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 | |
| Lee Jack Shiang | | Ryan Suryanto | |
| Date issued | Completion date | | Submitted on |
| 05-May-2022 | 17-May-2022 | | 17-May-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: 17-May-2022 |

Table of Contents

[1. Project Overview 3](#_Toc61938196)

[2. Project Technical Environment 4](#_Toc61938197)

[3. Activity 1 5](#_Toc61938198)

[4. Activity 2 5](#_Toc61938199)

[5. Activity 3 8](#_Toc61938200)

[6. Activity 4 9](#_Toc61938201)

[7. Activity 5 9](#_Toc61938202)

[8. Activity 6 11](#_Toc61938203)

[9. Activity 7 13](#_Toc61938204)

[10. Activity 8 14](#_Toc61938204)

[11. Activity 9 14](#_Toc61938204)

[12. Activity 10 16](#_Toc61938204)

[13. Annexure - 01 18](#_Toc61938205)

[14. Annexure - 02 29](#_Toc61938205)

[15. Annexure - 03 33](#_Toc61938205)

[16. Annexure - 04 34](#_Toc61938205)

[17. Annexure - 05 37](#_Toc61938205)

[18. Annexure - 06 39](#_Toc61938205)

# 1. Project Overview: Describe the Project along with Project Outcomes (Explain the Project in your own words in 15 – 20 lines) For this project the mtcars dataset will be used for the study. The data of mtcars is contained in a csv file. These data are derived car manufacturer information published by the Department of Transport. In this project, linear regression model will be utilized in both R program and Azure Machine Learning (AML) environment.

# This project serves as Summative Assessment in the Module ‘Basics in R Programming’ of the NICF Diploma in Infocomm Technology (Data) course.

# It assesses the skills acquired to write a program, create visualization plots and build a linear model using the R programming language.

# A Project Brief document, “M2\_NICF\_Basic\_R\_Programming\_SF\_\_Project\_3”, was provided to specify the purpose of the project, its outcomes and deliverables.

# The data set, “mtcars.csv”, was provided. The data was from the 1974 Motor Trend US Magazine and comprises fuel consumption in miles per gallon (mpg) and 10 aspects of automobile design and performance for 32 automobiles (1973-74 models).

# A project report template, “Basic\_R\_Programming-Project\_Report\_Template.docx”, was provided to specify the structure and content required of this Project Report.

# The objectives of the project are to:

# - develop an R-script using RStudio to create a linear model to forecast the mpg of automobiles; and - perform the same activities on Microsoft Azure Machine Learning Studio

# Information for the mtcars dataset:

# A dataset with 32 observations (rows) and 11 variables (columns):

# mpg Miles/(US) gallon

# cyl Number of cylinders

# disp Displacement (cu.in.)

# hp Gross horsepower

# drat Rear axle ratio

# wt Weight (1000 lbs)

# qsec 1/4 mile time

# vs Engine (0 = V-shaped, 1 = straight)

# am Transmission (0 = automatic, 1 = manual)

# gear Number of forward gears

# carb Number of carburetors

# where mpg is the dependent (target) variable for this project.

# 2. Project Technical Environment: (Describe the project with Tools and algorithm used)

Tools used for this project are as follows:

* RStudio 2022.02.0 Build 443
* R version 4.1.2 (2021-11-01) – “Bird Hippie”
* R libraries used: ggplot2, car, caret, corrplot
* Microsoft Azure Machine Learning Studio (classic)
* Microsoft Word
* Microsoft Excel

Algorithm used in this project:

* Linear Regression using Ordinary Least Square

R is an open-source programming language which is optimized for statistical analysis and data visualization. It was developed in 1992 and has a rich ecosystem with complex data models and elegant tools for data reporting.

Microsoft Azure Machine Learning Studio is a GUI-based web portal on Azure which provides an integrated development environment for building and implementing Machine Learning workflow.

For this project, linear regression (ordinary least square) algorithm is used for building the model and for the prediction of mpg values.

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

# ■ *Explain the process of the data flow, retrieval , combination of both R and AML and its usage* Project Workflow: The project workflow is as per the activities (Activity 2 till Activity 10) outlined in the Project Brief (M2\_NICF\_Basic\_R\_Programming\_SF\_\_Project\_3) and the evidence for the works done are provided in each of the activity below. Moreover, during the project execution, reverse engineering (refer Annexure - 02) was done for finding the best fitted model based on the following parameters: - Column selection: Based on the parameters like correlation, multicollinearity, coefficients/slopes, data types (numerical / categorical) and p values of feature variables

# - Row selection: Reduction of rows after detection of outliers via Exploratory Data Analysis (EDA) - Metrics: R-Squared, Adjusted R-Squared, Coefficient of Determination, Mean Absolute Error (MAE), Root Mean Squared Error (RMSE)

# 4. Activity 2: Create R code for effective data loading, storage and utilization

# *■ Install packages ggplot2, car, caret, corrplot and use library to use them* Codes and Outputs:

# 

# 

# *■ Load the MTCARS file from the R studio default dataset* Codes and Outputs: Note: Outliers w.r.t mpg are removed based on EDA (refer Annexure – 01) Text, letter Description automatically generated

# *■ Understand the dataset by using standard function str, head and summary* Codes and Outputs: Graphical user interface Description automatically generated with low confidence

# *■ Mention the number of records and fields from MTCARS in your project report* Codes and Outputs:

# 

# 5. Activity 3: Perform data handling standards and procedures

# *■ Factorise am, cyl, vs and gear fields* Codes and outputs:Text Description automatically generated

# *■ Dropping dependent variable for calculating Multicollinearity(mpg)* Codes and outputs:

# *■ Display the new data and check if mpg is displayed or not* Codes and outputs:A picture containing text Description automatically generated

# \*\* Exploratory Data Analysis (EDA) was carried out and it is provided in Annexure – 01. Each independent variable was checked w.r.t dependent variable (mpg).

# 

# 6. Activity 4: Identifying numeric variables

# *■ Identifying numeric variables using apply function and display* Codes and outputs: A picture containing text Description automatically generated

# *■ Write down in your project report all the numeric fields* Codes and outputs: Numerical fields are: "disp" "hp" "drat" "wt" "qsec" "carb"

# 7. Activity 5: Data management tools / standards for Correlation Matrix and Correlated attributes (Based on Multicollinearity)

# *■ Calculating Correlation* Codes and outputs: Graphical user interface, application Description automatically generated with medium confidence

# *■ Print correlation matrix and look at max correlation* Codes and outputs: Text Description automatically generated with medium confidenceMax correlation occurs between wt vs disp: 0.8754354

# *■ Visualize Correlation Matrix* Codes and outputs: Chart, bar chart Description automatically generated

# *■ Identifying Variable Names of Highly Correlated Variables* wt vs disp (consider the cutoff threshold for high correlation is 0.8)

# *■ Print highly correlated attributes* Codes and outputs: A picture containing text Description automatically generated

# *■ Write down in your project report all the highly correlated attributes* Codes and outputs: A picture containing graphical user interface Description automatically generated Highly correlated attributes:wt vs disp (Consider the cutoff threshold for high correlation is 0.8)

# *■ Remove highly correlated variables and create a new dataset* Codes and outputs: Note: For simplicity of subsequent activities reporting and presentation, the codes provided below are the results from reverse engineering for finding the best fitted model. Refer Annexure – 01 for details of the reverse engineering done. Selected features for the best fitted model: hp, drat, wt Graphical user interface, text, application Description automatically generated

# *■ Write down in your project report the dimensions of new dataset* Codes and outputs:Text Description automatically generated

# 8. Activity 6: Propose Model Creation

# *■ Build Linear Regression Model* Codes and outputs: Graphical user interface, text, application Description automatically generated

# *■ Check Model Performance using summary* Codes and outputs: Text Description automatically generated

# *■ Extracting Coefficients using summary* Codes and outputs:Text Description automatically generated

# *■ Take screenshot of summary coeff* Text Description automatically generated

# 9. Activity 7: Plot model

# *■ Plot the fit model in a 2\*2 matrix using par* Codes and outputs: Chart Description automatically generated

# *■ Take screenshot of fit* Codes and outputs:

# *■ Mention the R squared in the Project report* Codes and outputs:

# A picture containing graphical user interface Description automatically generated

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

# *■ Extracting R-squared value* Codes and outputs: Graphical user interface Description automatically generated with low confidence

# *■ Extracting Adjusted R-squared value* Codes and outputs:Logo Description automatically generated with medium confidence

# *■ Mention the above values in the Project report*

# R-squared: 0.8618913 Adjusted R-squared: 0.8459557

# 11. Activity 9: Predict mpg

# *■ Use function predict for the fit and dataframe from activity 5* Codes and outputs:

# *■ Use cbind to combine original mtcars and predicted values* Codes and outputs:

# *■ Print both actual and predicted mpg*

# Codes and outputs: Table Description automatically generated \*\* Please refer Annexure – 03 for plot of actual vs predicted mpg of the best fitted model \*\* Please refer Annexure – 04 for plots of actual mpg vs each selected feature of the best fitted model

# \*\* Please refer Annexure – 05 for plot of residuals (histogram) and root mean square error (RMSE) of the best fitted model

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

# *■ Login to AML studio and Upload the dataset in AML studio*

# *■ Use edit metadata to make cyl,vs,am,gear fields categorical*

# *■ Perform Compute Linear Correlation*

# *■ Use select columns in dataset and exclude disp, hp, wt as they are highly corelated*

# *■ Split the data to 80:20 and train the model using Linear regression to predict MPG field*

# *■ Score model and take screenshot of the predicted values*

# *■ Evaluate model*

AML Screenshot:

Diagram

Description automatically generated

Results of AML Model Scoring and Evaluation:

|  |  |  |
| --- | --- | --- |
| Definition | Analysis 7 (Best fitted model)  Features Selected: wt + qsec | Analysis 6 (Second best model)  Features Selected: wt + qsec + drat |
| Scoring |  |  |
| Evaluation |  |  |
| Actual vs Predicted |  |  |

The Coefficient of Determination of the best fitted model (Analysis 7): 0.851228

# 13. Annexure – 01 This Annexure shows the Exploratory Data Analysis (EDA) of this project, which was carried out after Activity 3. Each independent variable was checked w.r.t the dependent variable (mpg). Note: Plots shown in this Annexure are with outliers removed w.r.t mpg (mpg > 32).

# 1. mpg

# Text Description automatically generated

# Chart, histogram Description automatically generated Chart, bar chart, histogram Description automatically generated

# 2. mpg vs cyl

# 

Chart, box and whisker chart

Description automatically generated

# 3. mpg vs disp

# 

# Chart, scatter chart Description automatically generated

# 4. mpg vs hp

# A picture containing text Description automatically generated

Chart, scatter chart

Description automatically generated

# 5. mpg vs drat

# 

# Chart, scatter chart Description automatically generated

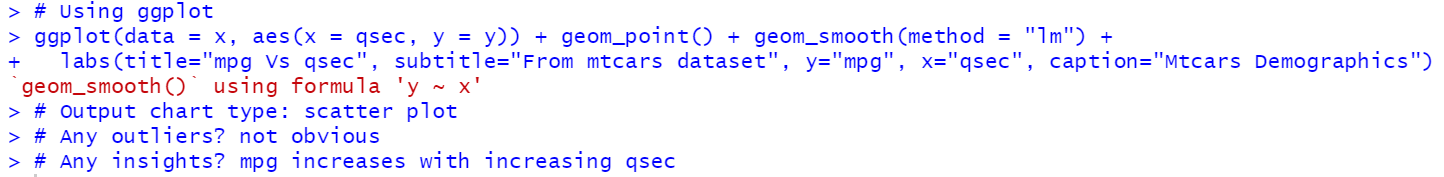
# 6. mpg vs wt

A picture containing logo

Description automatically generated  
Chart, scatter chart

Description automatically generated

# 7. mpg vs qsec



Chart

Description automatically generated

# 8. mpg vs Vs

A picture containing text

Description automatically generated

Chart, box and whisker chart

Description automatically generated

# 9. mpg vs am

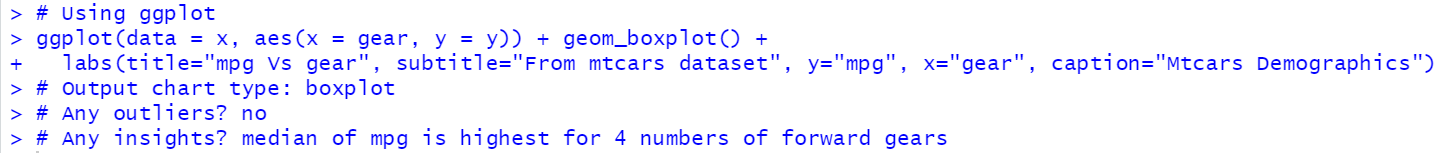
Text

Description automatically generated with low confidence

Chart, box and whisker chart

Description automatically generated

# 10. mpg vs gear



Chart, box and whisker chart

Description automatically generated

# 11. mpg vs carb

Text

Description automatically generated

Chart

Description automatically generated

# 14. Annexure - 02

This Annexure shows the blocks of codes used for reverse engineering for finding the best fitted model.

Text, letter

Description automatically generated

Notes:

1. Codes below are for checking of R2 and adjusted R2 of models with both numerical and categorical variables considered.
2. Different study cases are done based on different threshold values for high correlation (0.71, 0.8), and whether the outliers w.r.t mpg are removed or not.
3. The results of R2 and adjusted R2 for each study case are tabulated in the next page.

Graphical user interface, text, application

Description automatically generated

Notes:

1. Codes below are for checking of R2 and adjusted R2 of models with only numerical variables considered.
2. Different study cases are done based on different threshold values for high correlation (0.71, 0.8), and whether the outliers w.r.t mpg are removed or not.
3. The results of R2 and adjusted R2 for each study case are tabulated below.

Graphical user interface, text, application

Description automatically generated  
  
Table

Description automatically generated

Notes:

1. After getting the model in the previous page, next step is to check if removing variables with high p values can help improving adjusted R2 or not.
2. drop1() function is utilized for checking which insignificant variables to be removed sequentially.
3. Some variables with high p values (in descending order): carb, qsec, drat (cannot be further reduced after dropping drat)
4. The results of R2 and adjusted R2 for each case are tabulated in the next page.
5. Best fitted model is found in model with 3 features: hp, drat, wt

Graphical user interface, text, application, email

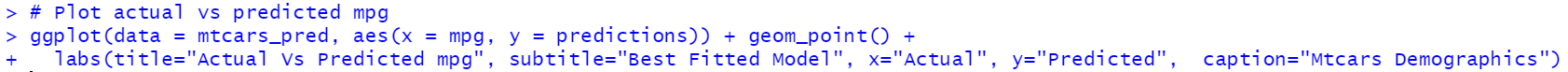
Description automatically generated  
  
Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, application, table

Description automatically generated

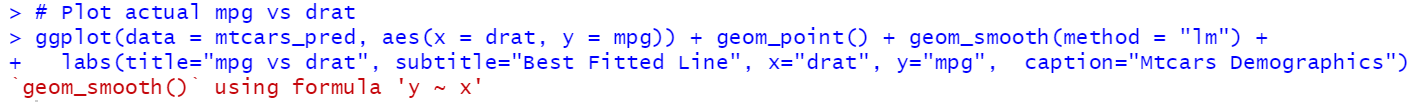
# 15. Annexure - 03

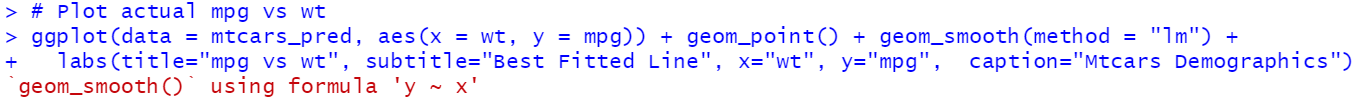
This Annexure shows the plot of actual vs predicted mpg of the best fitted model  
  
  
Chart, scatter chart

Description automatically generated

# 16. Annexure - 04

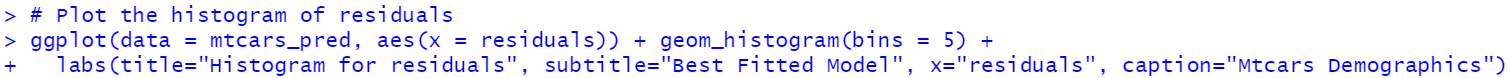
This Annexure shows the plots of actual mpg vs each selected feature (hp, drat, wt) of the best fitted model  
  
  
Chart, scatter chart

Description automatically generated  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
Chart, scatter chart

Description automatically generated  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
Chart, scatter chart

Description automatically generated

**17. Annexure – 05**This Annexure shows the plots of residuals (histogram) and root mean square error (RMSE) of the best fitted model.  
  
Table

Description automatically generated  
  
  
Chart, histogram

Description automatically generated  
  
Graphical user interface, text

Description automatically generated  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
**18. Annexure – 06**  
This Annexure provides the mtcars dataset, R codes and model in AML for this project for reference.  
mtcars dataset:

****

R Codes: ****  
  
Link for the AML model:  
  
https://gallery.azure.ai/Experiment/BRP-0322A-Lee-Jack-Shiang-Project-For-Reverse-Engineering