ELIF HACIOSMANOGLU

ERASMUS STUDENT

IOT AND BIG DATA FOR SMART SPACES

EFFECT OF HORSEPOWER ON FUEL CONSUMPTION- A COMPARISON OF REGRESSION MODELS

ACADEMIC YEAR 2021/2022

INTRODUCTION

- For many decades, engineers and innovators have strived for greater engine power of a car to increase its speed and its mile range and to lure customers into buying these cars.
- Yet, the customers and the car producers have just realized the hypothesis that this greater engine power of the car comes with a significant cost, which is higher fossil fuel consumption and therefore higher environmental degradation.
- Thus, the production and ownership of "IoT-assisted electrical vehicles" will help in preserving our nature and creating a sustainable city life.
- To empirically test the correlation between engine power and fuel consumption, I will use various regression models to examine the relationship between car characteristics variables in KNIME and compare the results of them to find the most robust model.

DATASET

- Dataset contains 392 variables from different type of car brands and models that is from the year of 1983.
- There are 6 variables:

- Miles per Gallon (MPG), refers to the distance a car can travel per gallon fuel, also indicates the fuel efficiency of a car.

- Horsepower (HP), indicates the power of a car's engine.

- Cylinders, refers to the chamber where fuel is combusted, and power is generated (more cylinders, more performance).

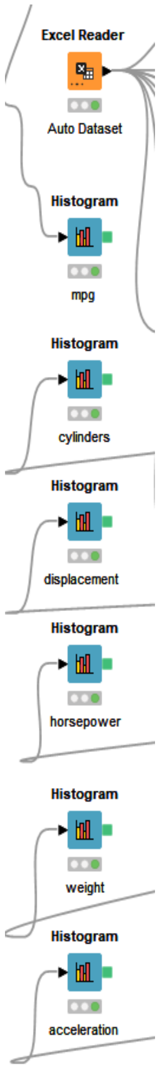
- Displacement, represents the total volume of all cylinders in a car's engine.

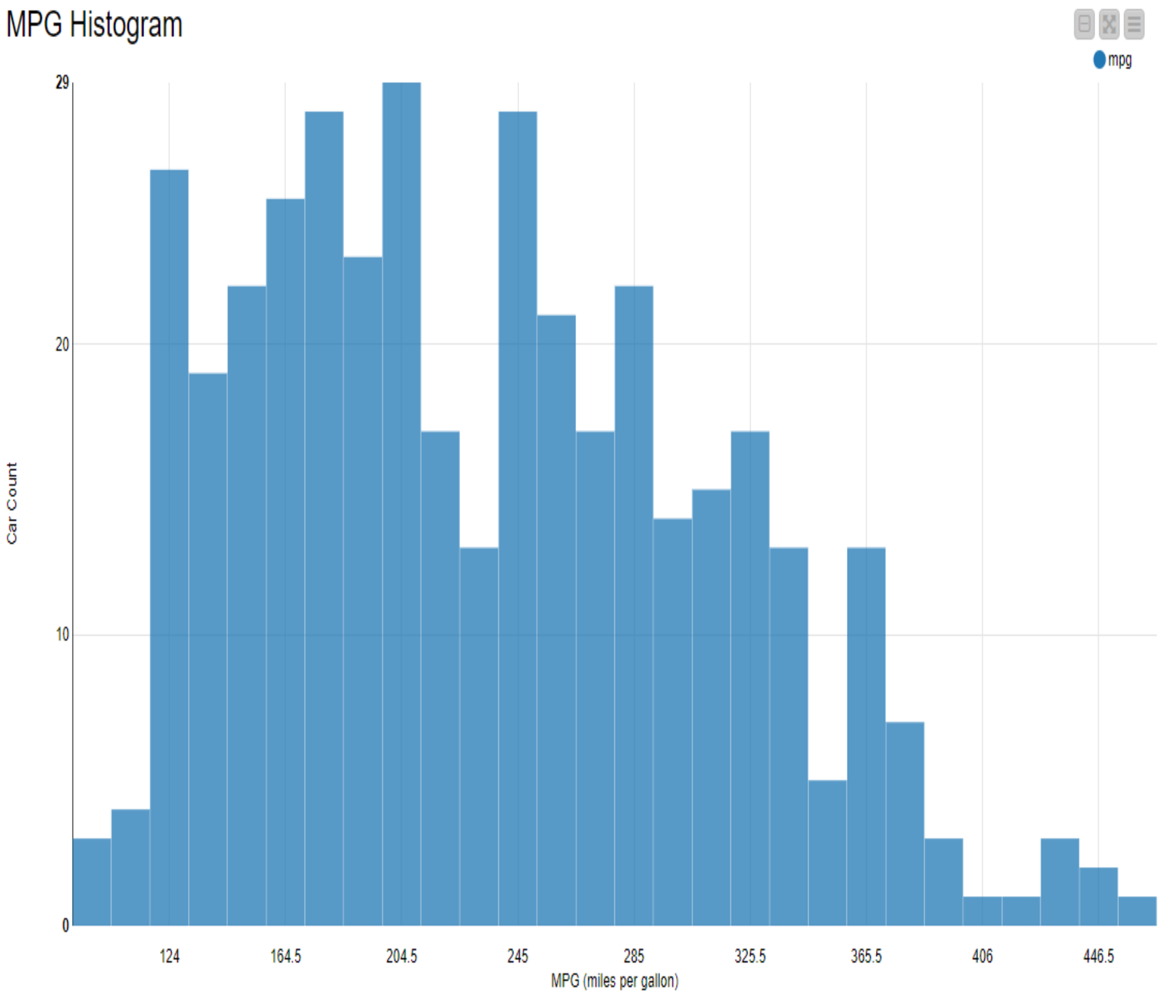
- Weight, shows the total weight of a car

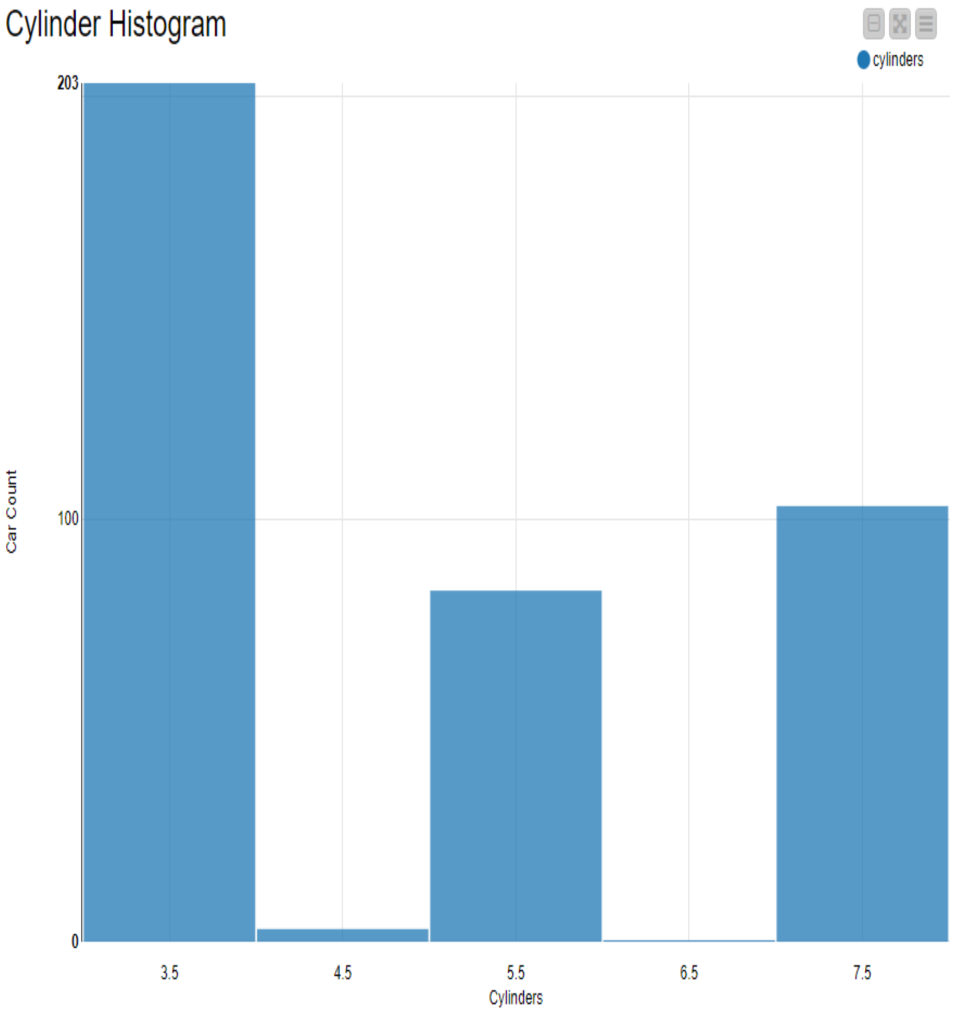
- Acceleration, refers to amount of time a car reaches to a velocity of 60 miles per hour (95.56 km/h).

The dataset is available on: https://archive.ics.uci.edu/ml/datasets/auto+mpg

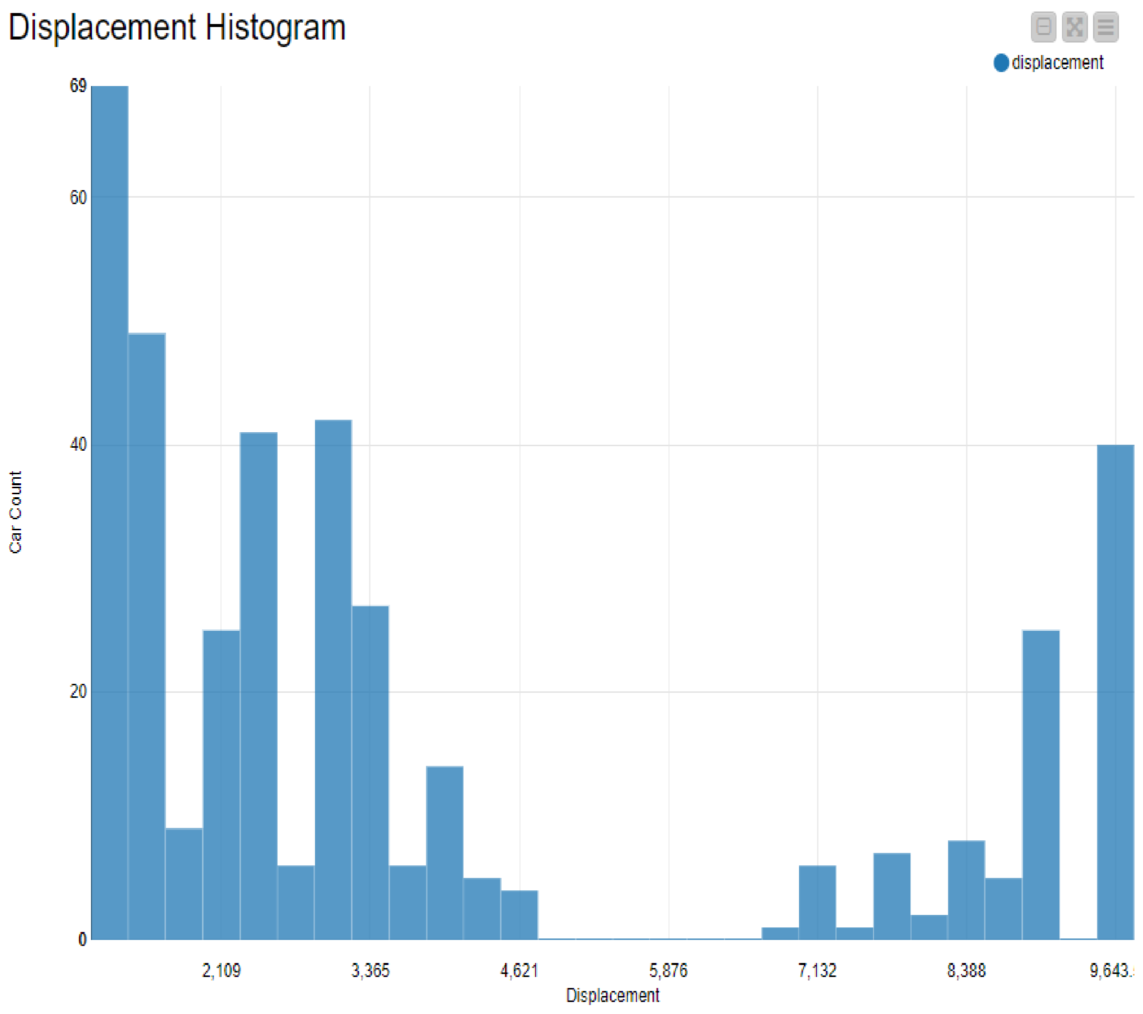
VISUALIZATION OF THE VARIABLES

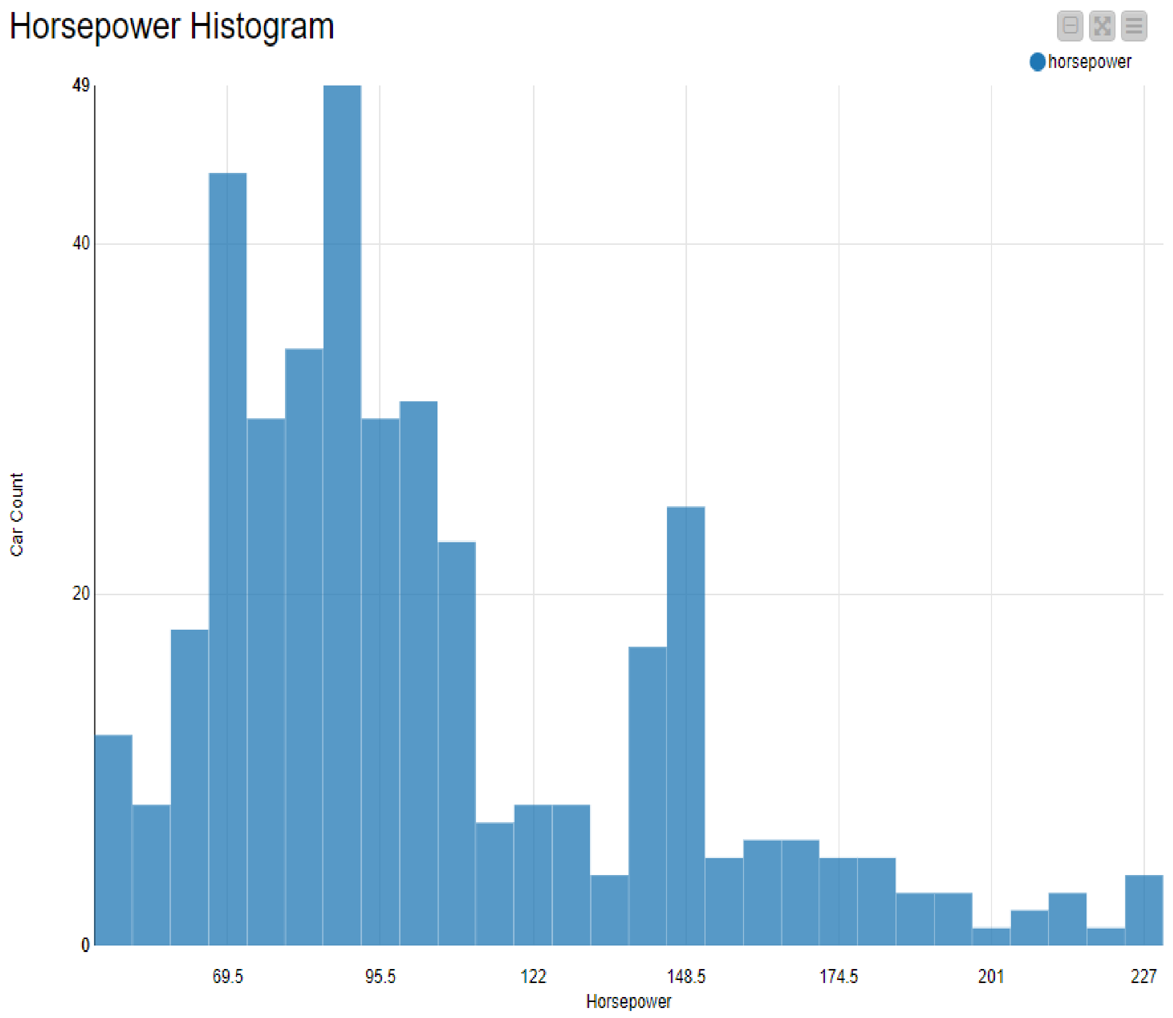




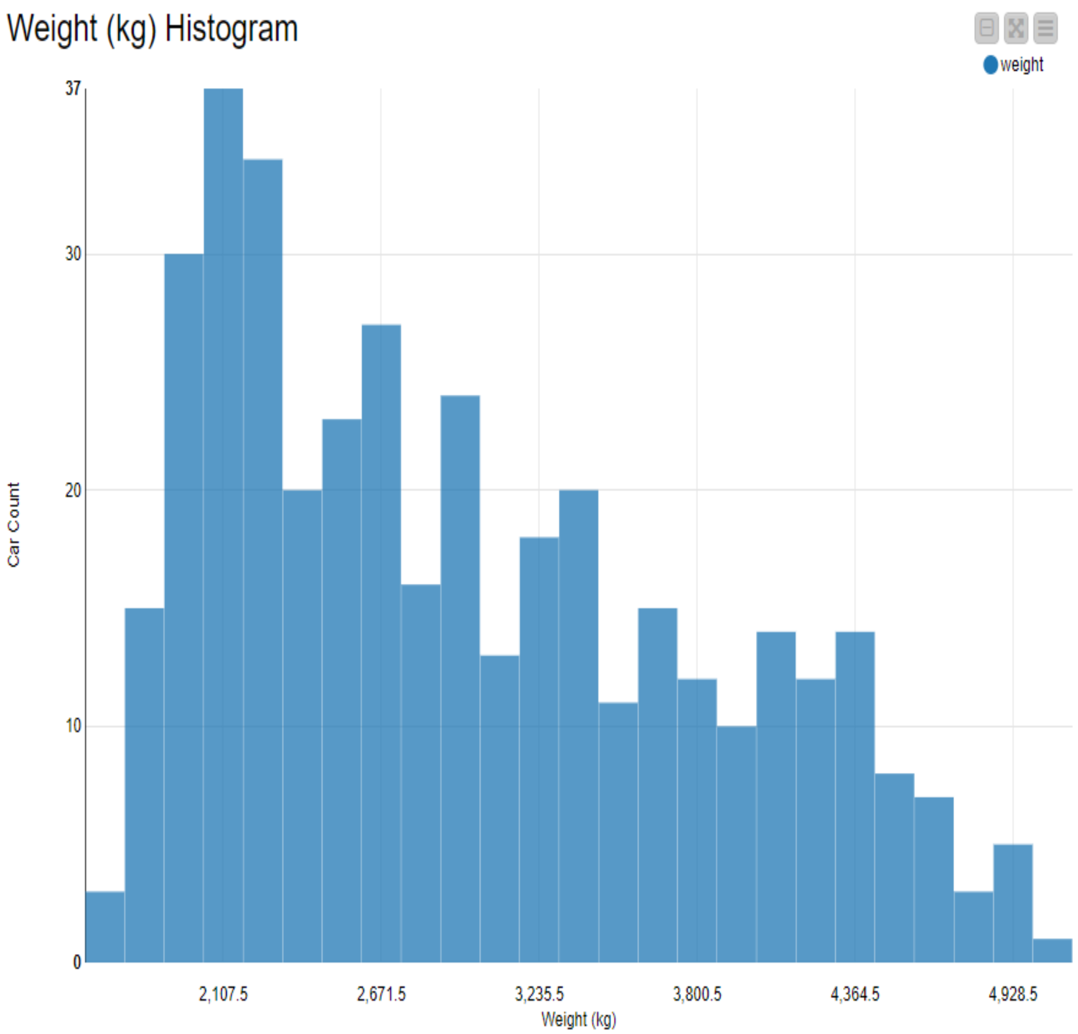


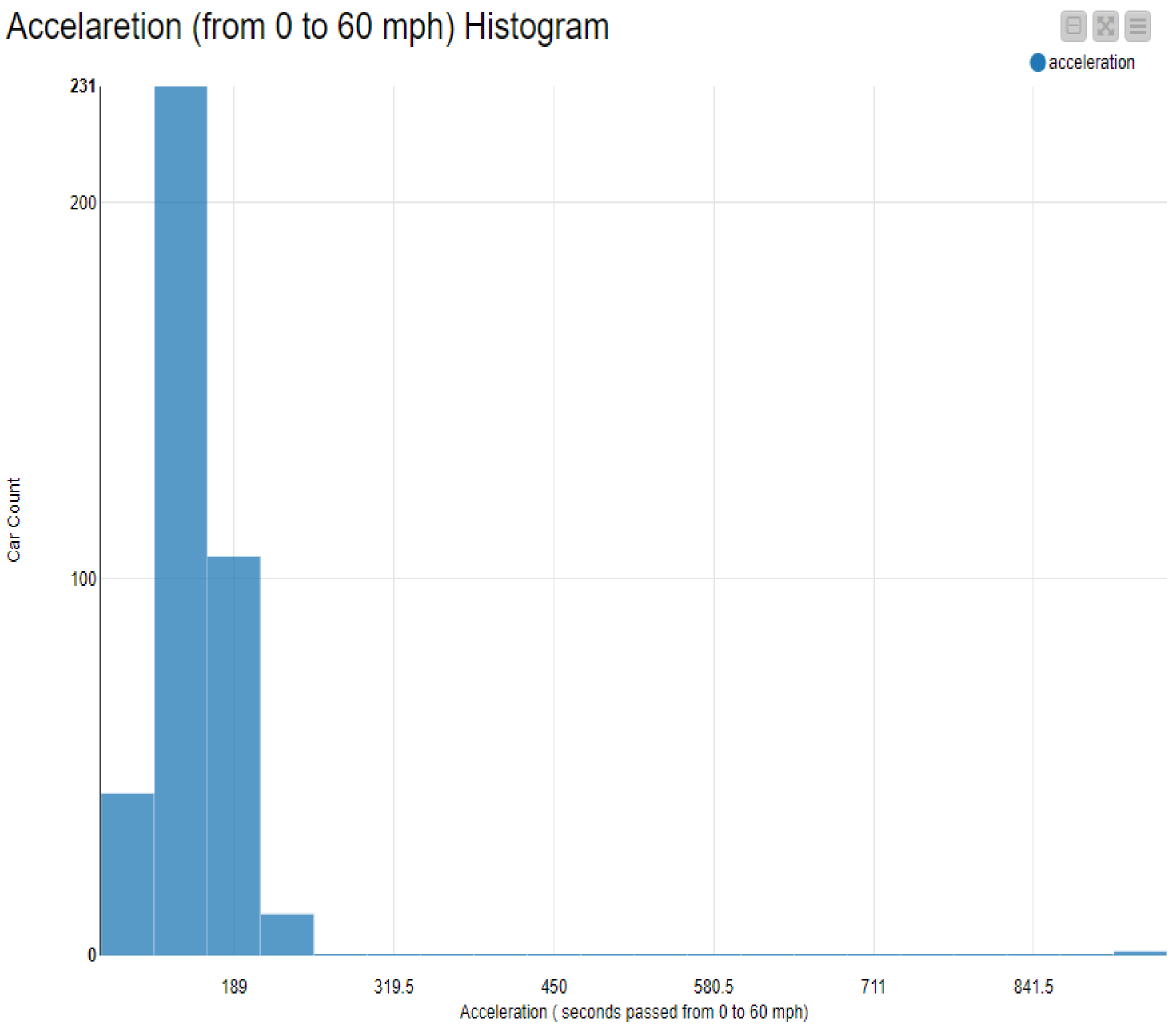
VISUALIZATION OF THE VARIABLES





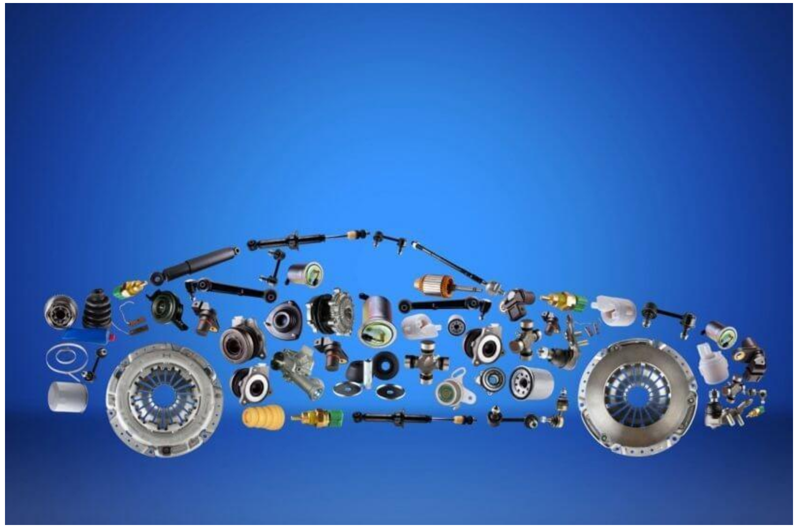
VISUALIZATION OF THE VARIABLES



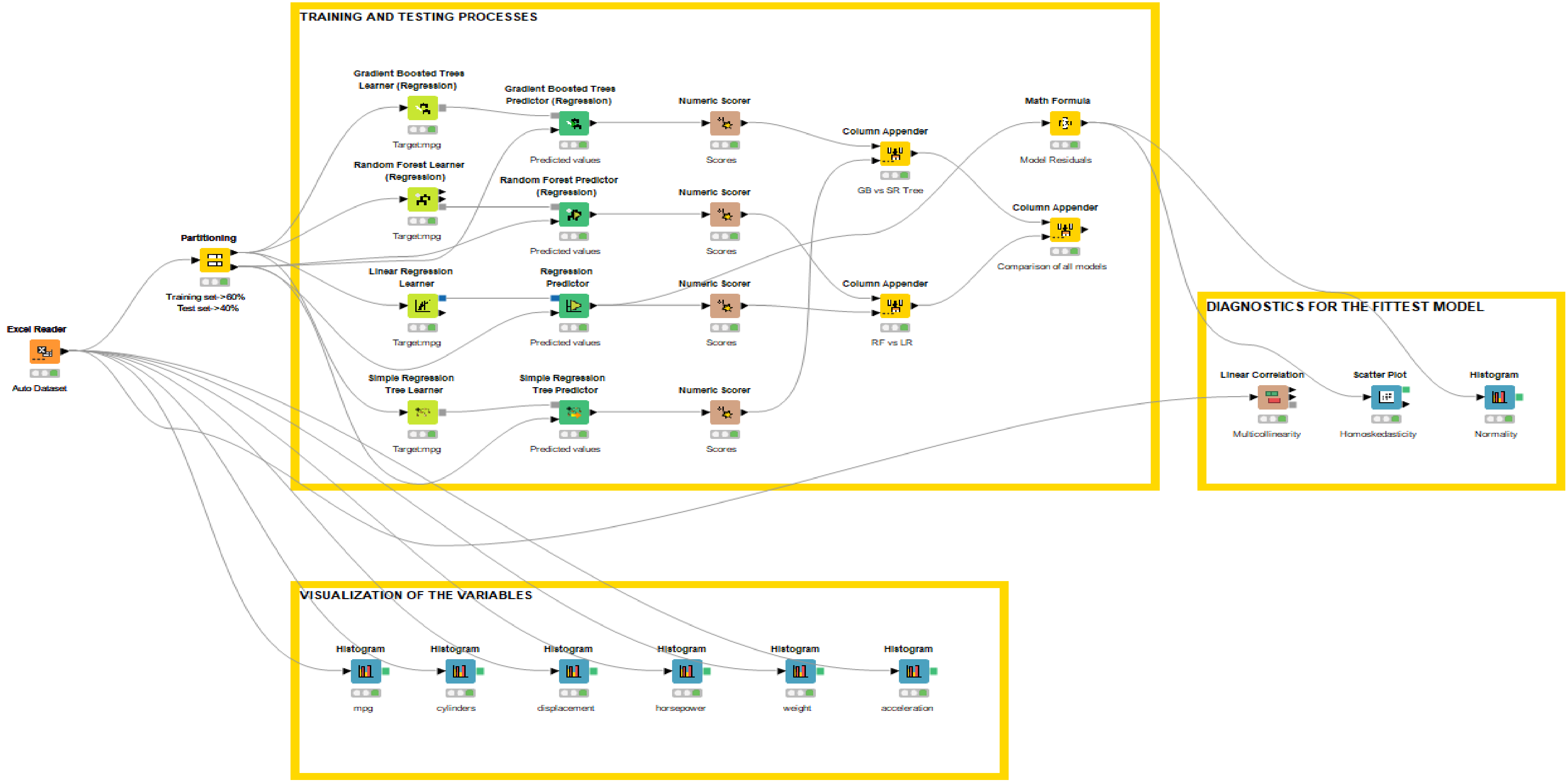


AIM OF THE WORK

Building models to find correlation between MPG as dependent variable and HP as independent interest variable by using cylinders, displacement, weight and acceleration as explanatory variables. Then, comparing the results of each model to find the most robust model that can explain the relationship.

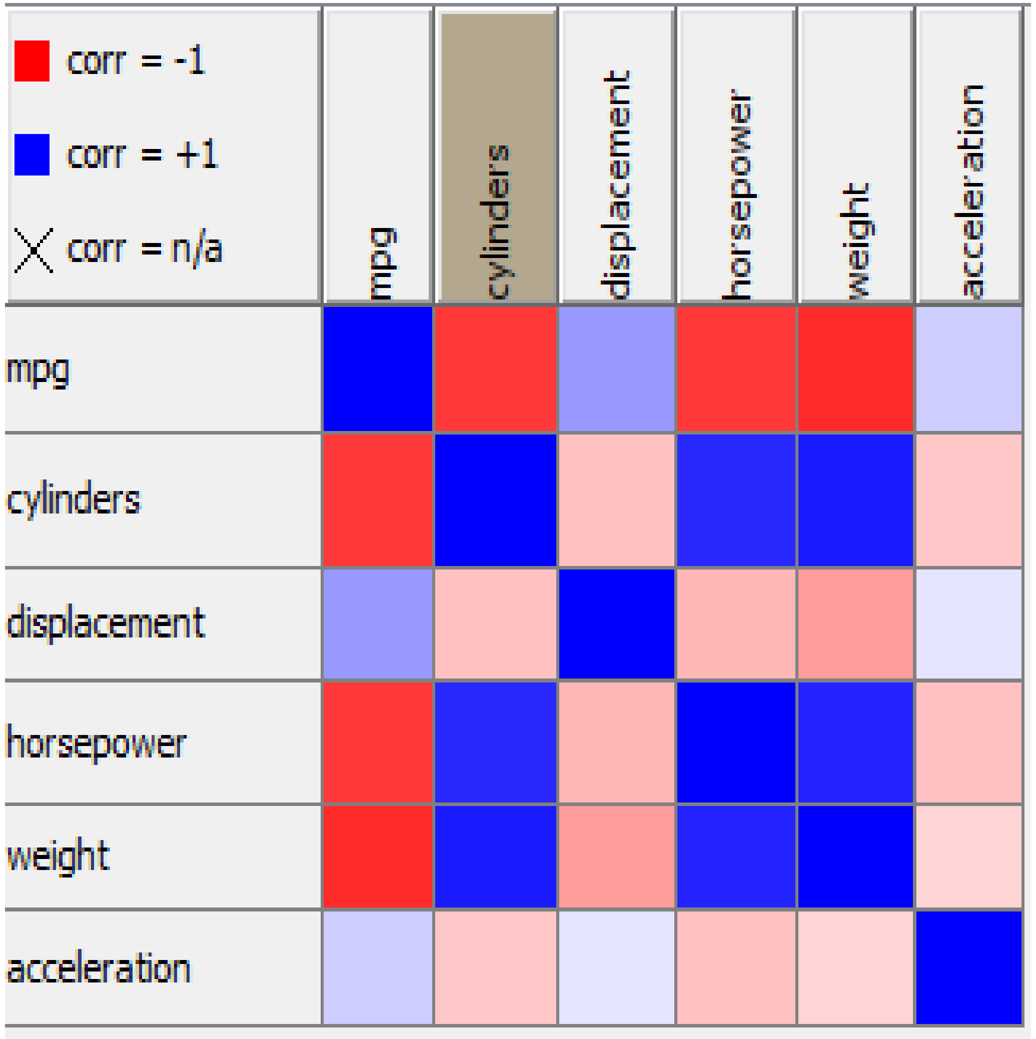


WORKFLOW



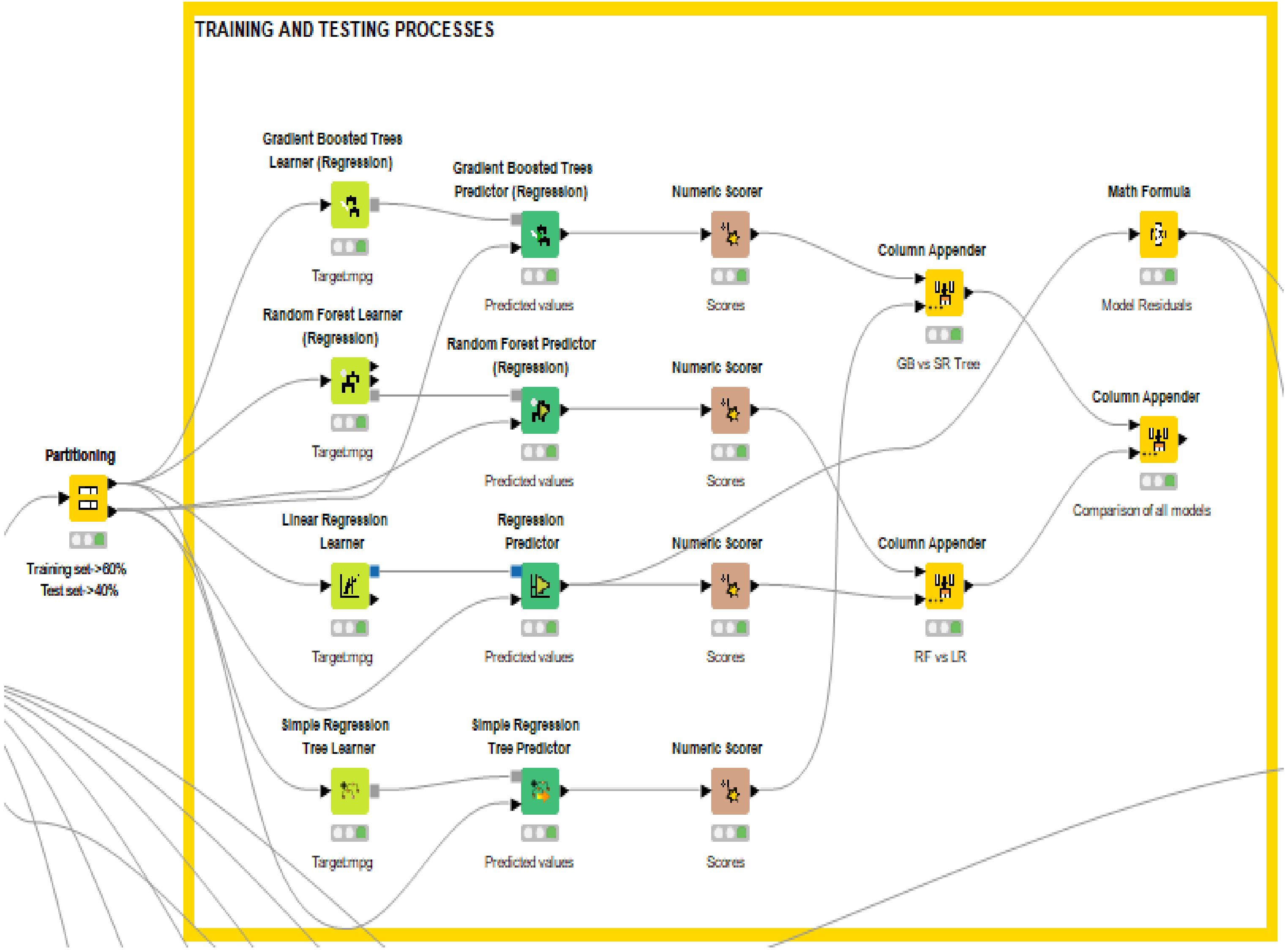
TRAINING AND TESTING PROCESSES
Gradient Boosted Trees Learner (Regression)
Target:mpg
Gradient Boosted Trees Predictor (Regression)
Predicted values
Numeric Scorer
Scores
Column Appender
GB vs SR Tree
Math Formula
Model Residuals
Random Forest Learner (Regression)
Target:mpg
Random Forest Predictor (Regression)
Predicted values
Numeric Scorer
Scores
Column Appender
Comparison of all models
Partitioning
Training set->60%
Test set->40%
Linear Regression Learner
Target:mpg
Regression Predictor
Predicted values
Numeric Scorer
Scores
Column Appender
RF vs LR
Excel Reader
Auto Dataset
Simple Regression Tree Learner
Target:mpg
Simple Regression Tree Predictor
Predicted values
Numeric Scorer
Scores
DIAGNOSTICS FOR THE FITTEST MODEL
Linear Correlation
Multicollinearity
Scatter Plot
Homoskedasticity
Histogram
Normality
VISUALIZATION OF THE VARIABLES
Histogram
mpg
Histogram
cylinders
Histogram
displacement
Histogram
horsepower
Histogram
weight
Histogram
acceleration

A QUICK TEST FOR MULTICOLLINEARITY BEFORE STARTING THE REGRESSIONS



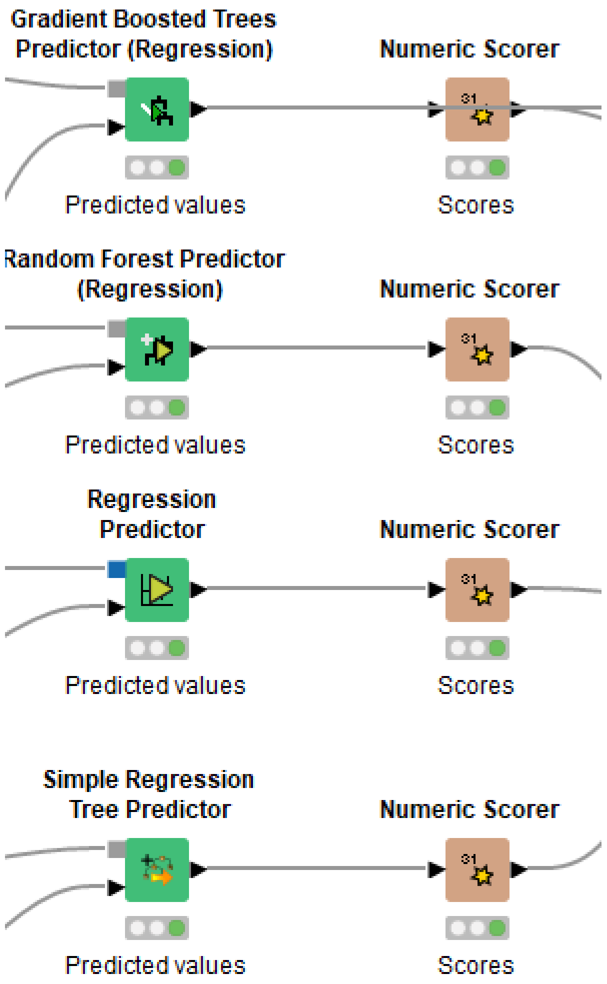
- Before proceeding to the next step, firstly, we should satisfy our first assumption in regression models that is having **no multicollinearity**. When we examine the collinearity table, we can see that **horsepower (HP)** and **weight** are strictly non-multicollinear with our dependent variable MPG. Thus, inclusion of HP and weight as independent variables are feasible in the models. Exclusion of **displacement** and **acceleration** is preferable.
- Also, **cylinders** are not included in the regression models as it is either 8 or 4, so it does create enough volatility for the proper conduction of the models.

SETTING TRAINING AND TESTING SETS



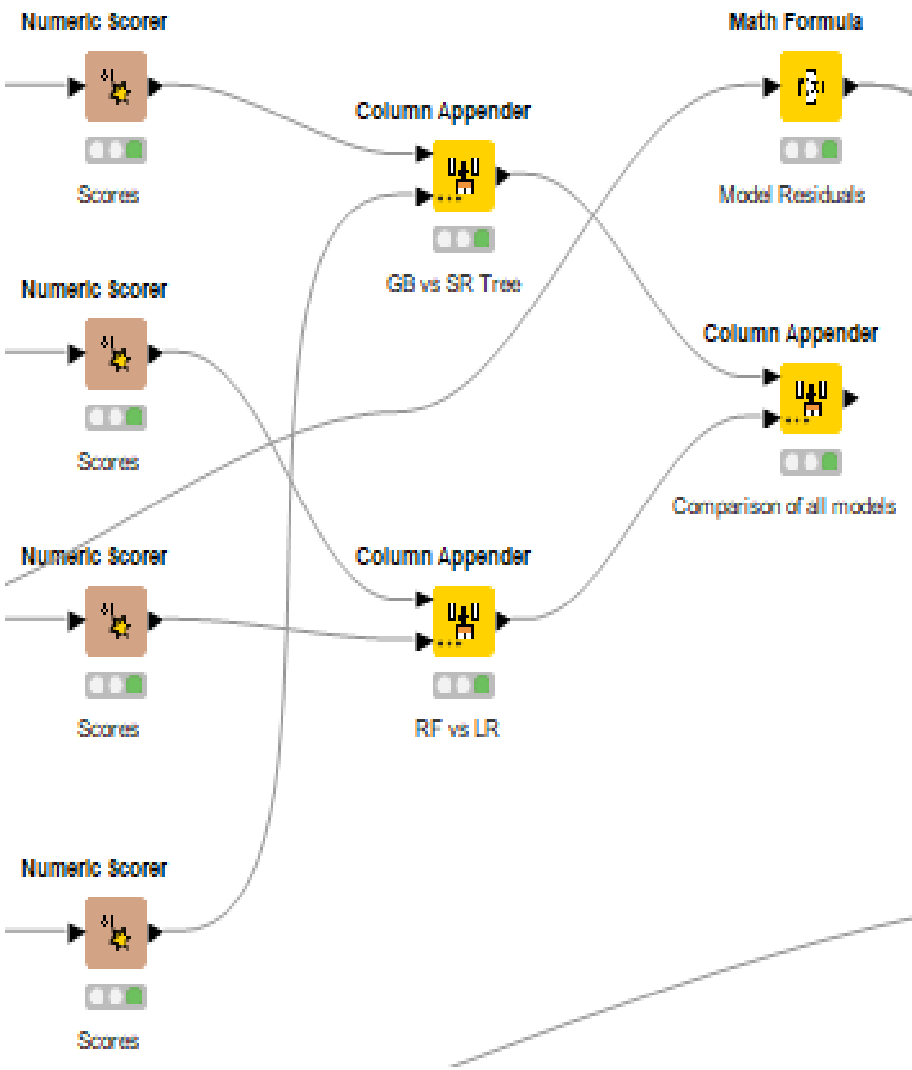
- The dataset partitioned into the training set and test set with the distribution of 60% and 40%, respectively. The ratio is chosen due to the limited number of observations.
- Models are conducted with learner functions using training set:
  - Gradient Boosted Trees (GB)
  - Random Forest (RF)
  - Linear Regression (LR)
  - Simple Regression (SR)
- Then test sets are conducted by using predictor functions of the same models.

USING THE NUMERIC SCORER FUNCTION



- Numeric scorers generated the key factors in regression models such as R-squared, adjusted R-squared etc.

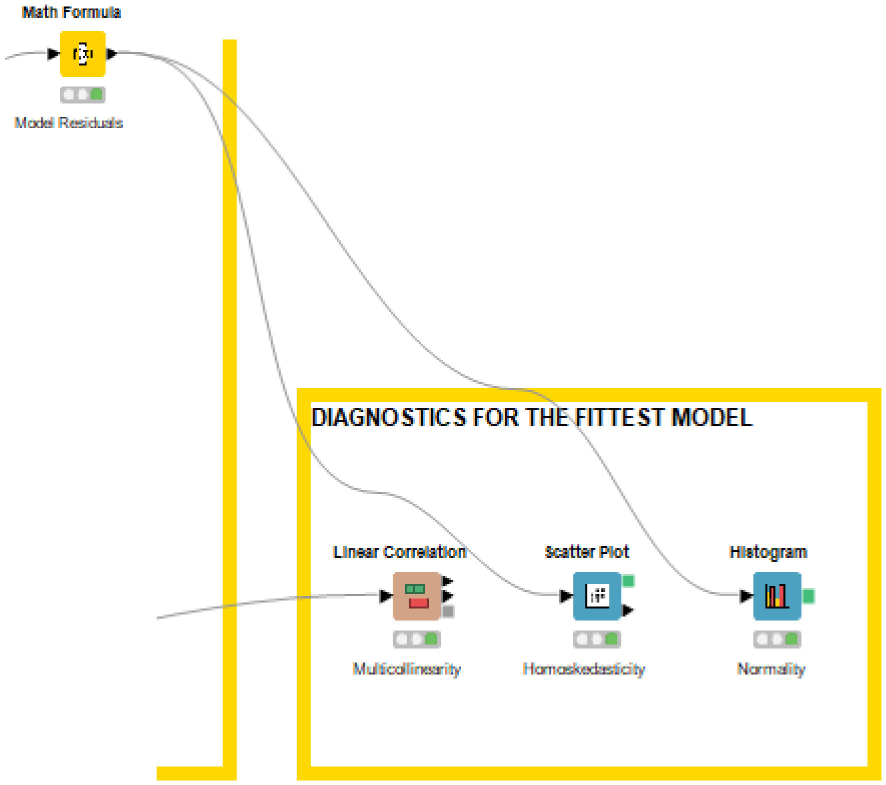
STORING AND COMPARING THE RESULTS



Row ID	D Prediction (mpg) GB	D Prediction (mpg) SR	D Prediction (mpg) RF	D Prediction (mpg) LR
R^2	0.671	0.501	-261.7	0.722
mean absolute error	32.975	41.656	789.673	32.143
mean squared error	1,947.963	2,957.758	1,555,970.22	1,649.105
root mean squared error	44.136	54.385	1,247.385	40.609
mean signed difference	-3.456	-3.019	771.993	-0.643
mean absolute percentage error	0.144	0.181	3.191	0.145
adjusted R^2	0.671	0.501	-261.7	0.722

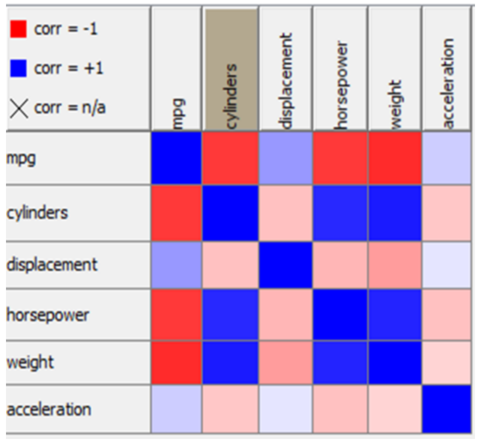
- By analyzing row by row, R-squared and adjusted R-squared is the highest in **Linear Regression** model, which means that HP and Weight variables can explain 72.2% of the sample variation in MPG.
- When we looked at the error factors, overall, **Linear Regression** has the lowest mean errors compared to GB,SR and RF. This means that prediction errors are low and predicted values are reliable.
- Ultimately, Linear Regression (LR) seems to explain more of the sample variation in MPG and has lower mean errors in model predictions.

DIAGNOSTICS FOR THE FITTEST MODEL

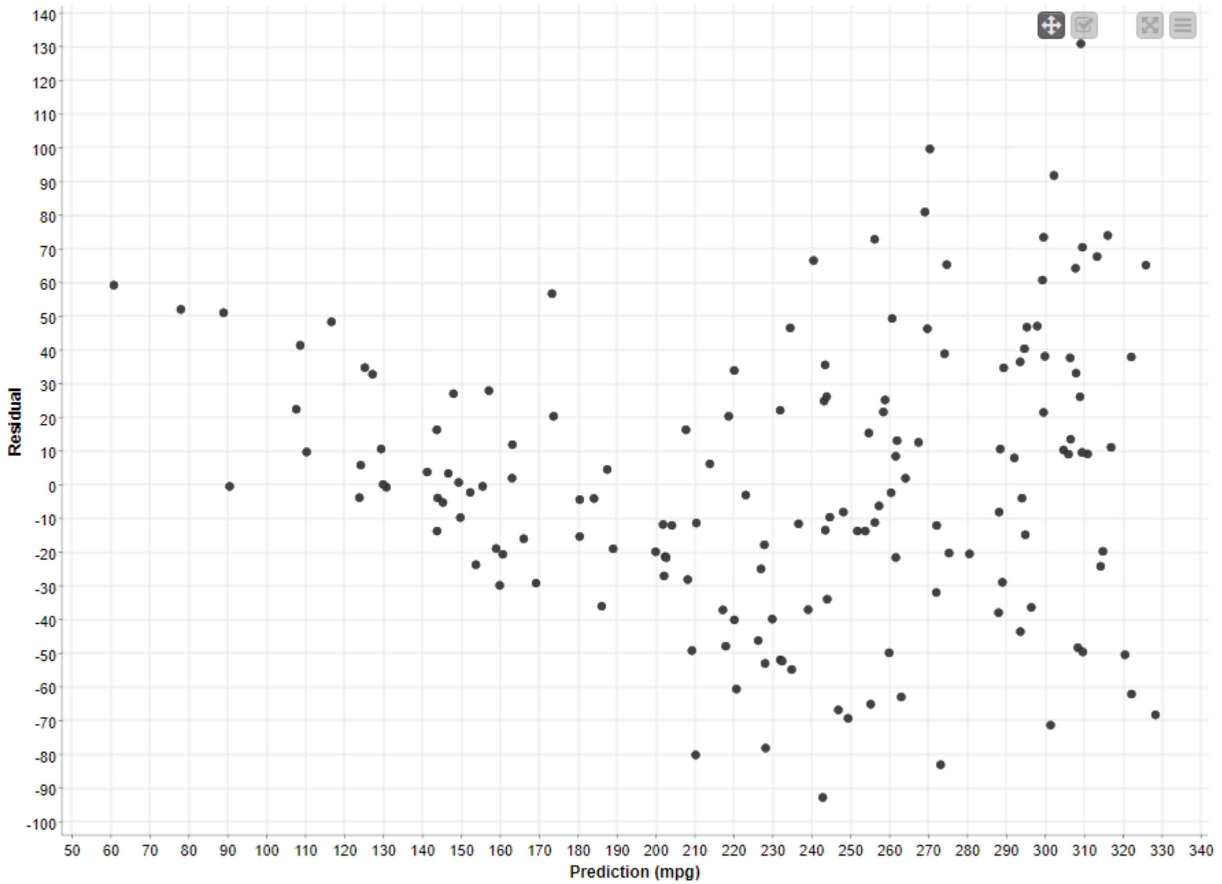


- To utilize the functions for the Linear Regression model, model residuals are computed by creating math function that subtracts the predicted values from observed values.

- For the **multicollinearity** assumption, it has shown in previous slides that regressing MPG, HP and Weight does not violate the assumption.



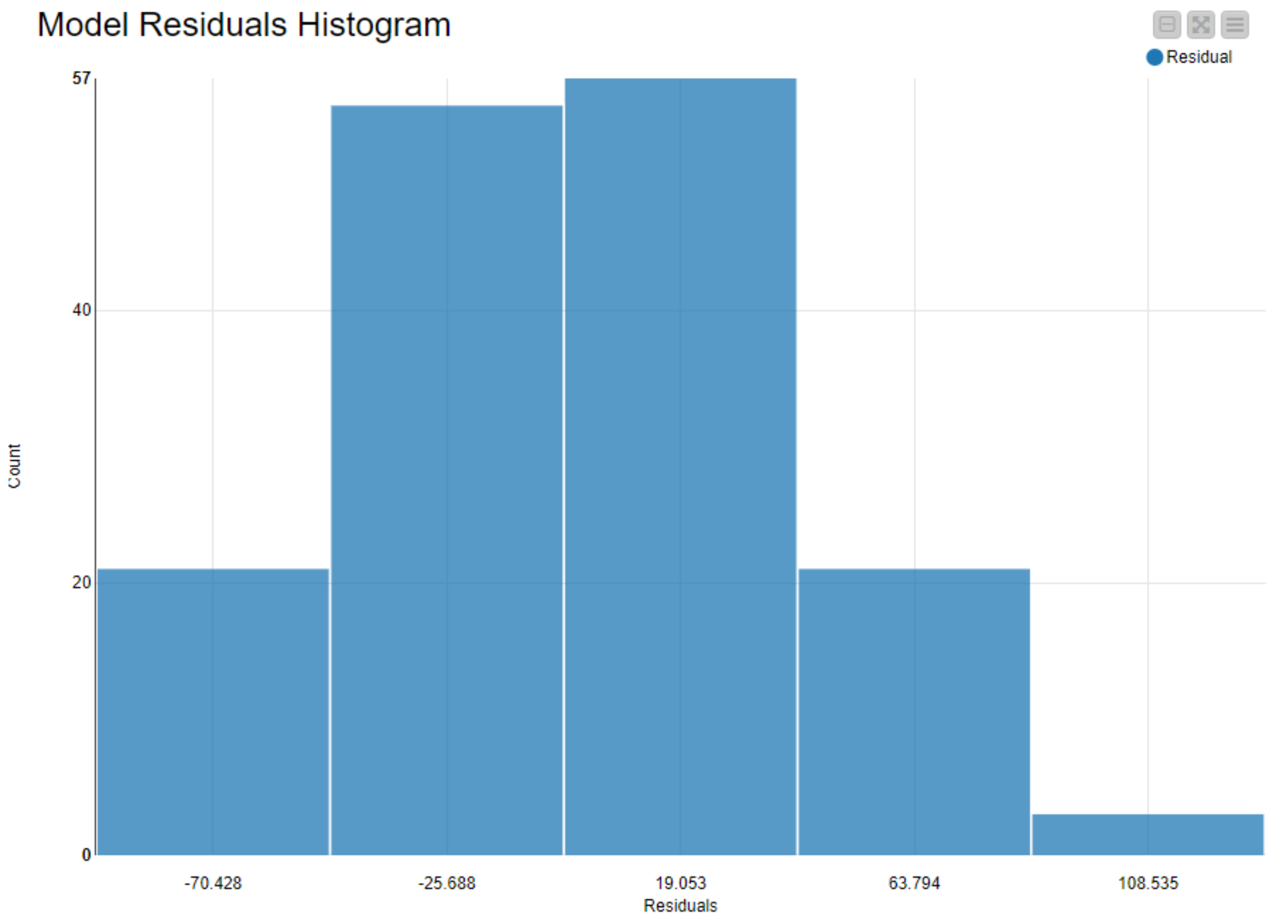
DIAGNOSTICS FOR THE FITTEST MODEL



- We can see from the graph that shows the model residuals and prediction of MPG values. The distribution of the plots indicates that there is **no heteroskedasticity** as the data points show no pattern.

- Thus, homoskedasticity assumption is also verified.

DIAGNOSTICS FOR THE FITTEST MODEL



- We can see that model residuals are very close to **normal distribution** as they follow the pattern of normally distributed values.
- Therefore, we can conclude that normality assumption is also satisfied.

LR MODEL COEFFICIENTS FOR TRAINING SET

Statistics on Linear Regression

Variable	Coeff.	Std. Err.	t-value	P>|t|
horsepower	-0,5623	0,1502	-3,7431	0,0002
weight	-0,0535	0,0067	-7,9916	6,24E-14
Intercept	452,3968	10,3147	43,8595	0.0

R-Squared: 0,6954
Adjusted R-Squared: 0,6928

- We can see that 1 unit increase in horsepower (HP) leads to decrease in Miles per Gallon by 0.56 units significantly as p value is lower then 5%. So, as expected, higher engine power, lesser miles travelled with constant fuel level.
- For the weight, 1 unit increase in car weight decreases the Miles per Gallon by 0.05 units in the sample data significantly as it is much lower than 5% significance level, which confers us that greater weight leads to shorter travel miles with constant fuel level.
- Finally, it is seen that Adjusted R-squared is different from our previous model result, which is 72.2%, it is because this results are only for the training set and does not include the test set.

CONCLUSION

- Among the models that I employed, the most parsimonious and robust model is found to be **Linear Regression Model**, it is done with 3 variables: **MPG** as dependent variable, **HP** as main independent variable and **Weight** as other explanatory variable.
- According to the model diagnostics, there is no bias in model coefficients and they are highly significant.
- As a limitation of this study, more recent car data needs to be shared with public with larger observation number. Also, the car data should include electrical cars as well to compare their miles per watt performance.
- The model supports the consensus of higher HP of internal-combustion engine cars and greater weight leads to lower MPG directly, and indirectly paves way to more demand for fossil fuels and more CO_2 emissions from the conventional cars.
- This study guides us that electrical vehicle cars will greatly diminish the CO_2 and save the future for generations.

THANK YOU FOR YOUR ATTENTION !

ELIF HACIOSMANOGLU

Erasmus Student