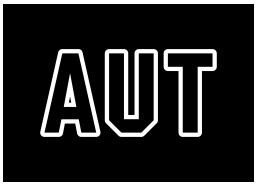

COMP824 Statistical Programming for Data Science
Semester 1, 2023
Assignment

Instructions

- Due date: Submit to Canvas by **Friday 5th May 2023, 11:59pm**
- This assignment is worth **25% of your final grade** and will be marked out of 100 marks.
- The assignment should be completed in **Rmarkdown**. This enables you to embed R code and text into the same document. You should use the file `COMP824_Assignment_template.Rmd` as a starting point.
- You should submit two files:
  - The Rmarkdown `.Rmd` file
  - The corresponding PDF file
- Both the source file and PDF will be marked.
- The source files (.Rmd) should be able to be compiled by the lecturers. When you compile your source file, the data file/s should be located in the **same directory** as your source file.
- Do **not** upload your submission as a zip file.
- For some questions an explanation should be provided, in addition to the relevant code and output.
- Submissions which contain large quantities of unnecessary code, output or text will be penalised.
- **Presentation:** 10 marks are awarded for the presentation of your assignment. This includes factors such as grammar, spelling, and code elegance.
- **Late Assignments:** Failure to submit the assignment on time will result in a penalty in accordance with the DCT late policy (5% per day up to a maximum of 5 days). If extenuating circumstances (e.g. illness) prevent the timely submission of your assignment you can apply for special consideration. You may also apply for special consideration if such circumstances result in your submission being incomplete. Applications for special consideration should be submitted via Canvas.
- **Originality:** This assignment is an **individual piece of work**. You are encouraged to discuss the assignment with your lecturers and classmates, however, the work you submit must be your own. Assignments that show similarities to work submitted by other students will be investigated for **plagiarism** and treated very seriously. Plagiarism software, such as TurnItIn, may be used to electronically compare submissions to those of other students and to documents on the internet. Talk to the lecturer if you have any questions about this requirement.

Question:	1	2	3	4	5	Total
Marks:	6	24	18	42	10	100
Score:						

Overview

You work for an data science consultancy which has been employed by a large retail chain to analyse various aspects of their sales data. The dataset `COMP824_sales_data.csv` contains historical sales data. Data is available from multiple different stores and for all their products (SKUs). [1] Further information about the variables in the dataset is provided on page 3.

Each student has been assigned a different product based on their student ID number (see page 4). This means every student will work with a unique dataset and will get slightly different results. For all questions you should use the data for the product (`sku_id`) which you have been assigned.

1. **Load and extract the data** Total for Question 1: 6 *marks*

   (a) Download the file `COMP824_Assignment_template.Rmd` from Canvas. Rename it by replacing the word "template" with either your name or student ID. (0 *marks*)

   (b) Download the file `COMP824_sales_data.csv` from Canvas and save it in the same directory as `COMP824_Assignment_YOURNAME.Rmd`. Do **not** rename the .csv file. (0 *marks*)

   (c) Load the data into RStudio using the command `readr::read_csv()` [2] and extract the records for your product using the `dpylr::filter` command. (6 *marks*)

   Use the command `print` with the appropriate options to print the top 5 rows of `sales`.

```
library(tidyverse)
sales_all <- read_csv("COMP824_sales_data.csv")
# sales <- #add filter command here
# print()
```

2. **Explore the Sales Data** Total for Question 2: 24 *marks*

   (a) What is the date range of the sales data? Provide R code and output required to determine this and write your answer in a sentence. (6 *marks*)

   (b) How many different stores sell your product (`sku_id`)? Provide R code and output required to determine this and write your answer in a sentence. (6 *marks*)

   (c) Compute some summary statistics and construct a histogram using the `ggplot2` package to analyse the variable `total_price` for your product. (6 *marks*)

   (d) Write 2-3 sentences describing your findings about the `total_price` variable from part (c). (6 *marks*)

3. **Analysis of Monthly Sales Data** Total for Question 3: 18 *marks*

   (a) Compute the total monthly sales for your product from 1st January 2011 – 30th June 2013. Print a tibble showing the 6 months with the highest total monthly sales. (6 *marks*)

   (b) Use `ggplot2` to present total monthly sales for your product in an appropriate plot. Ensure your graph has appropriate titles, labels, scales etc. (6 *marks*)

   (c) Write 2 – 3 sentences describing the plot in part (b). (6 *marks*)

4. **Analysis of Store Performance**

   The GM Sales wants to know which stores are performing well, in terms of product sales. You should analyse the data for the product (`sku_id`) which has been assigned to you. Total for Question 4: 42 *marks*

   (a) Use appropriate `tidyverse` functions to compute the total sales per store. Print a tibble showing total sales by store, sorted by total sales in decreasing order. (8 *marks*)

[1] The original dataset is available via Kaggle. The data has undergone some cleaning for the purposes of this assignment. More information the original dataset is available here: `https://www.kaggle.com/aswathrao/demand-forecasting?select=train_0irEZ2H.csv`

[2] The syntax `readr::read_csv()` means that the function `read_csv()` is part of the R package/library `readr`

Question 4 continues . . .

(b) Create an appropriate plot using `ggplot2` to visualise the total sales per store from part (a). (8 *marks*)
Hint: you may need to use a function like `as_factor` to ensure the store id is visualised correctly.

(c) Compute another performance metric (different to total sales in part (a)) in order to investigate the performance of stores. Print a tibble showing the results. (8 *marks*)
Note: for full marks, students should show creativity in the choice and computation of the performance metric.

(d) Create an appropriate plot using `ggplot2` to visualise the performance metric in part (c). (8 *marks*)

(e) Write 1 - 2 paragraphs for the GM Sales discussing your findings from parts (a – d). (10 *marks*)

5. **Presentation and Formatting** (10 *marks*)

Requirements for full marks (in order of importance):

- Any resources used should be correctly referenced. [3]
- Assignment should use the template provided.
- Assignment should be professionally presented and contain accurate spelling and grammar. [4]
- All data wrangling and analysis should be reproducible.
- The PDF file is able to be compiled by the lecturer.
- The PDF file should show appropriate code and output.
- R code should adhere to 'good practice' guidelines for R scripts.
- Results should be reported in-text using "inline" R commands, rather than hard-coded.[5]
- Code is elegantly written and makes extensive use of packages in the `tidyverse`.

Further information about dataset

Sales Data

Variable	Description
record_ID	Number of record in original dataset
week	Start of weekly sales period
store_id	ID number of retail store
sku_id	Stock-keeping unit ID number (ID of product)
total_price	Total price (dollars)
base_price	Base price (dollars)
is_featured_sku	Was the product featured during the sales period? 1 = Yes, 0 = No
is_display_sku	Was the product on display during the sales period? 1 = Yes, 0 = No
units_sold	Number of units sold during the sales period.
year	year of variable week
month	month of variable week
day	day of variable week
weekday	day of the week corresponding to the variable week
start_of_month	the start of the month corresponding the variable week
end_of_month	the end of the month corresponding the variable week

[3] For guidance on referencing refer to the AUT Library `http://aut.ac.nz.libguides.com/APA6th`

[4] For additional guidance on proof reading, grammar, and referencing, please refer to the Student Learning Centre. `https://www.aut.ac.nz/student-life/student-support/student-hub`

[5] See `https://rmarkdown.rstudio.com/lesson-4.html`

Allocation of products

Each student has been assigned a product (`sku_id`) to analyse for this assignment. The following table shows the last three digits of your student ID number and the corresponding `sku_id`.

Last 3 digits of Student ID number	Sales data product: sku_id
037	219009
049	216233
180	219029
270	223245
277	222087
280	217390
430	222765
431	216418
574	216419
591	216425
612	223153
690	300021
748	245387
785	320485