BUS111 Business Analytics
Fall 2018

Final Project: Investigating Data & Producing Managerially Relevant Decisions

Interim Report – I

Introduction: The first interim report for the team project involves introducing yourself to the dataset, and performing basic exploratory analyses. This would include creating summary statistics tables and key variable visualizations. The objective is that your team understands the broad managerial goal for the project you have chosen and familiarizes with the dataset provided.

Datasets: Your team has chosen one of these four datasets –

- Synthetic Financial Data from PaySim for Fraud Detection (https://www.kaggle.com/ntnu-testimon/paysim1/home)
- Analyzing Black Friday Purchases for customers and product categories (https://www.kaggle.com/mehdidag/black-friday/home)
- Analyzing the viability of Kickstarter projects (https://www.kaggle.com/kemical/kickstarter-projects/home)
- 2015 Flight Delays and Cancellations – which airline to fly on?(https://www.kaggle.com/usdot/flight-delays/home)

For detailed description of these datasets, please read the Team Project description document and the associated Kaggle webpages.

Assignment: The data assignment and tasks depend on the project chosen. Your task is to describe the basic patterns in your chosen dataset. In terms of report structure, your report should have four components –

1. **Introduction –** This should briefly talk about the data problem, why is it interesting to look at this problem (i.e. managerial objective), and the broad goals of your project.
2. **Data Description –** This should "introduce" the dataset to the reader. It should cover the following points –
   a. Describe the "conceptual" measure types of the different variables in your data.
   b. Data Cleaning – Mention all the steps you took to clean the data. This could include changing the computer data-types of the variables (type coercion), dealing with missing data, filtering out observations, selecting variables, etc. At the end of this process, you would have ONE cleaned dataset.
3. **Summary statistics and Data Visualizations** – This would describe the basic patterns in your CLEANED dataset. Specifically, apart from showing the code and resulting output, you should explain the following -

a. Why did you select the variables (and summaries) that you chose?
b. Why did you select the type of visualization for these variables? (*Hint:* Relate tp the conceptual measure-types for these variables)
c. How did you improve the graph from its initial ggplot2 output? (*Hint:* Specify some choices of aesthetics, facets or themes that helped improve the visualization of your plot)
d. What question are you trying to answer with each summary table or visualization?
e. What conclusion do you draw about the answer to your question?

Summary statistics can include-

a. One-way frequency tables
b. Two-way frequency tables
c. Summary tables
d. Any other summary tables as appropriate

Visualizations can include

a. Histograms, bar-graphs, density plots
b. Box-plots
c. Scatter-plots, line-plots
d. Any other plots as appropriate

4. **Conclusion -** Discuss what the data patterns indicate, and what this could mean for your firm's managers

While the above points represent basic guidelines, teams have autonomy in choosing which details are most interesting to include in the report. These also involve choosing the key variables for analysis. Brevity is appreciated.

Reports:

There is a LATTE Assignment page, and each group must submit a zipped folder with –

a. The RMarkdown file (.Rmd)
b. The knitted RMarkdown PDF file (see instructions below).

Page limit – 10 pages.

Specific requirements and guidelines:

- You are permitted to supplement your data with other publicly available data if you deem it appropriate.
- You must create a Project in your Team's private GitHub repository, and use RStudio (and/or SourceTree) to track versions and contributions of different team members. That project will contain your R Markdown code and knitted files.

- If you are using one of the Kaggle-supplied kernels, you must first get the original kernels to run, and cite them in the report. As part of your report and presentation, you should interpret the output of the code.
- You must, in some fashion, make a *substantial modification* to the original kernel (if any) to better respond to the supplier's questions. This might involve selecting different explanatory features, transforming data to accommodate non-linearity, using a different modeling approach, creating a more informative visualization, or other similar changes. Improvements should go beyond simple cosmetic changes or dropping a variable from the model. If you're uncertain whether a particular change is substantial, speak with me.
- Your R Markdown file should contain relevant code as well as narrative explaining the different parts of your investigation, explaining what each code chunk does and why you created the chunk of code. As with earlier assignments, knit the markdown file to a Word doc, then **resave** that document as a pdf for upload to LATTE.
- You **should not** be making any changes to the knitted word document. When I compile the RMarkdown code, I should see the same output as your final PDF document.

Under no circumstances should any individual or team in the class copy or share their code with any other individual or team. Code sharing, however minimal, will result in a score of 0 points for all members of any team involved.