SMC 516 FINAL PROJECT

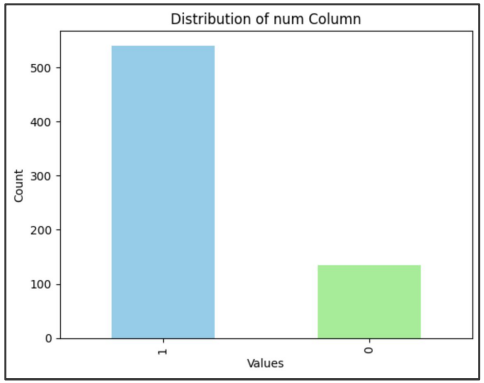
Presented by: Ritika Anand, Ayush Kaurav,
Daniel Farr and Karteek Attaluri

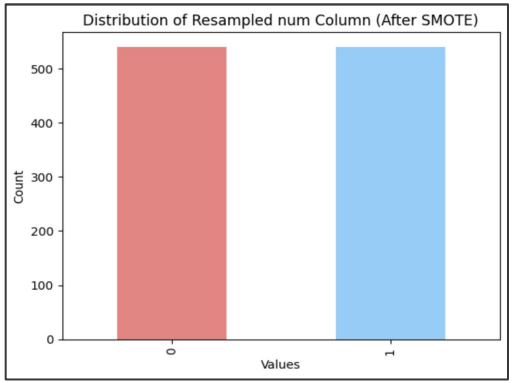
HEART DISEASE DATASET

Use predictive modeling to understand which factors contribute most to instances of heart disease?

BALANCING THE DATA SET

- Converted instances of heart disease (previously on scale of 0 to 4 instances) to 0 and 1





- Balanced data set using SMOTE, since for this instance, it was the best for creating synthetic samples to make up the difference.
- 540 instances of 1 (heart disease), 135 instances for 0 (no heart disease)
    - Resampled data gave 540 instances for both

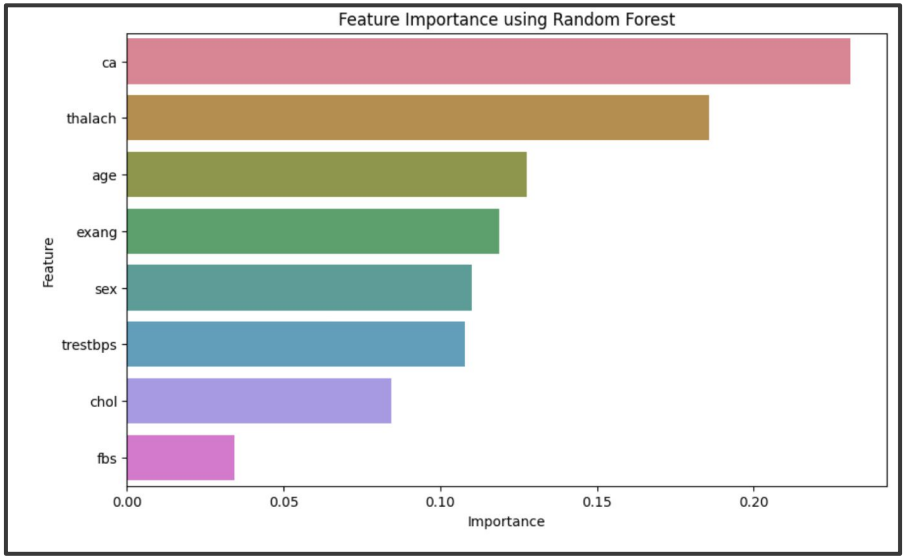
CONDITIONAL PROBABILITIES

- Looked into 3 key features out of the 9 features
  - Sex, resting blood pressure, fasting blood sugar
- Sex = male (marked by value of 1)
  - **P(heart disease = 1 | sex = 1) = 84.69%**
- Sex = female (marked by value of 0)
  - **P(heart disease = 1 | sex = 0) = 43.11%**
- Resting Blood Pressure
  - **P(heart disease = 1 | resting blood pressure > 130) = 43.12%**
- Fasting Blood Sugar
  - **P(heart disease = 1 | fast blood sugar > 120 mg/dl) = 0.0**

FEATURE IMPORTANCE: NB

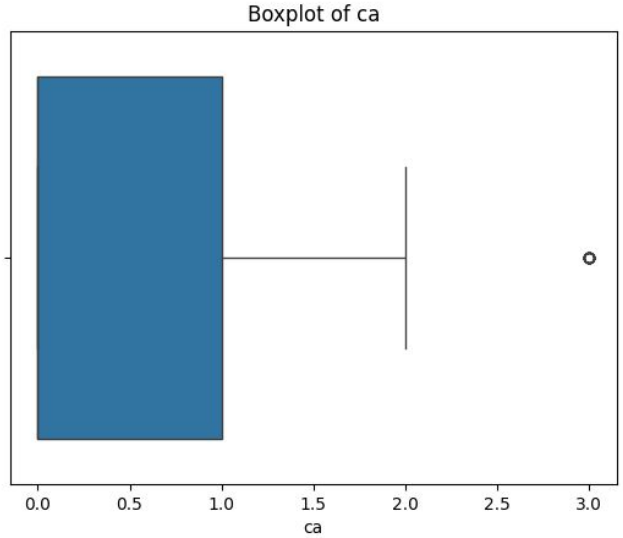
- Decided to look at 2 features:
  - Blood Pressure, threshold of 130 mmHg
  - Cholesterol
- Blood Pressure:
  - `P(heart disease = 1 | trestbps > 130) = 0.859`
  - `P(heart disease = 1 | trestbps <= 130) = 0.713`
  - `0.859-0.713 = 0.146 or 14.6%`
- Cholesterol:
  - `P(heart disease = 1 | chol > 200) = 0.809`
  - `P(heart disease = 1 | chol <= 200) = 0.744`
  - `0.809-0.744 = 0.065 or 6.5%`
- Feature more important when comparing both?
  - Blood Pressure

FEATURE IMPORTANCE: RANDOM FOREST

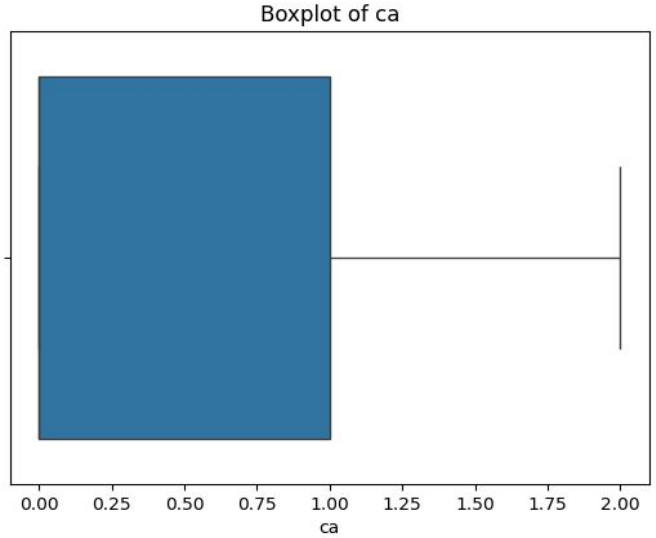


- Features (top to bottom): major blood vessels colored by fluoroscopy, max heart rate, age, exercise induced angina, sex, resting blood pressure, cholesterol and fasting blood sugar

Removing outliers



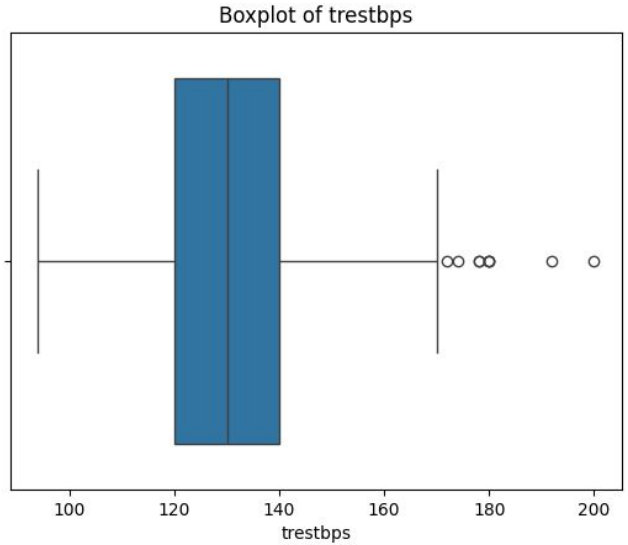
Before



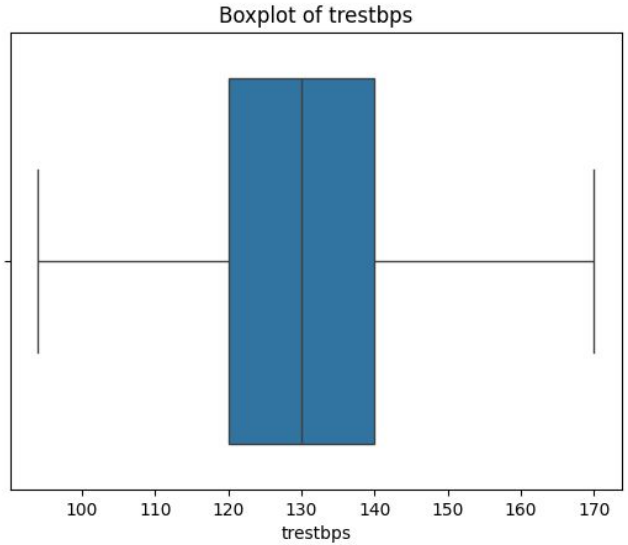
After

REMOVING OUTLIERS: CONTINUED

Before

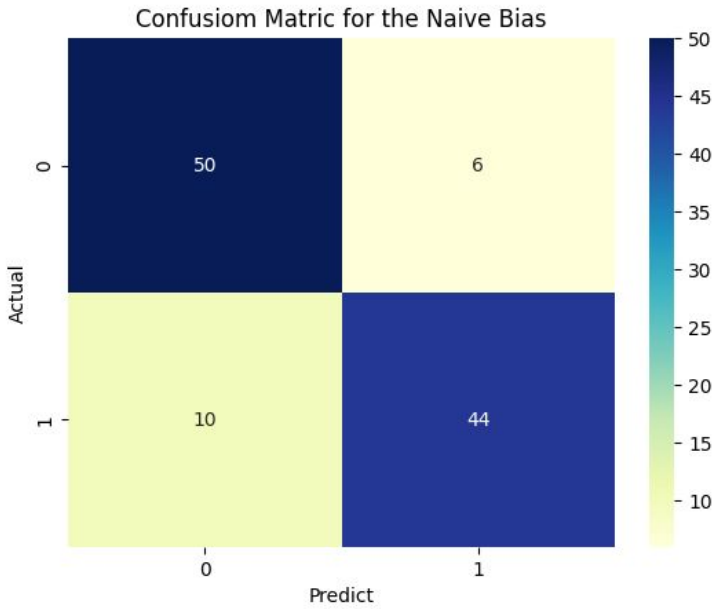


After



CONFUSION MATRIX PLOT

For Naive Bayes



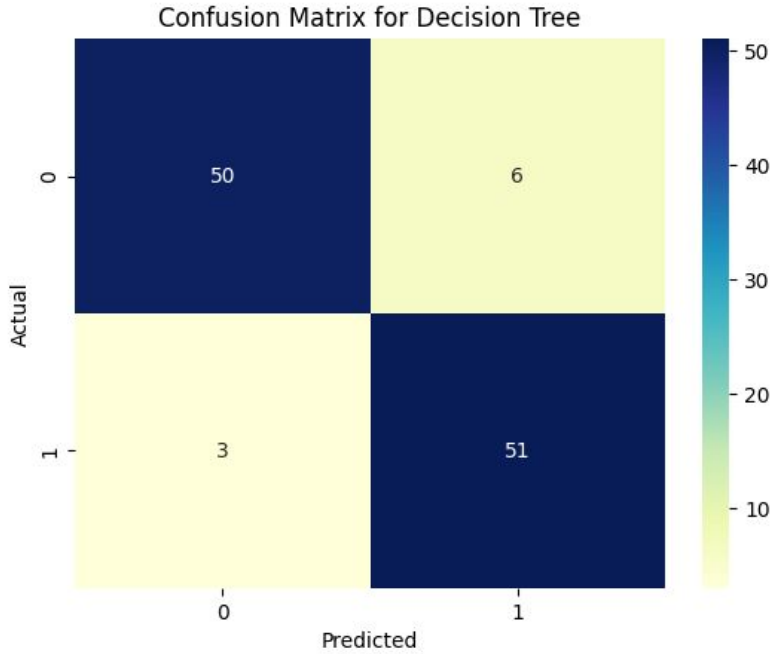

```
Model: Naive Bayes
Confusion Matrix:
[52  6]
 [ 3 45]]

Accuracy: 0.90
Precision: 0.90
PRecall: 0.90
F1 Score: 0.90
```

Code cell output actions

CONFUSION MATRIX PLOT

For Decision Tree

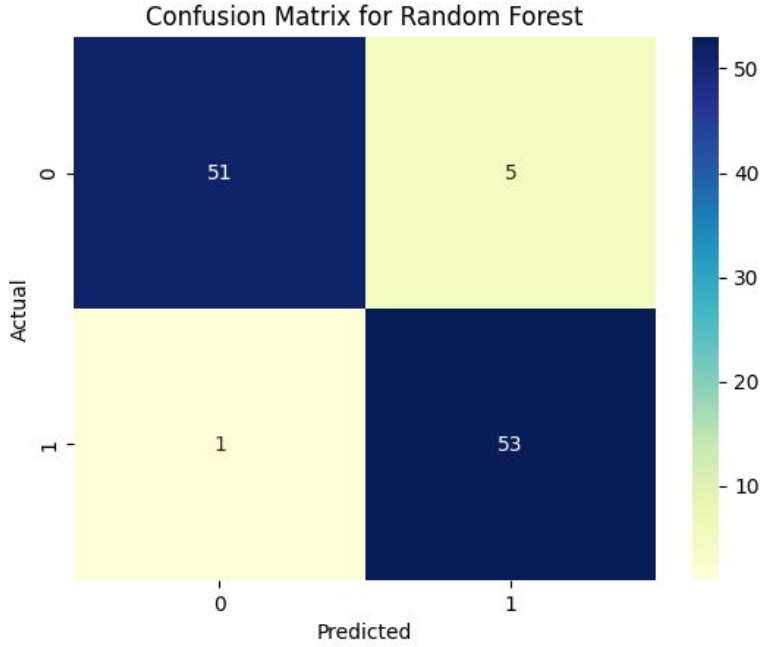



```
Model: Decision Tree
Confusion Matrix:
[[57  3]
 [ 4 44]]

Accuracy: 0.94
Precision: 0.94
PRecall: 0.94
F1 Score: 0.94
```

CONFUSION MATRIX PLOT

For Random Forest

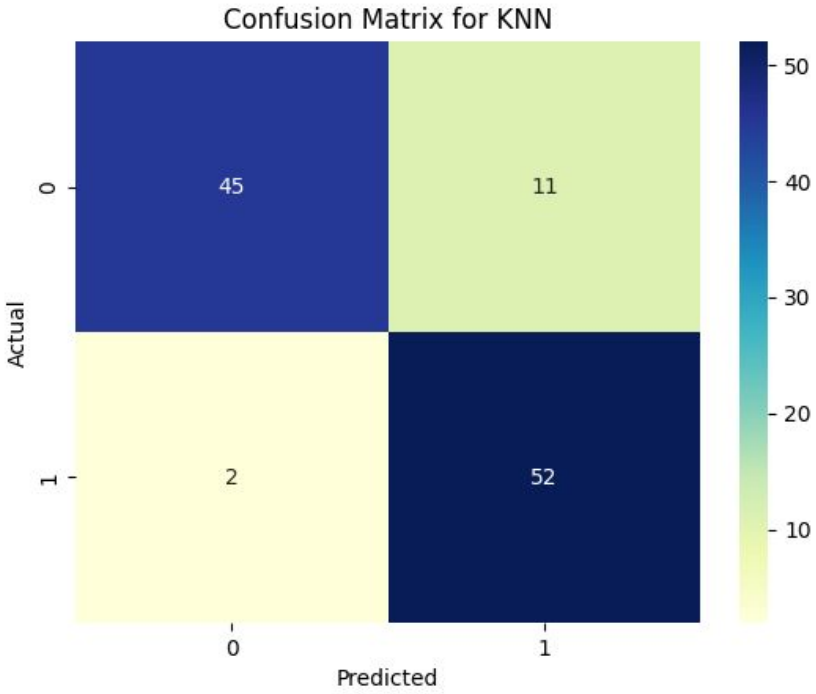



```
Model: Random Forest
Confusion Matrix:
[[58  2]
 [ 0 48]]

Accuracy: 0.98
Precision: 0.98
PRecall: 0.98
F1 Score: 0.98
```

CONFUSION MATRIX PLOT

For KNN




```
Model: KNN
Confusion Matrix:
[[49 11]
 [ 0 48]]

Accuracy: 0.90
Precision: 0.92
PRecall: 0.90
F1 Score: 0.90
```

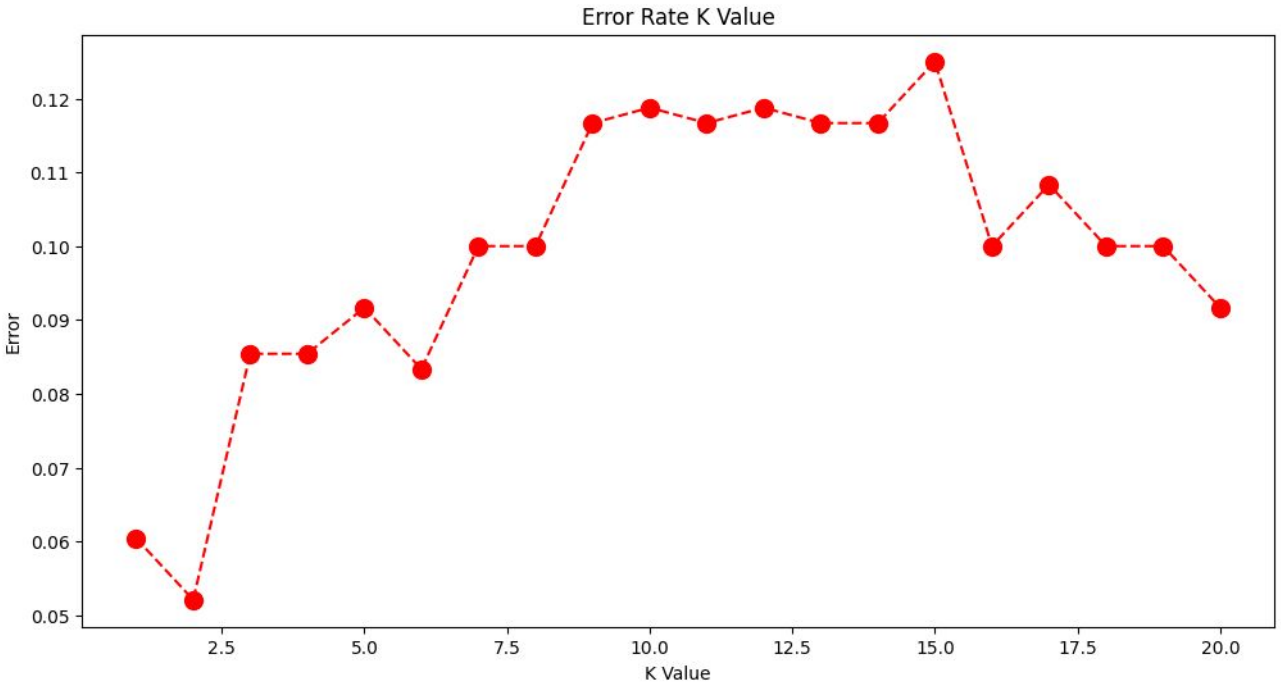
ANALYSIS

FOR BEST PERFORMANCE BY THE MODELS PROVIDED IN PREVIOUS SLIDE

	precision	recall	f1-score	support
0	1.00	0.97	0.98	60
2	0.96	1.00	0.98	48
accuracy			0.98	108
macro avg	0.98	0.98	0.98	108
weighted avg	0.98	0.98	0.98	108

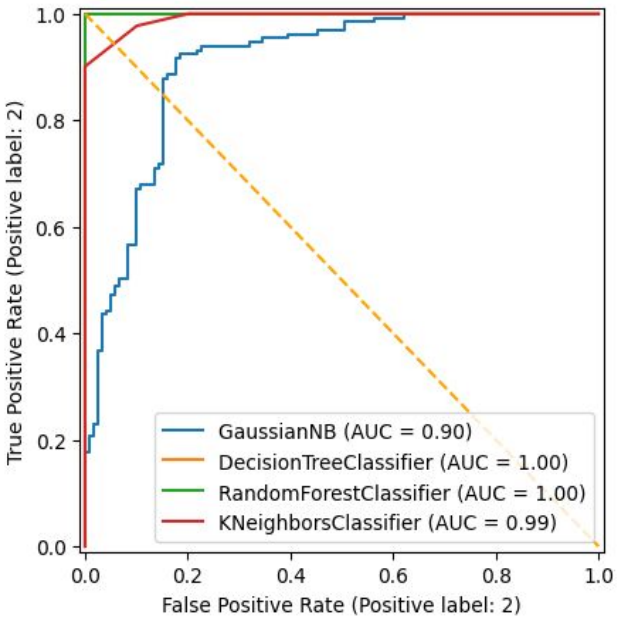
- The Random Forest model is performing very well, with a high overall accuracy of 98%.
- Both precision and recall are high for both classes, although Class 2 has a slightly lower precision, indicating more false positives.
- The F1-scores for both classes are above 0.9, suggesting a good balance between precision and recall.
- The dataset appears balanced, and the model isn't biased towards either class, as shown by the similar macro and weighted averages.

ERROR RATE IN K-VALUE

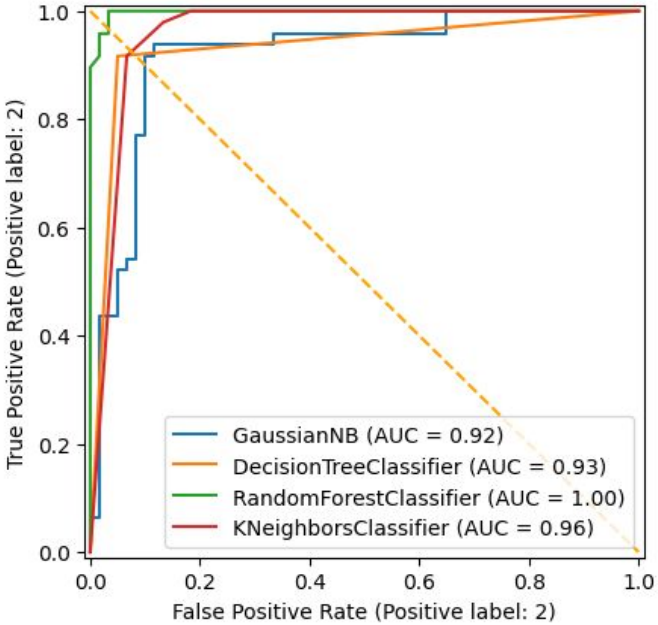


ROC Curve

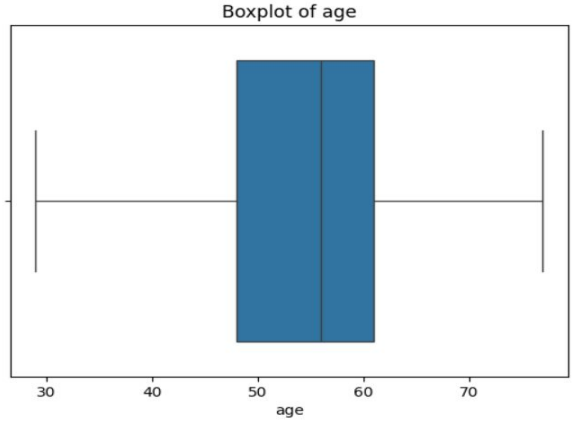
For Training Data

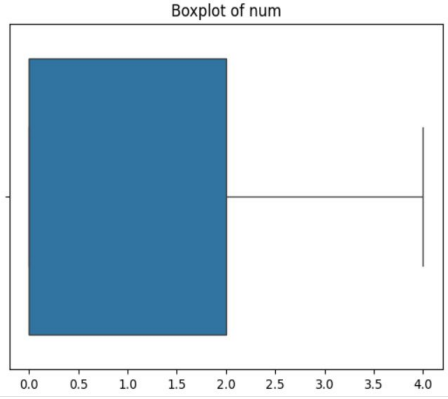


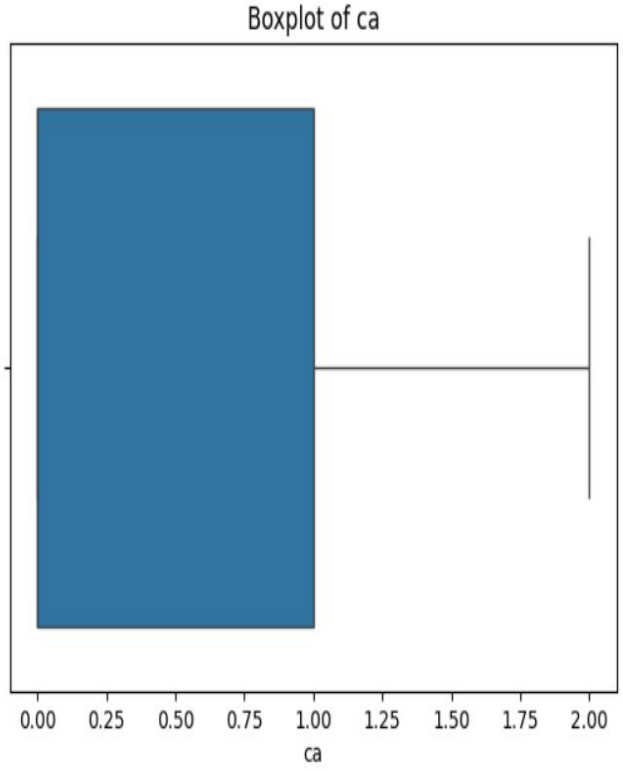
For Testing Data

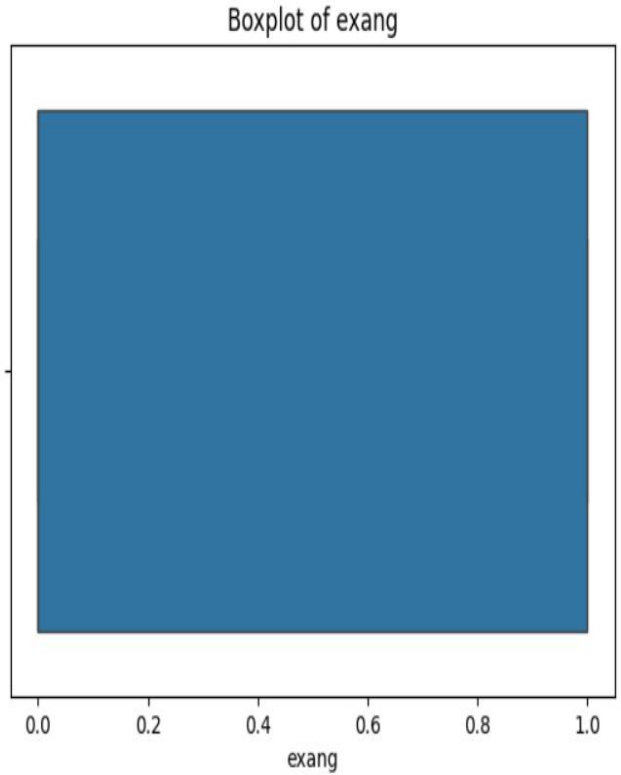


BOX PLOTS

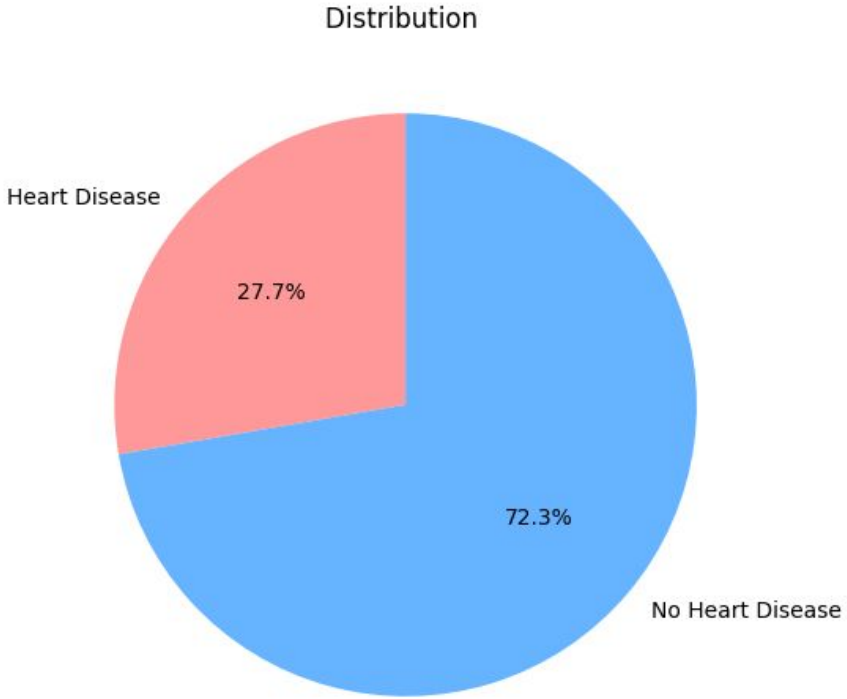

Boxplot of age
30
40
50
60
70
age


Boxplot of num
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0

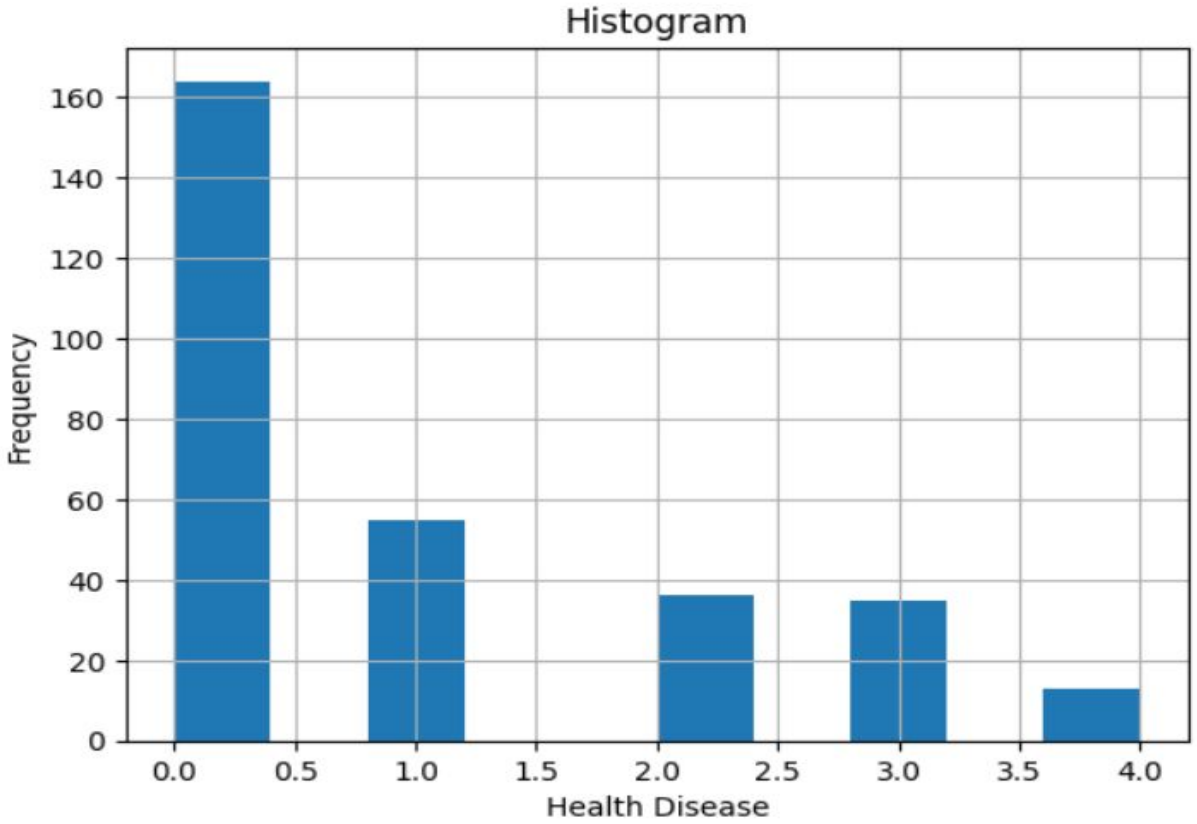

Boxplot of ca
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
ca


Boxplot of exang
0.0
0.2
0.4
0.6
0.8
1.0
exang

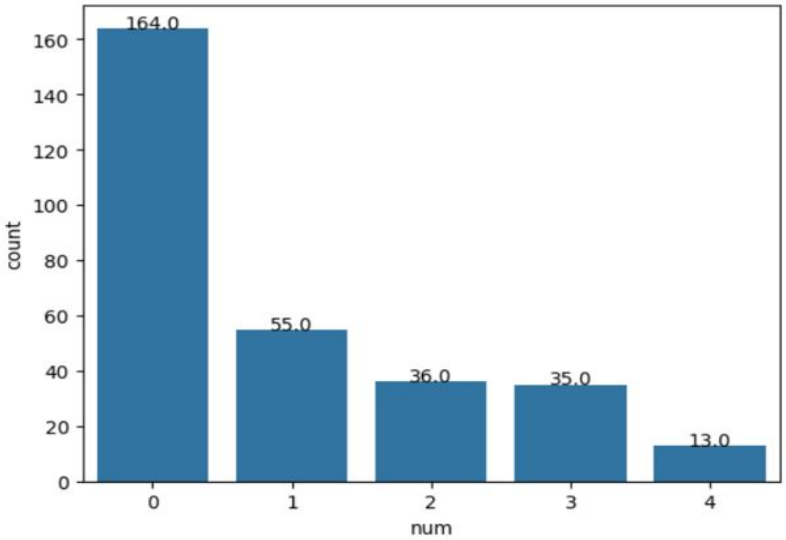
Pie chart

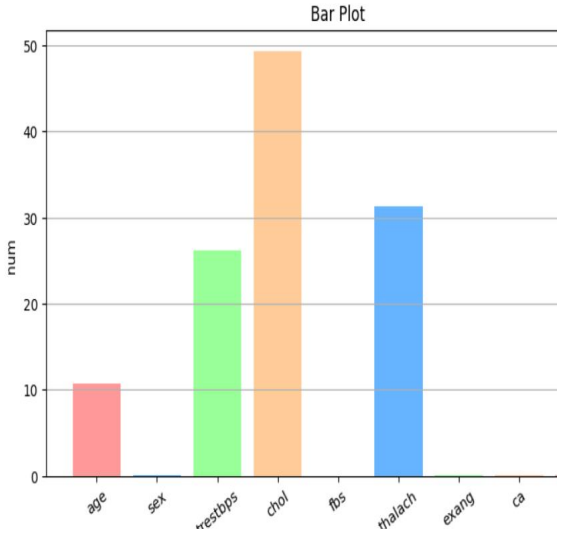


HISTOGRAM

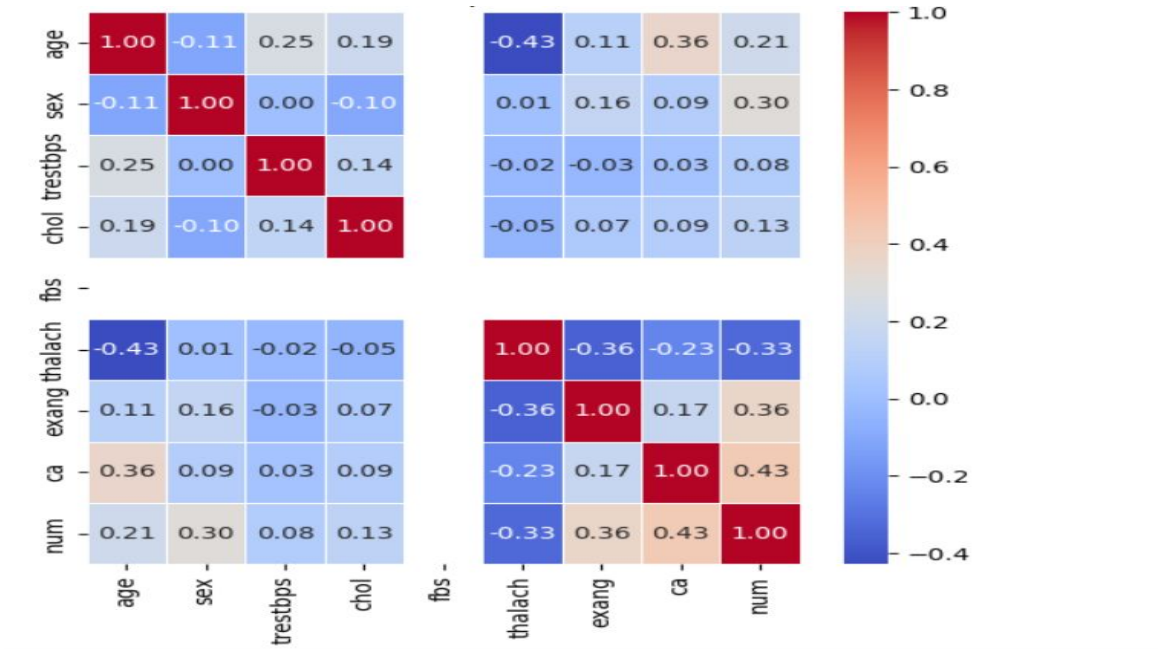

Histogram
Frequency
160
140
120
100
80
60
40
20
0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Health Disease

BAR PLOT


164.0
55.0
36.0
35.0
13.0
count
num
0
1
2
3
4
0
20
40
60
80
100
120
140
160


Bar Plot
num
0
10
20
30
40
50
age
sex
trestbps
chol
fbs
thalach
exang
ca

MATRIX PLOT


age
sex
trestbps
chol
fbs
thalach
exang
ca
num
1.00 -0.11 0.25 0.19 -0.43 0.11 0.36 0.21
-0.11 1.00 0.00 -0.10 0.01 0.16 0.09 0.30
0.25 0.00 1.00 0.14 -0.02 -0.03 0.03 0.08
0.19 -0.10 0.14 1.00 -0.05 0.07 0.09 0.13
-0.43 0.01 -0.02 -0.05 1.00 -0.36 -0.23 -0.33
0.11 0.16 -0.03 0.07 -0.36 1.00 0.17 0.36
0.36 0.09 0.03 0.09 -0.23 0.17 1.00 0.43
0.21 0.30 0.08 0.13 -0.33 0.36 0.43 1.00
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

Regression

Dataset Used: 'Auto MPG' from UC Irvine

This dataset contains various car features and corresponding data of 398 different cars from 1970 to 1982 that can be used to predict the city-cycle miles per gallon of a given vehicle.

Has 6 independent variables:

- Displacement
- Cylinders
- Horsepower
- Weight
- Acceleration
- Model Year
- Origin

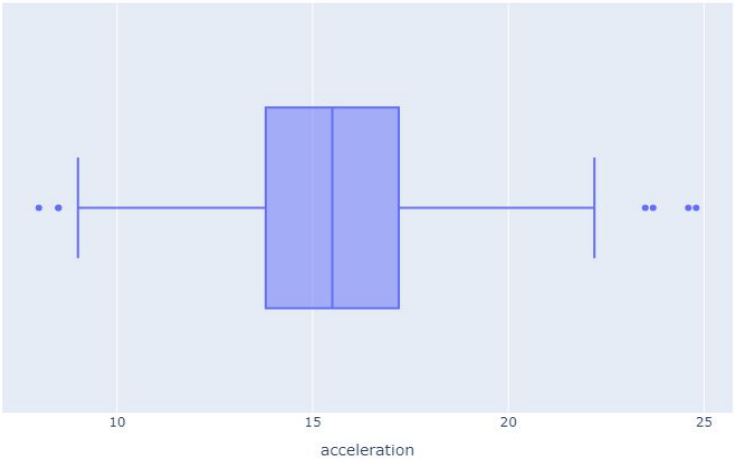
And 1 dependent variable:

- Miles per gallon (mpg)

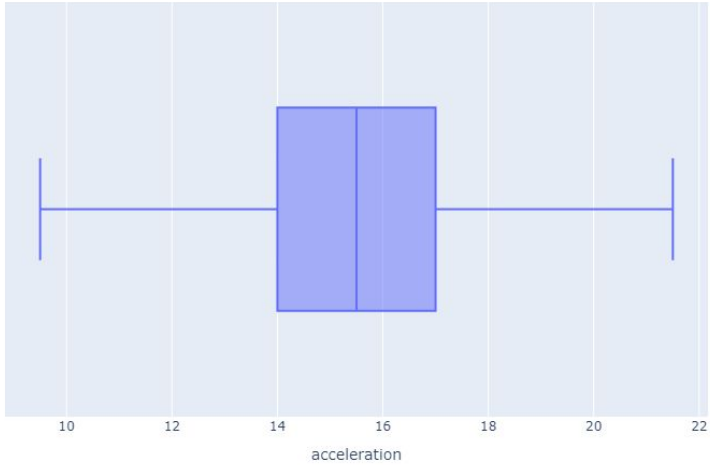
Goal: Use multivariate and linear regression to predict mpg

Removing Outliers

Before

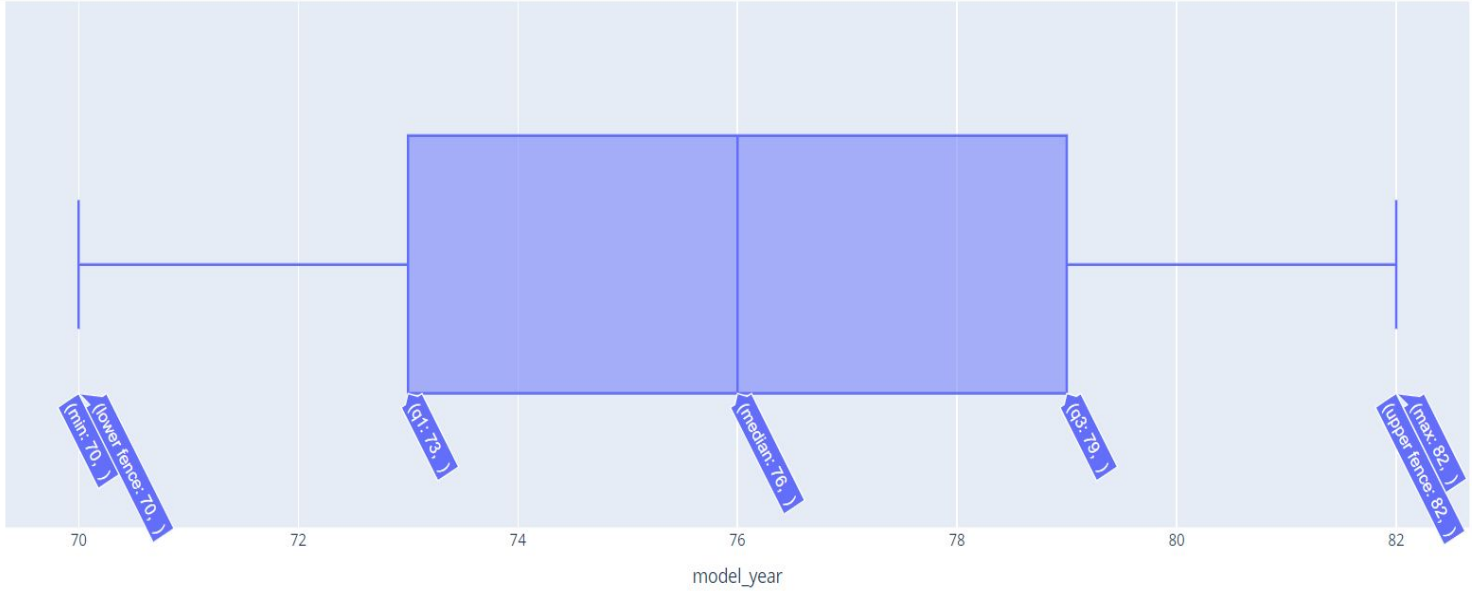


After

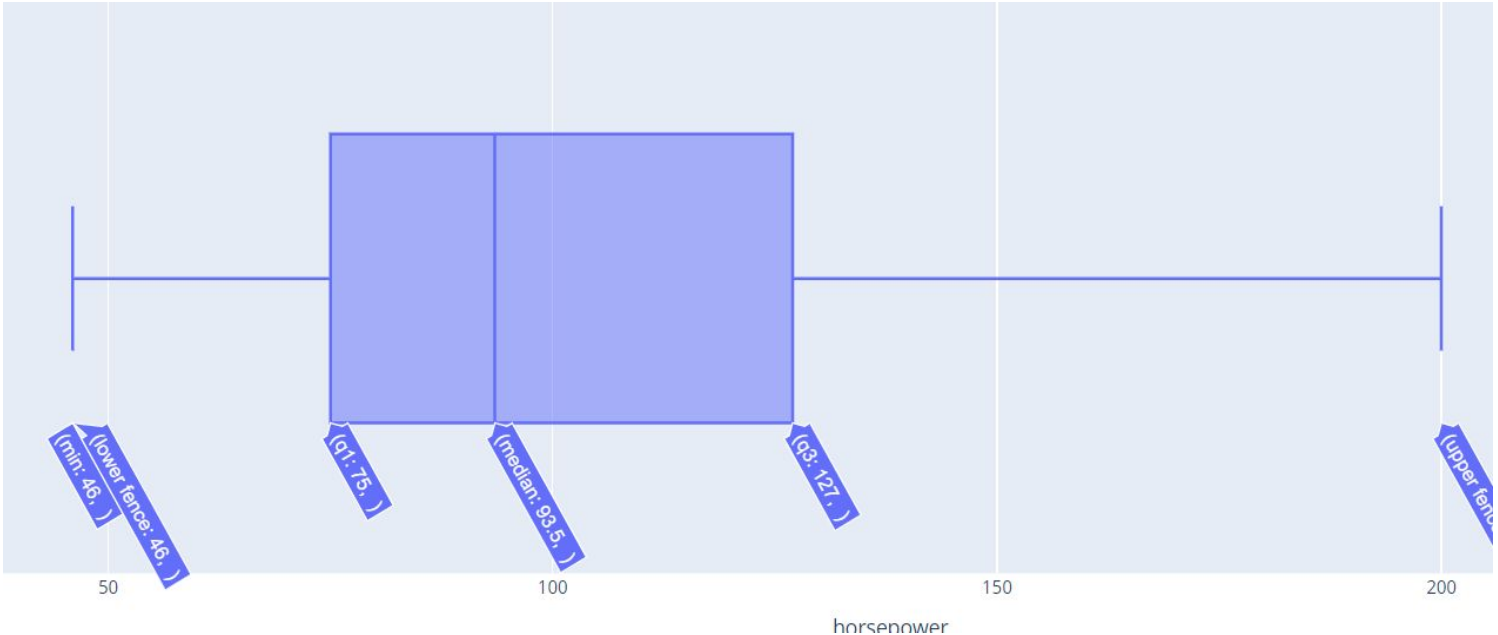


BOX PLOT

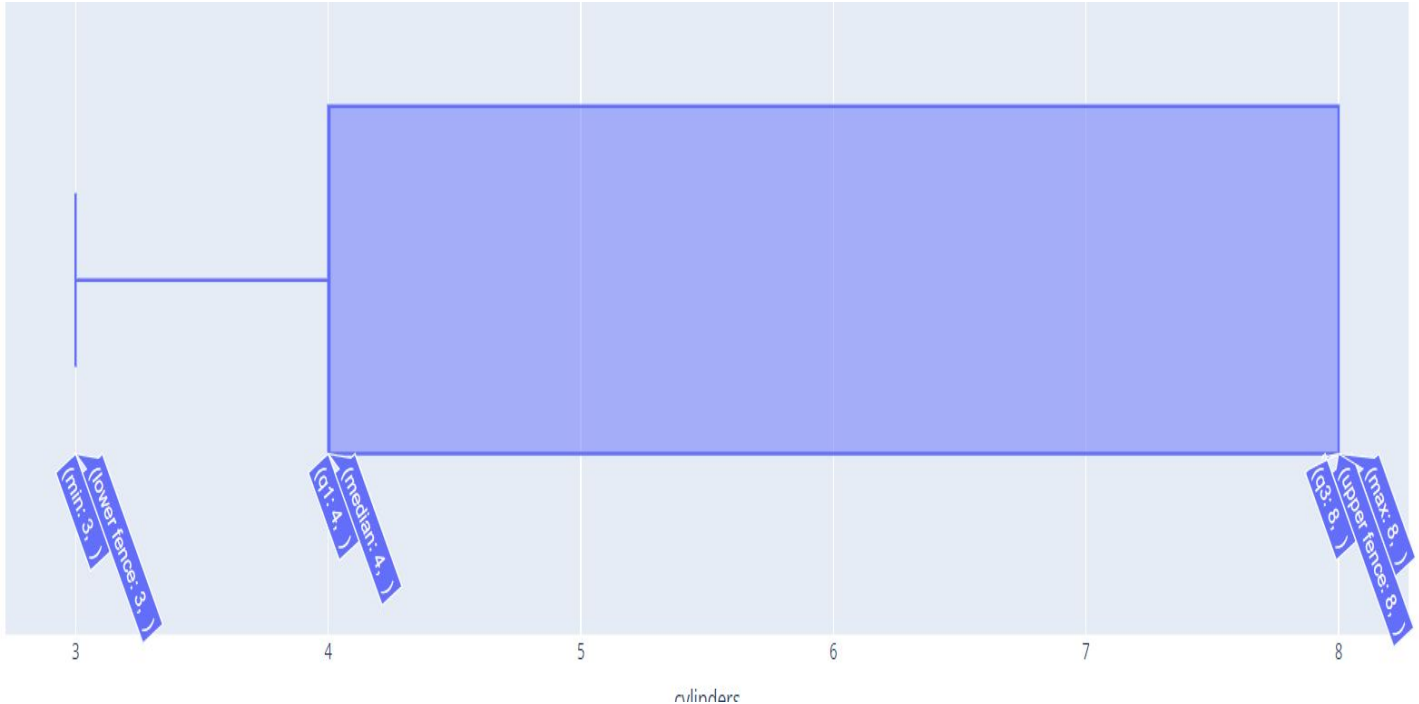
Model Year



HORSE POWER


(min: 46,)
(lower fence: 46,)
(q1: 75,)
(median: 93.5,)
(q3: 127,)
50
100
150
200
horsepower

CYLINDER


(lower fence: 3,)
(min: 3,)
(q1: 4,)
(median: 4,)
(q3: 8,)
(upper fence: 8,)
(max: 8,)
3
4
5
6
7
8
cylinders

DESCRIPTIVE STATISTICS

	displacement	cylinders	horsepower	weight	acceleration	model_year	origin
count	398.000000	398.000000	398.000000	398.000000	398.000000	398.000000	398.000000
mean	193.425879	5.454774	97.275126	2970.424623	15.453769	76.010050	1.572864
std	104.269838	1.701004	28.464797	846.841774	2.391552	3.697627	0.802055
min	68.000000	3.000000	46.000000	1613.000000	9.500000	70.000000	1.000000
25%	104.250000	4.000000	76.000000	2223.750000	14.000000	73.000000	1.000000
50%	148.500000	4.000000	93.500000	2803.500000	15.500000	76.000000	1.000000
75%	262.000000	8.000000	110.000000	3608.000000	17.000000	79.000000	2.000000
max	455.000000	8.000000	170.000000	5140.000000	21.500000	82.000000	3.000000

	mpg
count	398.000000
mean	23.455276
std	7.729446
min	9.000000
25%	17.500000
50%	23.000000
75%	29.000000
max	44.600000

Initial Results

OLS Regression Results

Dep. Variable:	mpg	**R-squared:**	0.831
Model:	OLS	**Adj. R-squared:**	0.827
Method:	Least Squares	**F-statistic:**	273.0
Date:	Thu, 03 Oct 2024	**Prob (F-statistic):**	4.98e-146
Time:	03:30:17	**Log-Likelihood:**	-1024.9
No. Observations:	398	**AIC:**	2066.
Df Residuals:	390	**BIC:**	2098.
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-12.2775	4.142	-2.964	0.003	-20.422	-4.133
displacement	0.0004	0.007	0.060	0.952	-0.013	0.014
cylinders	0.2430	0.329	0.738	0.461	-0.405	0.891
horsepower	-0.0531	0.010	-5.407	0.000	-0.072	-0.034
weight	-0.0054	0.001	-9.160	0.000	-0.007	-0.004
acceleration	-0.1835	0.088	-2.079	0.038	-0.357	-0.010
model_year	0.7462	0.047	15.750	0.000	0.653	0.839
origin	1.0481	0.258	4.057	0.000	0.540	1.556

Omnibus:	20.015	**Durbin-Watson:**	1.360
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	33.886
Skew:	0.334	**Prob(JB):**	4.38e-08
Kurtosis:	4.264	**Cond. No.**	7.98e+04

Multivariate Regression Results

OLS Regression Results

Dep. Variable:	mpg	**R-squared:**	0.830
Model:	OLS	**Adj. R-squared:**	0.828
Method:	Least Squares	**F-statistic:**	382.7
Date:	Thu, 03 Oct 2024	**Prob (F-statistic):**	2.37e-148
Time:	03:30:17	**Log-Likelihood:**	-1025.6
No. Observations:	398	**AIC:**	2063.
Df Residuals:	392	**BIC:**	2087.
Df Model:	5		
Covariance Type:	nonrobust		

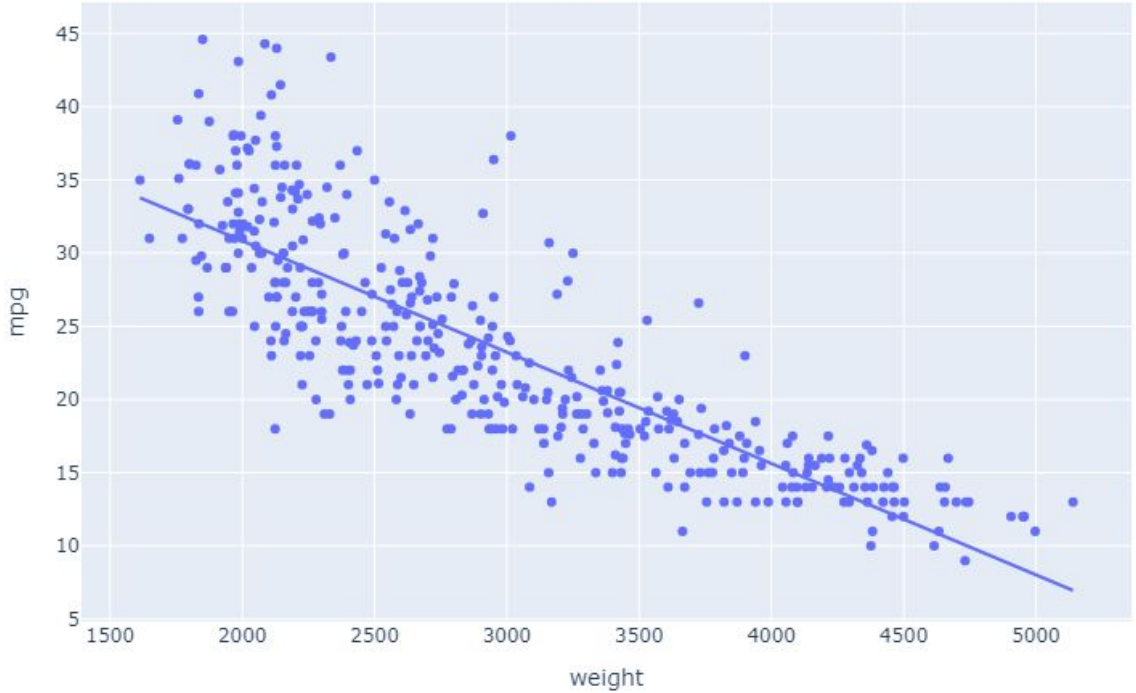
	coef	std err	t	P>|t|	[0.025	0.975]
const	-11.2319	4.004	-2.805	0.005	-19.104	-3.359
horsepower	-0.0510	0.009	-5.581	0.000	-0.069	-0.033
weight	-0.0050	0.000	-15.721	0.000	-0.006	-0.004
acceleration	-0.2035	0.080	-2.537	0.012	-0.361	-0.046
model_year	0.7384	0.046	15.997	0.000	0.648	0.829
origin	1.0063	0.247	4.075	0.000	0.521	1.492

Omnibus:	22.023	**Durbin-Watson:**	1.352
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	37.965
Skew:	0.361	**Prob(JB):**	5.70e-09
Kurtosis:	4.330	**Cond. No.**	7.70e+04

Multivariate Regression Equation

$$y = -0.051(x_1) - 0.005(x_2) - 0.204(x_3) + 0.738(x_4) + 1.006(x_5) - 11.232$$

Regression Plots



R-squared = 0.694

Linear Regression Results

OLS Regression Results

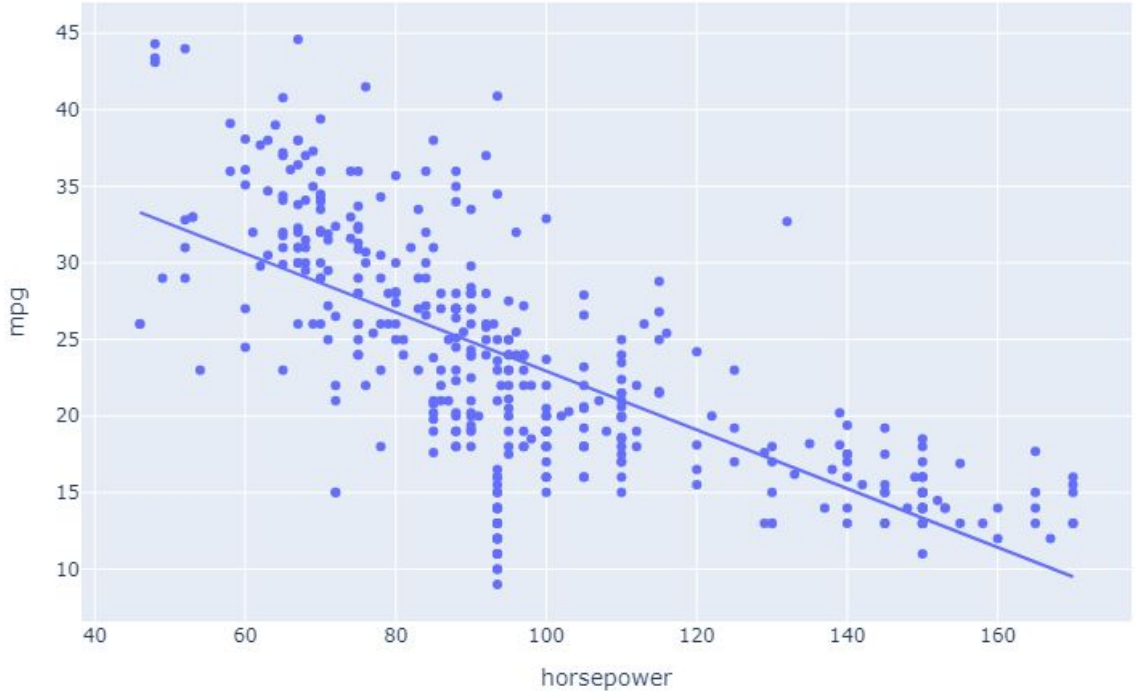
Dep. Variable:	mpg	**R-squared:**	0.694
Model:	OLS	**Adj. R-squared:**	0.694
Method:	Least Squares	**F-statistic:**	899.3
Date:	Thu, 03 Oct 2024	**Prob (F-statistic):**	5.94e-104
Time:	04:31:33	**Log-Likelihood:**	-1142.3
No. Observations:	398	**AIC:**	2289.
Df Residuals:	396	**BIC:**	2297.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	46.0462	0.783	58.788	0.000	44.506	47.586
weight	-0.0076	0.000	-29.989	0.000	-0.008	-0.007

Omnibus:	33.469	**Durbin-Watson:**	0.797
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	43.268
Skew:	0.650	**Prob(JB):**	4.02e-10
Kurtosis:	3.959	**Cond. No.**	1.13e+04

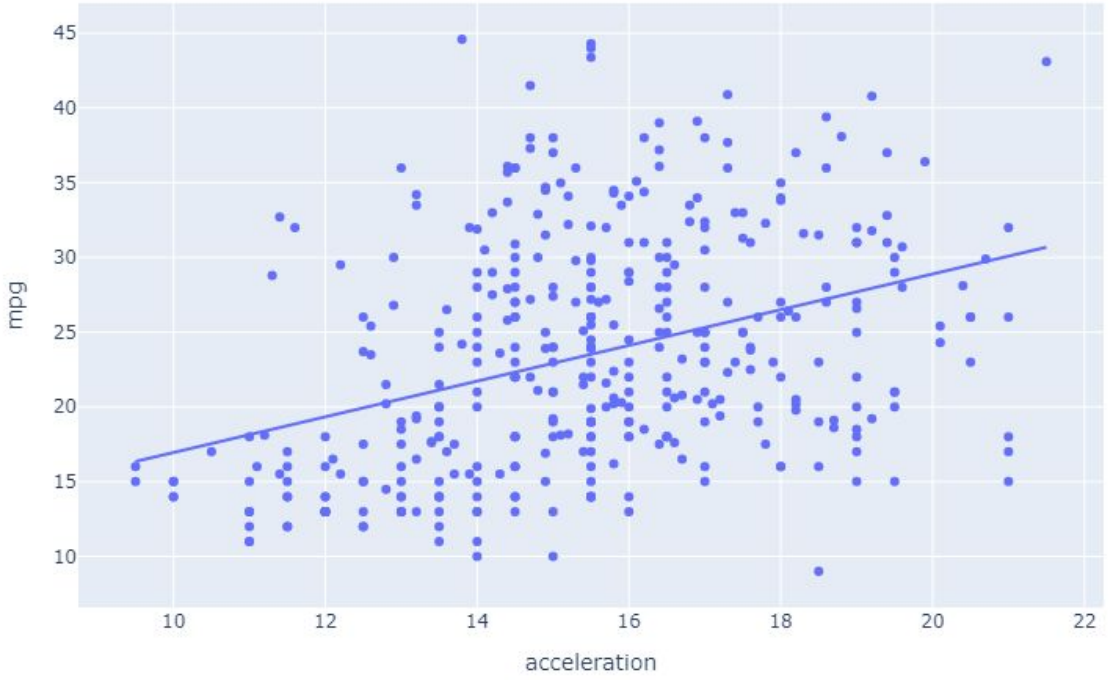
	Metrics	Value
0	Root Mean Square Error (RMSE)	200.4643
1	Mean Squared Error (MSE)	18.2187
2	Mean Absolute Error (MAE)	8.1914
3	Mean Absolute Percentage Error (MAPE)	0.3948

Regression Plots



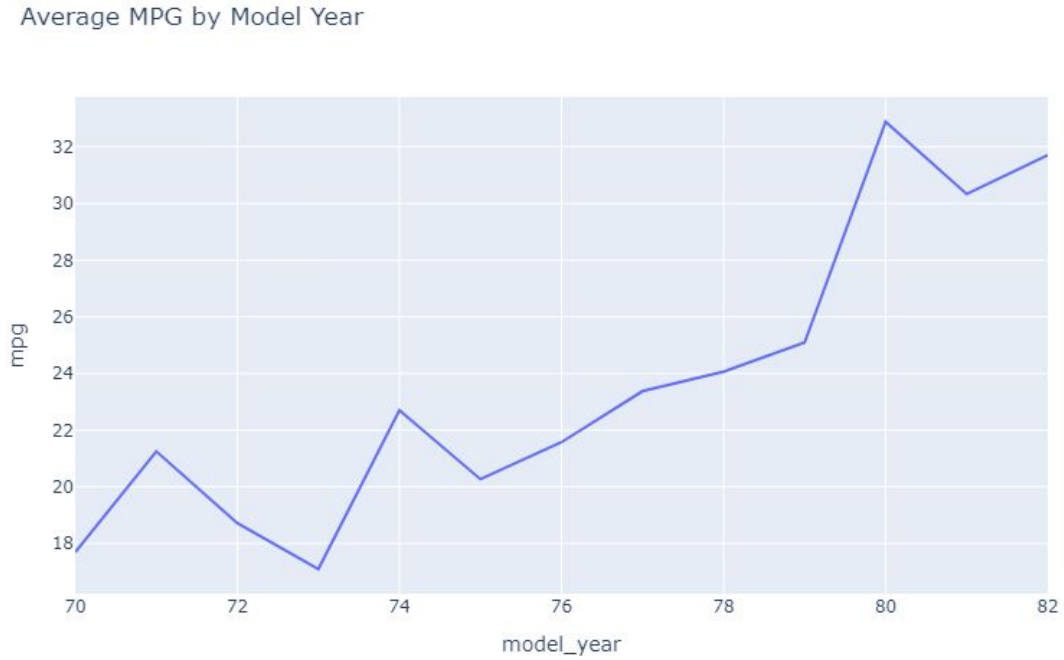
R-squared = 0.499

Regression Plots

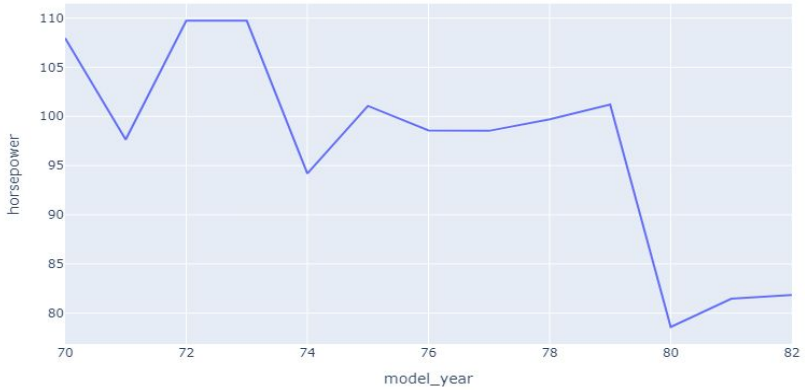


R-squared= 0.137

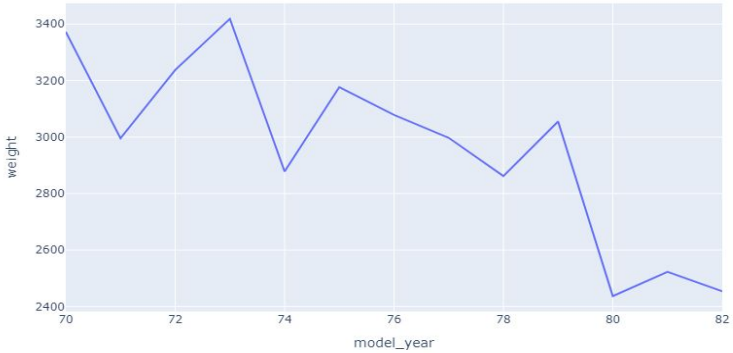
WE CAN SEE THESE INSIGHTS IN ACTION



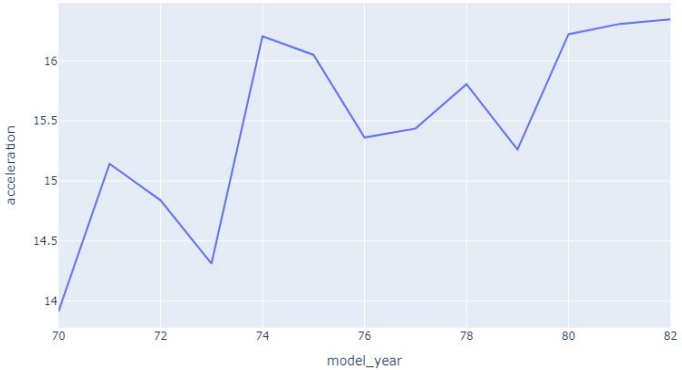
Average Horsepower by Model Year



Average Weight by Model Year



Average Acceleration by Model Year



THANK YŌU!