Data engineering assignment

Kory

2024-03-24

# Mongo DB

## Looking at the dataset

## [1] "number of missing values in entire dataset: 0"

There are no missing values in this dataset

## student\_id rows count type rows count score rows count   
## 400000 400000 400000   
## class\_id rows count   
## 400000

## student\_id unique rows count type unique rows count   
## 10000 3   
## score unique rows count class\_id unique rows count   
## 400000 501

## [1] "the range of scores (difference between the min and max scores): "

There are 400,000 rows and 4 columns in total

* There are 10,000 students with a unique student ID
* There are 501 classes with a unique class ID
* There are 3 assignment types, which is as we’d expect
* There are 400,000 unique scores, which is unexpected but makes sense as there are at least 7 decimal places in the scores, which range from ~ 0 to ~ 100. There’d be less unique scores if the score values were integers

## [1] "mean: 49.9107237106642"

## [1] "median: 49.9121492863691"

## [1] "maximum: 99.9997434060823"

## [1] "minimum: 0.000262781006732382"

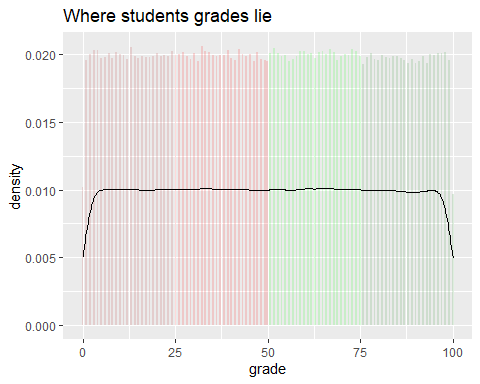
## [1] "1st quartile: 24.9377978071595"

## [1] "3rd quartile: 74.798609800433"

## [1] "standard deviation: 28.8425037321742"

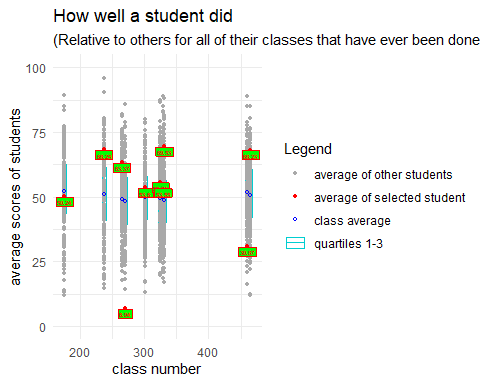
For all the scores in the dataset, we can derive the mean, median, minimum, maximum, 1st quartile, 3rd quartile, and standard deviation values

## Warning: The dot-dot notation (`..density..`) was deprecated in ggplot2 3.4.0.  
## ℹ Please use `after\_stat(density)` instead.  
## This warning is displayed once every 8 hours.  
## Call `lifecycle::last\_lifecycle\_warnings()` to see where this warning was  
## generated.



We can also show the distribution of scores for all of the assignments (quiz, exam, homework) that students have ever done

## Looking at subsets of the dataset



We can look up a student (e.g. the student with a ‘student\_id’ of 1) and see how well that student has done compared to other students enrolled in the same classes.

* To calculate the score each student achieved in each of their classes, an average of all of their scores for all of the assessment items in that class was used.
* To calculate the class average for each class, an average of all of the average scores a student received for each class was used
* To calculate quartile 1 and 3 values, I obtained the value at the 25% and 75% percentile for all of the average scores a student received for each class

# OALAD

Excel does not read the data right (e.g. student % can be shown as dates for student\_ID 74372 in studentInfo.csv) unless it’s loaded in via Data > Get Data > From File / CSV

Correcting column names (e.g. student\_id not id\_student)

Dealing with missing values

Renaming values (e.g. age 55 <= to like <= 55)