Default Question Block

Q14. The following 20 questions test a few concepts that we have covered so far in HUDK4050. You can use any resources you want to answer the questions. Good luck!

D1. First Name

D2. Last Name

D3. What is your TC email address? (Please double check that this is correct or you will not receive your results)

Block 1

Q1.
Download the Open University Learning Analytics dataset MOOC data set from:

https://analyse.kmi.open.ac.uk/open_dataset

Did you successfully download the data?

○ Yes
○ No

Q2. Review the documentation on the webpage. What is being "summed" in the variable "sum_click"?

○ The number of clicks that occured in the course

○ The number of clicks that each student made in the course

○ The number of correct clicks a student made in each unit of the course

○ The number of times a student interacts with the material in a day

Q3.
Extract the data into a folder on your computer. Upload the datasets "studentInfo.csv" and "studentAssessment.csv" into R as dataframes named D1 and D2.

Did you successfully create D1 and D2?

○ Yes

○ No

Q4. What is the average score for the assessment id "37426"?

○ 82.1

○ 69.6

○ NA_real_

○ 75.8

Q5. What is the code to quickly view a histogram of the range of the number of credits studied in the studentInfo dataset?

○ plot(D1$studied_credits)

○ ggplot(D1, studied_credits) + geom_histogram()

○ hist(D1$studied_credits, breaks = 100)

○ hist(D2$score)

Q6. What is the code to quickly plot a boxplot of age against number of credits?

○ boxplot(D1$age_band, D1$studied_credits)

○ plot(D1$age_band, D1$studied_credits)

○ boxplot(D1$studied_credits, D1$age_band)

○ boxplot(D1$studied_credits, col = D1$age_band)

Q7. What is the code using "dplyr" to create a new data frame from D1 called D3 that includes only the variables id_student, gender, highest_education and age_band?

○ D3 <- read(D1, id_student, age_band, studied_credits)

○ D3 <- select(D1, id_student, age_band, studied_credits)

○ D3 <- filter(D1, id_student, age_band, studied_credits)

○ D3 <- subset(D1, id_student, age_band, studied_credits)

Q8. Reduce D3 to the first 200 rows, what is the student ID of the 200th row?

○ 414117

○ 17244

○ 115182

○ 579211

Q9. What would the code be to create a new dataframe called D4 from D3 that has student ID as a variable and then a variable for each age band with the credits in the cells?

○ D4 <-gather(D3, age_band, studied_credits, 2:3)

○ D4 <- spread(D3, age_band, studied_credits)

○ D4 <- spread(D3, age_band, studied_credits, 2:3)

○ D4 <- gather(D3, age_band, studied_credits, 2:3)

Q10. What is the code to combine the three age columns in D3 into a single column?

○ D4<-spread(D4, id_student, 2:4)

○ D4 <- gather(D4, id_student, "measure", 2:4

○ D4 <- gather(D4, 2:4)

○ D4 <- unite(D4, "0-35", "35-55", "55<=", sep = "~")

Block 2

Q11. What is the R command to generate a vector called "studentID" from the student ID variable but also make sure that no ID appears more than once?

○ studentID <- unique(D1$id_student)

○ studentID <- no.repeat(D1$id_student)

○ studentID <- na.rm(D1$id_student)

○ special(D1$id_student)

Q12. What does the "date_submitted" variable represent?

○ The number of seconds since January 1, 1970

○ The number of months in the Zoroastrian calendar

○ The number of days since the start of the semester

○ The number of days since the start of the module-presentation

Q13. What is the R code to select only those rows from the "studentAssessment" dataset that correspond to students whose results have been rolled over from a previous presentation?

○ D5 <- select(D2, is_banked == 1)

○ D5 <- filter(D2, is_banked == 1)

○ D5 <- ifelse(D2$is_banked == 1, TRUE, FALSE)

○ D5 <- sample(D2, is_banked == 1)

Q14. What are the average scores in the "studentAssessment" dataset for the following student IDs in this order:

25261, 42818, 70827, 610573

○ 76.5, 42.7, 84.3, 71.0

○ 75, 75, 75, 75

○ 78, 39, 62, 45

○ 35, 72, 54, 42

Q15. What is the difference between a matrix and a data frame?

○ A data frame can incorporate columns of differing types, a matrix cannot

○ A matrix is multi-dimensional, a data frame is not

○ A data frame contains homogeneous data types, a matrix does not

○ A data frame is bigger than a matrix

Q16. How many rows are there in the data frame if you join matching rows **from** the studentAssessment **to** the studentInfo dataset according to student ID?

○ 213166

○ 2017319

○ 265788

○ 207319

Q17. What is the R code to generate a new column from the studentInfo dataset that subtracts 10 from the studied_credits variable for every time a student has previously attempted the module and also deletes the original columns?

○ mutate(D1, attempt.credits = studied_credits - num_of_prev_attempts)

○ transmute(D1, attempt.credits = studied_credits - num_of_prev_attempts*10)

○ mutate_each(D1, attempt.credits = studied_credits - num_of_prev_attempts)

○ mutate(D1, attempt.credits = num_of_prev_attempts + 10)

Q18. What is the R code to take the combined dataframe from Q16 to calculate the mean score and mean credits for each region?

○ D7 %>% group_by(region) %>% generate(av.credits = mean(studied_credits), av.score = mean(score, na.rm = TRUE))

○ D7 %>% group_by(region) %>% transmute(av.credits = mean(studied_credits), av.score = mean(score, na.rm = TRUE))

○ D7 %>% mutate(av.credits = mean(studied_credits), av.score = mean(score, na.rm = TRUE))

○ D7 %>% group_by(region) %>% mutate(av.credits = mean(studied_credits), av.score = mean(score, na.rm = TRUE))

Q19. What is the R code to generate a scatter plot using ggplot of the average regional score by the average regional credits and coloring the points by region?

○ ggplot(D9$av.credits, D9$av.score)

○ ggplot(D9, aes(av.credits, av.score, col = region)) + geom_point()

○ ggplot(D9, v.credits, av.score) + geom_point(col = region)

○ ggplot(D9$av.credits, D9$av.score, geom_point())

Q20. How would you remove the legend from your plot in Q19?

○ + theme_minimal()

○ + no.legend()

○ + geom_legend(NA)

○ + theme(legend.position="none")

Powered by Qualtrics