# Introduction

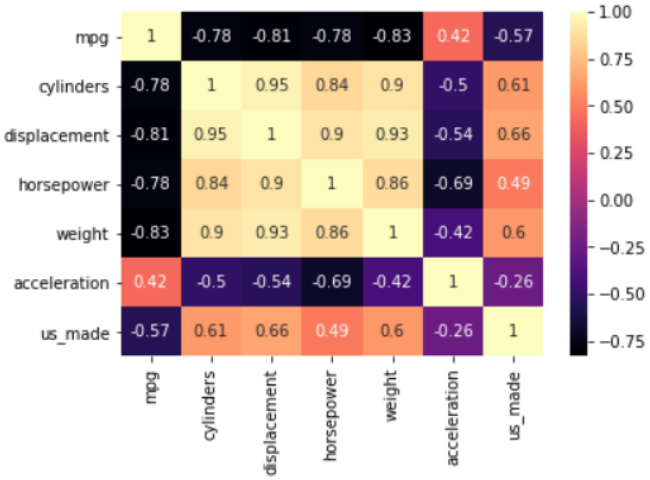
The dataset contains information about the cars from year from 1970 till 1982. There are 398 rows of observations with 8 attributes of the cars such as MPG, cylinders, horsepower, weight etc. the objective of this project is to use linear regression techniques to predict the best set of attributes that can be utilized in such a way that we create the most fuel-efficient car for the future. I'll be using various python packages and libraries such as scipy, statsmodels, sklearn for regression analysis in this report.

# Data cleaning

The first step in data cleaning is to convert the columns names into snake format i.e., lowercase names with underscore replacing the spaces. Next step is to drop the duplicates from the dataset. Now there are a few outliers in the dataset, but I've not dropped them as it seems they we'll important for analysis to build a fuel-efficient car. There are some special characters in the column horsepower '?', So I've replaced the special characters with null values and removed those rows from our dataset. Next after checking the data types of all the attributes, I've modified the data type for horsepower column into integer. As model year does not signify any relationship with fuel efficient future car I've dropped this column also. Now Finally we've only 392 rows of information and 7 columns.

# Exploratory data Analysis

Now after performing some exploratory data analysis, I can say that there is almost direct relation between weight and MPG i.e., as the weight of a car decreases the miles per gallon increases as displayed in the graph in appendix. Similarly, there is a direct relation between the horsepower of a car and the mpg, as the horsepower of a car increases the mpg decreases and vice versa. Now to verify this I've plotted a correlation matrix as displayed below:




A negative relation in the correlation plot is displayed by the darkness of the color and the minus sign. A yellow bar is considered as directly related to each other as the color of the box changes from yellow to orange and blue and dark the correlation among variables decreases. The highest positive correlations exist between cylinders and the displacement. Since both are parts of an engine, displacement is the total distance a piston travels inside the cylinder of the engine of a car. There is a negative correlation between displacement and the mpg variable.

# Linear Regression models

Linear regression model is used to predict the values of a continues dependent variable, for creating the model in this report I've taken mpg as the dependent variable as we'll be predicting a fuel-efficient car with best mpg and rest other columns as independent variables. I've split the data into train and test in 75:25 ratio. Next using statsmodels library in python I've created a linear regression model. The r squared value of the model is 72% and the AIC score is 1666. The constant coefficient of the linear equation is 51.6787. Apart from the displacement all other variables have a negative coefficient, meaning that increasing the cylinders, weight, acceleration or horsepower would result in a lower MPG and in turn lower fuel-efficient car.

Taking a confidence interval of 95% for this analysis, I can say that a variable having p value greater than 0.05 can be ignored. The model output also suggests that there is a high collinearity among independent variables. Next after checking the VIF scores, I can say that this is not the best model for a fuel-efficient car as the VIF values are quite high for all the variables.

# Optimization

Now, I'll try to optimize the model by using the backward selection techniques by dropping the least important variables one by one so that we only have the variables of importance only for the model. As the p value for cylinders column was more than 0.05 value, I've dropped this column and run the regression model again. This time the r squared value of the model is still 72% and the AIC score is also 1666 i.e., both remained unchanged. This time the p value of displacement is more than 0.05 so we'll drop this variable and run the model again to check if there is any change in the r square or AIC values.

Now after removing the variables cylinders, displacement from the independent variables list I run the linear regression model again and this time the r squared is 72% and the AIC score is 1665. The AIC score decreased slightly for this model as compared to the last two models. The p values of all variables except acceleration is less than 0.05 so we'll drop acceleration this time and again run the model.

Running the linear regression model with only weight, horsepower and US made attributes as independent variables and mpg as dependent variable. This time the r square value has dropped a little to 71% and the AIC score is 1666. While all the variables have a p value of less than 0.05 the constant coefficient is 44.2993. As the r squared value decreased after dropping the acceleration and the AIC remained unchanged, we can say that this is our best model for creating a fuel efficient future car.

The variables of importance are weight of a car, horsepower of a car, whether it's USA made or not. All of these variables have a negative coefficient meaning that with a single unit increase in the variables the mpg will drop by the coefficients value.

So, in order to build this car for the future we need to make sure the car is USA made and it should be light weight as we saw in the weight vs mpg graph that as the weight of the car decreases the MPG increases. Also cars with less horsepower tend to have a higher MPG so we should also consider keeping the horsepower to a minimum number for most fuel efficient car.

If we compare the variables, we can say that the US made is most significant as the coefficient is highest for this variable followed by horsepower and the weight.

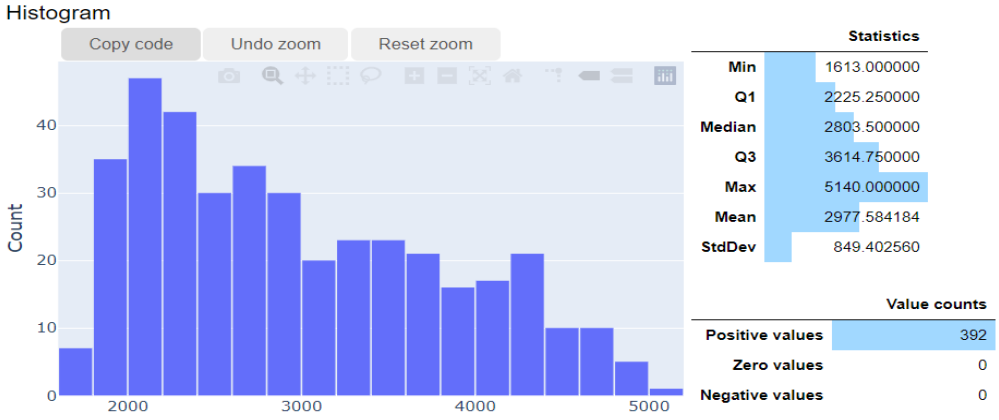
# References

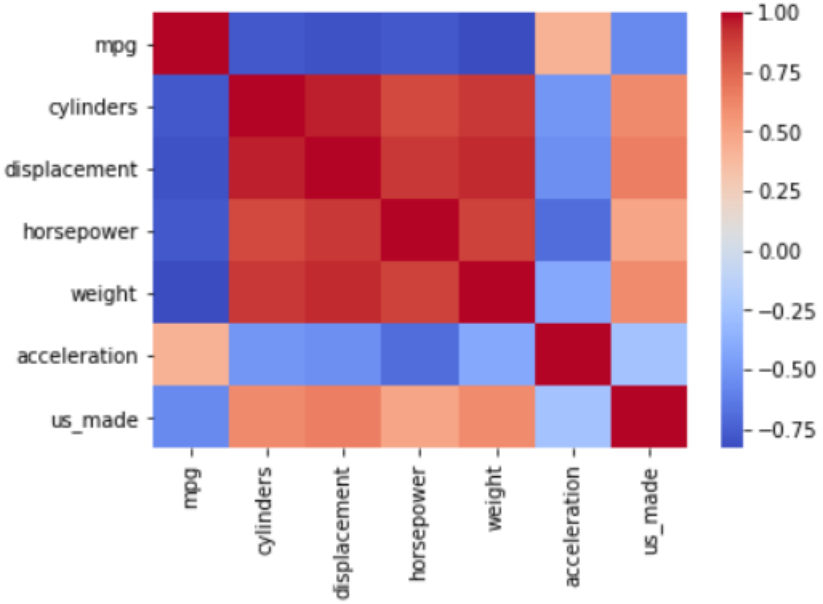

Frost, J. (2023, February 9). How to Interpret P-values and Coefficients in Regression Analysis. Statistics by Jim. Retrieved March 4, 2023, from https://statisticsbyjim.com/regression/interpret-coefficients-p-values-regression/

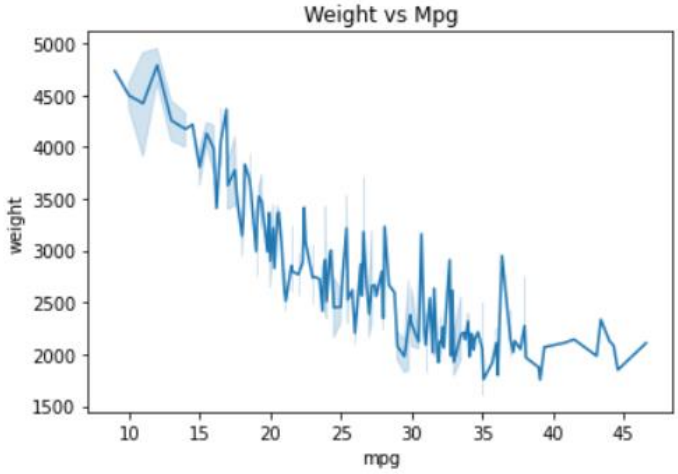
Linear Regression in Python using Statsmodels. (n.d.). Retrieved March 4, 2023, from https://datatofish.com/statsmodels-linear-regression/

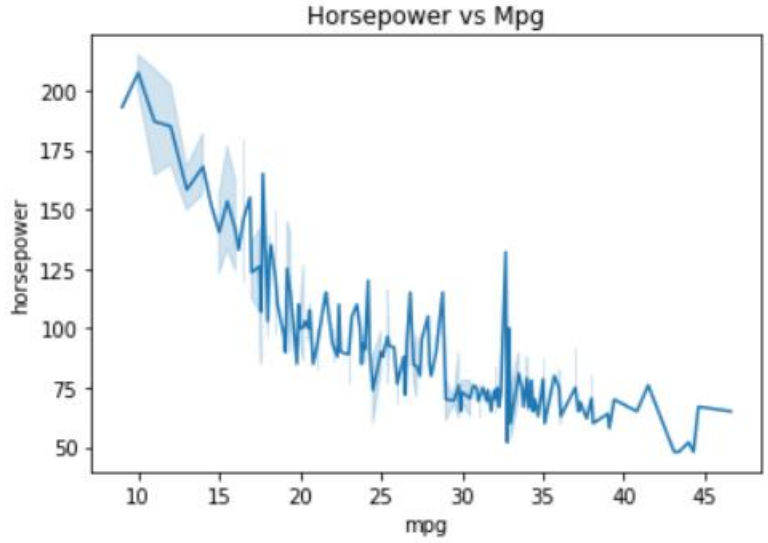
GeeksforGeeks. (2022, December 30). Stepwise Regression in Python. Retrieved March 4, 2023, from https://www.geeksforgeeks.org/stepwise-regression-in-python/

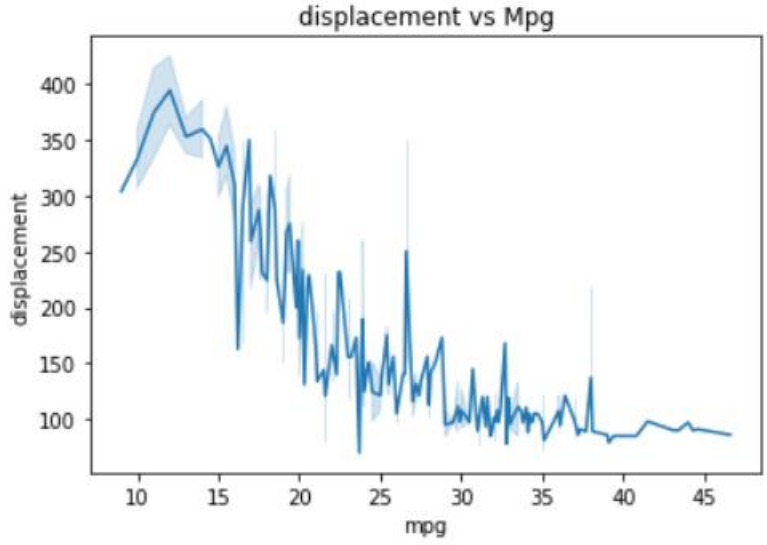

# Appendix

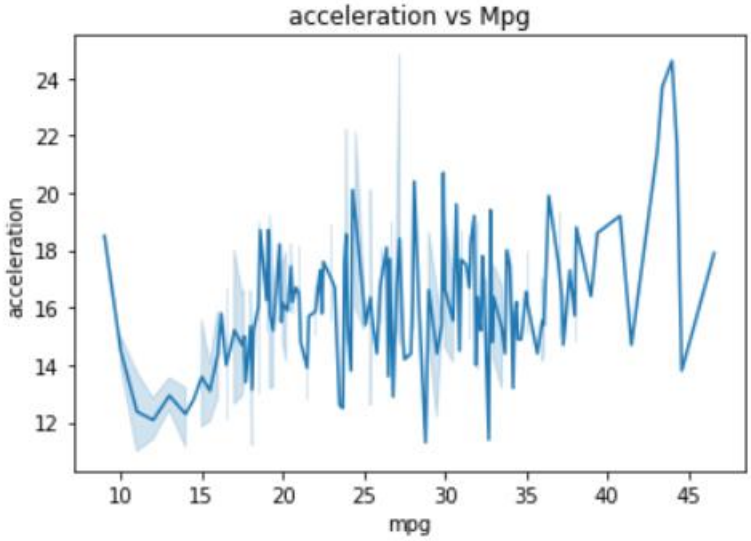



mpg
cylinders
displacement
horsepower
weight
acceleration
us_made
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75


Weight vs Mpg
weight
mpg
5000
4500
4000
3500
3000
2500
2000
1500
10
15
20
25
30
35
40
45


Horsepower vs Mpg
horsepower
200
175
150
125
100
75
50
10
15
20
25
30
35
40
45
mpg


displacement vs Mpg
displacement
400
350
300
250
200
150
100
10
15
20
25
30
35
40
45
mpg


acceleration vs Mpg
acceleration
mpg
24
22
20
18
16
14
12
10
15
20
25
30
35
40
45

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | mpg | R-squared: | 0.723 |
| Model: | OLS | Adj. R-squared: | 0.718 |
| Method: | Least Squares | F-statistic: | 125.1 |
| Date: | Sat, 04 Mar 2023 | Prob (F-statistic): | 4.30e-77 |
| Time: | 19:00:30 | Log-Likelihood: | -826.08 |
| No. Observations: | 294 | AIC: | 1666. |
| Df Residuals: | 287 | BIC: | 1692. |
| Df Model: | 6 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | 51.6787 | 3.161 | 16.351 | 0.000 | 45.458 | 57.900 |
| cylinders | -0.5470 | 0.449 | -1.218 | 0.224 | -1.431 | 0.337 |
| displacement | 0.0176 | 0.011 | 1.605 | 0.110 | -0.004 | 0.039 |
| horsepower | -0.0818 | 0.019 | -4.217 | 0.000 | -0.120 | -0.044 |
| weight | -0.0049 | 0.001 | -5.579 | 0.000 | -0.007 | -0.003 |
| acceleration | -0.2753 | 0.144 | -1.911 | 0.057 | -0.559 | 0.008 |
| us_made | -2.2624 | 0.697 | -3.244 | 0.001 | -3.635 | -0.890 |

| | | | |
|---|---|---|---|
| Omnibus: | 17.449 | Durbin-Watson: | 2.150 |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 19.558 |
| Skew: | 0.539 | Prob(JB): | 5.66e-05 |
| Kurtosis: | 3.658 | Cond. No. | 4.16e+04 |

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | mpg | R-squared: | 0.717 |
| Model: | OLS | Adj. R-squared: | 0.715 |
| Method: | Least Squares | F-statistic: | 245.5 |
| Date: | Sat, 04 Mar 2023 | Prob (F-statistic): | 2.91e-79 |
| Time: | 19:43:19 | Log-Likelihood: | -829.23 |
| No. Observations: | 294 | AIC: | 1666. |
| Df Residuals: | 290 | BIC: | 1681. |
| Df Model: | 3 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | 44.2993 | 0.919 | 48.225 | 0.000 | 42.491 | 46.107 |
| horsepower | -0.0446 | 0.012 | -3.805 | 0.000 | -0.068 | -0.022 |
| weight | -0.0051 | 0.001 | -8.628 | 0.000 | -0.006 | -0.004 |
| us_made | -1.6686 | 0.615 | -2.714 | 0.007 | -2.879 | -0.458 |

| | | | |
|---|---|---|---|
| Omnibus: | 19.251 | Durbin-Watson: | 2.178 |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 21.948 |
| Skew: | 0.572 | Prob(JB): | 1.71e-05 |
| Kurtosis: | 3.694 | Cond. No. | 1.23e+04 |