Homework 1

Covid-19 Cases from Kaggle

**Project Background**:

The first reported case of coronavirus disease (covid-19) occurred in Wuhan China in December of 2019; since then, it expanded to over 200 countries around the world. The writer believes that the covid-19 pandemic is an important use case for the US department of health to enable it to formulate vaccination strategies for the US in order to prevent unnecessary deaths caused by covid-19 infections. The expansion of covid-19 cases around the world is the primary area of concern for this project as it is a pandemic, which will expand to the US. Understanding the general characteristics of the covid-19 infections abroad, the relationship between vaccinations by manufacture and the volume of infections abroad will enable the data analysis project to determine effective projections for the US as it relates to most effective vaccines by manufacture to prevent covid-19 infections.

The data for this project will come from the Kaggle public database, which offers details such as country of covid-19 infections; longitude and latitude of cities; date of infection; the day of the year a case confirms covid-19 infection. Moreover, the year; length of infection in days; normalized cases; number of people vaccinated; number of people fully vaccinated; vaccines manufacture used for specific cases, and total vaccinations in each country based on manufacture.

**Problem Formulation**:

Understanding the effectiveness of covid-19 vaccines by manufacture to protect the population from covid-19 infections is the goal of this project for the US department of health. The information gained from this project will enable the US department of health to obtain the best manufacture vaccines into making them available at distribution points.

**Data Strategy Plan**:

The data strategy plan for this project will include data visualizations to understand volume of infections in countries abroad to focus the data analysis where most infections are occurring. Second, conduct regressions to understand the relationship between the number of vaccinations and number of covid-19 cases by manufacture. Third, creating scatterplots to understand the clustering of covid-19 cases by types of vaccines. Finally, because the results of predictive models can be miss leading the writer will use a combination of predictive models to project the number of covid-19 cases expected in the US. The proposed analysis will attempt to determine the effectiveness of specific vaccine manufactures. Moreover, the writer will use several predictive models depending on the results of the regression analysis; because the results of regressions can be miss leading, the writer will use logistic regression to address issues of data normalization such as multi collinearity. The variables of interest will include number of vaccinations by manufacture, length of infection and number of covid-19 cases by location. Furthermore, the writer will use visualizations to improve understanding of the analysis results.