**Group Assignment**

*EBIS 4043 Big Data Analytics*

The goal of this assignment is to conduct big data analytics for a business company/event and suggest design improvements for the company/event, using the visualization and statistical techniques you have learned so far. Remember that you are taking a business course - you are learning how to connect the use of the techniques of analytics with real business issues. The ability to observe the changes in your dataset, interpret the data analysis results and formulate the solutions to business problems is of particular importance.

**Group Size:**

4 to 6 students

**General Guideline:**

You are asked to use R to do the analysis.

1. Introduction: The first important step is to propose the business problem you want to solve.
2. Explain your variables. In this step, you need to explain where you get the data and then build up the data dictionary to detail your variables.
3. Process, transform, and describe data (You may need to transform you data and deal with the missing values in this step).
4. Perform below analysis on your dataset to get some findings.

4a. If all your data are structured and numerical, you may conduct statistical analysis:

* Use descriptive statistical measures to draw some general but conclusions about your dataset.
* Use univariate statistical data analysis techniques, such as A/B testing and correlation, to investigate the relationship between different variables and in particular the effect of explanatory variables on the dependent variables.
* Build a model (e.g., multiple regression, logistic regression, or time series analysis) to identify significant factors affecting the the problem you concern.
* You may consider the model building, variable selection and diagnostic checking techniques, and your procedures/arguments in arriving/choosing the final model.

4b. If you are interested into semi-structured or unstructured data, you may perform machine learning/data mining techniques to identify patterns and relationships within data

* Use descriptive statistical measures to draw some general but conclusions about your dataset
* Build a model (e.g., decision tree, cluster analysis) to solve a business problem you concern

4c. If you are interested into textual data, you can also conduct semantic analysis.

1. Obtain plot graphs (e.g., timeseries plot, heatmap, and network graphs), if appropriate. Explain how these graphs assist you in understanding relationship between variables.
2. Discuss the implications of your findings, determine how the results can help you predict the dependent variable and how the results can affect the business strategy in the market? Are there any limitations in the interpretation of these results?

**Submission:**

1. Deadline: by 11:59AM, Week 12. Only submit the soft copy (in word/pdf format) of your reports.
2. One group write and submit a *Group* Project Data Report (up to 18 pages) to disseminate your findings. If appropriate, provide related statistical outputs obtained from the data in an Appendix (free from the page limit).
3. Submit all the codes and the original data you have.
4. Note that plagiarism is a severe offense. Any group that committed plagiarism will get a zero for the project.
5. Please name the file of the soft copy using the group number and section number, e.g., “Group1\_1001.docx”.
6. The project/proposal must be written in English Language only. Follow the APA guidelines for references and in-text citations. Provide a complete and accurate list of "References." Each claim should be supported by at least one citation.
7. Late submissions: Assignment submissions after the due date will be considered late. Late submissions will be subjected to a penalty of 30% of the total points. No submission will be accepted after one week beyond the due date.

**Formatting:**

1. Total Page Length of Group report: 18 – 20 pages
2. Spacing: Double Spacing
3. Font Size: 12
4. Margin: 2.54 cm per Side

**Presentations**

1. Time: 12-15 minutes+ 5 minutes Q&A
2. Sequence: decided by a lucky draw
3. Print a hard copy of your slides for me, please.

**FAQ**

1. How to find the dataset?

You can choose the dataset I prepared for you. You can also find data everywhere you like. Model Whale (和鲸社区) and GitHub are strongly recommended. You can also use the dataset (e.g., World Bank e-library, RESSET) that our university library have had. Large sample size is welcome.

1. What is a good data analysis report?

A good data analysis report is a logical, decisive, and informative one. Please do not waste time trying those difficult but not applicable methods for your data analysis. Just choose the most appropriate analysis method for your business problem and clarify why this method is applicable. Multiple methods or meaningless methods will not help you win the bonus.

1. How to present my work?

Please focus on how to interpret your data analysis results and what implications you can draw from the results. Explanations are more important than the results! You have to justify your answer with data analysis - explain your results and codes during the presentation.

1. Shall I review the literature to support my findings?

You are welcome to review the literature to: define your variables, explain your model, and illustrate the methods you used.

1. Can I use other statistics software/programming languages to analyze the data?

You are asked to use R to do the analysis.