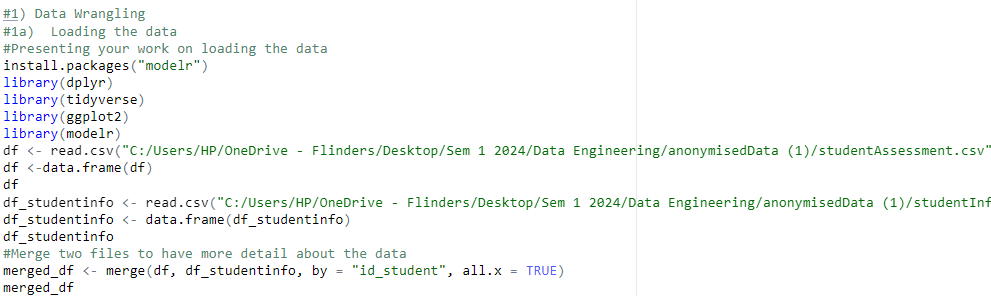
**Project report for the topic Data Engineering Semester 1 2024**

|  |  |
| --- | --- |
| Student name: | Thai Quang Le |
| FAN ID: | Le0424 |
| Email address: | Le0424@flinders.edu.au |
| Class enrolled | COMP8031 |
| Tutorial class: | COMP8031 |
| Group name: | Naruto |
| Names of other group members | 1) Thai Quang Le  2) Minh Quan Do Bui  3) Kimchheng Lim  4) Kazuya Tsubone |

1. **Data Wrangling**
   1. **Loading the data**

****

In this step, I used the read.csv function to load the Assessment.csv user data. A data frame named df is used to store data. I loaded libraries like dplyr, tidy, and ggplot2 and modelr for data manipulation and presentation. To get more information about the data, I decided to merged 2 dataset to make it more manageable.

* 1. **Handling missing data**

****

Next, I cleaned the dataset by removing the rows containing missing values using the na.omit function. This resulted in a cleaned dataset named merged\_df with no missing values.

* 1. **Tidying the data.**

**A white background with black text

Description automatically generated**

In this step, I tidied the data by converting the binary variable is\_banked to a more readable format. We used the if-else function to replace the binary values (1 and 0) with "Yes" and "No," respectively. The same technique have been used for the two other variables.

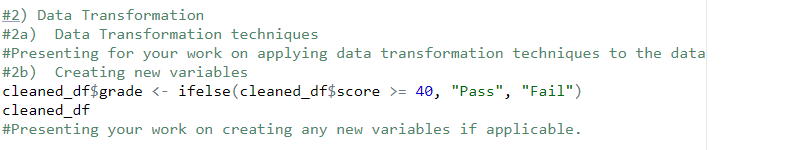
1. **Data Transformation**
   1. **Data Transformation techniques**

**A white background with black text

Description automatically generated**

To narrow the volume of the dataset, I will filter the range of the date\_submitted >0

* 1. **Creating new variables**

 One of the critical data transformation techniques I applied was creating a new variable grade based on the score variable. I used the if else function to categorize the scores into "Pass" or "Fail" based on a threshold of 40.

1. **Data Analysis**
   1. **Statistical analysis or exploratory data analysis**

**A computer screen shot of a computer code

Description automatically generated**

**A graph of a person in a suit

Description automatically generated**

In this step, I conducted statistical analysis and exploratory data analysis (EDA) to uncover patterns, trends, and relationships within the dataset.

I first created a histogram of the score variable to visualize the distribution of scores. The histogram revealed a right-skewed distribution, indicating that the median score is greater than the mean score.

Next, I calculated the mean and median scores to understand the central tendency of the data further.

A close-up of a computer code

Description automatically generated

**A computer code with blue text

Description automatically generated**

* 1. **Data visualisation**

I applied various data visualization techniques to further explore the dataset.

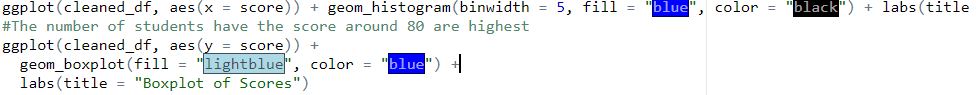
I created a scatter plot to visualize the relationship between the date submitted and the score, suggesting that more time spent submitting may affect students' scores.



A screen shot of a computer screen

Description automatically generated

I then created a histogram to show the distribution of scores, indicating that the highest number of students scored around 80.



A graph of a graph

Description automatically generated with medium confidence

Additionally, a boxplot was created to visualize the distribution of scores and identify outliers below a score of 25.

A white background with black text

Description automatically generated

A screenshot of a graph

Description automatically generated

And then I created two types of facet graph: 1 variable and 2 variables



A screenshot of a computer

Description automatically generated



A screenshot of a computer screen

Description automatically generated

Lastly, a bar chart was generated to show the number of students in each grade category (Pass/Fail).

A black rectangle with white text

Description automatically generated  
A screenshot of a computer

Description automatically generated

1. **Data Modelling**

In this step, I developed several linear models to analyse the relationship between variables in the dataset.

* 1. **Simple linear model**

**A close-up of a white background

Description automatically generated**

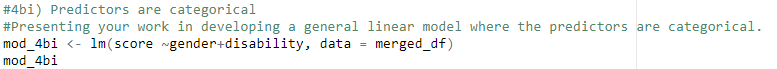
I first developed a simple linear model to understand the relationship between date\_submitted and score.

**A close-up of a computer screen

Description automatically generated**

I created a simple linear model based on the data of score and date\_submitted. In this case, the score is the dependent variable, and the date\_submitted is the dependent variable.

* 1. **General linear model**
     1. **Predictors are categorical.**

****

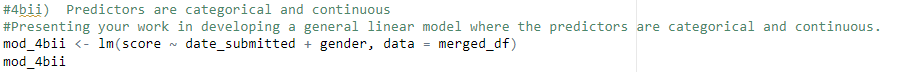
Next, I developed a general linear model where the predictor are gender and disability categorical and the predicted variable is the score.

**A screen shot of a computer

Description automatically generated**

* + 1. **Predictors are categorical and continuous.**

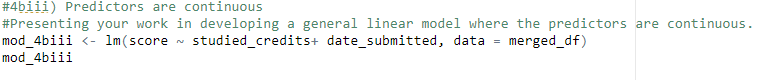
We then created a general linear model where the predictors date\_submitted and gender are both categorical and continuous.

****

**A screen shot of a computer

Description automatically generated**

* + 1. **Predictors are continuous.**

****

* 1. **Model evaluation**

**A white background with black text

Description automatically generated**

**A close-up of a computer code

Description automatically generated**

RMSE of part 4a: 19.08101

A black text on a white background

Description automatically generated

A close-up of a computer code

Description automatically generated

RMSE of part 4bi: 19.08056

A black text on a white background

Description automatically generated

A close-up of a computer code

Description automatically generated

RMSEof part 4 bii: 19.08013

**A black text on a white background

Description automatically generated**

**A computer screen with text

Description automatically generated**

RMSE of part 4biii: 19.05115

* 1. **Model Interpretation**

A close-up of a computer screen

Description automatically generated

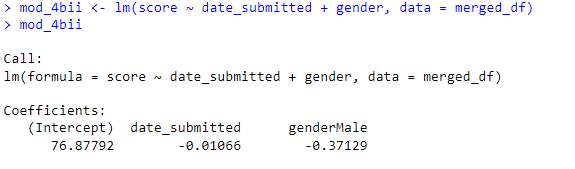
In this part, the dependent variable is score and the independent variable is date submitted. According to the results of simple linear model show us that date\_submitted has a negative impact to the score of students. These results mean that when the date submitted of students increase, they tend to get a lower score.

A white background with black text

Description automatically generated

Coefficients for genderMale (-0.45) and disabilityY (-2.63): Compared to the reference group, being male (genderMale) is associated with a decrease in score of approximately 0.45, and having a disability (disabilityY) is associated with a decrease in score of approximately 2.63, holding other variables constant.

Both Male and have disability has a negative impact to the score but disability will affect stronger than the gender.



#In this model, both two independent variables have negative impact on the dependent variable. So, when the students increase the date submitted and gender\_Male, the score will tend to get lower but gender will have more affect to the score.

A screen shot of a computer code

Description automatically generated

Coefficients for studied\_credits (-0.03) and date\_submitted (-0.01): For each additional studied\_credit, we expect the score to decrease by approximately 0.03, holding date\_submitted constant. For each unit increase in date\_submitted, we expect the score to decrease by approximately 0.01, holding studied\_credits constant.