INFO 659 Introduction to Data Analytics

Assignment #1 on
Problem Understanding and Data Exploration (**10 points**)

Purpose

The purpose of the assignment is for you to follow the data analytics lifecycle/process and start with understanding the business problem, data identification, preparation, and initial exploration. **The outcome** of this assignment should be a notebook (report) on the problem, data, and results from your initial analysis, with related code and data files.

Before You Start...

Please take note of the following:

- You need to have a **data set** with at least **3 numeric** attributes (columns) and at least **50 data** instances (rows), preferably in the *CSV* format.
- The final **notebook** (report) should contain your writing, R code (statements), and results (output/plots). While you can use a word processor and manually copy code and results into your writing, I *highly recommend* that you create a **R Notebook** in RStudio and submit the notebook for the assignment:
    - Go to *File -> New Project...* to create a *INFO659* project;
    - Go to *File -> New -> R Notebook* to create a new notebook *INFO659A1*.
    - After you are all done with the assignment, go to *File -> Knit Document* and the **INFO659A1.nb.html** (in the project directory) becomes the final report in HTML that you can use to submit with your data file.

Tasks and Steps

1. Think Business: Framing the business problem (2 points)

Please identify a specific business domain or topic suitable for data analytics, introduce the ***background*** to the problem you would like to resolve, and frame the problem (question) while thinking about what data you can access to solve the problem (or answer the question). Please present the **business problem** and state your **objectives** clearly, for which the following steps will be conducted.

2. Understand Data: Data source identification and understanding (2 points)

(Note that in certain situations **you might have had the data first**. In this case, you can examine your data to come up with the business problem or question in the first step.)

Now you can explore and identify a specific data source, data service or database which includes data relevant to the business problem. Learn about the data and reason on ***whether the data can help resolve the business problem*** (or answer the question). If not, find another data set or revised your business problem and objectives.

I encourage you to use data related to your background/work/interest. But you can visit kaggle for potential ideas and datasets: https://www.kaggle.com/datasets

Describe your data in terms of the following aspects:

- *Concept of learning*: From the machine learning point of view, what is the concept of learning or what do you want your analytic models/methods to discover and learn about from data? This is essentially converting your understanding of the business problem into a data analytics and machine learning problem.
- *Data attributes (***at least 3 numeric attributes/columns** among others*)*: What are available attributes (variables) in the data? What are the data types? Which ones are potentially relevant to the concept of learning, in what ways?
- Data instances (**at least 50 instances/rows**): Provide a few **examples** of your data, discuss the **data format (preferably CSV)** and explain the **meaning** with one or two data instances/examples.

3. Data in Action: Data preparation, visualization and exploration (4 points)

3.A. Data preparation and loading:

- Download your data and ***move*** *the file* to your R project (*INFO659*) directory;
- Make sure it is in the proper ***CSV*** format (comma delimited with header in first row);
- Insert an R statement to load the data into a variable with *read.table* or *read.csv*:

```
dataName <- read.table("DataFileName.csv", header=TRUE, sep=",")
```

- Double check whether data has been properly loaded into the variable;
- Show the first 5 rows of the data with *head()* function:

```
head(dataName, 5)
```

3.B. Data distribution and anomalies

For each of the numeric variables (***at least three***) relevant to the concept, produce its histogram (distribution) using the hist() function, for example:

```
hist(dataName$attributeName, breaks=10, xlab="X Label", main="Figure Label")
```

Adjust the *breaks* parameter until the distribution is visually clear and discuss the following for each variable:

1. Does the **range** of the variable make sense? Is it reasonable according to your knowledge of the data and domain?
2. Does the histogram/distribution look normal (a bell curve)? Do any parts of the distribution appear to be anomalies (e.g. sudden "spikes")? Should data be corrected (cleansed) given the anomalies if any?

3.C. Data distribution with log transformation

Examine the above distribution (histograms) and pay attention to the skewness of each distribution (i.e. heavy on one side vs. the other).

Pick the variable with the most skewed distribution (one variable only), produce another histogram based on logarithm of its values, using hist() and log10() function, for example:

```
hist(log10(dataName$attributeName), breaks=10, xlab="..", main="..")
```

Discuss the skewness of the distribution. Does it appear to be normal now?

3.D. Examining multiple variables and regression

Identify two variables relevant to the concept: one as the predictor variable (e.g. attributeX) and the other to be predicted (e.g. attributeY).

- Produce a scatter plot between attributeX and the attributeY variables;

  ```
  plot(dataName$attributeX, dataName$attributeY, xlab="...", ylab="...")
  ```

- Conduct linear regression on attributeY to be modeled by attributeX using the lm() function, and assign the result model to a variable:

  ```
  myline <- lm(dataName$attributeY ~ dataName$attributeX)
  ```

- Add the regression line to the scatter plot by:

```
points(dataName$attributeX, myline$coefficients[1] + myline$coefficients[2] * dataName$attributeX, type="l", col="red")
```

Now you should see a red regression line on the scatter plot. Does the linear regression line capture the relation between the two variables? To what degree? What do you think is the relation between the two?

4. Discussion, understanding, and planning (2 points)

Based on the above data exploration, visualization, and initial analysis, discuss the following issues:

- Given your problem, concept, and objectives, can the concept be learned from your data? And can your problem be resolved and objectives achieved with analytics on this data?
- Is the data of good quality? Are data within their normal ranges? Do the data need cleansing or correction?
- Is log transformation or any other transformation necessary for some of the numeric variables in the data?
- Is a linear model (linear regression) suitable for the modeling?
- Do you need more data instances (rows) for actual modeling?
- Do you need additional data (other variables and/or sources of data) for the project?
- What other insight do you gain from this exploratory analysis? Discuss any other observation you have so far.

Please submit your **notebook** (INFO659A1.nb.html) and **data** file (CSV) to Blackboard Learn.

Note: If you have a website, you can upload your notebook and data files to your website and submit the links only.