# Data engineering project

Report on the MongoDB + OULAD dataset

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

**Class:** COMP 2031

**Group (for presenting):** Group T2

**Tutorial:** Tutorial 03

**Page count (for pages that count towards the limit):** 9

**Word count (for sections that count towards the limit):** ~ 3,000words

**Page limit:** 10

**Word limit:** 3,000 words

(Limits exclude my bibliography, references, figures, title page, and appendices)

## What I’ve done

Understanding the code requires an understanding of all the libraries, functions, and operators I used, which I’ve listed in **appendices 01 to 03**.This report tries to provide a high-level overview of what I’ve done so the reader doesn’t need to know all the technical details

#### I. Loading the libraries in

I wrote instructions to check if the packages I want to use are installed and if not, to install them. I then loaded them so they’re available for use

#### II. Loading the data in

To help put the data into R, I developed a *“data loader”* class. **Within it, I wrote functions for:**

* Putting data retrieved from the account of a MongoDB user > from a cluster > from a MongoDB collection into a tibble
* Putting data from .csv files into a tibble

#### III. Data observations

To help understand the data, I wrote a *“data checker”* class. Within it, I wrote functions for:

* Checking for duplicate primary key IDs
* Checking that the combination of multiple foreign key IDs was unique in each row
* Telling me basic facts about the dataset like the number of rows, columns, and duplicate, missing, and unique values, as well as the data types of the columns in the dataset

This helped me see that the {400,000 row × 4 column} Mongo DB dataset contained 10,000 students and 501 classes with a unique ID, 3 assignment types (‘exam’, ‘homework’ and ‘quiz’), and 400,000 unique scores from ~ 0 to ~ 100. There were no missing values found and neither of the two ID columns (student and class) are PKs as these columns contain duplicate values

It also helped me see that there were 23,369 students, 188 assessments, and 6,364 learning materials with a unique ID in the OULAD dataset. The assessment ID in the assessmentsTable was a PK and VLE material ID in the VLETable was a PK

**To further help me understand the data, I began to look at the data tables.**

After viewing the MongoDB dataset table in R, I discovered nested scores

A screenshot of a graph

Description automatically generated

After viewing the OULAD dataset tables in Excel [Data > Get data > From text/CSV], I discovered fifteen issues with the data

1. Missing student scores *{student\_assessment.csv}*

A screenshot of a search box

Description automatically generatedA screenshot of a search box

Description automatically generated

1. Use of quantitative (‘0’/’1’ values) unnecessarily *{student\_assessment.csv}*

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

1. Column names that aren’t clear - People would have to look up their meaning on the OULAD site to understand them {all .csv files}
2. Abbreviations used {student\_info.csv}.

A green and white box with black text

Description automatically generated

1. Wrongly entered age value of ‘55<=’, which should be ‘>55’ *{student\_info.csv}*

A screenshot of a search box

Description automatically generated

A screenshot of a search box

Description automatically generated

1. Missing IMD band values and a wrongly entered IMD band value ’10-20’ with no ‘%’ sign at the end of it) *{student\_info.csv}*

A screenshot of a screenshot of a computer

Description automatically generated

A screenshot of a screenshot of a computer

Description automatically generated

1. Missing when student unregistered values *{student\_registration.csv}*

A screenshot of a search box

Description automatically generated

A screenshot of a search box

Description automatically generated

1. Missing when student registered values *{student\_registration.csv}*

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

1. Missing exam due dates {*assessments.csv}*

A screenshot of a test

Description automatically generated

1. Missing values for when students use VLE materials *{vle.csv}*

A screenshot of a search box

Description automatically generated

A screenshot of a search box

Description automatically generated

1. Missing values for when students stop using VLE materials *{vle.csv}*

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

1. The sum of all assessments for a module + presentation combination should add up to 100% and it does not. **This is because the assessment weightings are too low**

* As it is, a student could do well on every single assignment for a module + presentation combination and fail due to all their assessments being weighted very lowly, which is very counterintuitive

E.g. For the student with an ID of *‘465730’* doing module *‘BBB’* and presentation *‘2013J’*, I observed that this student had high scores for all their assessments but failed because the assessment weightings add up to ~ 47.91%, not 100%

A screenshot of a graph

Description automatically generated

1. The exam cannot be weighted at 100% and all other assessments altogether cannot be weighted at 100% for the case where a student does an exam and other assessments

* As it is, a student can obtain a weighted score between ≥ 101 and ≤ 200. This would make the student’s grades out of 200% as opposed to 100%, which no university grading system accounts for

1. There is no data on students from the Northeast region of England. And it was not clear from looking at maps of England, what the “North” and “South” regions referred to
2. Student grades are not based on their scores

#### IV. Data cleanup

I wanted to address the one problem I had with the MongoDB dataset (nesting) and many of the problems I had with the OULAD dataset. To help with this, I developed a *“data cleanup”* class. **Within it, I wrote functions for:**

* Unnesting a column
* Finding and replacing values
* Removing junk columns and rows
* Putting values in a column into a vector of length 1 with comma-separated terms and quotation marks between each term. I can use such vectors in conjunction with a SQL IN
* Setting the names of columns

**Outside of this class, I also performed data clean-up operations such as:**

* + Re-ordering the data

**When finding and replacing values, I:**

* Replaced misspelt values
* Replaced abbreviated terms like *“No Formal quals”* with their proper form so *“No recognized qualifications”* in this case
* Used nominal *‘no/’yes’* values in cases where *‘0’/’1’* was being used to refer to ‘no/’yes’ since the text version is clearer and I do no calculations on these values
* Replaced missing exam due dates with the last week of the course value
* **Replaced missing student scores with zero because:**

1. The default mark is zero until a student completes their assignment
2. If the student had completed the assignment and obtained a score > 0, the student would have contacted the marker and then their mark would not be missing

##### How I solved the problem of missing IMD values

I filled in the missing IMD values for a particular region with the most common IMD value for that region, which I calculated

##### How I solved the problem of module + presentation combinations not adding up to 100% due too low assessment weightings for a module + presentation

**There were two cases I needed to consider:**

* Case 1) Some of the assessment weightings for the module + presentation combo
* Case 2) All of the assessment weightings for the module + presentation combo

Making this change was important as now, students that do well for all their assessments will not fail. This means that compared to the original OULAD dataset, more students with a high average score will have a higher grade and higher cumulative GPA. **(In Appendices 10, 11, & 17, I have shown what grades and cumulative GPAs students have before & after this change. Notice in Appendix 10, how there are no longer module and presentation combinations where ALL students fail)**

#### V. Data additions

*Everything I do in this section is based on what Australian universities do*

I added a grades column to the MongoDB dataset, which I filled in with tertiary grades (‘F’, ‘P’, ‘Cr’, ‘D’, and ‘HD’) based on what score students received for their assessments

##### How I solved the problem of student grades being out of 200% if taking an exam and other assessments items

For the OULAD dataset, I addressed the issue with the assessment weightings for the **problematic 2nd case**

**Cases 1) and 2):**

1. Module has assessments but no exams.

* For this case, assessment weights should altogether add up to 100% so I leave the assessment weight column how it is

1. Module has assessments, including exams.

* For this case, I added a second assessment weight column that scales all the assessment values to a new weighting. I set exams to be worth 40% (can change this value depending on the grading decisions of the university) and other assessments to be worth 100% - exam % so that they both add up to 100%

##### How I solved the problem of student grades not being based on their scores

I added columns for the weighted score for each module + presentation combination, which will be equivalent to a grade (e.g. > 85% is a ‘HD’/’7’)

The weighted score (and grade) I decide to give a student is based on if their module + presentation had exams and if the student withdrew or not

**(Appendix 05 shows this in action)**

From here, I decided to calculate students’ cumulative GPA

#### VI. Data subsets and supersets

To help obtain a subset of the data, I wrote a *“data subsetter”* class with functions for:

* Getting a few rows of interest based on some selection criteria
* Getting a few columns of interest

To help join data together, I wrote a “data supersetter” class with functions for:

* Merging tables

I joined the OULAD tables that I was able to join (without R Studio freezing) into one table since I found that easier to work with **(Appendix 04)**Using a combination of subsetting and supersetting was helpful when, for example, I was working with the large *“student\_vle.csv”* file. I split it up into 8 parts, performed the operations I wanted to perform on each part, and then put the pieces of operated data together into a table

#### VII. Dataset querying

Sometimes I needed the dataset to be in a particular form and to help with this, I wrote a *“dataset querier”* class with functions for:

* Querying the data using SQL

#### VIII. Data analysis

To help analyse the data, I wrote a *“data analysis”* class with functions for:

* Working out statistical information (mean, mean of each group, median, maximum, minimum, quartile value, quartile value for each group, interquartile range, standard deviation, z-score, & range of values). **Here’s an example of me using these functions. Below, I’ve found the “mean of each group” where each group is a different class:**

A screenshot of a table

Description automatically generated

#### IX. Data visualization

To help visualize the data, I wrote a *“data visualization”* class with functions for:

* Getting the inverted colours for a list of colours (used for distinguishing the colours of labels from colours of the plot objects)
* **Graphing:**
  + Simple scatter plots
  + Complex scatter plots **(Appx. 07)**
  + Histogram w/ density plots **(Appx. 06, 30)**

**I graphed these plots outside of this class since I didn’t think I’d re-use them:**

* + Bar charts **(Appx. 08, 11)**
  + Bar charts side by side (**Appx. 09,** **10, 37)**
  + Violin plots **(Appx. 12 to 15, 18)**
  + Density plots **(Appx. 16)**
  + Pie charts **(Appx. 21)**
  + Box plot **(Appx. 20)**
  + Square pie chart **(Appx. 28)**
  + Maps (**Appx. 29)**
  + Correlation plots **(Appx. 31)**

#### X. Data modelling

To help model the data, I wrote a *“data modeller”* class with functions for:

* Extracting values from a linear model (the p value, adjusted r squared value, number of outliers, and root mean square value (rmse))

##### Linear models

A linear model is used to see if there’s a relationship between one or more predictors and an outcome. **It can be used when:**

* There is a continuous outcome variable
* There is a linear relationship
* The observations are independent
* The residuals follow a normal distribution
* The are no extreme outliers (the variance of the residuals should be the same).

**What I considered when evaluating which model was the better one:**

* The adjusted R-squared value. The closer this value is to one, the better the model
* The P-value. The closer this value is to zero, the more that this suggests that there is a relationship between variables with
* The root mean square error (RMSE) value. The closer this value is to zero, the better that this model fits the data
* The number of outliers. The more outliers there are, the worse the model is

Threshold can be used (e.g. p ≤ 0.05) but these can often be arbitrary, so it makes more sense to compare all the models and see which one is best

**What I looked at when interpreting a model:**

* If the line has a positive slope, the predictor has a positive association with the outcome

**What I considered when seeing if two variables were related:**

* The correlation value, which is defined as

Correlation can range between -1 and 1. Higher positive numbers mean there’s a closer relationship. Lower negative numbers mean there’s an inverse relationship. Numbers near 0 mean that there’s no relationship

##### Correlation plot

If there was a positive correlation between a variable and the cumulative GPA then I used this variable as a predictor in one’s linear models **(Appendix 31)**

##### Simple linear model

A simple linear regression model has 1 outcome (y) and 1 explanatory variable (x)

**Equation for a simple linear model**

##### Multiple linear model

A multiple linear regression model has 1 outcome (y) and 2 or more explanatory variables (Xn)

**Equation for a multiple linear model**

##### Linear models I did

###### Simple linear models

**Linear model 01 – Continuous predictor**

Relationship between the cumulative GPA of a student (y) and the average score of a student (x)

**(Appendix 19)**

**How I interpreted this model:**

* As the average score of a student increases, the cumulative GPA tends to increase with it

**Linear model 02 – Categorical predictor**

Relationship between the cumulative GPA of a student (y) and the education level of the student on entry to the module (x)

**(Appendix 32)**

**How I interpreted this model:**

* There was not a linear relationship or a clear relationship for that matter between how educated a student is and their cumulative GPA, it’s quite scattered all over the place

**Linear model 03 – Continuous predictor**

Relationship between the cumulative GPA of a student (y) and the total times a student has clicked onto VLE learning material (x)

**(Appendix 33)**

**How I interpreted this model:**

* It seems like engaging more with learning materials might boost the cumulative GPA of some students while others don’t need to engage with the material much to do well. Can’t predict GPA from learning material use

##### Multiple linear models

**Linear model 04 & 05 – Categorial and continuous predictor. First model has + between terms and the second \***

Relationship between the cumulative GPA of a student (y) and total VLE learning material clicks (x1) and IMD band (x2)

Where x1 is a continuous variable and x2 is a categorical variable

**(Appendix 34)**

**How I interpreted this model:**

* There was not a clear relationship between the number of times a student interacted with the learning material or their IMD band and their cumulative GPA

**Linear model 06 & 07 – Two continuous variables. One with \* and one with +**

Relationship between the cumulative GPA of a student (y) and average student quickness at handing in assignments, (x1) and total VLE learning material clicks (x2). Where x1 and x2 are both continuous variables **(Appendix 35)**

**How I interpreted this model:**

* The model is a poor fit

**Linear model 08 & 09 – Two categorical variables. One with \* and one with +**

Relationship between the cumulative GPA of a student (y) and IMD band (x1) and highest education level on entry to the module (x2). Where x1 and x2 are both categorical variables **(Appendix 36)**

**How I interpreted this model:**

* The model is a poor fit

##### Comparing the linear models I did

**How well these models performed:**

I created a bar chart for a visual comparison between the measures of model performance (**Appendix 37)** and from this, it was clear that linear model 01 was the best model of the bunch as it has the adjusted r^2 value closest to 1 and rmse value closest to 0. The first model did, however, have nine more outliers than all our other models. All the models had a p-value < 2.2e-16, which is the lowest value R can show. This finding makes sense to me as the GPA a student receives is based on their score and I think the scores a student historically had are the best predictor for what GPA they’ll have in future

#### XI. Data insightful findings

##### MongoDB dataset

**I deduced that the MongoDB score data was not based on real data for these reasons:**

1. Markers don’t tend to provide scores with upwards of 7 decimal places.
2. Marks don’t tend to be all different to one another when the number of assignments marked is high, which they are.
3. Marks tend to have a minimum of 0 as some students receive a 0 for not handing in their work. Here, the minimum score is ~ 0.0002.
4. Marks tend to have a maximum of 100 as some students receive full marks for their work. Here, the maximum is ~ 99.99.
5. Scores tend to have a more normal distribution. I’ve provided a visualization of the distribution of scores for all the assignments to show how uniform it is **(Appendix 06)**

Revealed that more students succeed with their assignments than fail at them by a small margin. This means that students are underperforming or that the markers are harsh (**Appendix 08)**

##### OULAD dataset

For the scores, the scores ranged from 0 to 100 with a mean of ~ 75.72 and a median of 80, The Q1 value was 65 and Q3 value 90 so the IQR was 25. The distribution of scores shows that the scores are positively skewed **(Appendix 30)** - Most students score > 50 for their assessments.

Both genders score about as good as each other

A screenshot of a computer

Description automatically generated**(Appendix 15)**

Students without a disability have a slight edge over students with one

A screenshot of a computer

Description automatically generated

**(Appendix 14)**

Students that are older have a slight edge over students that are younger

A screenshot of a computer

Description automatically generated

**(Appendix 13)**

Students that are less impoverished have a slight edge over students that are more impoverished (two exceptions can be seen. The 10-20% band outdoes the 20-30% band and the 40-50% band outdoes the 50-60% IMD band)

A screenshot of a computer

Description automatically generated

**(Appendix 12)**

Students that are more well-education on entry have an edge over less educated students

A screenshot of a computer

Description automatically generated

**(Appendix 18)**

Students from almost all over England are taking these courses – there’s still the potential for the institution to reach students from Northeast England **(Appendix 29)**

Most (~ 78.47% of) students come from England, 2nd most (~ 10.62% of) students come from Scotland, 3rd most (~6.81% of) students come from Wales, and 4th most (~ 4.1% of) students come from Northern Ireland **(Appendix 29)**

Students perform the best in module + presentation combo *“EEE, 2014J”* and the worst in module + presentation combo *“DDD, 2013B”*

A close up of a word

Description automatically generated

A close up of a screen

Description automatically generated

Students score the best for *‘CMA’* assessments, second best for *‘TMA’* assessments, and third worst for *‘Exam’* assessments

**(Appendix 20)**

~ 59.99% of students have a high enough {> 4.5} cumulative GPA for postgraduate studies **(Appendix 11)**

Most (~ 81.51% of) students completed their studies, but lots (~ 18.49%) did withdraw

**(Appendix 21)**

Students that are more impoverished are generally but not always likelier to withdraw

A screenshot of a computer

Description automatically generated

The gender balance is quite even as ~ 53% of students are men and ~ 47% women

**(Appendix 28)**

#### XII. Bibliography

**Sources I used to develop a deeper understanding of statistics, working with data, and R:**

* Chang, W, 2024, *“Introduction to R6 classes”*, <https://r6.r-lib.org/articles/Introduction.html>, viewed March 2024
* Cotton, R 2013, *“Learning R”*, <http://duhi23.github.io/Analisis-de-datos/Cotton.pdf>, viewed March 2024
* Freeman, M 2019, *“Programming skills for data science: Start writing code to wrangle, analyze, and visualize data with R”*, viewed March 2024
* Grolemund, G 2014, *“Hands-On Programming with R”*, <https://rstudio-education.github.io/hopr/>, viewed March 2024.
* Huynh, H 2019, *“R for graduate students”*, <https://bookdown.org/yih\_huynh/Guide-to-R-Book/>, viewed March 2024
* IBM 2024, *“R Programming Basics for Data Science”*, <https://www.edx.org/learn/r-programming/ibm-r-programming-basics-for-data-science>, viewed March 2024
* Irizarry, R 2019, *“Introduction to Data Science: Data Analysis and Prediction Algorithms with R”*, <https://rafalab.dfci.harvard.edu/dsbook/>, viewed March 2024
* Ismay, C; Kim, A; McConville, K, 2024, *“Statistical Inference via Data Science A ModernDive into R and the Tidyverse”*, <<https://moderndive.com/>>, viewed April 2024
* Lander, J 2013, *“R for everyone: Advanced analytics and graphs”*, <https://www.researchgate.net/profile/David-Booth-7/post/Best\_approach\_for\_model\_selection/attachment/62588e5933d0f000f76d1352/AS%3A1144874569072657%401649970776919/download/R+for+Everyone+Advanced+Analytics+and+Graphics+by+Jared+P.+Lander+%28z-lib.org%29.pdf>, viewed March 2024
* Long, J & Teetor, P 2019, *“R Cookbook, 2nd Edition”*, <https://rc2e.com/>, viewed March 2024
* Matloff, N 2017, “The Art of R Programming: A Tour of Statistical Software Design”, <https://diytranscriptomics.com/Reading/files/The%20Art%20of%20R%20Programming.pdf>, viewed March 2024
* Mecklin, C 2023, *“STA 440/441 Notes (Mathematical Statistics I)”*, <<https://bookdown.org/cmecklin/sta440/>>, viewed April 2024
* Monkman, M, 2024, *“The Data Preparation Journey Finding Your Way With R”*, <https://bookdown.org/martin\_monkman/DataPreparationJourney/>, viewed April 2024
* Nahhas, R 2024, *“Introduction to Regression Methods for Public Health Using R”*, <https://www.bookdown.org/rwnahhas/RMPH/>, viewed March 2024
* Neth, H, 2024, *“Data Science for Psychologists”*, viewed April 2024, <https://bookdown.org/hneth/ds4psy/>
* Ooms, J 2018, *“Mongolite user manual”*, <https://jeroen.github.io/mongolite/>, viewed March 2024
* Peng, R, Kross, S & Anderson, B, 2020, *“Mastering software development in R”*, <<https://bookdown.org/rdpeng/RProgDA/>>, viewed April 2024
* Roback, P & Legler, J, 2021, *“Beyond Multiple Linear Regression Applied Generalized Linear Models and Multilevel Models in R”*, < <https://bookdown.org/roback/bookdown-BeyondMLR/>>, viewed April 2024
* Scott, J, 2021, *“Data Science in R: A Gentle introduction”*, <<https://bookdown.org/jgscott/DSGI/>>, viewed April 2024
* Shang, Y, 2022, *“Making Sense of Data with R”*, viewed April 2024, <<https://bookdown.org/yshang/book/>>
* Wickham, H 2014, *“Advanced R”*, <https://adv-r.hadley.nz/index.html>, viewed March 2024.
* Wickham, H; Rundel, M; Grolemund, G 2023, *“R for data science (2e)”*, <https://r4ds.hadley.nz/>, viewed March 2024.
* Xie, Y; Altaire, J; Grolemund, G, 2023, *“R Markdown: The Definitive Guide”*, <<https://bookdown.org/yihui/rmarkdown/>>, viewed April 2024
* Zieffler, A, 2012, *“Statistical Thinking: A Simulation Approach to Modeling Uncertainty (STAT 216 edition)”*, <https://bookdown.org/frederick\_peck/textbook/>, viewed April 2024

#### XIII. References

**Where I obtained the datasets:**

* Kuzilek, J; Hlosta, M; Zdrahal, Z 2017, *“Open University Learning Analytics dataset”*, <<https://analyse.kmi.open.ac.uk/open_dataset>>, viewed March 2024
* MongoDB 2024, *“Sample training dataset | Grades”*, <https://www.mongodb.com/docs/atlas/sample-data/sample-training/#sample\_training.grades>, viewed March 2024
* Office For National Statistics 2014, *“Regions (December 2015) Full Extent Boundaries in England”*, <https://www.data.gov.uk/dataset/a2c0cb64-a37a-4ee0-a7b9-90679efb219d/regions-december-2015-full-extent-boundaries-in-england>, viewed April 2024

**Where I obtained information about university grading decisions:**

* Australian Education Info 2024, *“Grading system in Australia”,* <https://www.australiaeducation.info/Education-System/Grading-System.html>, viewed March 2024
* Rutgers 2024, *“How to calculate your cumulative GPA”*, <https://myrun.newark.rutgers.edu/how-calculate-your-cumulative-gpa>, viewed March 2024
* University of Southern Queensland 2023, *“Grades procedure”*, <https://policy.usq.edu.au/documents/20592PL>, viewed March 2024.
* University of Southern Queensland 2024, *“How do I work out my final grade?”*, <<https://usqassist.custhelp.com/app/answers/detail/a_id/2615/~/how-do-i-work-out-my-final-grade%3F>>, viewed March 2024.
* University of Sydney, *“How do I work out my final grade?”,* <https://usqassist.custhelp.com/app/answers/detail/a\_id/2615/~/how-do-i-work-out-my-final-grade%3F>, viewed March 2024
* University of Adelaide 2023, *“How do Withdraw No Fails (WNF) and Withdraw Fails (WF) affect my Grade Point Average (GPA)?”*, <https://student.ask.adelaide.edu.au/app/answers/detail/a\_id/1684/~/how-do-withdraw-no-fails-%28wnf%29-and-withdraw-fails-%28wf%29-affect-my-grade-point>, UofA, viewed March 2024

**Where I obtained information about a cut-off GPA for postgraduate study:**

* University of Adelaide 2023, *“Master of Artificial Intelligence and Machine Learning”*, <https://www.adelaide.edu.au/degree-finder/maiml\_maiml.html/>, UofA, viewed March 2024