Project Report

|  |  |
| --- | --- |
| **Course Name (NICF)** | NICF Diploma in Infocomm Technology (Data) |
| Product Name (Marketing & Sales) | Professional Diploma in Data Science |
| **Module Name (NICF)** | NICF Basic R Programming (SF) |
| Product Name (Marketing & Sales) | Basic R Programming |

|  |  |  |  |
| --- | --- | --- | --- |
| Student name | | Assessor name | |
| Tan Chee Kuang | |  | |
| Date issued | Completion date | | Submitted on |
| 11 Apr 2022 | 11 Apr 2022 | | 11 Apr 2022 |
|  | |  | |
| Project title | Design and deploy Forecasting Model | | |

|  |
| --- |
| Learner declaration |
| I certify that the work submitted for this assignment is my own and research sources are fully acknowledged.  Student signature: Date: 11 Apr 2022 |

Content

1. **Project Overview**

**Activity 1:** Define data management structures to align and streamline processes of data ownership, retrieval, combination and usage

1. **Activity 2:** Create R code for effective data loading, storage and utilization

**Activity 3:** Perform data handling standards and procedures

**Activity 4:** Identifying numeric variables

**Activity 5:**Data management tools/standards for Correlation Matrix and Correlated attributes

**Activity 6:**Propose Model Creation

**Activity 7:**Plot model

**Activity 8:** Establish internal processes to Calculating Model Performance ,monitor compliance of data with relevant metrics procedure

**Activity 9:** Predict mpg

**Activity 10:** AML rules and guidelines to ensure proper adoption and adherence of same R program in AML

**Project Overview:**

The Department of transport has provided a data set containing car manufacturers data. In the provided data, we can find 32 observations, each containing 11 variables.

Our goal of this project is to create a model each in R programing and, in Azure machine learning, that is capable of predicting the cars miles per gallon value(mpg) based on some of its 10 others remaining variables.

We would do so by first analyzing this given set of data structure and summary. We would then clean and re-class the data type as needed. With this cleaned data-set to run some tests to determine which of the independent variables were most suitable to be used for our linear model building. We would then build a few models each with some variations and choose the most appropriate model best on the statistics.

Finally, we would train the models to predict the mpg.

Project Technical Environment:   
In this project, we will be provided with a data-set “mtcars.csv”. This data-set will contain 32 observations each with 11 variables. We will be using 2 platforms:  
  
R programing (with corrplot, caret, ggplot library packages)  
Azure machine learning studio  
  
It is possible to create an AI predictive model with similar capability with or without programming skills. To demonstrate this, we would be performing the same task of producing a predictive model on each of the platforms.

1. Linear model

**Activity 1: Define data management structures to align and streamline processes of data ownership, retrieval, combination and usage**

1. Retrieve the data-set, mtcars within our learning portal.

2. Import the data-set, mtcars.csv into ‘R’ using the read.csv() function.

3. Install the required visualisation packages(ggplot2,caret, corrplot) using the install.package function

4. Analyze the data-set by using head(), tail(), str() and summary() functions

5. Factorize categorical independent variables and dropped defendant variable from data-set

6. Check correlation between dependent variables and remove the strongest related items.

7. Fit the remaining lowly correlated dependent variables into a linear model.

**Activity 2: Create R code for effective data loading, storage and utilization**

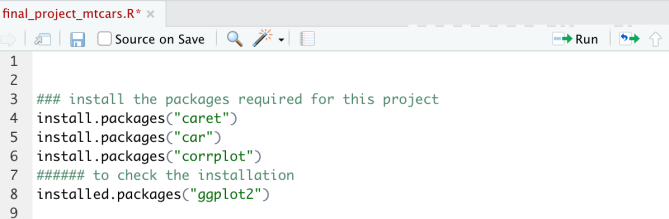
To start we will first Install the following packages:

1) Ggplot2,

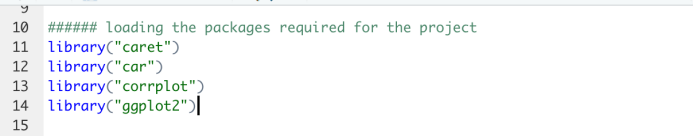
2) Car,

3) Caret, and

4) Corrplot.



We would then run the library command to load them:

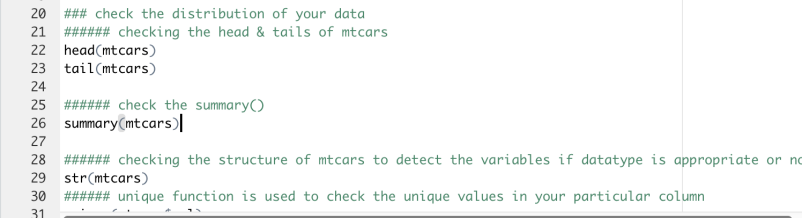


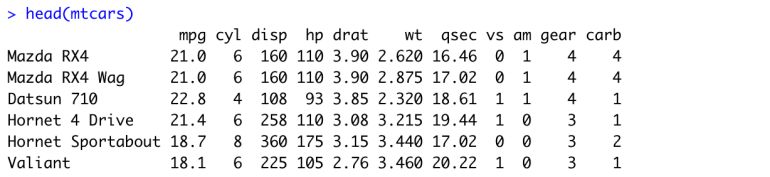
Next, we will load the data-set, “mtcars.csv”, which is provided by the Department of transport:

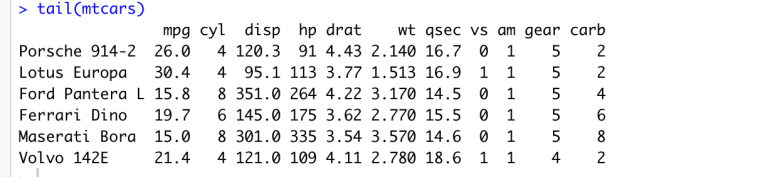


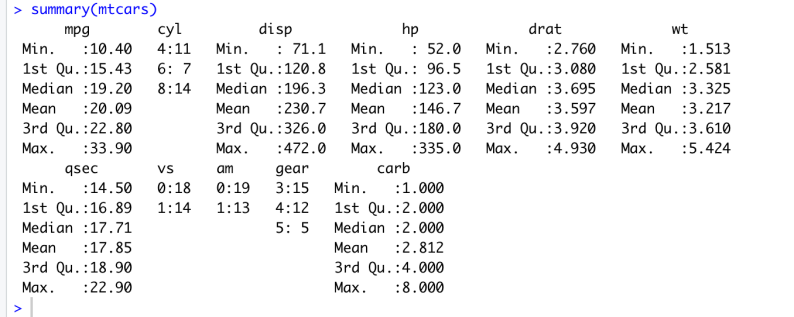
Upon loading the data-set, we will first use the functions head() and tail() to visualize the first and last few data observations.

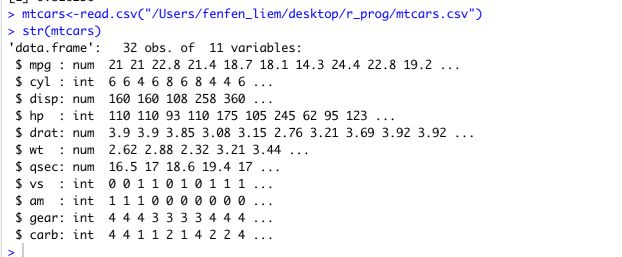
Next, we would use the summary() and str() functions to better understand the summary of data ranges and structural data types provided.











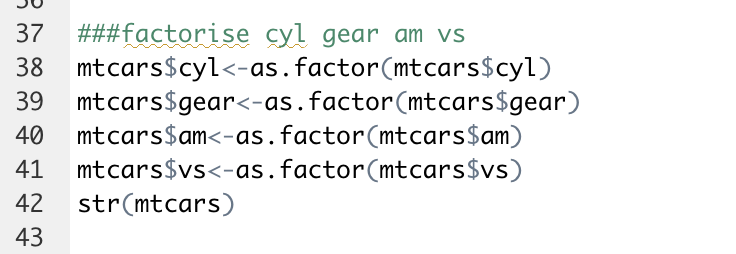
From the above functions we were able to determine that we have in our data set, 32 observations , each with 11 variables. Some of these data has been stored as numeric, while others were stored as integers.

**Activity 3: Perform data handling standards and procedures**

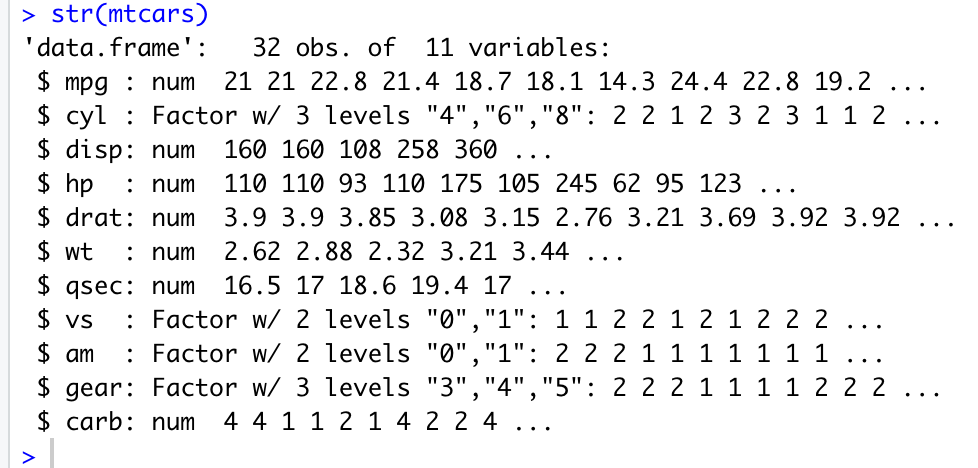
In order to continue processing our data from the last activity, we would need to perform some standard data cleaning procedures.

In mtcars, we have some fields such as cyl, gear, am and vs which were stored as integer. These variables have a fixed and known set of possible values. They can better represented as categorical variables rather than integers.

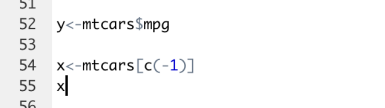
We would proceed to convert these 4 columns into factors:



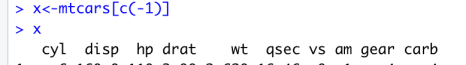
After the conversion, our data structure should look as follows:



Next, we will separate our data set into dependent(y) and independent(x) variables:



The result would show y contains only mpg, while x contains everything except for mpg:



Our next steps would be to fit the x variables into a linear regression model with y. An accurate linear regression model must fulfill all the following four assumptions:

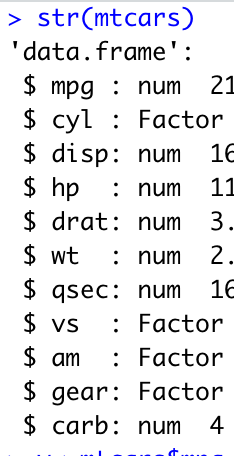
1. Existence of linear relationship between Dependent variable (MPG) and independent variable( Fitted values)
2. Independent variables have low correlation.
3. The residuals of the model are normally distributed.
4. Homoscedasticity of the residuals

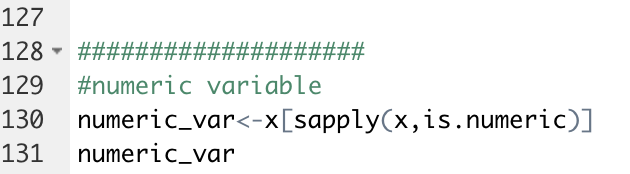
In our next activities 4 to 6, we will be taking steps to handle our data in such a way that rule 2(independent variables have low correlation) assumptions can be achieved. We will then create a model and check if the remaining 3 assumptions can be satisfied in activity 7.

**Activity 4: Identifying numeric variables**

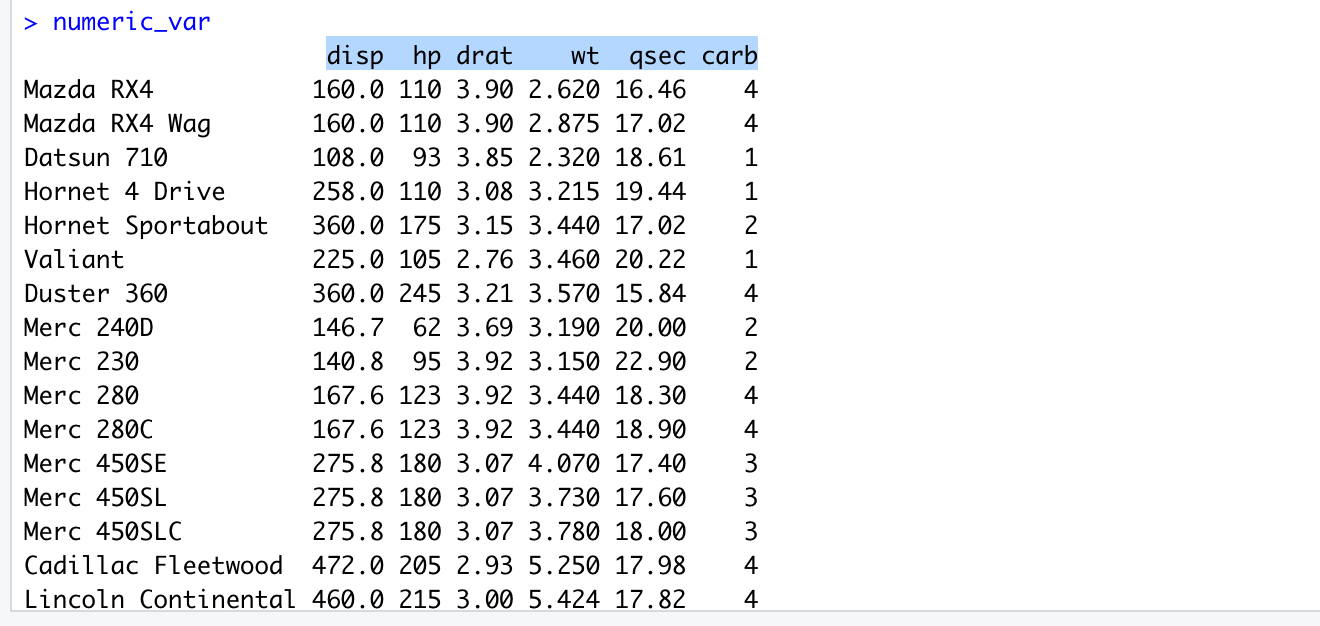
To fulfill the low correlation between independent variables assumption requirement, we will first identify the numeric variables which we are able to meaningfully compute the correlation value between them.

We use sapply function to assign all numeric variables into numeric\_var:



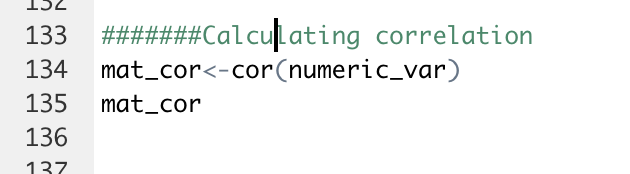


After this we had effectively removed our factor variables, should we have the remaining 6 numeric variables as follows:

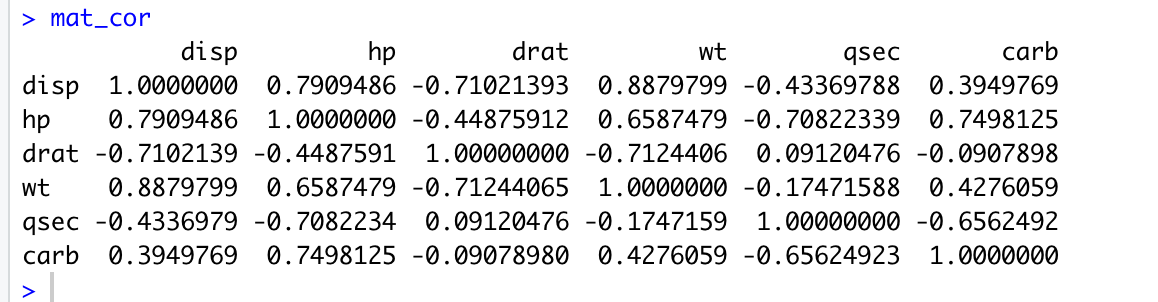


**Activity 5: Data management tools / standards for Correlation Matrix and Correlated attributes**

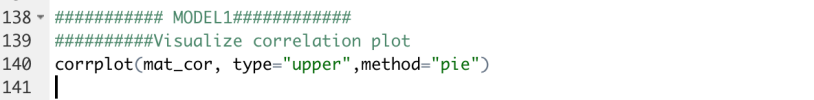
Having separated the numeric variables, we will next perform a correlation calculation by using the cor() function and assigned it to mat\_cor:



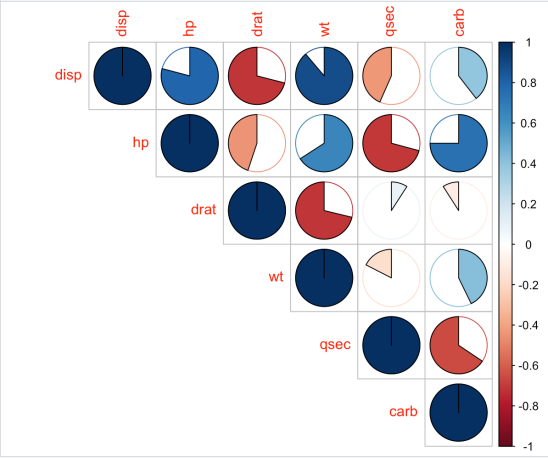
Our correlation matrix will look as follows:



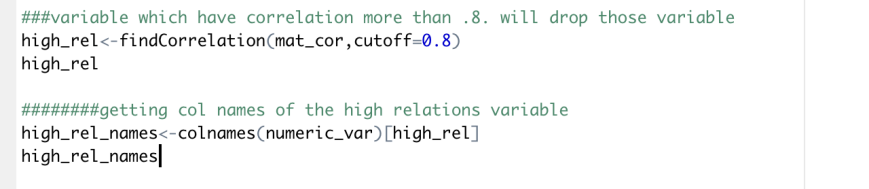
Next we will visualize the correlation using the plot() function.



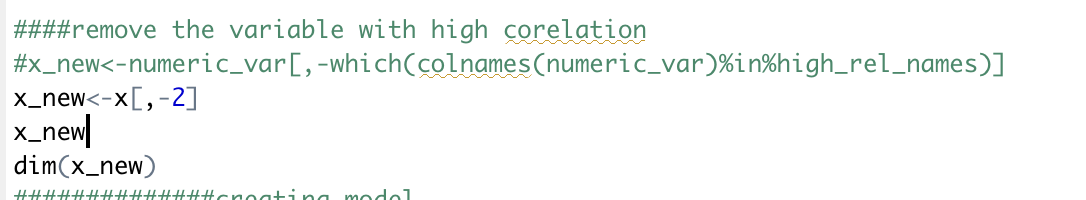
We have added the type=”upper” parameter to remove the duplicated bottom half of the graphical representation. Also we have added the method=”pie” parameter, to make it easier to see the strength of relationship:



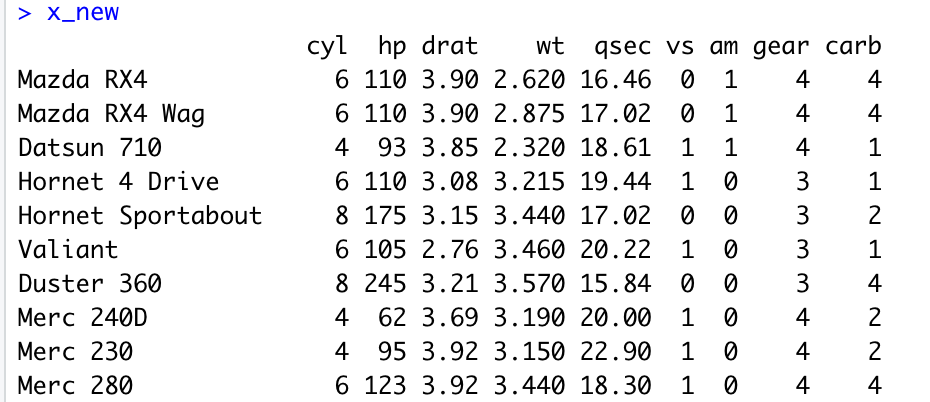
With the cor-relationship matrix plotted, we next use the find() function to set various cutoff points to identify and variables with high correlations and remove them:



In this case our variable “disp” was removed due to it having over 88% correlation to “wt” and more than 79% with “hp”



Our data of x variables are as follows after removing the “disp”:

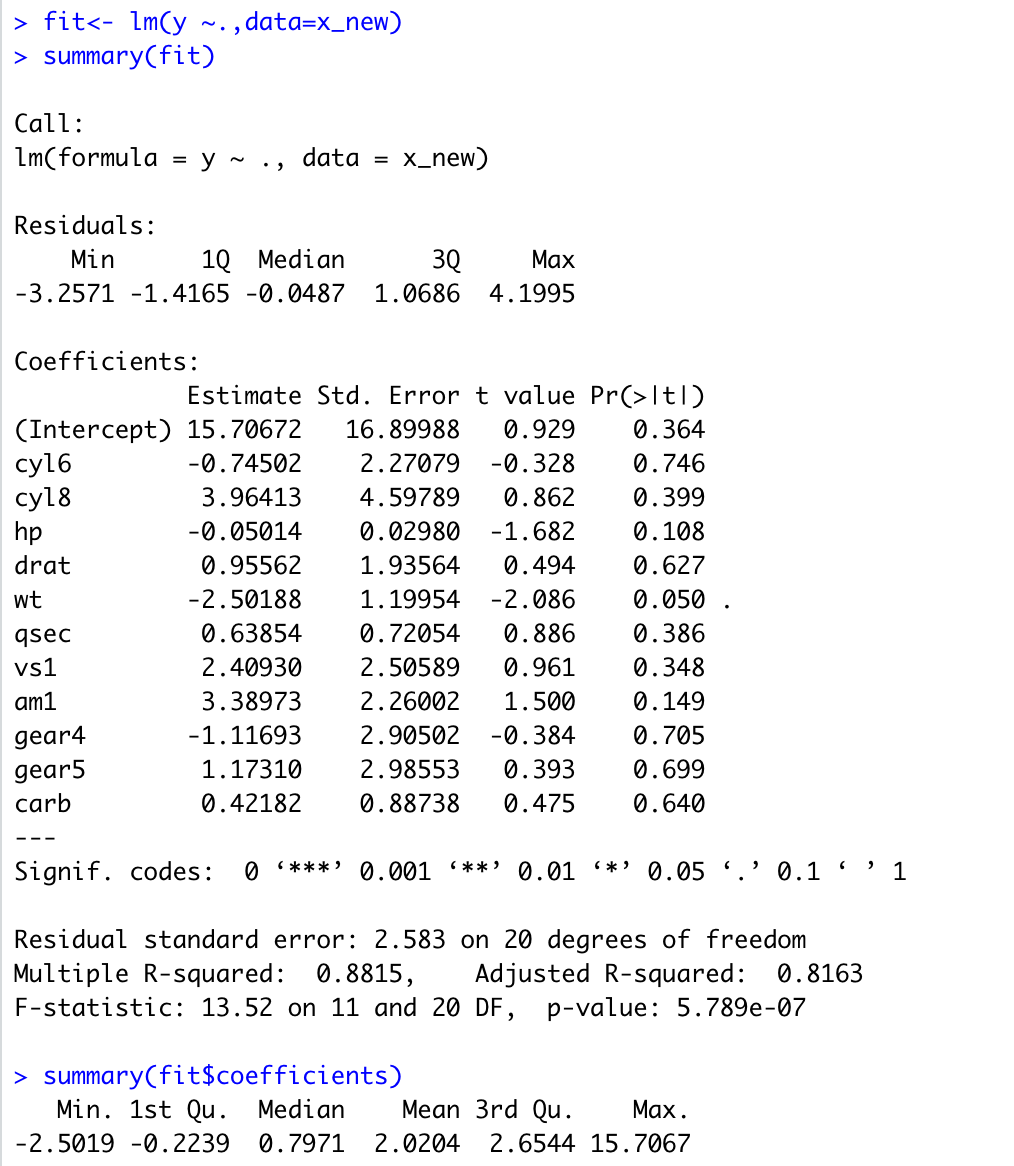


**Activity 6: Propose Model Creation**

Now that we have removed the most highly correlated independent variables, we can use the lm() function to find a linear relationship between mpg and the fitted values.

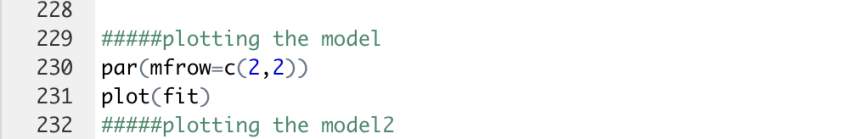


Our model was able to return a 81% accuracy in prediction with coefficient between -2.5 to 15.7, our residuals centers around 0:

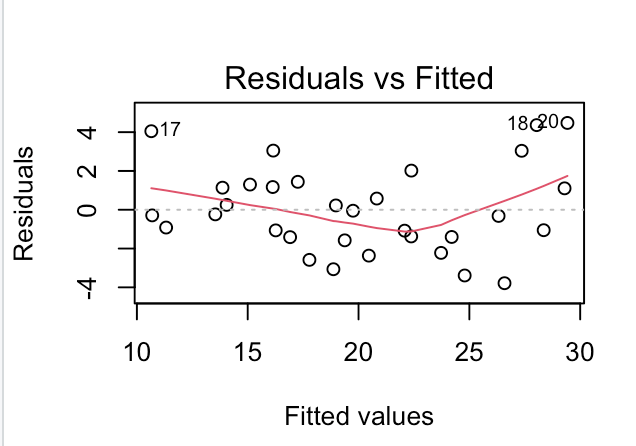


**Activity 7: Plot model**

Next we will set par() to display into a 2x2 matrix showing the suitabiity of linear model for our case:

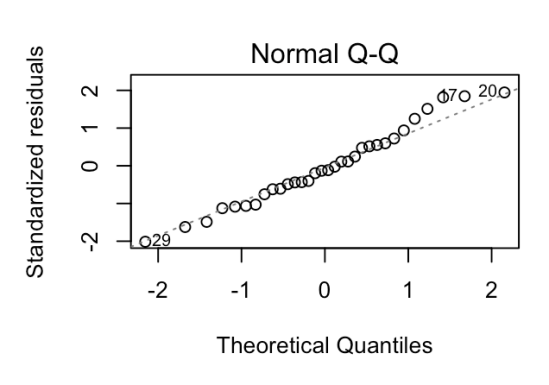


**Residuals vs fitted**



In our first chart, residuals vs fitted, we are trying to detect non linearity and outliers. The residual bounce randomly around zero line suggests that the linear relationship is reasonable. In this case, our assumption 1 has been fulfilled.

**QQ Plot**

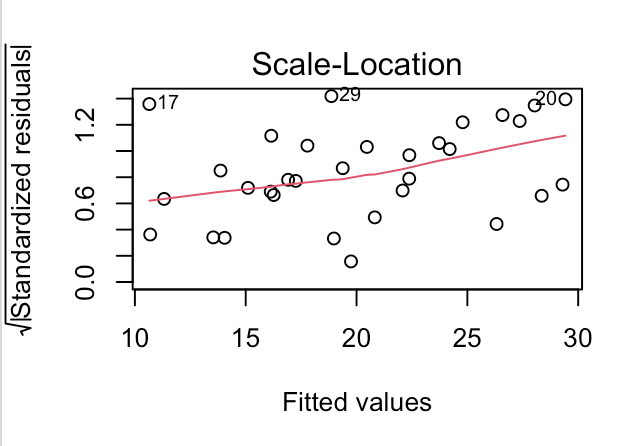
****

Our second chart is a QQ plot. This is a test that our data has a normal distribution.

The QQ plot tests the quantiles from mpg against the quantiles from estimated responses. The slight curved at right ends of our plot suggest that our data is a little skewed, but not far from normal distribution. But the slight flaw is very negligible.

We can conclude that assumption 3( normal distribution ) has also been met.

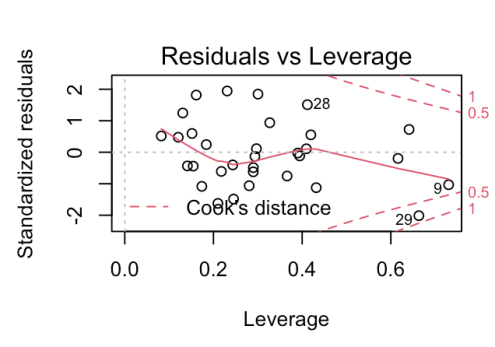
**Scale location plot**

****

Our third plot is a Scale Location plot.

The red line spread horizontally across the plot suggest that the homoscedastcity assumption (assumption 4) has been met for our regression model.

**Residual vs Leverage plot**

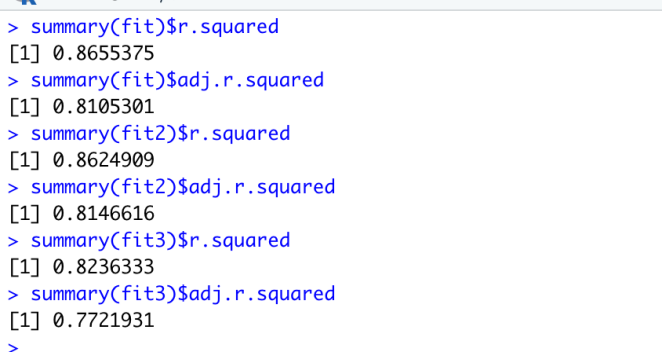


Our fourth plot is a residual vs leverage plot. The purpose of this plot is to identify influential observation. Observations appearing beyond the cooks distance suggest high leverage.

A high leverage observation,if removed, will greatly affect coefficient of the regression model.

**Activity 8: Establish internal processes to Calculating Model Performance, monitor compliance of data with relevant metrics procedure**

For the purpose making more accurate prediction, we have created 3 models (repeating activities 5 to 7 with cutoff 0.80 , 0.75 and 0.70).



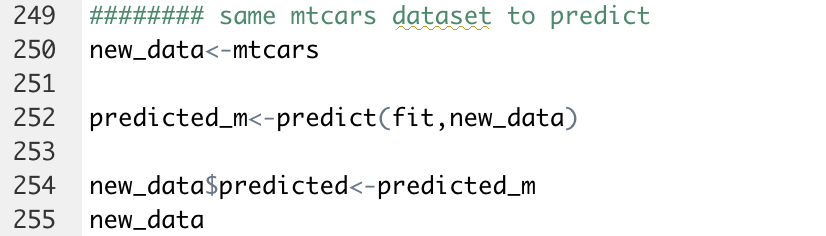
Of the 3 model we have decided to use our first model which has the following R square and adjusted R square values:

R square value = 0.8814555

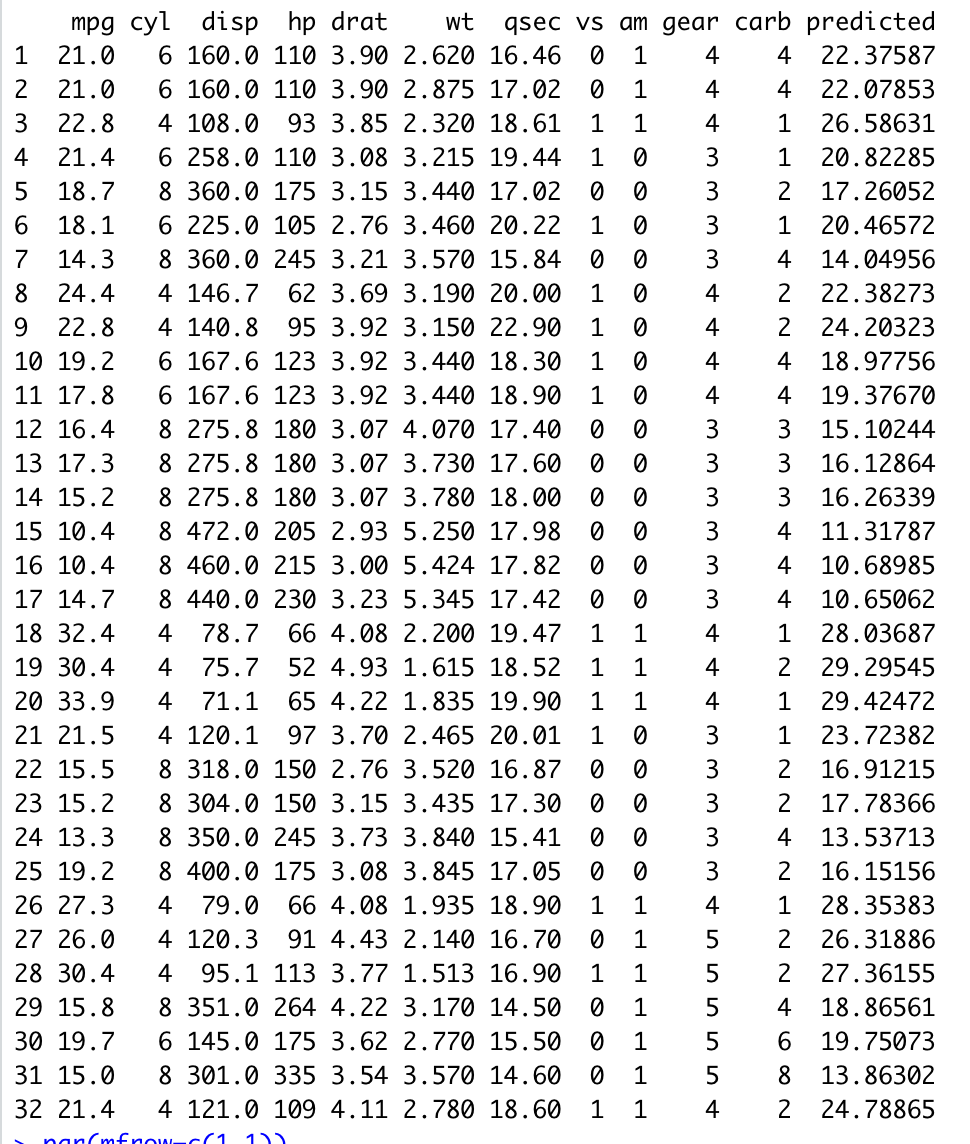
Adjusted R square value = 0.816256

**Activity 9: Predict mpg**

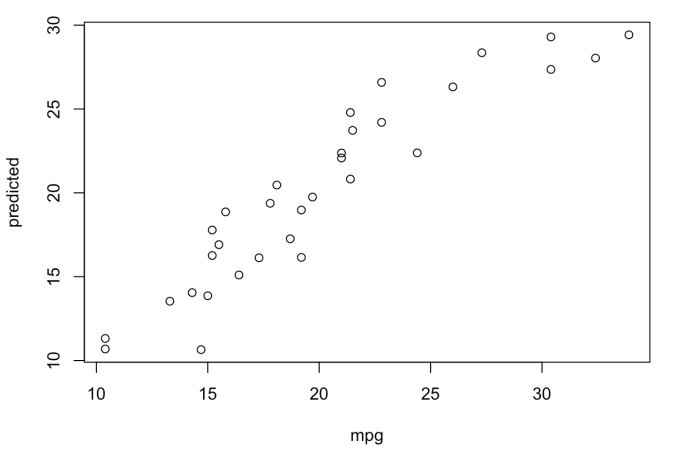
Next we will use the predict() function to predict our mpg values against our independent variables:



Results:



The final prediction vs actual can be represented as follows:

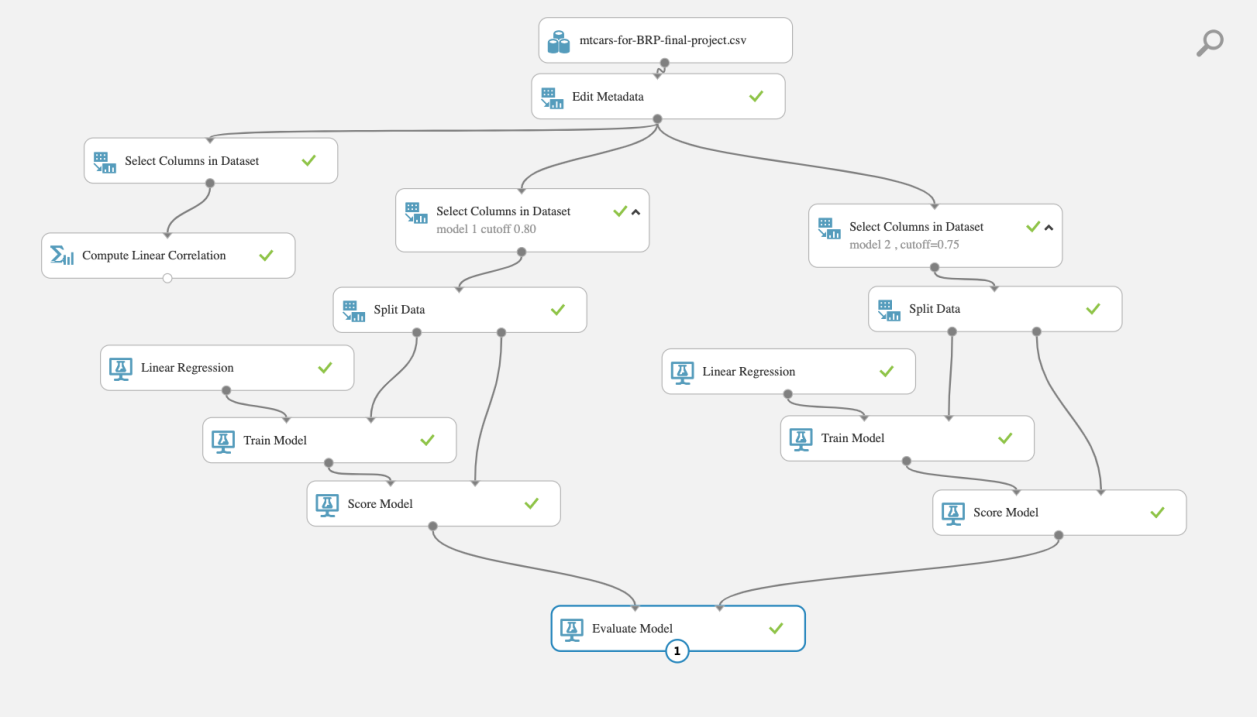


**Activity 10: AML rules and guidelines to ensure proper adoption and adhernece of same R program in AML**

With Azure machine learning, we are able to produce a similar predictive model without the use of programming language.

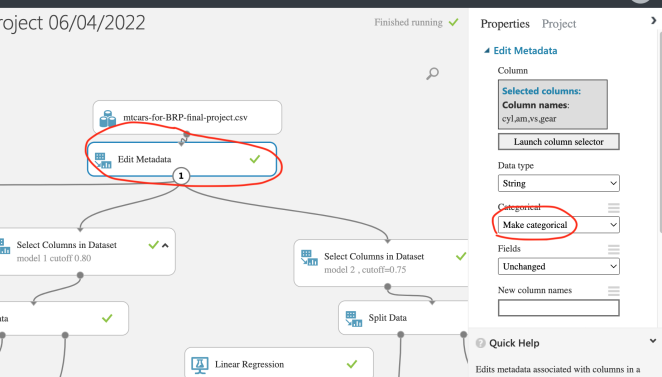
We will next create a the same predictive program using Azure machine learning studio as the tool.

Below is the overview of our setup:

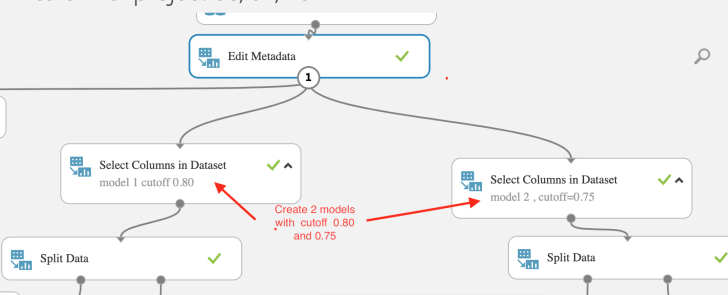


Our first step after importing the data-set is to Edit metadata.

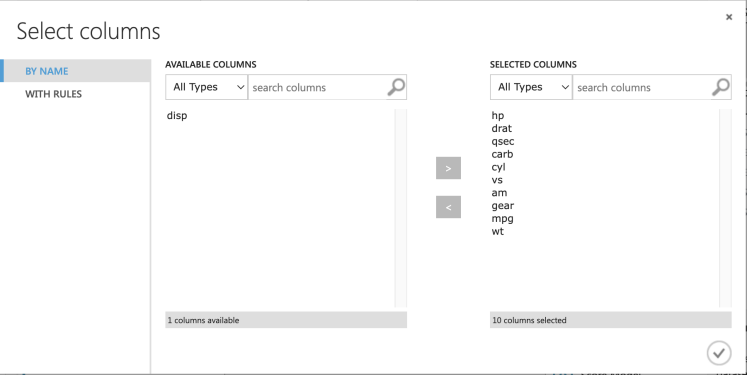
Here we would choose cyl, am, vs and gear to make them string data and make them categorical. This step is similar to what we did in activity 3, factorizing the data in R.



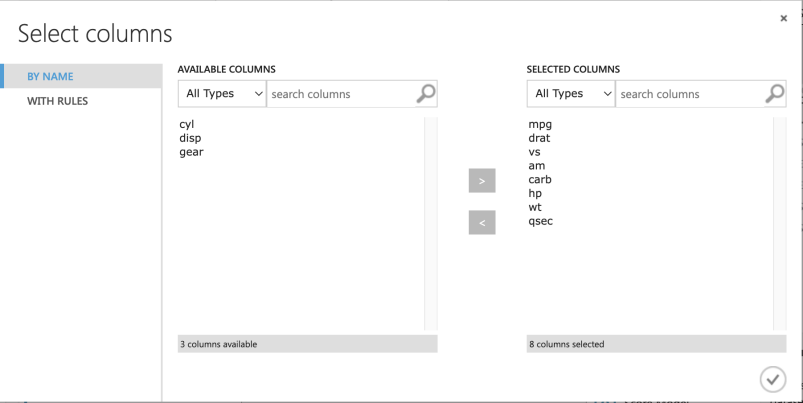
Our next step is to select columns in the data



For our first model, we have included everything except “disp” to replicate our model 1 in R:

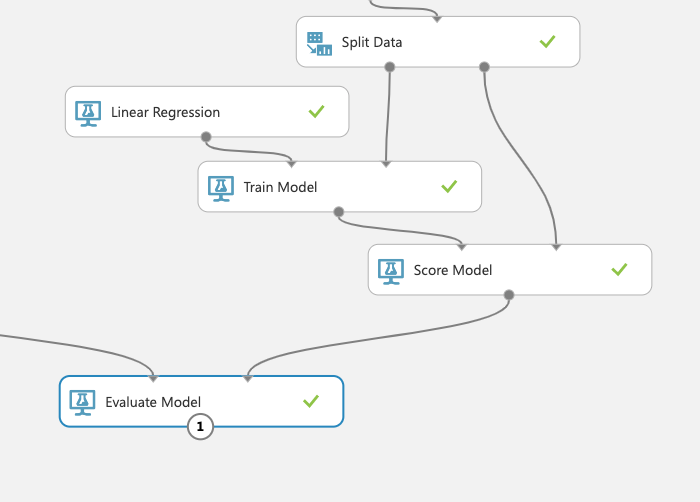


Similarly in our second model we have selected everything but exclude “cyl”, “disp” ,“gear” to replicate model 2 in R:

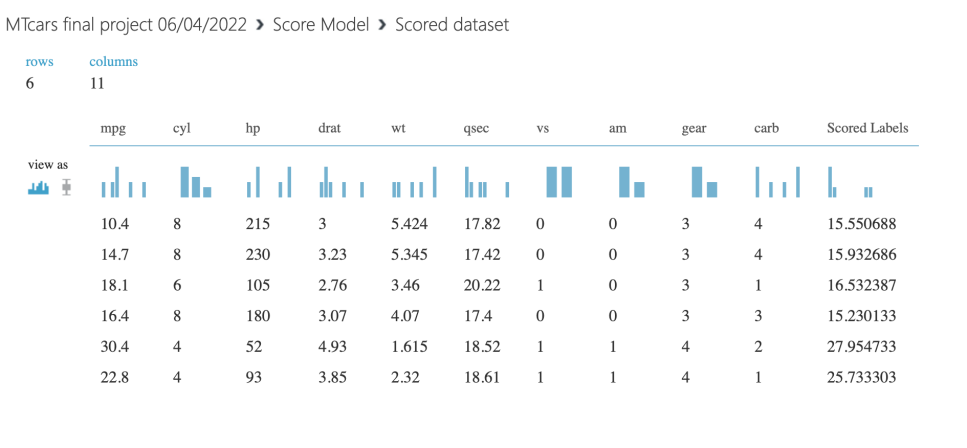


Next for both models we have split the data into 80% : 20% ratio

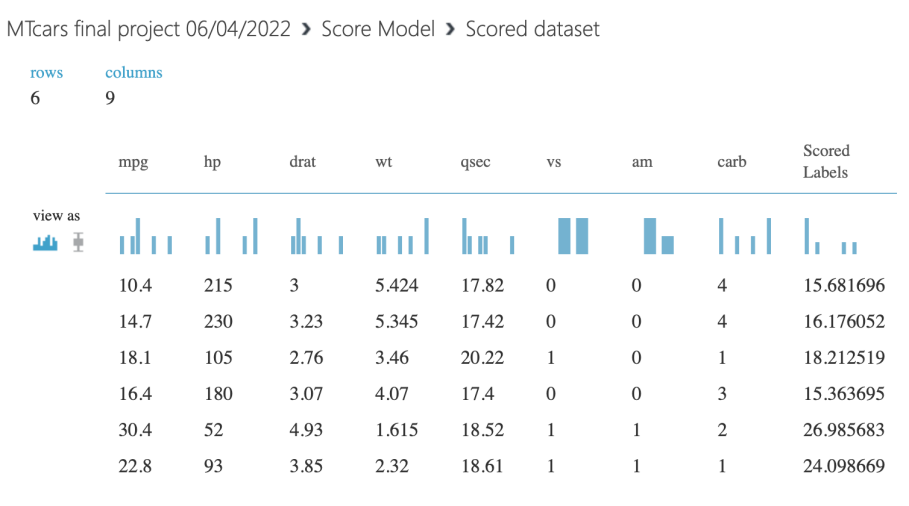
We used 80% of our data to train the AI with a linear regression logic and our 20% of our data was used to score our model



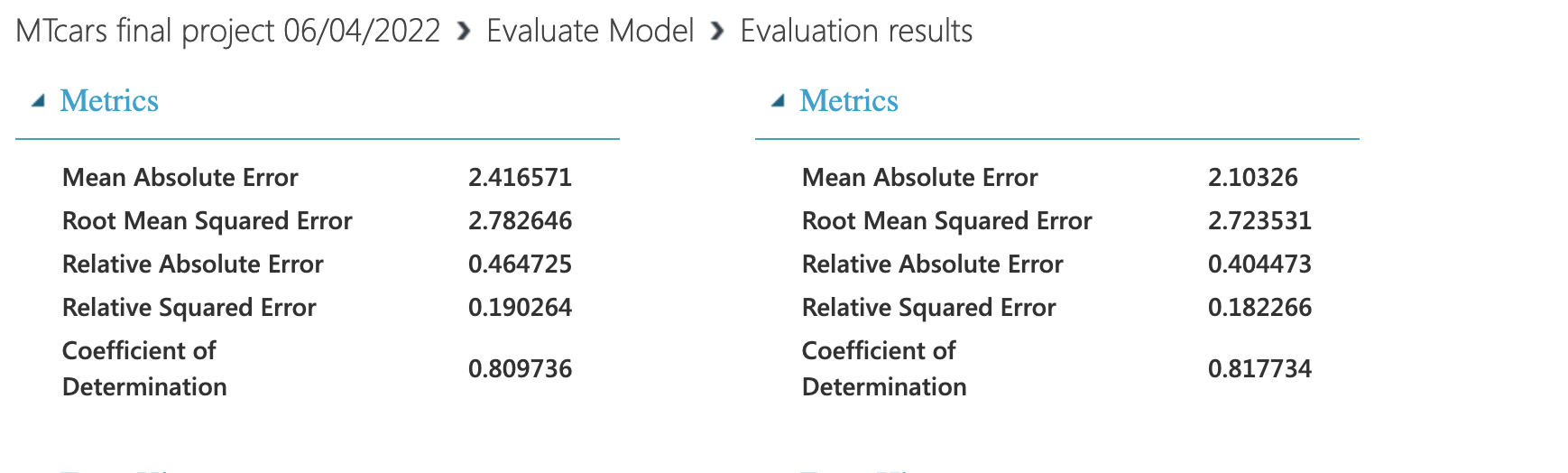
**Model 1 actual mpg vs predictions**

****

**Model 2 actual mpg vs predictions**

****

Our results shows model 2 is slightly more accurate than model 1



With Azure machine learning studio, we were able to replicate the predictive model in R with very similar result on a GUI that does not require programming language.