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.

**Summary of Data Cleansing and EDA:**

The cleansing for the exploratory data analysis involved the following steps:

In the first step, the process of getting the data ready for analysis from kaggle involved importing the following files: country\_by\_population\_2020, country\_vaccinations, country\_vaccinations\_by\_manufacturer and CovidCasesEurope. Once that was completed, the writer had to rename all the columns involving the country’s name to “country” in order to standardize the column name. Afterwards, the writer gave aliases to the column names in order to organize better which columns were from which data set before merging (e.g., date was renamed to date\_summary in one of the data sets).

In the second step, after completing the renaming the columns appropriately, some of the columns the writer dropped in order to have a clear data set. For example, in the country\_by\_population\_2020 data set, the writer dropped all columns except for the country’s entire population as of the year 2020. In the CovidCasesEurope data set, the writer dropped the column named “Unnamed: 0”. After the writer dropped all unnecessary columns, the writer then started the merging process for all data sets.

In the third step, the writer first merged both country\_vaccinations and country\_vaccinations\_by\_manufacturer by doing an outer join based on the column named country. After this merge completed, the writer merged the country’s population from the country\_by\_population\_2020 dataset with the rest of the datasets by doing an outer join based on the country column. In the final merged dataset, the writer cleansed null values with zeros as an appropriate value for analysis (e.g., the number of vaccinations for a particular day that had a null value the writer changed to 0). The writer decided to keep the CovidCasesEurope dataset separate for its own data analysis.

In the final step, the writer created total vaccination ratio columns by taking the total number of vaccinations of a country divided by the country’s population as of 2020 in order to get a vaccine ratio score from each country in order to do a comparative analysis.

The preliminary data analysis shows that the data for covid cases is between the year 2020 and 2022. In addition, the Europe covid cases dataset shows the total number of observation for length of sickness is 1.048575e+06; the mean is 1.352943e+01; the standard deviation is 3.021219e+00. Moreover, the min is 7.742285e+00; the 25th percentile is 1.074121e+01; the 50th percentile is 1.342631e+01; the 75th percentile is 1.614536e+01; and the max is 2.006000e+01.

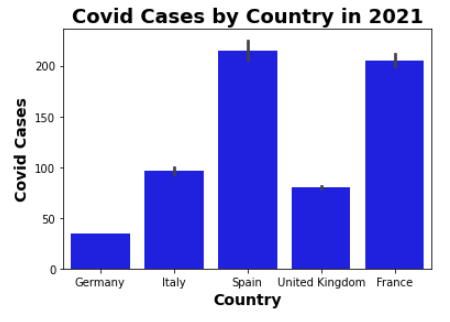
The total number of covid cases in Europe is 1.048575e+06; the mean is 8.923526e+01; the standard deviation is 3.724366e+02; the min is 0.000000e+00; the 25th percentile is 2.000000e+00; the 50th percentile is 1.200000e+01; the 75th percentile is 6.100000e+01; and the max is 3.490400e+04.

Additionally, the preliminary analysis for outliers shows that Spain has some outliers of covid cases beyond 25000. The writer will explore outliers in more detail as the analysis progresses so that outliers do not bias the results of the analysis. The writer believes that the US department of health would be interested to know the insights gained from Spain as the preliminary analysis shows that they had 35,000 covid cases followed by France with 23,000 covid cases. In addition, the US department of health would be interested to know that the goal of this project is to understand which vaccine manufactures are more effective for the prevention of the spread of covid-19.

**Summary of Dashboard and Data Visualizations:**

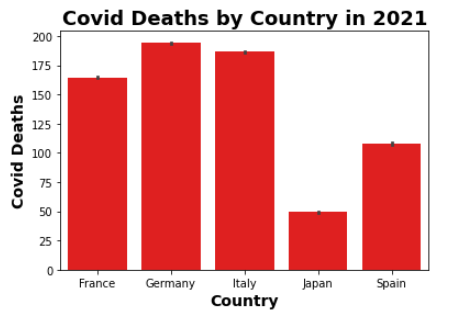
The focus of the analysis for this data science project to determine the effectiveness of vaccines for covid-19 by manufacture to advise the US Department of Health to determine which manufacture is more effective for the prevention of covid cases and deaths.

Figure 1



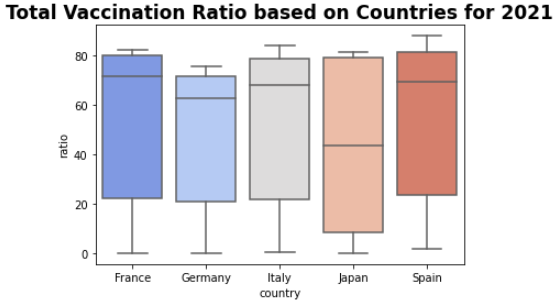
The analysis of Figure 1 appears to indicate that the greatest number of covid cases is in Spain with approximately 210k cases, followed by France with 200k cases, followed by Italy with 100k cases, followed by U.K. with approximately 85k cases, and finally Germany with approximately 40k cases, which indicates that Germany had the least number of covid cases.

Figure 2



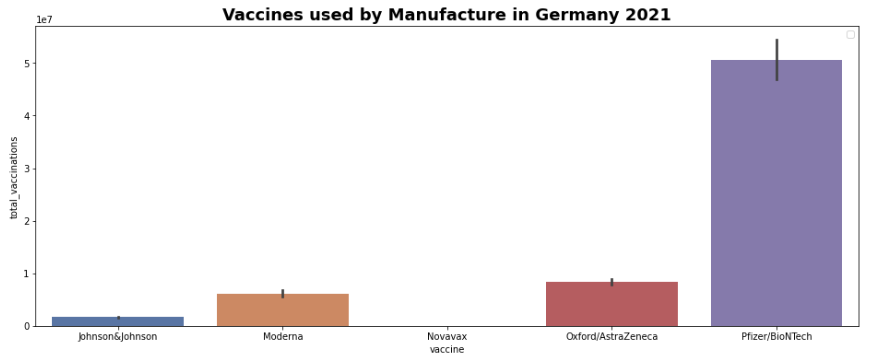
The analysis of Figure 2 appears to indicate that the greatest number of covid deaths is in Germany with approximately 190k deaths, followed by Italy with 185k deaths, followed by France with 160k deaths, followed by Spain with approximately 115k deaths, and finally Japan with approximately 50k deaths, which indicates that Japan had the least number of covid deaths. The writer believes that the reason for Japan having the least number of covid deaths is that Japan is an island nation and was able to manage the entry of people from other countries more effectively.

Figure 3



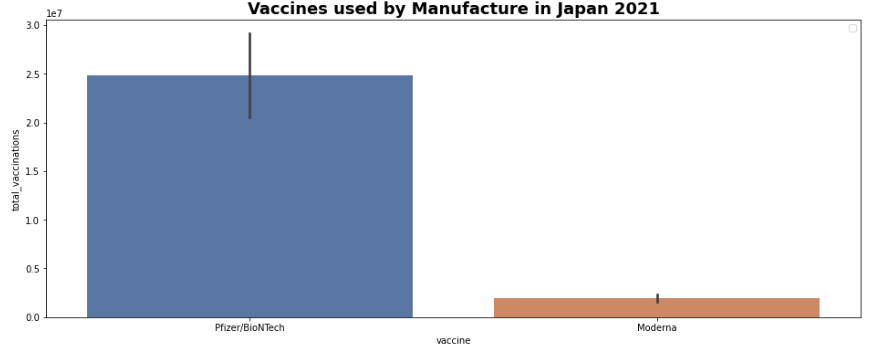
The analysis of Figure 3 appears to indicate that France has the greatest median vaccination ratio with a value of approximately 75k; followed by Spain with median value of approximately 73k; followed Italy with a median value of 70k; followed by Germany with a median value of approximately 65k. Additionally, Japan with a median value of approximately 45k, which suggests that Japan has the lowest median of vaccination ratio. Figure 3 also shows right skewed distributions of the vaccination ratios for all countries except Japan; Japan appears to be more normally distributed.

Figure 4



The analysis of Figure 4 appears to indicate that Pfizer vaccine manufacture has approximately 6k vaccinations; followed by Oxford/AstraZeneca with approximately 0.8k vaccinations; followed by Moderna with approximately 0.5k vaccinations; finally, johnson&johnson with approximately 0.1k vaccinations; there were no Novavax vaccinations in Germany, which seems to suggest that Pfizer had the greatest number of vaccinations.

Figure 5



The analysis of Figure 5 appears to indicate that Pfizer vaccine manufacture has approximately 2.5k vaccinations; followed by Moderna with approximately 0.3k vaccinations. Which indicates that Pfizer has the greatest number of vaccinations.

Consequently, the analysis seems to suggest that Germany had the lowest number of covid cases and largest number of vaccinations by manufacture is Pfizer. The writer believes that the Pfizer vaccinations are more effective at preventing the spread and death of covid-19. Additionally, the findings of this analysis in Japan confirms by the same results. The writer also believes that this information would be extremely beneficial for the U.S. Department of Health to focus implementing vaccination programs using the Pfizer to prevent the spread of covid in the future.

As a final note for the U.S. Department of Health, the writer used Jupiter notebook to create the analytic visualizations using data sources from kaggle.