Data Analytics Team Project

Purpose: This project allows students to go through the data analysis lifecycle and apply the knowledge and skills learned through the course to solve a real-world problem. Students will form teams of 4-5 members and complete a data-related analytics project. This involves identifying a real-world problem, finding or developing a dataset, identifying appropriate measures, conducting analyses, and preparing a report for decision-making.

Skills: This project will help you practice the following skills that are essential to your success in this course and your professional career beyond school:

- applying significance tests or regression models to problem-solving
- analyzing data and synthesizing the results
- reporting data analysis results and using the results for decision making
- Understanding and applying R programming in data analysis

Knowledge: This project will also help you become familiar with the following important content knowledge in the data analytics discipline:

- The data analysis lifecycle model
- Data collection and access
- Data preparation and filtering
- Data visualization
- Significance tests and/or preliminary data modeling

Team project proposal (Deliverable #1)

Problem definition and dataset

The proposal must be approved by the instructor before the team starts working on the second deliverable.

Tasks:

1. Identify a field (e.g., business, health, education, etc.) that is of interest to your team.
2. Find a publicly available free dataset in that field. There are many websites that hosts free datasets. I recommend you start with kaggle.com. Kaggle offers a variety of interesting externally shared data sets and an easy-to-use search function to find datasets from different fields. For this project, I recommend you find a small dataset or you can use a subset of a big dataset.
3. Study the data in that dataset and propose several questions that you might be

able to answer based on the data.

Deliverable:

1. The dataset as a csv file and the link to the original dataset on Kaggle.com
2. Provide a description about the dataset
3. List at least 3 questions that you might be able to answer based on the data

Team project mid-point (Deliverable #2)

Data clean-up, exploration and visualization

Tasks:

1. Clean-up and re-organize the data as needed
2. Explore the dataset
3. Identify all the variables that might be related to the questions that you are trying to answer
4. Identify all the potential relationships that you suspect exist between the variables
5. Use descriptive statistics to help understand the key characteristics of the variables
6. Generate different graphs to help the readers understand the variables and the possible relationships between the variables

Deliverable:

1. The updated dataset as a csv file
2. List all the related variables and provide a brief description to each variable
3. List all the potential relationships that you suspect exist between the variables
4. Provide the R code to run the descriptive statistics and provide the output
5. Provide the graphs and the R code to create the graphs. For each graph, provide a brief description about what we can learn from the graph.

Team project report (Deliverable #3)

Hypothesis testing/modeling and report

In this deliverable, you will demonstrate your knowledge and skills in one or both of the following 2 areas:

- significance testing comparing multiple means or
- predictive modeling using simple or multiple regression

Tasks:

1. Develop hypotheses for significance testing comparing multiple means or associative relationships.
2. Conduct significance testing to compare multiple means and regression analysis to test each of your hypotheses
3. Interpret the test results and use the result to propose possible suggestions for decision making
4. Develop a formal data analytics report

Deliverable 3.1: The data analytics report

Section 1. Problem definition

Section 2. Description of the dataset

Include descriptive statistics to illustrate key variables. Use graphs to illustrate the descriptive statistics.

Section 3. Hypotheses

Section 4. Description of the data analysis procedure and methods

Section 5. Test results

Make sure you use standardized language to report the results. Create graphs to illustrate the results.

Section 6. Interpretation of the results and suggestion for decision making

Deliverable 3.2: The presentation (see separate ELMS assignment for more details)