**IS6030 Homework #5**

**Instructions**

You will need to download a dataset from a website, or use a dataset from another resource (ex. from your work). Please do not share any data that has private information. Please use the following link to “signup” your dataset: https://docs.google.com/spreadsheets/d/1GlUJ3Kihl2z31hd\_CW7-1O5dzqeY0Jsp3ZBUtxgsrBg/edit?usp=sharing

Your dataset can be any format .txt, .csv, .xlsx, etc. you will need to import the data to your SQL Server.

After you import your data you should start to explore the dataset, using some of the following approaches: SQL Queries, R, Tableau, or any other tools that are business intelligence related, data analytics, or data visualization based (SAS, SPSS, Microstrategy, Splunk, Spotfire, SSAS, SSRS). You have to write some SQL for this project and you can use any assortment of tools. I suggest using at least 2 different tools in addition to using SQL.

After you have done your data exploration you will need to generate a report about the information/ knowledge you have obtained from your data. Your report should be at least **4 pages single spaced**, and you may include graphics showing any visualizations or models you may have built.

A quality submission will include the following:

* [0.5 point] A general description of the dataset (what kind of data is being stored) and where you found the data (include the URL).
* [0.5 point] A simple overview of the columns and values that are stored.
* [Bonus: 0.1 point] If the data is normalized or not, and if not include a write-up of how to normalized the data (what kind of tables could you separate the data into).
* [1 point] A description of any problems in the data (such as bad column names, inconsistent values, repeating data, etc.).
* [4 points] A few general statistics which you will use SQL to calculate (ex. number of records, total orders, average price, average wait time, records by region, order by customers). Think of using COUNT, Aggregation/GROUP BY (MAX, MIN, SUM, AVG), and other keywords to provide these numbers/tables.
* [3 points] Leverage software product(s) and what you have learned in other classes to provide some more useful output. Use R, Tableau, Excel, or any other product to create analytic models or visualizations. Provide descriptions/write-ups for any output from the software products and include any interpretations/findings of data models/visualization. You can also use an Entity Relationship program to create a ERD of your dataset.
* [1 point] Finish your analysis with a summary of any meaningful findings that you found, or any suggestions you might have to improve the process/data/decision making the data might support (ex. provide discounts on weekends to increase sales, increase marketing efforts with product A, the south region buys more of product B, product C is the best value for money, etc.).
* [Bonus: 0.1 point] You may share any challenge(s) you face while finishing the assignment and how you overcome the challenge.

**Submission Instructions:**

You will need to submit a Word document with the following filename: HW5\_LastName\_FirstName.docx

You also need to **include any code or files** that you wrote or built to support your analysis. That includes SQL code, R/SAS code, Tableau workbooks, etc. Please put your name in any file names that you submit, if you cannot send a file you must take screenshots and include them in an appendix in your word document (does not count towards the page total).

**Data Websites:**

<https://data.opendatanetwork.com/>  
<https://www.google.com/publicdata/directory>  
<https://data.cityofchicago.org/>  
<http://www.opendatacincy.org/>  
<http://catalog.opensandiego.org/>  
<https://www.data.gov/open-gov/>  
<http://data.worldbank.org/about/open-government-data-toolkit/knowledge-repository>

Hint: Use “Open Data” in your google search and you should be able to find resources. Also look at analytical websites for data resources or you can download any Tableau public workbook and have access to the data the person used in the workbook (you need to cite the link of the workbook and data). Also data/analytic contest/competitions post free datasets ([www.kaggle.com](http://www.kaggle.com)) which you could use and maybe participate in the contest with.

**Quick notes of some R/Analytics you could do:**

* Five number summary (boxplot)
* Histograms
* Scatterplots
* Discuss any anomalies (e.g. outliers) and/or noteworthy findings from data exploration
* Build a linear model and CART tree that fits the prediction problem presented by your data (classification vs. numeric prediction)
* Higher level modeling techniques (Neural net, random forest, K-nearest neighbor, SVM, etc.)
* Model fit (e.g. R sq, goodness of fit stats)
* Key assumptions satisfied (e.g. normality of errors)
* Accuracy/quality (classification accuracy or MSE) out of sample