SI 618 WN 2024 Project Part II: **Analysis**

Version 2024.03.06.1.CT

Now that you have successfully formed your team, selected your datasets, framed your research questions, and conducted basic manipulation to clean and merge your data, you are ready to dive into the data analysis component of this team project. The purpose of this section is to apply statistical and graphical analyses to extract insights and answer the research questions posed. You should plan to **focus on using only the statistical and graphical analysis techniques covered in the course** thus far and ensure you do **not** utilize any machine learning methods for this part. Your work will consist of the following sections, all of which should be included in one notebook that you will submit in the usual way via Canvas.

NOTE: We strongly encourage you to work on this over the next two weeks and not leave things until the last minute as this assignment requires considerable time and effort.

Descriptive Statistics

Provide a comprehensive summary of your combined dataset using descriptive statistics. This should include means, medians, modes, ranges, variance, and standard deviations for the relevant features of your data. The descriptive statistics should inform your guiding questions that you developed in Part I of the project, rather than merely providing an overview of your data. Interpret these results to draw preliminary conclusions about the data.

Inferential Statistics

Conduct appropriate hypothesis tests to investigate if there are significant differences or correlations within your data. This might involve regression analysis, ANOVA, and/or chi-squared tests.

Clearly state your null and alternative hypotheses, choose an appropriate significance level, and discuss your findings. Make sure to justify the choice of your tests.

Graphical Analysis

Create various types of plots to visualize relationships within your data. Use histograms, bar charts, scatter plots, box plots, and any other suitable graphical representations you've learned.

Be sure to use appropriate titles, labels, and legends to make your plots readable and informative.

Interpret the graphical representations to uncover patterns, trends, and outliers.

Comparative Analysis

Compare and contrast different subsets of your data. This can include comparisons over time, across different categories, or any other relevant segmentation. Note that for some projects, the nature of this comparative analysis will be obvious. For others, you will need to think about how you might subset your data.

Discuss any notable similarities or differences you have identified.

Multivariate Analysis

Perform multivariate analysis to understand the relationships among three or more variables in your dataset.

Use techniques like cross-tabulation, pivot tables, and multivariate graphs.

Synthesis

Synthesize the findings from your descriptive and inferential statistics along with your graphical analyses to answer your research questions.

Discuss how the combination of the datasets has provided added value in terms of insights or capabilities that would not be possible with the individual datasets in isolation.

Documentation

Outline the steps you took in your analysis, providing the rationale for the choice of tools and techniques.

Clearly document your code, analyses, and interpretations so that they are understandable and reproducible.

Reporting & Interpretation

Prepare markdown blocks that summarize your methodology, analysis, and findings. You should intersperse these with your code.

Your notebook should be well-structured, with clear sections, headings, and subheadings.

Reflection

Include a section (using one or more markdown blocks) at the end of your notebook in which you reflect on the process of analyzing the data. Discuss any challenges you encountered and how you overcame them.

Critically evaluate the limitations of your analysis and suggest areas for further research or improvement.

Submission Instructions

Submit your Jupyter Notebook with all the code, documentation, and interpretations used for your analysis by the project. Be sure to adhere to PEP-8 and general style guidelines shared earlier in the course.

Rubric

Grades (200 points in total) will be allocated as follows:

Descriptive statistics (30 points)
Inferential statistics (30 points)
Graphical analysis (30 points)
Comparative analysis (30 points)
Multivariate analysis (30 points)
Synthesis (25 points)
Reflection (25 points)

Note that "Reporting and Interpretation" and "Documentation" are **included** in each of the sections listed above. You are encouraged to use the rubric categories as headings in your notebook.

Similar to the previous project deliverable, each component will be graded on the following 5-point scale:

5: Excellent work, goes **beyond** stated requirements, shows innovation
4: Good work, meets stated requirements, does not offer particularly innovative approaches or scenarios
3: Fair work, meets some stated requirements, some errors and omissions
1-2: Inadequate work, meets only a few requirements, errors present
0: Work is absent

Good luck, and we look forward to seeing your analytical skills in action!