Name: Andrea Fox

Date: June 20, 2020

Class: DSC550-T301

Assignment: Term Project – Proposal

Title: Exploring Survival on the Titanic

Project Plan:

I decided to do my project on Titanic survival as my daughter and I love discussing the story of what happened to the Titanic. We have spent many hours going through where things went wrong as well as the possibility of a different outcome if they had hit the iceberg head on instead of trying to turn away from it. She also enjoys coding so I may have her do some of this with me as a fun activity. I plan on pulling data from a few different places to try and answer some of the following questions:

1. Did female passengers have a higher rate of survival then males? (I have heard and read that women and children were evacuated first but I want to form my opinion that is data driven)
2. Did the type of class fare play a role into survival rates?
3. Were families more likely to survive than a single passenger?
4. Did age of passenger play a role in survival?

Data Sources:

1. <https://www.kaggle.com/kandij/titanic-data>
2. <https://www.kaggle.com/spscientist/titanic>
3. <https://www.kaggle.com/azeembootwala/titanic>
4. <https://www.kaggle.com/heptapod/titanic>

These are the ones I have gone through already but depending on data I may use more or less. I will not know until I start putting everything together.

Proposed Evaluation:

So far, the data I have selected are csv files. I plan on reading in each file and creating an initial dataframe. I will then spend time cleaning up each dataframe to make sure I have eliminated null values, analyzed any outliers and made a decision whether they are accurate, remove columns if they do not fit into the narrative, and any other processes I may need to create clean copies of each dataframe. Before I combine each dataframe into a single dataframe I want to do some basic visualizations to double-check whether everything looks clean as well as do some variable comparison. I also want to perform analysis on each dataframe looking at correlation, mean, p-value, and running a few tests. Looking into some of the topics in this course I am toying with the idea of doing a decision tree I am curious whether something like ticket type would determine a different outcome (not sure it can do this, but from what I have seen it would). Some other modes of evaluation that look to be possible to use are logistic regression, Naïve Bayes, and Random Forest if I go with the decision tree as it allows me to build multiple decision trees and merge them together. There may be more added as we continue through the class and learn more techniques and algorithms.