# Data engineering project

Appendices for the report

**By** **Kory Frankee** (fran0618)

## What I worked with

###### Appendix 01

**R version** 4.3.3 | **R Studio version** 2023.12.1

**Packages I used in one’s code and for what:**

* mongolite {for connecting to a Mongo DB}
* tidyverse
  + ggplot2 {for visualizing data}
  + tibble {for storing data}
  + dplyr {for data manipulation}
  + tidyr {for rectangling}
  + readr {for reading .csv files}
  + stringr {for working with strings}
  + forcats {for factoring data}
* waffle {for a square pie chart}
* ggrepel {for repelling overlapping text}
* R6 {for object-oriented classes}
* ggcorrplot {for correlation plots}
* sqldf {for data manipulation}. I used dplyr and SQL to show that I understood and can do what I wanted to do with them both. Deciding on which one to use is a matter of preference
* colourvalues {for converting from an RGB value to a HEX value}
* car {for finding extreme observations}
* devtools {for getting packages from GitHub}
* sf, rnaturalearth, & rnaturalearthhires {for working with spatial data}
* modelr {for adding predictions and/or residuals to a data frame}. I was able to work out the rmse() with & without this package
* ggiraphExtra {for visualizing multiple linear regression models}

## What functions I used

###### Appendix 02

This list excludes custom functions I wrote like is.integerOrCharacterOfLength0()



## What operators I used

###### Appendix 03

A black background with colorful text

Description automatically generated

## What schema I worked with

##### OULAD dataset schema

###### Appendix 04

|  |
| --- |
| Shows the original schema |
| A screenshot of a computer screen  Description automatically generated |

|  |
| --- |
| Shows the new schema after I merged some tables together |
|  |

## What explanations I’ve provided to help understand the code

##### OULAD dataset changes made to the data

###### Appendix 05

|  |
| --- |
| Showing how if student has (Orange) or has not (Blue/Purple) taken an exam affects how I calculate the grade  Showing how if student has not (Yellow) finished the course affects how I calculate the grade |
| A screenshot of a computer  Description automatically generated |

## What visualizations I did

##### MongoDB dataset visualizations

###### Appendix 06

|  |
| --- |
| Visualization of the distribution of student scores |
| A graph of a graph  Description automatically generated with medium confidence |

###### Appendix 07

|  |
| --- |
| Here, I’ve how the selected student has performed compared to other students in the same classes using an average of their scores for all their assessment items as the basis of comparison. |
| Showing it in action...  With student ID of 0  A screenshot of a computer screen  Description automatically generated  With student ID of 1  A screen shot of a graph  Description automatically generated |

###### Appendix 08

|  |
| --- |
| Visualization of the least to most common grades students receive for their assessment items. |
| A graph with different colored bars  Description automatically generated |

###### Appendix 09

|  |
| --- |
| Visualization of student scores across two different classes. **Useful insights could be gathered from this side by side analysis if:**   * We could compare the grades for a given class with previous years to see if there has been an improvement in student performance + teaching quality |
| A graph of students in different classes  Description automatically generated |

##### OULAD dataset visualizations

###### Appendix 10

|  |
| --- |
| Visualization of student grades for the module and presentation combinations |
| Before change to assessment weightings so that the assessments for a module and presentation pairing add up to 100%  A screenshot of a graph  Description automatically generated  After change to assessment weightings so that the assessments for a module and presentation pairing add up to 100% |

###### Appendix 11

|  |
| --- |
| Visualization of cumulative GPA |
| Before change to assessment weightings so that the assessments for a module and presentation pairing add up to 100%  A graph with different colored bars  Description automatically generated  After change to assessment weightings so that the assessments for a module and presentation pairing add up to 100%  A graph of different colored rectangular bars  Description automatically generated |

###### Appendix 12

|  |
| --- |
| Visualization of what affect poverty has on student scores |
|  |

###### Appendix 13

|  |
| --- |
| Visualization of what affect the age of a student has on student scores |
| A screenshot of a computer screen  Description automatically generated |

###### Appendix 14

|  |
| --- |
| Visualization of what affect being disabled has on student scores |
| A line drawing of a bird  Description automatically generated |

###### Appendix 15

|  |
| --- |
| Visualization of what affect gender has on student scores |
| A drawing of a funnel  Description automatically generated |

###### Appendix 16

|  |
| --- |
| Visualization of how fast students are at handing in their assignments |
| A graph of a person  Description automatically generated with medium confidence |

###### Appendix 17

Before change to assessment weightings so that the assessments for a module and presentation pairing add up to 100%

A screenshot of a group

Description automatically generated

A bar code with text

Description automatically generated

After change to assessment weightings so that the assessments for a module and presentation pairing add up to 100%

A screenshot of a computer

Description automatically generated

A screen shot of a bar code

Description automatically generated

###### Appendix 18

|  |
| --- |
| Visualization of how education levels on entry to the module affected student scores |
| A chart of different colors  Description automatically generated with medium confidence |

###### Appendix 19

|  |
| --- |
| Visualization of linear model 01 with the regression line in blue and prediction line in red |
|  |

**A close-up of numbers

Description automatically generated**

###### Appendix 20

|  |
| --- |
| Visualization of student performance based on assessment type |
|  |

###### Appendix 21

|  |
| --- |
| Visualization of students that followed through their course + module combination to completion |
|  |

###### Appendix 28

|  |
| --- |
| Visualization of the number of male students and number of female students |
|  |

###### Appendix 29

|  |
| --- |
| Visualization of where students come from |
| Students in the UK    Students in England    Note that I excluded students from the {North region} and {South region} in the OULAD dataset from this plot because I couldn’t find a map of England that featured these regions |

###### Appendix 30

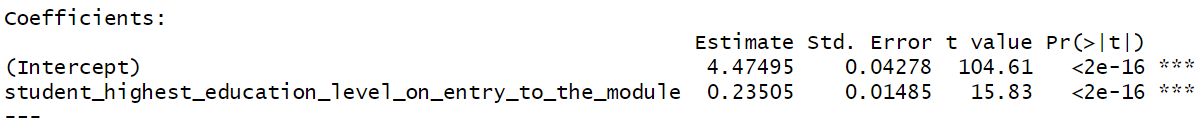
|  |
| --- |
| Visualization of the distribution of student scores |
|  |

###### Appendix 31

|  |
| --- |
| Visualization of whether some variables have / do not have some connection to the cumulative GPA a student received |
|  |

###### Appendix 32

|  |
| --- |
| Visualization of linear model 02 with the regression line in blue and prediction line in red |
|  |



###### Appendix 33

|  |
| --- |
| Visualization of linear model 03 cumulative GPA to total VLE learning material clicks with the regression line in blue and prediction line in red |
|  |

A close-up of numbers

Description automatically generated

###### Appendix 34

|  |  |
| --- | --- |
| Visualization of linear model 04 & 05 cumulative GPA to total VLE learning material clicks and IMD band | |
| **Linear model 04 (+ between terms)** | **Linear model 05 (\* between terms)** |
|  |  |

###### Appendix 35

|  |  |
| --- | --- |
| Visualization of linear model 06 & 07 cumulative GPA to total VLE learning material clicks and student quickness at handing in assignments | |
| **Linear model 06 (+ between terms)** | **Linear model 07 (\* between terms)** |
|  |  |

###### Appendix 36

|  |  |
| --- | --- |
| Visualization of linear model 08 & 09 cumulative GPA to student highest education level and IMD band | |
| **Linear model 08 (+ between terms)** | **Linear model 09 (\* between terms)** |
|  |  |

###### Appendix 37

|  |
| --- |
| Visualization of values we can use to compare all our linear models |
| A grid of white boxes with different colored squares  Description automatically generated with medium confidence |