INDIAN INSTITUTE OF TECHNOLOGY
JODHPUR

Data and Computational Science

Automobile Price Prediction

M22MA007
POTHULA AKASH

Problem Statement

"Perform data analysis and predictive modeling on automobile pricing data to understand the factors influencing the prices of automobiles. Explore the dataset to identify key features that affect car prices, build predictive models to estimate car prices based on these features, and evaluate the model's performance."

Colab_link:

https://colab.research.google.com/drive/1DIZcKFf_gj8J-VZz9Q8KCh5b2NKPmoz-?usp=sharing

Summary of the Project

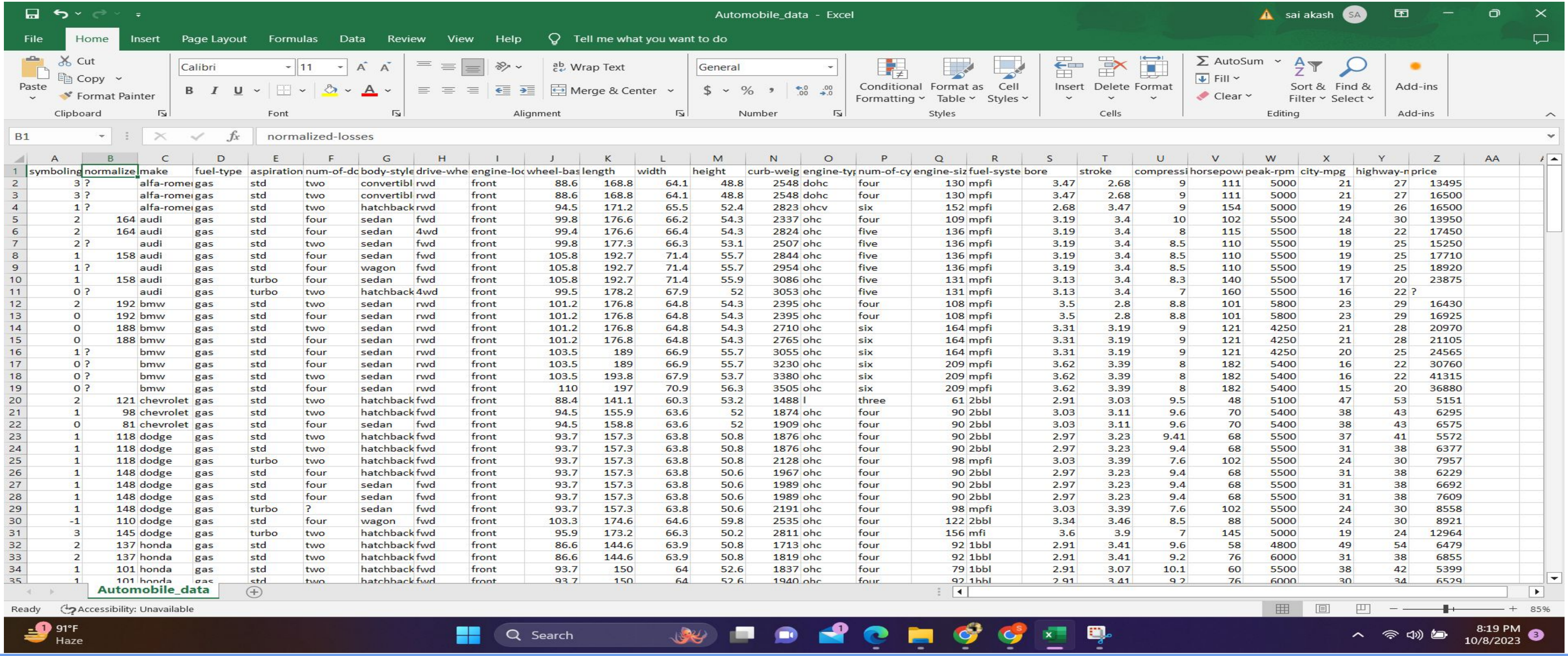
Data understanding and cleaning : examine the Data types of each column , check for the missing values, check the Summary statistics for numeric columns, deal with missing data, converting all the columns to similar datatype

Exploratory Data analysis: Drawing box plots ,scatter plots, find correlation,performed statistical tests like ANOVA, Correlation Analysis(finding pearson correlation analysis).

Machine Learning Models: Applied Linear Regression and Multiple Linear Regression.

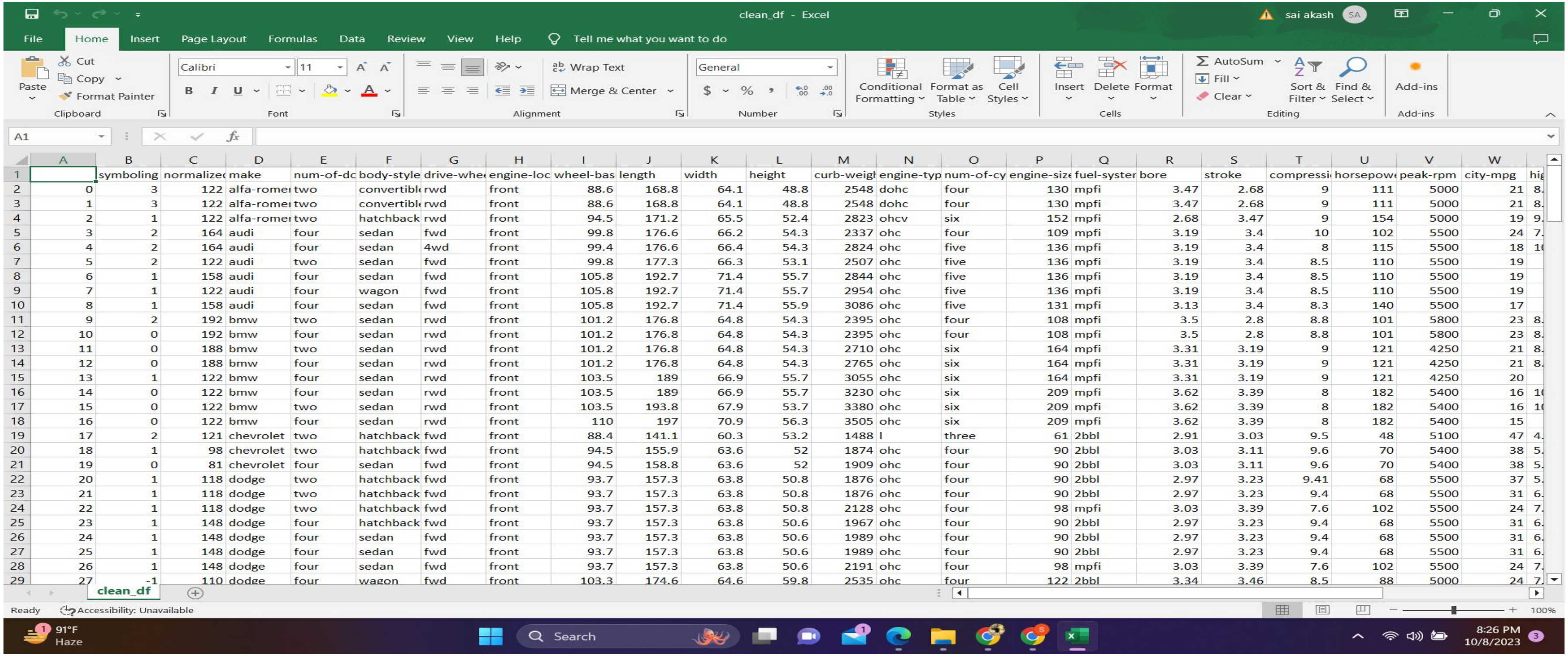
Evaluation: Mean Squared Error , R-squared error, F-test, T-test

Data before cleaning



	symboling	normalize	make	fuel-type	aspiration	num-of-dc	body-style	drive-whe	engine-loc	wheel-bas	length	width	height	curb-weig	engine-typ	num-of-cy	engine-siz	fuel-syste	bore	stroke	compressi	horsepow	peak-rpm	city-mpg	highway-n	price
2	3	?	alfa-romei	gas	std	two	convertibl	rwd	front	88.6	168.8	64.1	48.8	2548	dohc	four	130	mpfi	3.47	2.68	9	111	5000	21	27	13495
3	3	?	alfa-romei	gas	std	two	convertibl	rwd	front	88.6	168.8	64.1	48.8	2548	dohc	four	130	mpfi	3.47	2.68	9	111	5000	21	27	16500
4	1	?	alfa-romei	gas	std	two	hatchback	rwd	front	94.5	171.2	65.5	52.4	2823	ohcv	six	152	mpfi	2.68	3.47	9	154	5000	19	26	16500
5	2	164	audi	gas	std	four	sedan	fwd	front	99.8	176.6	66.2	54.3	2337	ohc	four	109	mpfi	3.19	3.4	10	102	5500	24	30	13950
6	2	164	audi	gas	std	four	sedan	4wd	front	99.4	176.6	66.4	54.3	2824	ohc	five	136	mpfi	3.19	3.4	8	115	5500	18	22	17450
7	2	?	audi	gas	std	two	sedan	fwd	front	99.8	177.3	66.3	53.1	2507	ohc	five	136	mpfi	3.19	3.4	8.5	110	5500	19	25	15250
8	1	158	audi	gas	std	four	sedan	fwd	front	105.8	192.7	71.4	55.7	2844	ohc	five	136	mpfi	3.19	3.4	8.5	110	5500	19	25	17710
9	1	?	audi	gas	std	four	wagon	fwd	front	105.8	192.7	71.4	55.7	2954	ohc	five	136	mpfi	3.19	3.4	8.5	110	5500	19	25	18920
10	1	158	audi	gas	turbo	four	sedan	fwd	front	105.8	192.7	71.4	55.9	3086	ohc	five	131	mpfi	3.13	3.4	8.3	140	5500	17	20	23875
11	0	?	audi	gas	turbo	two	hatchback	4wd	front	99.5	178.2	67.9	52	3053	ohc	five	131	mpfi	3.13	3.4	7	160	5500	16	22	?
12	2	192	bmw	gas	std	two	sedan	rwd	front	101.2	176.8	64.8	54.3	2395	ohc	four	108	mpfi	3.5	2.8	8.8	101	5800	23	29	16430
13	0	192	bmw	gas	std	four	sedan	rwd	front	101.2	176.8	64.8	54.3	2395	ohc	four	108	mpfi	3.5	2.8	8.8	101	5800	23	29	16925
14	0	188	bmw	gas	std	two	sedan	rwd	front	101.2	176.8	64.8	54.3	2710	ohc	six	164	mpfi	3.31	3.19	9	121	4250	21	28	20970
15	0	188	bmw	gas	std	four	sedan	rwd	front	101.2	176.8	64.8	54.3	2765	ohc	six	164	mpfi	3.31	3.19	9	121	4250	21	28	21105
16	1	?	bmw	gas	std	four	sedan	rwd	front	103.5	189	66.9	55.7	3055	ohc	six	164	mpfi	3.31	3.19	9	121	4250	20	25	24565
17	0	?	bmw	gas	std	four	sedan	rwd	front	103.5	189	66.9	55.7	3230	ohc	six	209	mpfi	3.62	3.39	8	182	5400	16	22	30760
18	0	?	bmw	gas	std	two	sedan	rwd	front	103.5	193.8	67.9	53.7	3380	ohc	six	209	mpfi	3.62	3.39	8	182	5400	16	22	41315
19	0	?	bmw	gas	std	four	sedan	rwd	front	110	197	70.9	56.3	3505	ohc	six	209	mpfi	3.62	3.39	8	182	5400	15	20	36880
20	2	121	chevrolet	gas	std	two	hatchback	fwd	front	88.4	141.1	60.3	53.2	1488	l	three	61	2bbl	2.91	3.03	9.5	48	5100	47	53	5151
21	1	98	chevrolet	gas	std	two	hatchback	fwd	front	94.5	155.9	63.6	52	1874	ohc	four	90	2bbl	3.03	3.11	9.6	70	5400	38	43	6295
22	0	81	chevrolet	gas	std	four	sedan	fwd	front	94.5	158.8	63.6	52	1909	ohc	four	90	2bbl	3.03	3.11	9.6	70	5400	38	43	6575
23	1	118	dodge	gas	std	two	hatchback	fwd	front	93.7	157.3	63.8	50.8	1876	ohc	four	90	2bbl	2.97	3.23	9.41	68	5500	37	41	5572
24	1	118	dodge	gas	std	two	hatchback	fwd	front	93.7	157.3	63.8	50.8	1876	ohc	four	90	2bbl	2.97	3.23	9.4	68	5500	31	38	6377
25	1	118	dodge	gas	turbo	two	hatchback	fwd	front	93.7	157.3	63.8	50.8	2128	ohc	four	98	mpfi	3.03	3.39	7.6	102	5500	24	30	7957
26	1	148	dodge	gas	std	four	hatchback	fwd	front	93.7	157.3	63.8	50.6	1967	ohc	four	90	2bbl	2.97	3.23	9.4	68	5500	31	38	6229
27	1	148	dodge	gas	std	four	sedan	fwd	front	93.7	157.3	63.8	50.6	1989	ohc	four	90	2bbl	2.97	3.23	9.4	68	5500	31	38	6692
28	1	148	dodge	gas	std	four	sedan	fwd	front	93.7	157.3	63.8	50.6	1989	ohc	four	90	2bbl	2.97	3.23	9.4	68	5500	31	38	7609
29	1	148	dodge	gas	turbo	?	sedan	fwd	front	93.7	157.3	63.8	50.6	2191	ohc	four	98	mpfi	3.03	3.39	7.6	102	5500	24	30	8558
30	-1	110	dodge	gas	std	four	wagon	fwd	front	103.3	174.6	64.6	59.8	2535	ohc	four	122	2bbl	3.34	3.46	8.5	88	5000	24	30	8921
31	3	145	dodge	gas	turbo	two	hatchback	fwd	front	95.9	173.2	66.3	50.2	2811	ohc	four	156	mfi	3.6	3.9	7	145	5000	19	24	12964
32	2	137	honda	gas	std	two	hatchback	fwd	front	86.6	144.6	63.9	50.8	1713	ohc	four	92	1bbl	2.91	3.41	9.6	58	4800	49	54	6479
33	2	137	honda	gas	std	two	hatchback	fwd	front	86.6	144.6	63.9	50.8	1819	ohc	four	92	1bbl	2.91	3.41	9.2	76	6000	31	38	6855
34	1	101	honda	gas	std	two	hatchback	fwd	front	93.7	150	64	52.6	1837	ohc	four	79	1bbl	2.91	3.07	10.1	60	5500	38	42	5399
35	1	101	honda	gas	std	two	hatchback	fwd	front	93.7	150	64	52.6	1940	ohc	four	92	1bbl	2.91	3.41	9.2	76	6000	30	34	6529

Data after cleaning



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	
1		symboling	normalizec	make	num-of-dc	body-style	drive-whe	engine-loc	wheel-bas	length	width	height	curb-weigl	engine-typ	num-of-cy	engine-size	fuel-syster	bore	stroke	compressi	horsepowe	peak-rpm	city-mpg	hig
2	0	3	122	alfa-romei	two	convertibl	rwd	front	88.6	168.8	64.1	48.8	2548	dohc	four	130	mpfi	3.47	2.68	9	111	5000	21	8.
3	1	3	122	alfa-romei	two	convertibl	rwd	front	88.6	168.8	64.1	48.8	2548	dohc	four	130	mpfi	3.47	2.68	9	111	5000	21	8.
4	2	1	122	alfa-romei	two	hatchback	rwd	front	94.5	171.2	65.5	52.4	2823	ohcv	six	152	mpfi	2.68	3.47	9	154	5000	19	9.
5	3	2	164	audi	four	sedan	fwd	front	99.8	176.6	66.2	54.3	2337	ohc	four	109	mpfi	3.19	3.4	10	102	5500	24	7.
6	4	2	164	audi	four	sedan	4wd	front	99.4	176.6	66.4	54.3	2824	ohc	five	136	mpfi	3.19	3.4	8	115	5500	18	1(
7	5	2	122	audi	two	sedan	fwd	front	99.8	177.3	66.3	53.1	2507	ohc	five	136	mpfi	3.19	3.4	8.5	110	5500	19	
8	6	1	158	audi	four	sedan	fwd	front	105.8	192.7	71.4	55.7	2844	ohc	five	136	mpfi	3.19	3.4	8.5	110	5500	19	
9	7	1	122	audi	four	wagon	fwd	front	105.8	192.7	71.4	55.7	2954	ohc	five	136	mpfi	3.19	3.4	8.5	110	5500	19	
10	8	1	158	audi	four	sedan	fwd	front	105.8	192.7	71.4	55.9	3086	ohc	five	131	mpfi	3.13	3.4	8.3	140	5500	17	
11	9	2	192	bmw	two	sedan	rwd	front	101.2	176.8	64.8	54.3	2395	ohc	four	108	mpfi	3.5	2.8	8.8	101	5800	23	8.
12	10	0	192	bmw	four	sedan	rwd	front	101.2	176.8	64.8	54.3	2395	ohc	four	108	mpfi	3.5	2.8	8.8	101	5800	23	8.
13	11	0	188	bmw	two	sedan	rwd	front	101.2	176.8	64.8	54.3	2710	ohc	six	164	mpfi	3.31	3.19	9	121	4250	21	8.
14	12	0	188	bmw	four	sedan	rwd	front	101.2	176.8	64.8	54.3	2765	ohc	six	164	mpfi	3.31	3.19	9	121	4250	21	8.
15	13	1	122	bmw	four	sedan	rwd	front	103.5	189	66.9	55.7	3055	ohc	six	164	mpfi	3.31	3.19	9	121	4250	20	
16	14	0	122	bmw	four	sedan	rwd	front	103.5	189	66.9	55.7	3230	ohc	six	209	mpfi	3.62	3.39	8	182	5400	16	1(
17	15	0	122	bmw	two	sedan	rwd	front	103.5	193.8	67.9	53.7	3380	ohc	six	209	mpfi	3.62	3.39	8	182	5400	16	1(
18	16	0	122	bmw	four	sedan	rwd	front	110	197	70.9	56.3	3505	ohc	six	209	mpfi	3.62	3.39	8	182	5400	15	
19	17	2	121	chevrolet	two	hatchback	fwd	front	88.4	141.1	60.3	53.2	1488	l	three	61	2bbl	2.91	3.03	9.5	48	5100	47	4.
20	18	1	98	chevrolet	two	hatchback	fwd	front	94.5	155.9	63.6	52	1874	ohc	four	90	2bbl	3.03	3.11	9.6	70	5400	38	5.
21	19	0	81	chevrolet	four	sedan	fwd	front	94.5	158.8	63.6	52	1909	ohc	four	90	2bbl	3.03	3.11	9.6	70	5400	38	5.
22	20	1	118	dodge	two	hatchback	fwd	front	93.7	157.3	63.8	50.8	1876	ohc	four	90	2bbl	2.97	3.23	9.41	68	5500	37	5.
23	21	1	118	dodge	two	hatchback	fwd	front	93.7	157.3	63.8	50.8	1876	ohc	four	90	2bbl	2.97	3.23	9.4	68	5500	31	6.
24	22	1	118	dodge	two	hatchback	fwd	front	93.7	157.3	63.8	50.8	2128	ohc	four	98	mpfi	3.03	3.39	7.6	102	5500	24	7.
25	23	1	148	dodge	four	hatchback	fwd	front	93.7	157.3	63.8	50.6	1967	ohc	four	90	2bbl	2.97	3.23	9.4	68	5500	31	6.
26	24	1	148	dodge	four	sedan	fwd	front	93.7	157.3	63.8	50.6	1989	ohc	four	90	2bbl	2.97	3.23	9.4	68	5500	31	6.
27	25	1	148	dodge	four	sedan	fwd	front	93.7	157.3	63.8	50.6	1989	ohc	four	90	2bbl	2.97	3.23	9.4	68	5500	31	6.
28	26	1	148	dodge	four	sedan	fwd	front	93.7	157.3	63.8	50.6	2191	ohc	four	98	mpfi	3.03	3.39	7.6	102	5500	24	7.
29	27	-1	110	dodge	four	wagon	fwd	front	103.3	174.6	64.6	59.8	2535	ohc	four	122	2bbl	3.34	3.46	8.5	88	5000	24	7.

Dealt with missing values using the following methods

Replacing Missing Values with Averages: In several instances, we replaced missing values with the average (mean) value of the respective columns. This approach was used for columns such as "normalized-losses," "bore," "stroke," "horsepower," and "peak-rpm."

```
# Replace missing values in these columns with their respective averages
columns_to_impute = ["normalized-losses", "bore", "stroke", "horsepower", "peak-rpm"]

for column in columns_to_impute:
    # Calculate the average of the current column
    avg_value = df[column].astype("float").mean(axis=0)

    # Replace missing values in the current column with the calculated average
    df[column].replace(np.nan, avg_value, inplace=True)

# Display the calculated average values
for column in columns_to_impute:
    avg_value = df[column].astype("float").mean(axis=0)
    print(f"Average of {column}: {avg_value}")
```

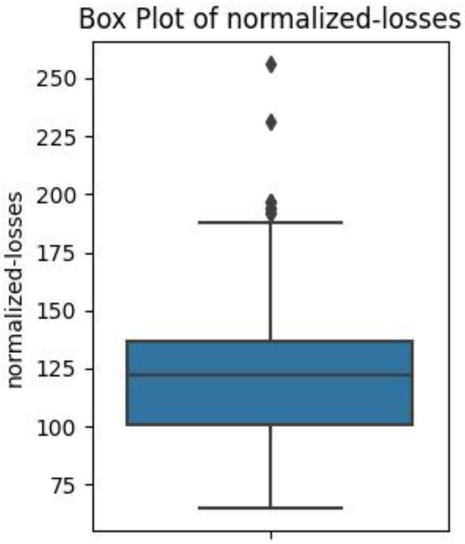
Filling Missing Values with Most Frequent Values: For the "num-of-doors" column, missing values were replaced with the most frequent value (mode) in the column.

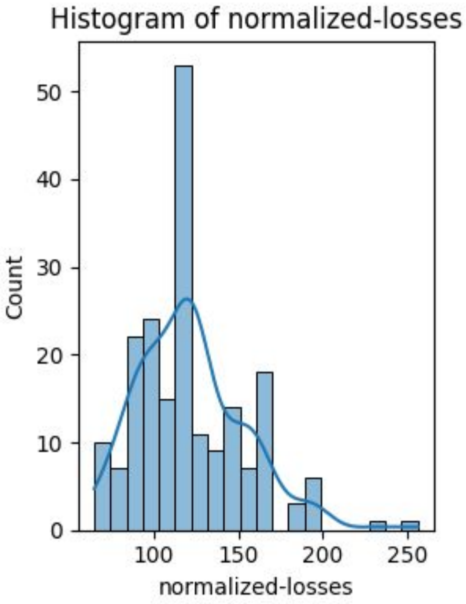
Resetting Index: After dropping rows with missing values, the index was reset using the reset_index method

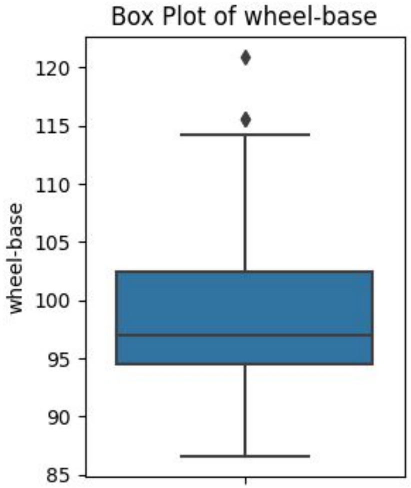
Dropping Rows with Missing Values: In the "price" column, rows with missing values were simply dropped using the dropna

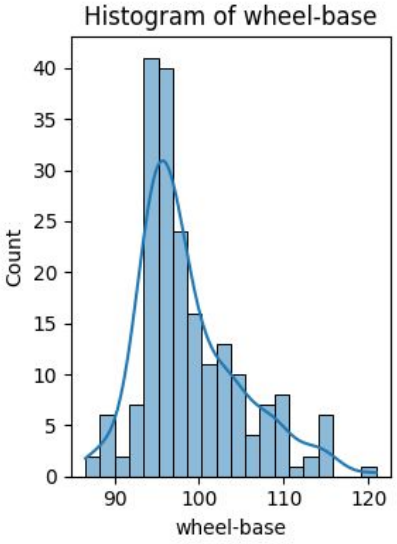
```
[11] # Replace missing values in the 'num-of-doors' column with the most frequent value ('four')
     df['num-of-doors'].replace(np.nan, "four", inplace=True)

     # Drop rows with missing values in the 'price' column
     df.dropna(subset=["price"], axis=0, inplace=True)

     # Reset the DataFrame index after dropping rows
     df.reset_index(drop=True, inplace=True)
```

Drawing boxplots and histograms









Pearson correlation coefficient

The Pearson Correlation Coefficient between wheel-base and price is 0.5846418222655081 with a P-value of 8.076488270732885e-20

Correlation Strength: The positive value of 0.585 suggests that as the "wheel-base" increases, the "price" of the automobile tends to increase as well. However, the strength of this relationship is moderate, not extremely strong.

Significance: The p-value associated with the correlation coefficient is very close to zero (8.076e-20), indicating that the observed correlation is statistically significant.

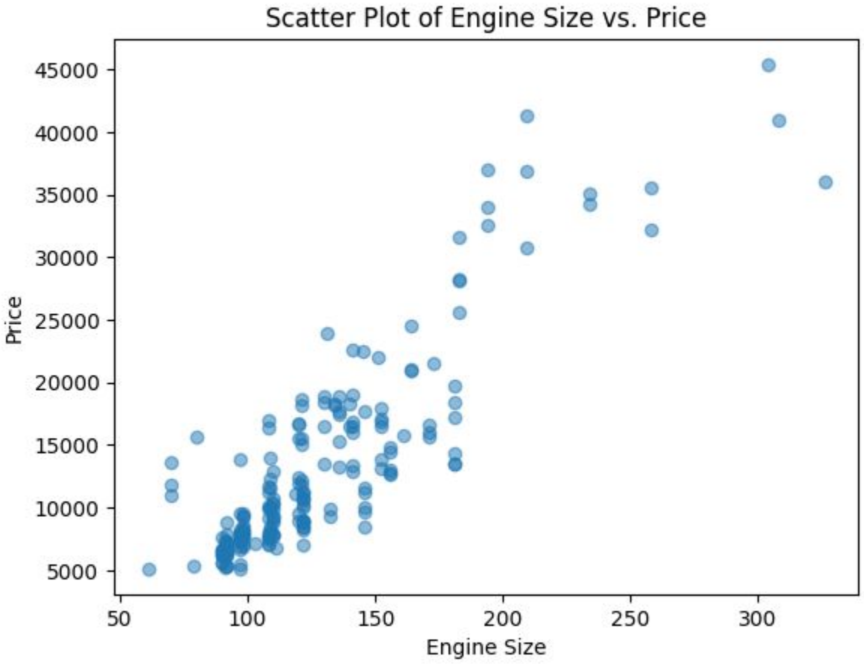
```
from scipy import stats

numeric_columns = ['wheel-base', 'horsepower', 'length', 'width', 'curb-weight', 'engine-size', 'bore', 'city-mpg', 'highway-mpg']

for column in numeric_columns:
    pearson_coef, p_value = stats.pearsonr(df[column], df['price'])
    print(f"The Pearson Correlation Coefficient between {column} and price is {pearson_coef} with a P-value of {p_value}")
```

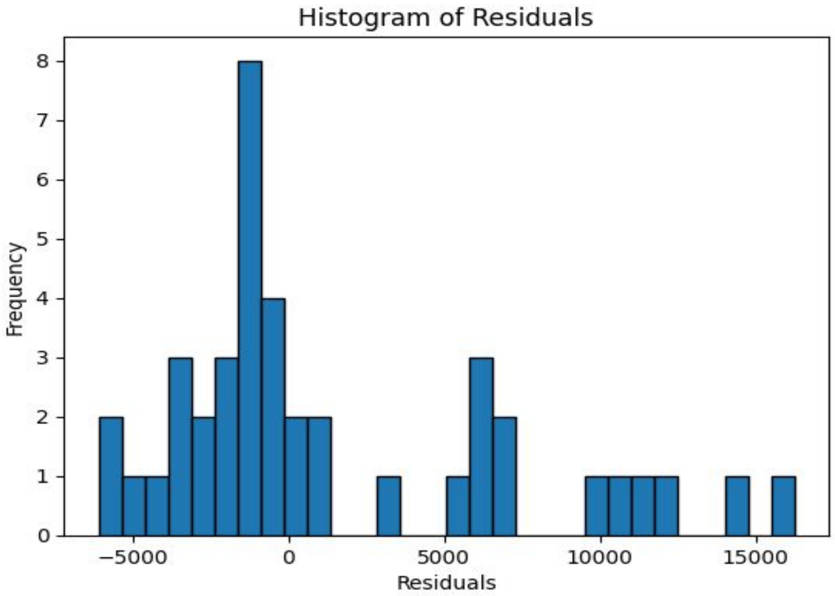
Calculating ANOVA

ANOVA results for rwd: F=130.5533160959111, P=2.2355306355677845e-23

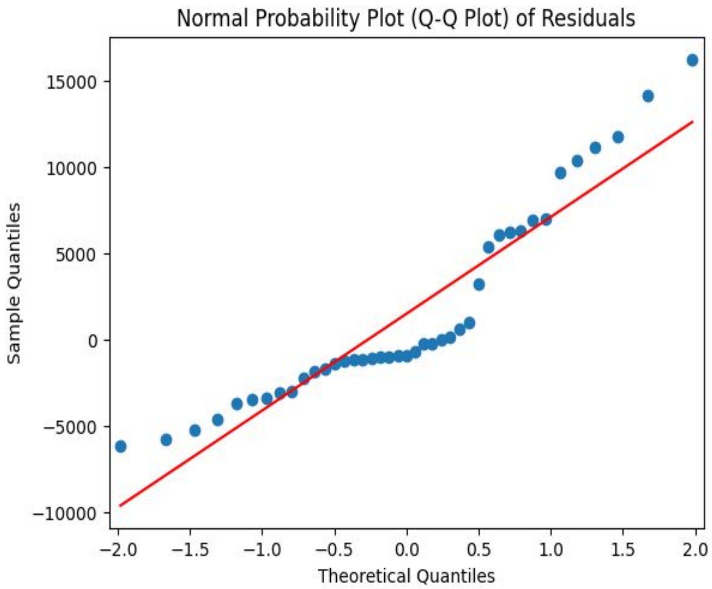
```
# List of unique drive-wheels values
drive_wheels = df['drive-wheels'].unique()

# Initialize empty lists to store results
f_values = []
p_values = []

# Calculate ANOVA for each drive-wheels group
for drive_wheel in drive_wheels:
    group = grouped_test2.get_group(drive_wheel)['price']
    f_val, p_val = stats.f_oneway(grouped_test2.get_group('fwd')['price'], group)
    f_values.append(f_val)
    p_values.append(p_val)

# Print ANOVA results
for i, drive_wheel in enumerate(drive_wheels):
    print(f"ANOVA results for {drive_wheel}: F={f_values[i]}, P={p_values[i]}")
```

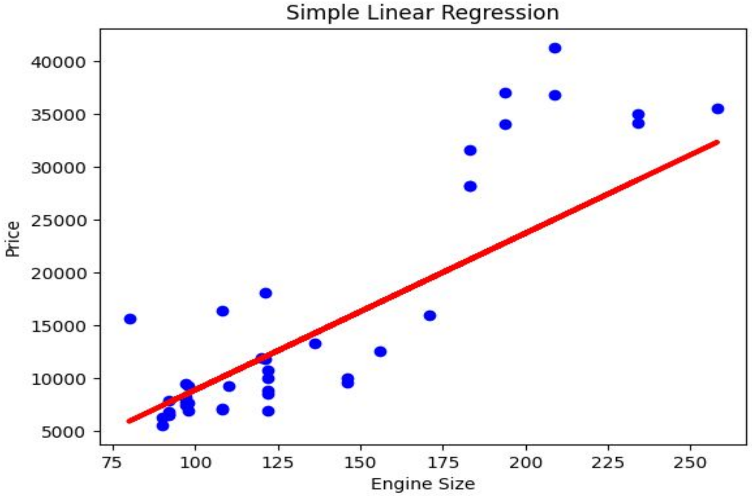
Scatter plot and histogram of residuals



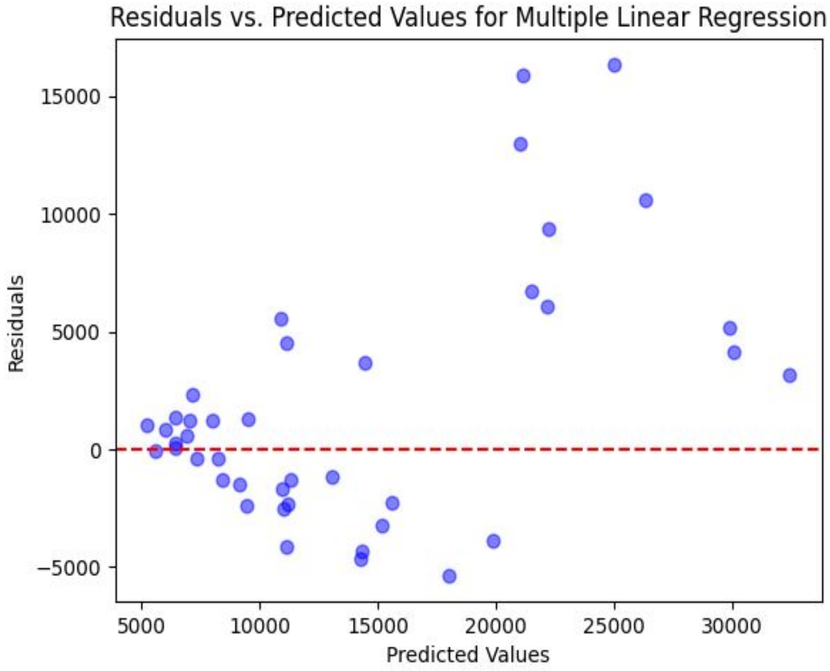


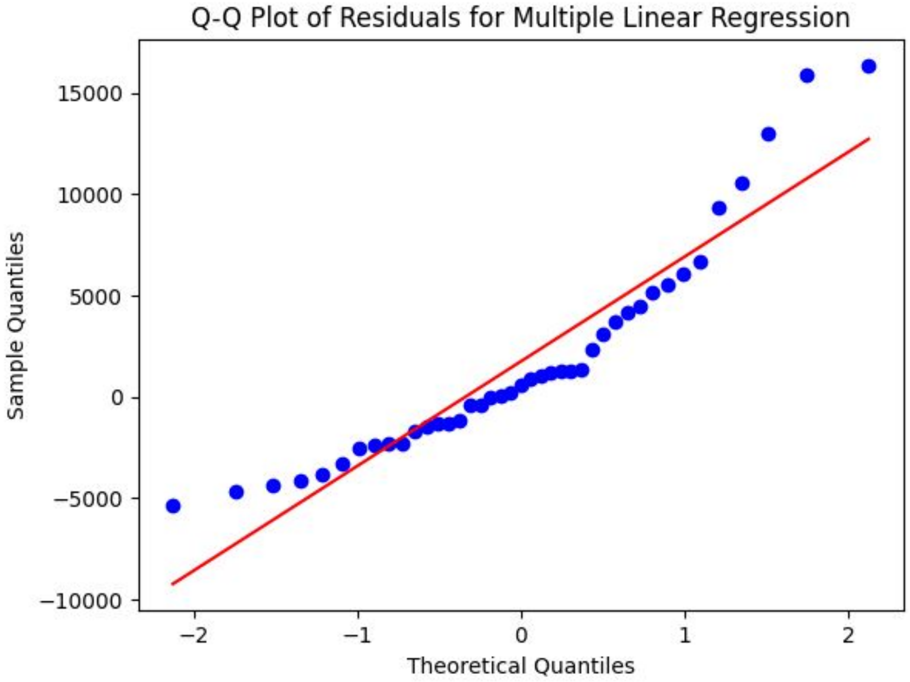
Simple Linear Regression and Q-Q plot





Multiple Linear Regression





Evaluation of multiple linear regression and Linear Regression

Mean Squared Error (Multiple Linear Regression): 30393323.64

R-squared (Multiple Linear Regression): 0.75

Mean Squared Error(Linear Regression): 33696986.98421676

R-squared(Multiple Linear Regression): 0.724578048646674

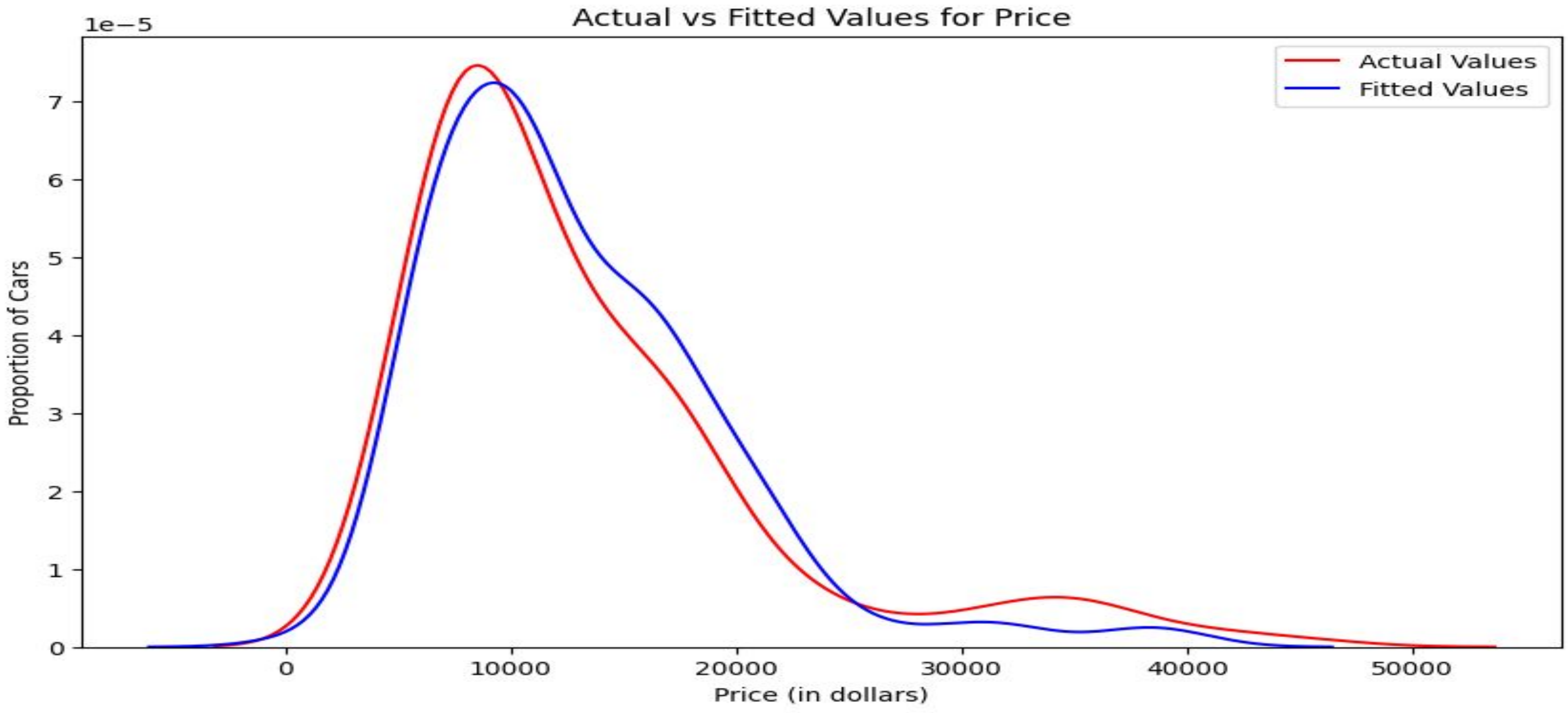
Overall Model Significance (F-test):

```
Feature: engine-size
F-statistic: 633.5267598010946
P-value: 9.265491622197996e-64
This feature is statistically significant.
==============================
Feature: horsepower
F-statistic: 378.5870228443837
P-value: 6.273536270652618e-48
This feature is statistically significant.
==============================
Feature: curb-weight
F-statistic: 456.138858276953
P-value: 2.189577238897131e-53
This feature is statistically significant.
==============================
Feature: highway-mpg
F-statistic: 356.53919541614164
P-value: 3.0467845810501095e-46
This feature is statistically significant.
==============================
```

Individual Coefficient Significance (t-test):

```
Feature: const
T-statistic: -10.25497271186749
P-value: 4.9895128140540304e-20
-------------------------------
Feature: engine-size
T-statistic: 6.485355428291303
P-value: 7.044275273003609e-10
-------------------------------
Feature: horsepower
T-statistic: 2.6067283758896402
P-value: 0.0098438986362382
-------------------------------
Feature: curb-weight
T-statistic: 3.28273539858447
P-value: 0.0012172841623101253
-------------------------------
Feature: highway-mpg
T-statistic: 1.6674304790368502
P-value: 0.09702580553167357
-------------------------------
```

Actual vs Fitted Values for price



Thank you