CS601C Spring 2025
Professor Paul Dantzig
Due 05/07/25

CS601C Final / Project

Complete the exercises and answer the questions outlined below. Your submission must be a combination of what the problem statement is you are working on, # R comments, R code, and screen captures of your plots, and use bold print to indicate your findings. Your code must run successfully and you must answer all the questions to get credit. I.e. The objective is to make it look as closely as possible to real statistics paper / data mining analysis. Do not just give me an RMD file. I want either a Word, HTML, or PDF document with all the R code, graphs, etc. as you would see in a real statistics paper.

Rubik: I have labeled this a combination of final and project. This course because it is fully online doesn't really provide an appropriate environment for standard exams or projects. This activity is going to be about 15% of your grade. Just because it has simple requests like "numerical summary of the variables / quantitative columns" doesn't mean you identify one column and give me the mean, median, standard deviation, etc. It means looks at the columns carefully and indicate which columns should be used for your study and why. If you just do the minimum you will end up with a grade like a B. If you want an A each question should be answered in detail. The better your discovery and justification aspects the better the grade.

1. This exercise relates to the dataset "Cybercost.csv"

   The dataset has only a few columns.
   1. Csecurity
   2. Attacks
   3. Databases
   4. Priv Users
   5. Users
   6. Cost P/M

A. Total number of Cybersecurity Personnel 2) Number of monthly cyberattacks 3) Databases 4) Number of Privilege Users 5) Total number of Users 6) Cost P/M
   a. Just for your information this is not real data collected from customers. There are 1219 records in the dataset which is the full customer base.

B. Read the data into R. Make sure that you have the directory set to the correct location for the data or use file.choose(). Note: when you convert it to a RMD must be a specific location.

C. Produce a descriptive statistical summary of quantitative attributes in the data set.

D. Scatter Plot, Box Plot, correlate, and linear regression plot all the quantitative attributes in the data set.

E. Now comes the interesting part of the study. Your job is determining how is cost P/M determined. You are to use all your skills that you have learned during this course to determine what attribute(s) are best. Note: one variable (attribute) may not be enough to determine how the cost is determined.

F. This dataset is the complete customer list. This means what you determined above is the real mean. Now we want to use your other skills you learned during this semester (i.e. sampling, inference, confidence intervals) to figure out the mean without using the entire customer list.

G. Take multiple random samples of the customer list. First try with 30, 50, 100, etc. till you can determine how big a sample you need so that with 95% confidence level that you are able to determine a sample means that are similar to the real mean. Plot the sample means, confidence levels, real means, etc. on the same plots as they did in your textbook for each random sample.

H. Prove or disprove the following hypothesis "Companies that are smaller (<1000) are being overcharged (i.e. number of employees))". Note: hypothesis many times are proven incorrect. See if you can make up another hypothesis that you can prove or disprove. In this case since you have the entire customer list you can actually verify if the hypothesis calculations done with the sample are correct.

I. Obviously, the idea here is to show off what you have learned during the semester. So, the more the project looks like a real paper that a "statistician/data analyst" would make the better the grade.

2. This exercise relates to the dataset "Hypertension-risk-model-main.csv"

   The dataset has a few columns.
   1. male 1 / female 0
   2. Age
   3. Current Smoker yes 1 / no 0
   4. Cigarettes per day
   5. Takes Blood Pressure Pills yes 1 / no 0
   6. Has diabetes yes 1 / no 0
   7. Total Cholesterol
   8. Systolic Blood Pressure Reading
   9. Diastolic Blood Pressure Reading

10. Body Mass Index (BMI)
11. Resting Heart Rate
12. Glucose
13. Risk (What you are trying to predict based on the above information)

A. Read the data into R. Make sure that you have the directory set to the correct location for the data or use file.choose(). Note: When you convert it to an RMD file must be a specific location.

B. Need to read a little about what are good values for blood pressure, BMI, heart rate, glucose. All available from either Wikipedia, WebMD, etc.

C. Produce a descriptive statistical summary of quantitative attributes in the data set.

D. Scatter Plot, Box Plot, correlate, and linear and multiple linear regression plot all the quantitative attributes in the data set.

E. This dataset is easier than the Cybercost example above to find attributes that will predict risk. Clearly you will find many attributes that will predict risk. The idea here is if you could choose three attributes which three attributes would you choose that give you the best prediction of risk and in what order. Example: (male, age, smoker). Why did you choose those three (i.e. justify your answer tables, graphs)

F. Make up two hypotheses about the above data. Example: If a person smokes, he is at risk. (don't use this one)

G. Prove or disprove your two hypotheses. (i.e. tables and graphs)

H. Obviously, the idea here is to show off what you have learned during the semester. So, the more the project looks like a real paper that a "statistician/data analyst" would make the better the grade.

3. This exercise is now that your experienced working with text. Use the file smsspamcollection. This file contains 1000s of SMS texts. Do all the usual text basics (i.e. descriptive statistics) from your last homework and determine what the SMS texts are about. The readme gives you the format of the SMS texts and indicates whether they are real (ham) or fake (spam). Separate the spam from the ham. Anyway to use our statistical sampling, confidence, and hypothesis to predict anything about the spam emails.

   This problem requires out of the box thinking unlike the first two problems which just show off what you learned from the book and my lectures.

Obviously, the idea here is to show off what you have learned during the semester. So, the more the project looks like a real paper that a “statistician/data analyst” would make the better the grade.