AUT University
School of Engineering, Computer and Mathematical Sciences
Assignment
COMP828: Statistical Programming for Data Science

The purpose of this assignment is to assess your analytical and computing skills on the material covered.

Total Possible Marks: 100 marks, which contribute 25% towards your final grade in this paper.

Deadline: 11:59pm, Friday, May 3, 2024

Report/Assignment:

- 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 `COMP828_Assignment_template.Rmd` as a starting point.
- For some questions an explanation should be provided, in addition to the relevant code and output.

Submission:

- You should submit two files: 1) the Rmarkdown `.Rmd` file and 2) the corresponding `.pdf` file.
- Your filename must include 1) your lastname, 2) your firstname, and 3) your student id, e.g., if Jane Doe submits her assignment, her .pdf file must be named "Doe_Jane_123456789".
- 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.
- Submissions which contain large quantities of unnecessary code, output or text will be penalized.

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 assignment 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/Plagiarism: 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.

Overview

- You work for a data science consultancy which has been employed by a large retail chain to analyze various aspects of their sales data. The dataset `COMP828_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 below.
- **Each student has been assigned a different product based on their student ID number (see last page). 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.**

Question 1 Load and Extract the Data (6 marks)

Load the data into RStudio using the command `readr::read_csv()` and extract the records for your product using the `dpylr::filter` command. Use the command `print` with the appropriate options to print the top 5 rows of sales. (6 marks)

```
library(tidyverse)
sales_all <- read_csv("COMP828_sales_data.csv")
# sales <- #add filter command here
# print()
```

Question 2 Explore the Sales Data (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 analyze the variable `total_price` for your product. (6 marks)

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

Question 3 Analysis of Monthly Sales Data (18 marks)

(a) Compute the total monthly sales. **Here, sales means number of units sold** (`units_sold`). 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)

[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

Question 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. (42 marks)

(a) Use appropriate tidyverse functions to compute the total sales per store. **Here, "sales" refers to the number of units sold** (`units_sold`). Print a tibble showing total sales by store, sorted by total sales in decreasing order. (8 marks)

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

(c) Compute another performance metric (different to total sales in part (a)) in order to investigate the performance of stores. **Any assumptions made about the meaning of the variables in the dataset should be reasonable and clearly stated.** Print a tibble showing the results. *Note: for full marks, students should show creativity in the choice and computation of the performance metric.* (8 marks)

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

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

Question 5 Presentation and Formatting (10 marks)

Requirements for full marks (in order of importance):

- Any resources used should be correctly referenced.
- Assignment should use the template provided.
- Assignment should be professionally presented and contain accurate spelling and grammar.
- 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.
- 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

Variable	Description
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

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
973, 045, 380	219009
870, 398, 425	216233
958, 747, 418	219029
865, 089, 104	223245
999, 573, 919	222087
406, 114, 415	217390
908, 858, 790	222765
655, 187	216418
420, 537	216419
520, 838	216425
914, 567	223153
389, 098	300021
352, 616	245387
859, 245	320485