## Assessment 2: Section 1: 15 % Mark

Updated Version

Focus: Basic R, RMarkdown and `tidyverse` data wrangling

1. Setup your Assessment 2 project:

- create a RMarkdown document and save it in this project using assessment2.rmd

- Delete parts that you do not need in the RMarkdown template generated by default and copy the tasks below there. You need to leave the `meta data` section and initial R chunk.

- The title of your document should be Assessment 2

- The Author is your name and student ID

- The date is the date you created the document

You will need to answer each task using the text/graph narration and include R chunks to show how you get them.

2 Create the heading of the first level and call it RMarkdown editing

3. Write 2-3 sentences about RMarkdown (you can google this information or use resources recommended in class - no need to reference)

4. In the above use bold and italics for editing.

5. Review the suggested documentation on how to insert links in the .rmd file and include an in-text link to https://rmarkdown.rstudio.com/lesson-8.html

6. Insert an R chunk and create a variable with this dataset **fastfood_calories.csv** located on [redacted]

The name of the variable should be `fastfood`

7. Display the first 10 rows of the dataset using head() and kable().

Display the first 10 rows of the dataset and the first 5 variables

Use Help and the link below to read more about those functions https://bookdown.org/yihui/rmarkdown-cookbook/kable.html

Save your file as .rmd

8. Display the observations that has more than 1000 calories

9. Arrange observations with more than 40 in total_fat and more than 80 in total_carb in the descending order (PLEASE USE THE VARIABLE OF YOUR CHOICE TO ORGANISE THE DESCENDING ORDER) and save them to a new variable (dataset) called `dont_eat_this`

10. Using the initial dataset variable, use `case_when()` to create a new variable `heavy_food` which is equal to "heavy" when total calories (`calories`) are greater than 500, "low" when total calories are less than 250 and "average" for all other cases. Count the number of observations in each created category.

11. Display the types of variables in the dataset using `skimr` package

12. Present the count observations from each restaurant in a descending order

Show the number of distnct items on a menu in the dataset

13. Using groupings (group_by()), summarise and display the average number of calories for each restaurant.

14. Add variables to the dataset, which:

- calculates the average calories per type of restaurant and call it `average_calories`
- calculates the maximum total_fat per type of restaurant and call it `max_fat`
- calculates the minimum cholesterol per type of restaurant and call it `min_cholesterol`

15. Display the data vis of total fat per each type of restaurant. Write a narration (2-3 sentences) why you believe this type of data viz presents such information best.

16. Add a variable to the dataset, which calculates the sum of cholesterol and sodium and call it `cholesterol_sodium`.

Remove the variable `salad`

17. Use observations for Mcdonalds to plot sugar variable against protein with `geom_point()`

Save your file as .rmd

****Focus**: ggplot2, factors, strings, dates**

18. Identify variable(s) which should be factors and transform their type into a factor variable.

19. Create a new variable:

Read about `cut_number()` function using Help and add a new variable to the dataset `calories_type`. Use `calories` variable for `cut_number()` function to split it into 3 categories `n=3`, add labels `labels=c("low", "med", "high")` and make the dataset ordered by arranging it according to calories.

Do not forget to save the updated dataset.

20. Create a dataviz that shows the distribution of `calories_type` in food items for each type of restaurant. Think carefully about the choice of data viz. Use facets, coordinates and theme layers to make your data viz visually appealing and meaningful. Use factors related data viz functions.

21. Add a new variable that shows the percentage of `trans_fat` in `total_fat` (`trans_fat`/`total_fat`). The variable should be named `trans_fat_percent`. Do not forget to save the updated dataset.

22. Create a dataviz that shows the distribution of `trans_fat` in food items for each type of restaurant. Think carefully about the choice of data viz. Use facets, coordinates and theme layers to make your data viz visually appealing and meaningful.

23. Calculate and show the average (mean) `total_fat` for each type of restaurant. No need to save it as a variable.

24. And create a dataviz that allow to compare different restaurants on this variable (`total_fat`). You can present it on one dataviz (= no facets).

Think carefully about the choice of data viz. Use coordinates and theme layers to make your data viz visually appealing and meaningful.

Save your file as .rmd

## Assessment 2 Section B 20% Mark

### R libraries to use:`tidyverse`

### Dataset: thanksgiving_meals.csv

Tasks:

See the definition of variables in a separate section “Data dictionary”

To import the data use dataset thanksgiving_meals.csv located in [redacted]

----

1. Use the same (=clone a repository) R Project for the assignment2 as you created for Assignment1. Create a new .rmd document "Assignment2.rmd"

2. Use the provided csv file to complete the tasks below. The file needs to be uploaded to your project. You can use the variable name of your choice.

3. For each question below record your answer in the markdown document that will show the question, your code and the results.

Your explanation of the data insights is VERY important as well as your code

----

### Section2 Part 1: formatting RMarkdown document

Create an Rmarkdown document with webpage as output (same as in setup)

At the start of the output document include your name in italic font and

your student id in bold font as level 2 heading

Separate with a solid line

Include the title “Assignment 2” as level 1 heading

Separate with a solid line

List all tasks in the assignment as headings of the third level and include your results (=output) below each task showing your R code.

### Section 2 Part 2: Data Wrangling and visualization

For all tables below, you need to use the RMarkdown functionality to present tables (`kable`).

1. Display the first 10 rows of the dataset using `kable()` function

2. Using `skim()` display the summary of variables.

Think about the task to predict a family income based on their menu: what variables may be useful? Are all of them correct type?

Write 2-3 sentences with your explanation.

Think about the task to predict a community type or US_region based on their menu: what variables may be useful? Are all of them correct type?

3. Use `fct_reorder` and `parse_number` functions to create a factor variable `family_income`

4. What is the number of people who celebrate?

5. What are categories and insights for each main dish served and the method it is prepared?

6. Create 3 different data viz showing insights for main dish served and the method. Provide your own legend and use themes.

Write 2-3 sentences with your explanation of each insight.

7. How many use cranberry sauce? How many use gravy?

8-9. What is the distribution of those who celebrate across income ranges. Create a data viz.

Write 2-3 sentences with your explanation of each insight.

10. Use the following code to create a new data set

```
select(id, starts_with("side"),
       starts_with("pie"),
       starts_with("dessert")) %>%
  select(-side15, -pie13, -dessert12) %>%
  gather(type, value, -id) %>%
  filter(!is.na(value),
       !value %in% c("None", "Other (please specify)")) %>%
  mutate(type = str_remove(type, "\\d+"))
`
```

Write 2-3 sentences with your explanation of what it does.

11. Intall package `widyr` and use `pairwise_cor()` function https://www.rdocumentation.org/packages/widyr/versions/0.1.3/topics/pairwise_cor

Write 2-3 sentences with your explanation of what it does.

Use this code for the new dataset

pairwise_cor(value, id, sort = TRUE)

Write 1 sentence with your explanation of what insights it shows.

12. Use `lm()` or randomForest() function to build a model that predict a family income based on data in the dataset.

Compare 3 models using different set of input variables. Use different number of variables.

Explain your choice of variables (3 sentences)

Write 2 sentences explaining which model is best.

#### Resources:

RMarkdown tutorial https://rmarkdown.rstudio.com/lesson-1.html

ggplot2: Elegant Graphics for Data Analysis https://ggplot2-book.org/

1/8/2022