AM01 Group Project Sep 2022

Data: sales.csv contains weekly sales by department for 45 stores of a US retailer; details.csv and stores.csv contain additional information. A data dictionary is provided in the appendix.

Submission: You need to organise your work as an R Project and push your answers to a private repo on Github. You must add me as collaborator so I can check your work.

On Canvas, you need to upload a knitted HTML file which will be the result of knitting your Rmd – *study_group_number.HTML*
N.B. Do not make the mistake of printing the dataframe inside your Rmd. The resulting HTML will be hundreds of pages long and your Rmd will take forever to knit.

Part A: Exploratory Data Analysis (EDA)

In the **R4DS** Exploratory Data Analysis chapter http://r4ds.had.co.nz/exploratory-data-analysis.html the authors state:

"Your goal during EDA is to develop an understanding of your data. The easiest way to do this is to use questions as tools to guide your investigation... EDA is fundamentally a creative process. And like most creative processes, the key to asking quality questions is to generate a large quantity of questions."

Conduct a thorough EDA. Recall that an EDA involves three things:

1. Looking at the raw values, identifying any missing or wrong values
2. Computing summary statistics of the variables of interest
3. Creating informative visualizations with ggplot. In all cases, please think about the message your plot is conveying. The title shouldn't just be "This is my X-axis, this is my Y-axis", but rather what's the key message, the ***so what*** of the plot.

As a minimum, any EDA should address the following questions:

1. How many variables/columns? How many rows/observations? Any NAs? How do we handle them?
2. Which variables are numbers? Which are categorical or factor variables (numeric or character variables with variables that have a fixed and known set of possible values?
3. What are the correlations between variables? Does each scatterplot support a linear relationship between variables? Do any of the correlations appear to be conditional on the value of a categorical variable?

Specific questions for our data

- How many **more** weeks of data are there in details.csv compared to the sales.csv data?
- What is the date range covered by sales.csv (in terms of first day of sales to the last day of sales)?
- There is variation in the number of departments by store. What is the **maximum** number of departments within a single store?
- What can you infer about the different Types of Store?
- Which store had the **most** sales between weeks starting 07/01/2011 and 30/12/2011? What was its average sales per week?
- What kind of a distribution do store sales follow? Can you create a useful graph to show spread of sales by store, ranked by median weekly sales?
- Which numerical variable in details.csv has the highest correlation with sales? What is the value of this correlation?
- Are there any stores with **mean** sales during holiday periods lower than normal trading weeks?
- Are there any stores with **median** sales during holiday periods lower than normal trading weeks?

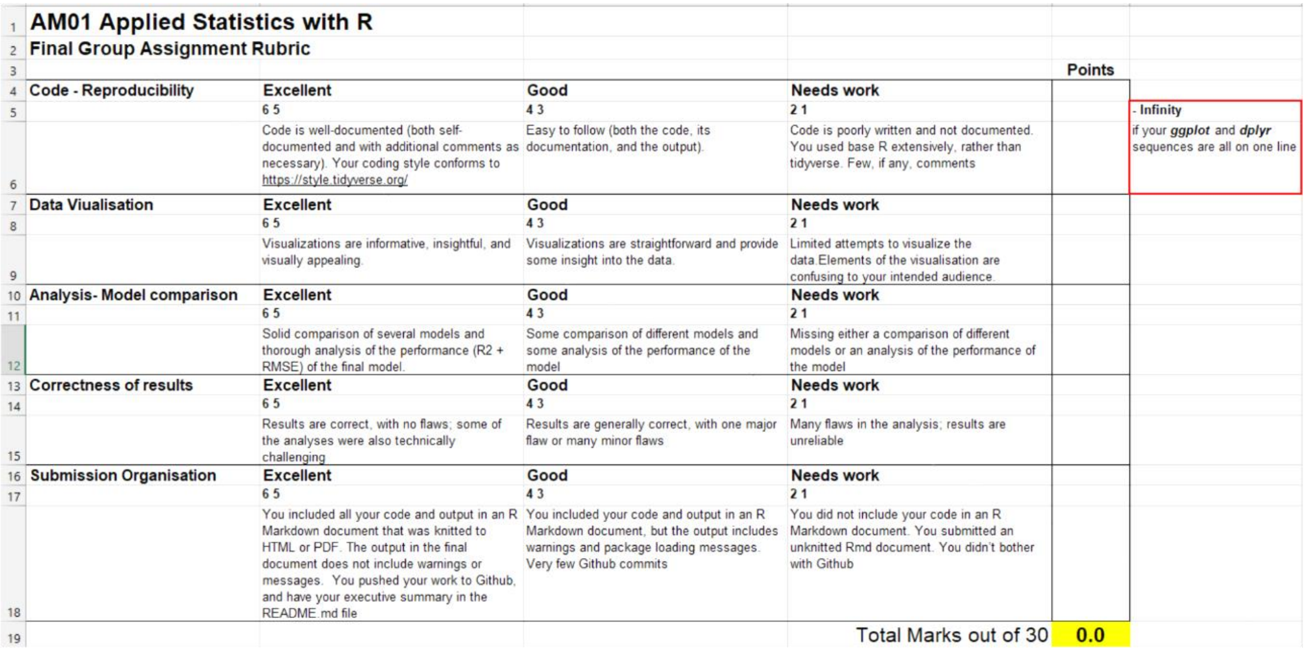
Part B: Inferential Statistics

- We are interested in understanding the difference in impact on sales of the different holiday weeks. The ***IsHoliday*** variable indicates the week of occurrence for 4 different holidays through the course of the year: Christmas, the Super Bowl, Thanksgiving & Labor Day. Given that Christmas occurs in December, the Super Bowl occurs in February, Thanksgiving in November and Labor Day in September:
  - Calculate which holiday ($, in total for all stores) generates the highest average weekly sales
  - How much more value ($, in total for all stores) is generated during the average week in the holiday given in a) compared to Labor Day?
- There is a known increase in sales during Thanksgiving; we want to know whether there is a significant difference is sales during other holiday periods and normal trading weeks. After filtering out Thanksgiving from the data, test whether the difference in weekly sales on holiday weeks compared to non-holiday weeks is significantly different. What is the p-value of this test?

Part C: Regression

- Fit a linear regression model where the response variable is weekly sales. Start by running a model with all explanatory variables. Are all predictors coded appropriately? Do you have to remove any variables from your model? Why?
- For your best regression model
  a) What impact on sales is associated with a unit change in temperature? Fuel Price?
  b) On average, how much more do we expect to sell on Labor Day weeks compared to Super Bowl weeks?

The grading rubric for the final group project is attached.



AM01 Applied Statistics with R

Final Group Assignment Rubric

				Points	
Code - Reproducibility	**Excellent**	**Good**	**Needs work**		
	6 5	4 3	2 1		- **Infinity**
	Code is well-documented (both self-documented and with additional comments as necessary). Your coding style conforms to https://style.tidyverse.org/	Easy to follow (both the code, its documentation, and the output).	Code is poorly written and not documented. You used base R extensively, rather than tidyverse. Few, if any, comments		if your ***ggplot*** and ***dplyr*** sequences are all on one line
Data Viualisation	**Excellent**	**Good**	**Needs work**		
	6 5	4 3	2 1		
	Visualizations are informative, insightful, and visually appealing.	Visualizations are straightforward and provide some insight into the data.	Limited attempts to visualize the data.Elements of the visualisation are confusing to your intended audience.		
Analysis- Model comparison	**Excellent**	**Good**	**Needs work**		
	6 5	4 3	2 1		
	Solid comparison of several models and thorough analysis of the performance (R2 + RMSE) of the final model.	Some comparison of different models and some analysis of the performance of the model	Missing either a comparison of different models or an analysis of the performance of the model		
Correctness of results	**Excellent**	**Good**	**Needs work**		
	6 5	4 3	2 1		
	Results are correct, with no flaws; some of the analyses were also technically challenging	Results are generally correct, with one major flaw or many minor flaws	Many flaws in the analysis; results are unreliable		
Submission Organisation	**Excellent**	**Good**	**Needs work**		
	6 5	4 3	2 1		
	You included all your code and output in an R Markdown document that was knitted to HTML or PDF. The output in the final document does not include warnings or messages. You pushed your work to Github, and have your executive summary in the README.md file	You included your code and output in an R Markdown document, but the output includes warnings and package loading messages. Very few Github commits	You did not include your code in an R Markdown document. You submitted an unknitted Rmd document. You didn't bother with Github		
			Total Marks out of 30	**0.0**	

Appendix: Data Dictionary

Dataframe: sales.csv

Contains: Weekly sales data by department by store

Variable Name	Description
Store	Store ID
Dept	Department ID
Date	Date of the starting day of the week to which the data is attributed
Weekly_Sales	Total sales for that week in US Dollars
IsHoliday	Identifier of whether that week is one of the 4 weeks of the year which falls on a significant American holiday (Christmas, Thanksgiving, Labor Day, Super Bowl)

Dataframe: details.csv

Contains: Additional data related to the store and its region over time

Variable Name	Description
Store	Store ID
Date	Date of the starting day of the week to which the data is attributed
Temperature	Average temperature in the region (Degrees Fahrenheit)
Fuel_Price	Price of fuel in the region (US Dollars per Gallon)
Markdown1-5	Anonymized data related to different types of promotional markdowns. If there's no value, assume there was a zero markdown
CPI	Consumer Price Index
Unemployment	Unemployment rate, %
IsHoliday	Identifier of whether a week is during a significant American holiday (Christmas, Thanksgiving, Labor Day, Super Bowl)

Dataframe: stores.csv

Contains: Anonymised information about the stores in the data

Variable Name	Description
Store	Store ID
Type	Anonymous tag, related to the format of the store
Size	Size of the store, related to the number of products it sells