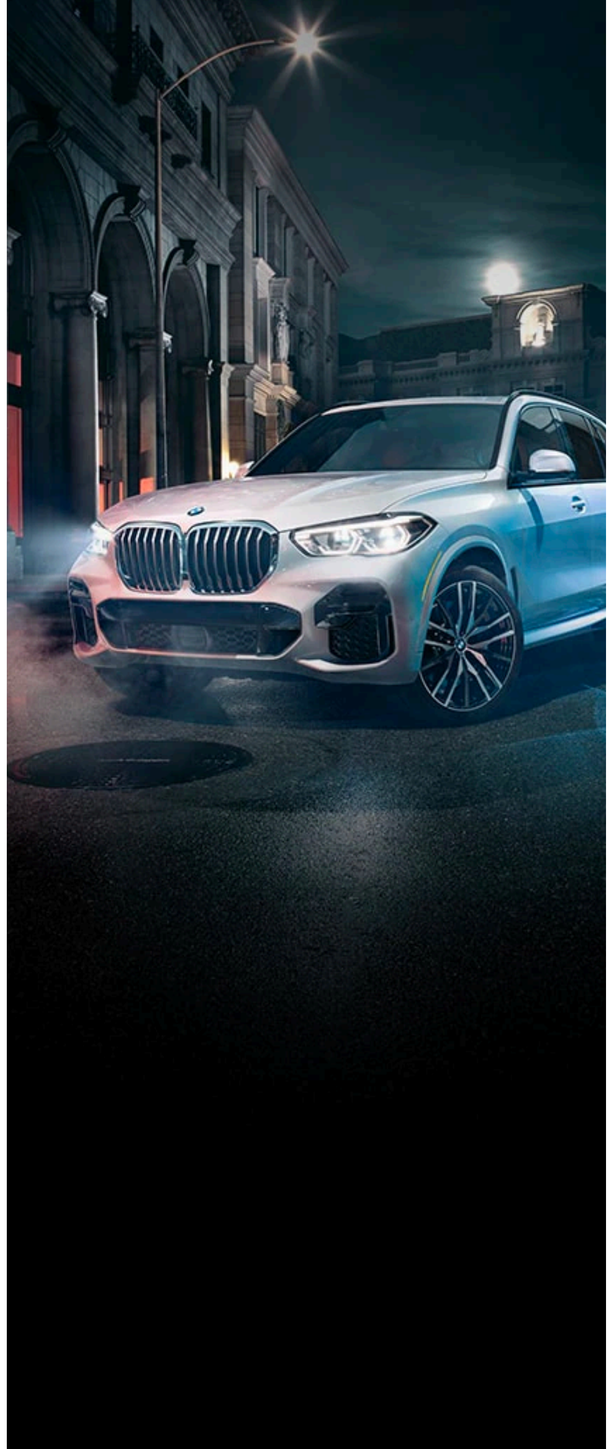
```
0.7545613519593195
```

9.3

Is the intercept for the model for when the predictor variables and response variable are equal to zero.

```
lm.intercept_
```

```
9.313134239130532
```



	Features	Coefs
0	Mileage	-0.004824
1	EngineV	0.229320
2	Brand_BMW	0.056329
3	Brand_Mercedes-Benz	0.062536
4	Brand_Mitsubishi	-0.510781
5	Brand_Renault	-0.532696
6	Brand_Toyota	-0.165598
7	Brand_Volkswagen	-0.182318
8	Body_hatch	-0.567574
9	Body_other	-0.375428
10	Body_sedan	-0.392562
11	Body_vagon	-0.453398
12	Body_van	-0.457476
13	Engine Type_Gas	-0.360217
14	Engine Type_Other	-0.182109
15	Engine Type_Petrol	-0.319495
16	Registration_yes	1.075893

ANALYSIS FINDINGS

FACTORS WITH GREATEST INFLEUNCE

CONCLUSIONS THAT CAN BE SUPPORTED BY THE MODEL

MILEAGE

The more mileage a car has, the less expensive

BRAND

BMW and Mercedes cars tend to be more expensive

BODY

Sedans tend to more expensive, over hatch's and vans.

ENGINEV

The bigger the engine volume, the higher the price.

BUSINESS GOAL

RECOMMENDATIONS

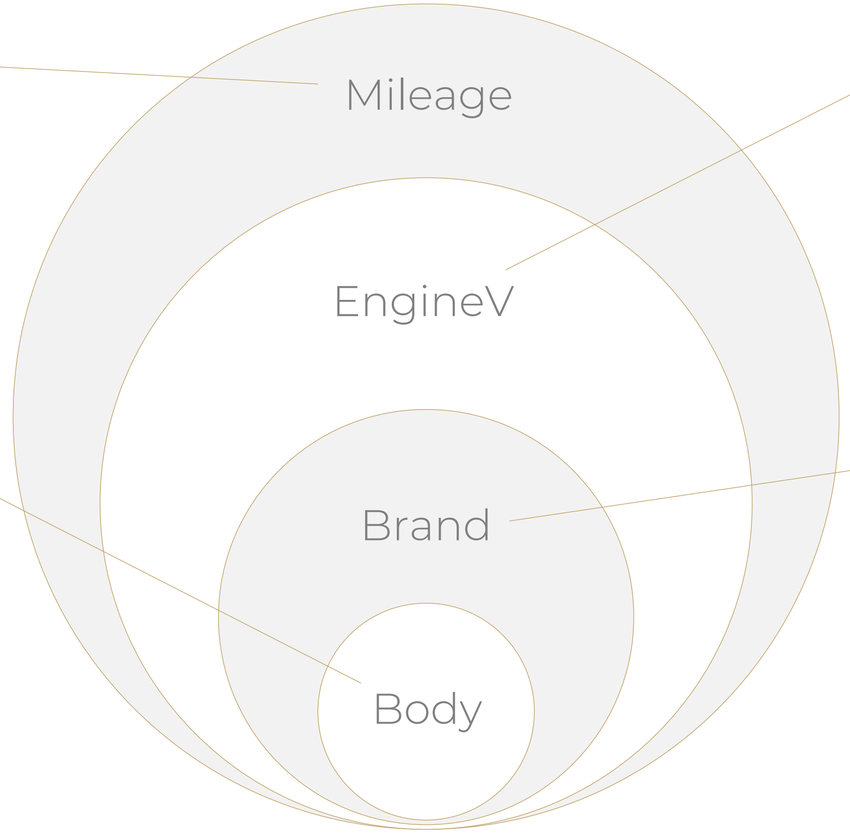
HELP DECISION MAKERS AND CAR MANUFACTUERS DETERMINE CAR PRICES FOR BUSINESS GROWTH AND MARKET MANAGEMENT

THE BIGGEST FACTOR

Buyers want the best bang for their buck. Lower mileage cars are less likely to break down than cars with higher mileage. Consumers will spend premium prices for low mileage cars.

IMPACTED BY GROWTH

Luxury Sedans are likely to sell better than any other body type, however it's important to consider current trends such as family size and how household numbers have increased since the pandemic. Families with new additions will soon be searching for bigger body types to accommodate.



GAS PRICES

The bigger the engine, the more expensive the car, however the more fuel it will consume. With gas prices also experiencing record inflation, consumers may be less interested in cars with large engine volume.

DESIRE FOR LUXURY

Study the local market you plan to sell in. Are you in an area that values luxury, brand and affluence over budgeting? What restaurants in your target area are most busy on a Friday night? What type of shopping bags do you see consumers carrying? Keep an eye out for signs of luxury spend to support your decision to sell top auto brands.

END

THANK YOU