**Assignment 1**

**(Individual Assignment)**

**IT 418/518**

**Due Date:** 03/08/2020; 11 pm Central time

**Points:** 140 Total (20 points for each question)

**Submission:**

Submit a Word document, Tableau file, and R file (along with the screenshot on a word document) in D2L dropbox. Name your folder as *Assignment1\_Part 1\_IT418\_YourLastName.zip.*

**Purpose of the Assignment:**

The purpose of this assignment is to evaluate learning on the following key objectives:

* Identify and assess the needs of an organization for a data science task.
* Develop analytical thinking using scenarios that will be helpful in group project.
* Identify and analyze social, legal, and ethical issues in data science.
* Apply visual analytics concepts and techniques.

**Grading:**

This assignment will be graded using the following criteria:

* Attention to instructions: follows instructions
* Clarity: clear, succinct writing; substantiated arguments
* Creativity: high degree of insight, originality, and creativity
* Problem solving: apply analytical thinking steps and logical approach to frame/reframe and solve the problem
* Concise and well-commented code

**Exercises**

1. Jimmy’s Best Burgers (a fictitious company) is a nationwide, fast food chain. The company has over 2000 franchise throughout the country, which sell a variety of fast food options such as hamburgers, cheese burgers, fries, soda, sandwiches, etc. The company spends 40% of franchise fee on nationwide marketing efforts include print media, social media, and radio/TV ads. The marketing mix of the company has been focused on its unique value on four dimensions: low price, product, promotion, and place. The company has kept the prices of its products lower than competitors in the belief that its target market is general masses including low income families. It has always put an emphasis on developing a menu customers want. For example, it recently started offering some vegetarian options in many large cities to cater to the needs of growing vegetarians. Jimmy’s Best Burgers has been selective about the locations of its chains (for example, most of its locations are in attractive, high density areas that are also close to its distribution centers and the locations have a drive-through for convenience of customers). Yet, the company has suffered loss of sales and revenue in each of the last six quarters. The top leaders of Jimmy’s Best Burgers are worried about declining revenue and fear that this might be a long-term trend that could severely hurt the company. But they are not sure why the sales and revenue have declined.

The Chief Marketing Officer of the company believes that their marketing strategy needs to be analyzed and adjusted to optimize the value of money invested on advertisement. He has hired you as a data scientist to perform marketing mix analysis.

In order for conducting marketing mix analysis, you will need to clearly define the problem and investigate its nature, scope, and implications. To solve the problem through marketing mix analysis, you would like to follow an analytical thinking framework that include the following steps:

* Problem recognition and framing
* Review of previous findings
* Modeling
* Data collection
* Data analysis
* Presentation and action

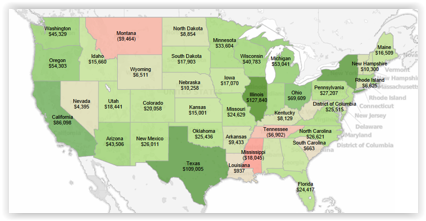
Submit the answers to the following questions:

* What would be important elements for each of the above steps?
* What information would you collect for each of the following steps?

(NOTE: Many details in this exercise are intentionally not mentioned. Please make and state your assumption that you feel necessary to answer the questions.).

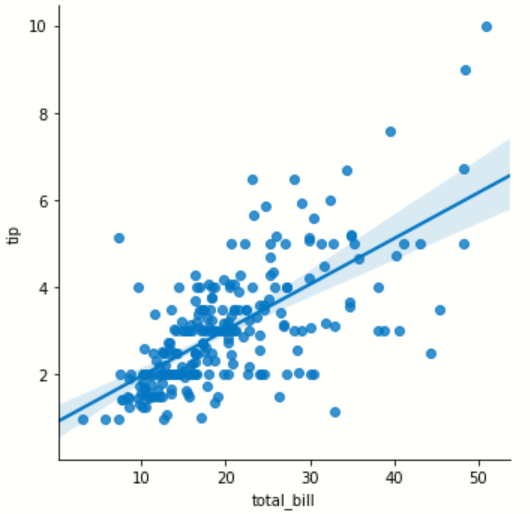
1. Review the article, “How Companies Learn Your Secrets” (<http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html>). Discuss the ethical, social, legal issues related to data collection and practices described in the article.
2. Using the “Sample – Superstore Subset” Example Data Set in Tableau, create the following map visualizations:
   1. Are there any US states in the dataset that have lost money? Show the profitability of the each state using a color scale. For example, most profitable states should be darker green in color than less profitable states. Similarly, use scales of red to show non-profitable states.
   2. Which state has the highest sales in the dataset? Using a color scale, show the total sales for each US state. Label each state with state name. (30 points)

Hint: The “Sample – Superstore Subset” dataset is a built-in sample dataset in Tableau. Your map visualizations should be similar to the following image:

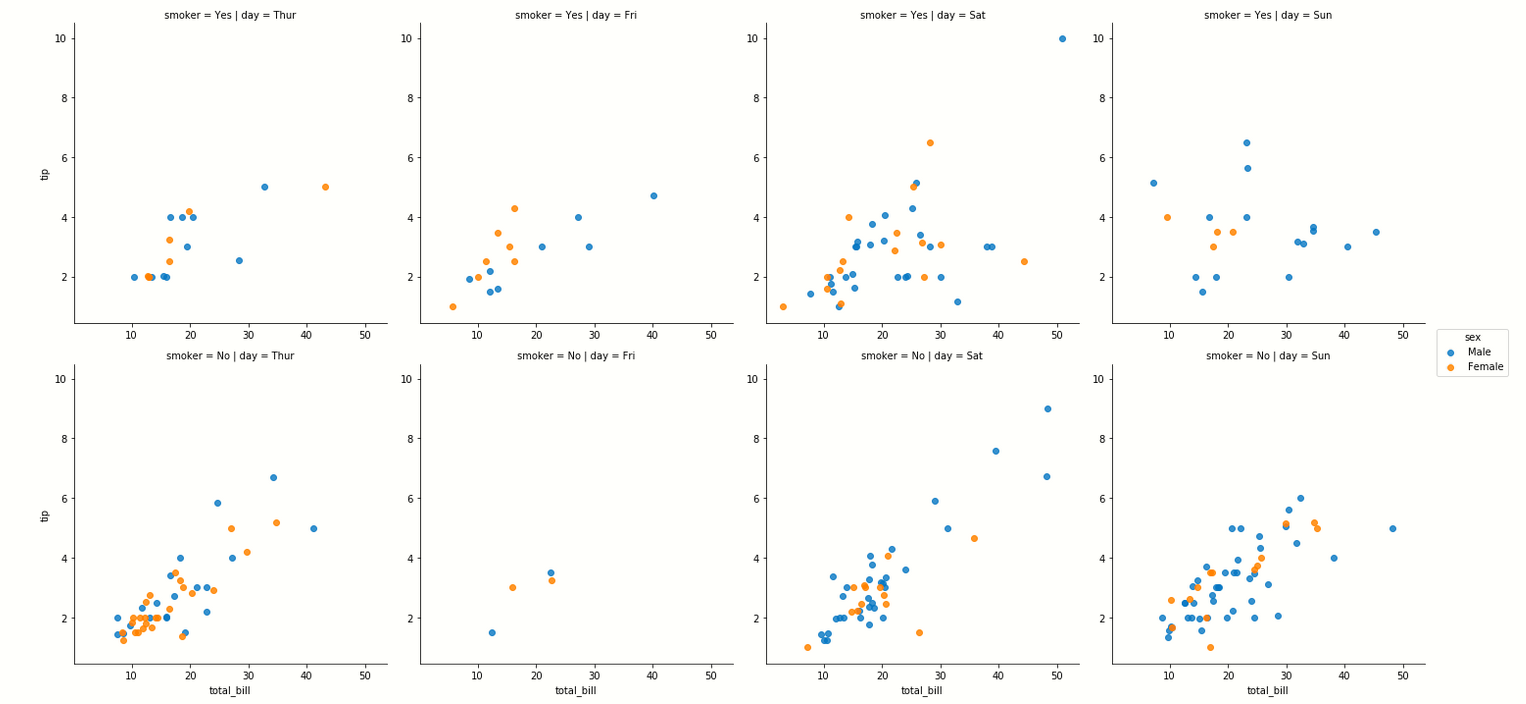


Use SUM to calculate the total sales for each state. For additional help, review the Tableau training videos at <http://www.tableau.com/learn/training>.

1. Perform the following operations in R Studio with the sample “Diamonds” dataset: (40 points)
2. A histogram of carat using ggplot2. Explain your findings.
3. A scatterplot mapping carat (x) and price (y) using ggplot2. Use the color aesthetic. Explain your findings.
4. A boxplot of diamond carats using ggplot2. Explain your findings.
5. A violin plot of diamond carats by cut using ggplot2. Explain your findings.
6. Using summary() and head() functions, provide an overview of the dataset.
7. In R Studio, create a word cloud for the term “Minnesota” using at least 1000 most recent tweets from Twitter. Show 25 most frequent words in your word cloud. Take a screenshot of the word cloud in R Studio and paste on Word document for submission. Also, submit your R syntax as an R file.
8. Using the Seaborn library in Python, complete the following exercises. Submit the PDF exported from Jupyter Notebook. It should have your code and output.
   1. Import Seaborn and Matplotlib.
   2. Load the tips dataset. Tips is a built-in dataset in Seaborn.
   3. Examine the first few records using head() and info().
   4. Create a scatterplot using the Tips dataset. x= total\_bill; y= tip



* 1. Explain the visualization.
  2. Create individual scatterplots as shown below. x= total\_bill; y= tip; column = day; row = smoker



* 1. Explain your finding based on the above visualizations.

1. Perform the following data cleaning exercise in Python Pandas.
   1. Load the Ebola dataset (provided)
   2. Examine the first few records using head() and info().
   3. Identify missing values for each attribute.
   4. Fill median for missing values for Cases\_Guinea.