STAT702 Industrial and Business Analytics Project

Instructions

- Due date: All deliverables should be submitted to Blackboard by **Wednesday 26th May 2021, 10pm**
- This project is worth **30% of your final grade** and will be marked out of 100 marks.
- This is a group project and should be completed by groups of 2 students.
- **Summary of deliverables:**
  - Each group should submit two files to Blackboard:
    - The Rmarkdown `.Rmd` file containing your report
    - The corresponding PDF file
  - Each individual should complete:
    - Self and Peer evaluation form (via Blackboard)
    - Individual Contribution Statement (included within the report)
- **Requirements for the report**
  - The report should be written in R Markdown.
  - The report should be no more than 15 pages in length (single spaced, including the title page and any appendices).
  - All code should be included within the R Markdown file, but hidden in the PDF file.
  - The relevant results and graphs should be appropriately formatted and shown within the PDF file. For example:
    - Tables should be formatted using `xtable` or equivalent.
    - Numerical results referred to in-text should be referenced rather than hard-coded (e.g. `` `r mean(x)` ``)
    - Raw R output should not be included in the PDF file.
  - The R Markdown file should be able to run on any computer when it is located within the same folder as the data files.
  - The report should include the completed Group Assessment Cover Sheet. Ideally it should be included within the Rmd and PDF documents (see Rmd examples on Blackboard), but can be submitted separately. It should not be included within the page limit for the report. The form is available from Blackboard[1].
- **Late Assignments:** Failure to submit the assignment on time will result in a penalty in accordance with the policy outlined in the STAT702 Study Guide. 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 Blackboard.
- **Originality:** Reports that show similarities to work submitted by other students or material available on the internet 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:	25	40	25	10	0	100
Score:						

[1] Blackboard/ Assessment/ Assessment Policies, Regulations, Guides and Forms/ Forms and Coversheets/ Group Assessment Cover Sheet

Overview

You work for an analytics consultancy which has been employed by a large retail chain to analyse various aspects of their sales data, as detailed below. The results of your analysis should be written in a report which will be presented to the management of this retail chain. Two datasets have been provided for this analysis. Further information about both datasets is available on page 6.

- The dataset `STAT702_project_sales_data.csv` contains historical sales data. Data is available from multiple different stores and for all their products (SKUs). This dataset should be used for questions 1 and 2. [2]
- The dataset `STAT702_project_reviews_data.csv` contains customer reviews on a range of products from Amazon.com. This dataset should be used for question 3. [3]

---

1. **Analysis of Sales Data**

   (a) For the product (`sku_id`) which has been assigned to your group (see page 6), compute the total monthly sales from January 2011 – September 2013. Present your results in an appropriate plot and write 2 – 3 sentences describing your results. (5 *marks*)

   Hint: This will require some "wrangling" of the variable `week`. To do this, format `week` as a date and then use the appropriate `lubridate` function to extract the month.

   Marking Criteria
   - Total monthly sales have been correctly computed and are displayed in an appropriate plot.
   - Description of results/plot is correct and provides useful insights.
   - Plot is constructed using `ggplot2` and has appropriate titles, labels, scales etc.

   (b) The GM Sales wants to know which stores are performing well and which are not, in terms of product sales. For the product (`sku_id`) which has been assigned to your group, use appropriate summary statistics and plots to investigate sales performance across the stores and write 2 – 3 paragraphs summarising your findings. (20 *marks*)

   Hint: You will need to decide what it means for a store to be "performing well" and how you will evaluate this using the data.

   Marking criteria
   - Sales performance is clearly defined.
   - Written summary includes relevant and appropriate summary statistics and plots.
   - Plot/s are constructed using `ggplot2` and have appropriate titles, labels, scales etc.
   - Descriptions of results and plots are correct and provides useful insights.

2. **Analysis of Inventory**

   The retail chain has a central warehouse from where it supplies products to all its stores. In this part of the project you will analyse data from 2012, i.e. data from weeks which began in 2012.

   (a) The Operations Manager is interested in studying an EOQ model for product 216233, based on sales in 2012. The setup and holding costs are known to be \$130 per order and \$1.50 per unit per year, respectively.

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

[3] The original dataset is available here: `https://nijianmo.github.io/amazon/index.html`. The data has undergone some cleaning for purposes of this project. For more information on original dataset refer to: Jianmo Ni, Jiacheng Li, Julian McAuley, Empirical Methods in Natural Language Processing (EMNLP), 2019.

i. Determine the best order quantity in such a way that the costs are minimised. Write 1 – 2 paragraphs summarising your findings. (10 *marks*)

Marking criteria
- Number of orders during a year, number of days between orders, and the total annual inventory cost are correctly computed and included in the findings.
- The paragraphs clearly explain your findings.
- Assumptions of the EOQ model are clearly stated.

ii. The Operations Manager is also interested in studying a model in which backorders are permitted. According to its estimates, the cost of backorders is approximately 5% of the total price (price per unit). Determine the best order quantity in the sense that inventory costs are minimised. Write 1 – 2 paragraphs summarising your findings and plot the first two inventory cycles. (10 *marks*)

Marking criteria
- The optimum order quantity, maximum level of stock, optimum time between orders, proportion of time the company have to take backorders, and total annual inventory cost are correctly computed and included in your answer.
- The paragraphs clearly explain your findings.
- Assumptions of the model are clearly stated.
- The first two inventory cycles are correctly plotted.

iii. Plot the inventory cycles associated with the model in part ii and compare with the observed inventory levels in 2012, assuming actual demand during 2012, and the order frequency and order quantity from the model. Write 2 – 3 sentences describing your plot. (5 (bonus))

Marking criteria
- The inventory levels from the model and data are correctly plotted.
- Accurate and insightful comments are made about the plot.
- Note: This is a bonus question. The maximum mark that could be awarded for this project is 100.

(b) The Operations Manager is considering the option of a multi-period inventory model. The company, as a policy, is not willing to tolerate more than 5% chance of a stock-out. The Operations Manager has estimated that the annual holding cost is $6.50 per unit and the ordering cost is $20.50 per order.

i. Calculate a multi-period inventory model for product 216425, based on the 2012 sales data. Create plot/s of the weekly average demand of this product. Use the costs stated in part (b) above. Write a paragraph explaining the results of your model and the plot/s. (10 *marks*)

Hint: Use the weekly demand to estimate the demand during a one-week lead time.

Marking criteria
- The optimal order quantity, safety stock, expected annual cost, orders per years are correctly computed and included in your answer.
- The paragraph clearly explains your findings.
- The assumption of normality for the demand during a one-week lead time is discussed.
- The weekly average demand of this product is correctly plotted and discussed.

ii. Investigate the use of a multi-period inventory model for the product which has been assigned to your group, based on the 2012 sales data. Create plot/s of the weekly average demand of this product. Use the costs stated in part (b) above. Discuss the assumptions of the model and suggest a solution, in case of finding any problems. Write a paragraph explaining the results of your findings and the plot. (10 *marks*)

Marking criteria

- The optimal order quantity, safety stock, expected annual cost for this system, order per years are correctly computed and included in your answer.
- The paragraph clearly explains your findings.
- The assumption of normality for the demand during a one-week lead time is discussed.
- Assumptions of the model are clearly stated and assessed for the selected product.
- Implications of your findings are discussed.
- The weekly average demand of this product is correctly plotted and discussed.
- Sensible solutions are proposed in the event that the assumptions are not met.

3. **Analysis of Customer Reviews** The General Manager – Sales wants to expand the company's product range, but before doing so wants to gather some information about how the potential products have been perceived by consumers. The potential products are currently sold via Amazon, so there is a wealth of data available through the customer reviews. Using review data from Amazon, the GM sales wants you to analyse the ratings and written text reviews to determine issues that have lead to customer satisfaction/dissatisfaction.

   (a) For the product (`asin`) that has been assigned to your group, use summary statistics and plots to analyse the overall review rating (`overall`). Write a paragraph describing your findings to the General Manager - sales. (5 *marks*)

   Marking Criteria

   - Summary statistics for the overall review rating have been correctly computed and are displayed in appropriate plot/s.
   - Descriptions of results and plots are correct and provide useful insights.
   - Plot/s are constructed using `ggplot2` and have appropriate titles, labels, scales etc.

   (b) Using the review text (`reviewText`) and any other variables you think are relevant, investigate the customer sentiment towards, and satisfaction/dissatisfaction with, the product that has been assigned to your group. Your answer should include a word cloud and a sentiment analysis. (20 *marks*)

   Marking Criteria

   - Appropriate methods are used to tidy the text data.
   - Correctly construct and interpret a word cloud of the `reviewText` variable.
   - Correctly perform and interpret a sentiment analysis of the `reviewText` variable.
   - Correctly perform and interpret some additional analysis of the `reviewText` variable, incorporating at least one other variable from the dataset.
   - Interpretations of analyses are correct, provide insight and are written at an appropriate level for a manager.

4. **Presentation and Formatting:** Requirements for full marks (in order of importance): (10 *marks*)

   - All resources used are correctly referenced. [4]
   - Report should be professionally presented and contain accurate spelling and grammar. [5]
   - Figures should have appropriate titles and labels.
   - All data wrangling and analysis are reproducible.
   - The PDF file should not show code or raw output.
   - Numerical results should be rounded appropriately.
   - 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.[6]

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

[5] 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`

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

- Code is elegantly written and makes extensive use of packages in the `tidyverse`.

5. **Appendix A: Individual Contribution Statement** (0 *marks*)

   Together the group should agree on a percentage allocation of the work undertaken and should state this at the start of Appendix A. This information, along with the *Individual Contribution Statement* and the *Self and Peer Evaluation*, will be used to determine the marks allocated to each group member. The percentages must sum to 100%.
   Example statement: *Person A 45%, person B 55%.*

   Each group member should write 75 - 150 words detailing their individual contribution to this project. These statements should be collated and included in the report. Marks are not awarded for this statement but if it is absent it will result in a deduction of marks for the corresponding group member.

Question 5 continues . . .

Allocation of products

For some of the questions above you need to analyse the product that has been assigned to your group. Look on Blackboard to find your group number and then use the table below to identify the product you need to analyse.

Group	Sales data product: sku_id	Reviews data product: asin
1	216418	B00000JRRD
2	216419	B00005249G
3	219009	B00006IAKM
4	219029	B00006IEEV
5	219844	B00006IE7J
6	222087	B00006IFAY
7	222765	B00006IFEU
8	223153	B00006IFI5
9	223245	B00006IFJ0
10	245338	B00006JNNS
11	245387	B00006JNJD
12	300021	B0008G8G8Y

Further information about datasets

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.

Review Data

Variable	Description
title	name of the product
brand	brand name
main_cat	list of categories the product belongs to
price	price in US dollars (at time of crawl)
asin	ID of the product, e.g. 0000031852
document.id	id variable
overall	rating of the product
reviewerID	ID of the reviewer, e.g. A2SUAM1J3GNN3B
verified	is the review verified
reviewTime	time of the review (raw)
reviewerID	ID of the reviewer
reviewerName	name of the reviewer
reviewText	text of the review
summary	summary of the review
unixReviewTime	time of the review (unix time)