Investigating the Effect of Diet on Death due to Covid-19

STAT 3080: Project Part 1

Description of Chosen Data Set

Background

In the wake of the Covid-19 pandemic, research has been performed to analyze the effect of numerous variables on mitigating transmission and death rates. Such variables concern different methods of self-isolation and quarantine restrictions imposed by different countries, age and gender data for differing populations, differences in wealth and the healthcare of nations, and more. This data set and project utilize nutritional data about the diet composition of different countries to investigate the effects of diet on Covid-19 death. As recommended by the USDA Center for Nutrition Policy and Promotion, a diet should be comprised of 40% vegetables, 30% grains, 20% protein, and 10% fruits. (Reference 1) This branch of analytics is powerful because understanding what nutrient makeup for one's diet proves most effective against death due to Covid-19 can improve the health of all nations during this difficult time. (Reference 1)

Explanation of Rows and Variables

This data set combines information about diet composition with other data sets concerning the population, death, confirmed cases, and recovery data for 170 countries. There are 170 row entries (one for each country studied) and 32 columns (one for each variable measured). The variables measured concerning diet composition are presented as percents (in kilograms) of the diet attributed to that food group. These percentages were alcoholic beverages, animal fats, animal products, aquatic products, cereals, eggs, fish/seafood, fruits, and meat. Other variables measured covered smaller food categories such as oil crops, spices, sweeteners, and tree nuts. The variables merged with the data set displayed health data and were also

projected as percentages for each country:obesity, undernourished, confirmed cases, deaths, recovered, active, and population (as counts). Merging the diet composition data with health data provides a convenient comparison to investigate the effects of diet on health and death resulting from Covid-19.
(Reference 1)

Explanation of Data Collection Methods

The data set employed for this project combines nutritional diet composition data with Covid 19 data (deaths, confirmed cases, recovered, etc.) on a population level for 170 countries around the globe (Reference 1).
More specifically, the variables for the percentages of the diet attributed to each food group, the Obesity variable, and Undernourished variable for each country were collected from the Food and Agriculture Organization of the UN (Reference 3). The FAO employs several methods to collect their population data, including questionnaires to citizens, and direct data provided from food group providers like fisheries and traders in the form of ledgers and tabular records. The Population variable was extracted from the PRB source (Reference 4). Population data is collected and provided as counts in the data, not as percentages. The John Hopkins Coronavirus Resource Center (Reference 5) provided population data for the Confirmed, Deaths, Recovered, and Active percentage variables within the data for each of the countries.

Potential Issues in Data

One potential issue involved creating the scatter plot, which removed "10 rows with missing values". The presence of missing values in the data is an issue that prevents consistent comparison of the countries to one another. Data cleaning should be employed to alter the data set and control for the missing values. Missing values were typically seen in the merged health data regarding the percentages of recovered people, confirmed cases, and deaths but not in the health composition data. Secondly, the data set description on Kaggle indicates that the Covid health data is updated regularly. However, if the food and diet data set for each country is not updated alongside the Covid data then a correlation cannot be made effectively as to how diet may impact Covid deaths and transmission for each country. Both data sets must be continually updated. (Reference 1) With respect to data collection errors, the sources cited in the data collection section are all population-based and so relied on other repositories and data sets monitoring Covid around the world. Using data that itself

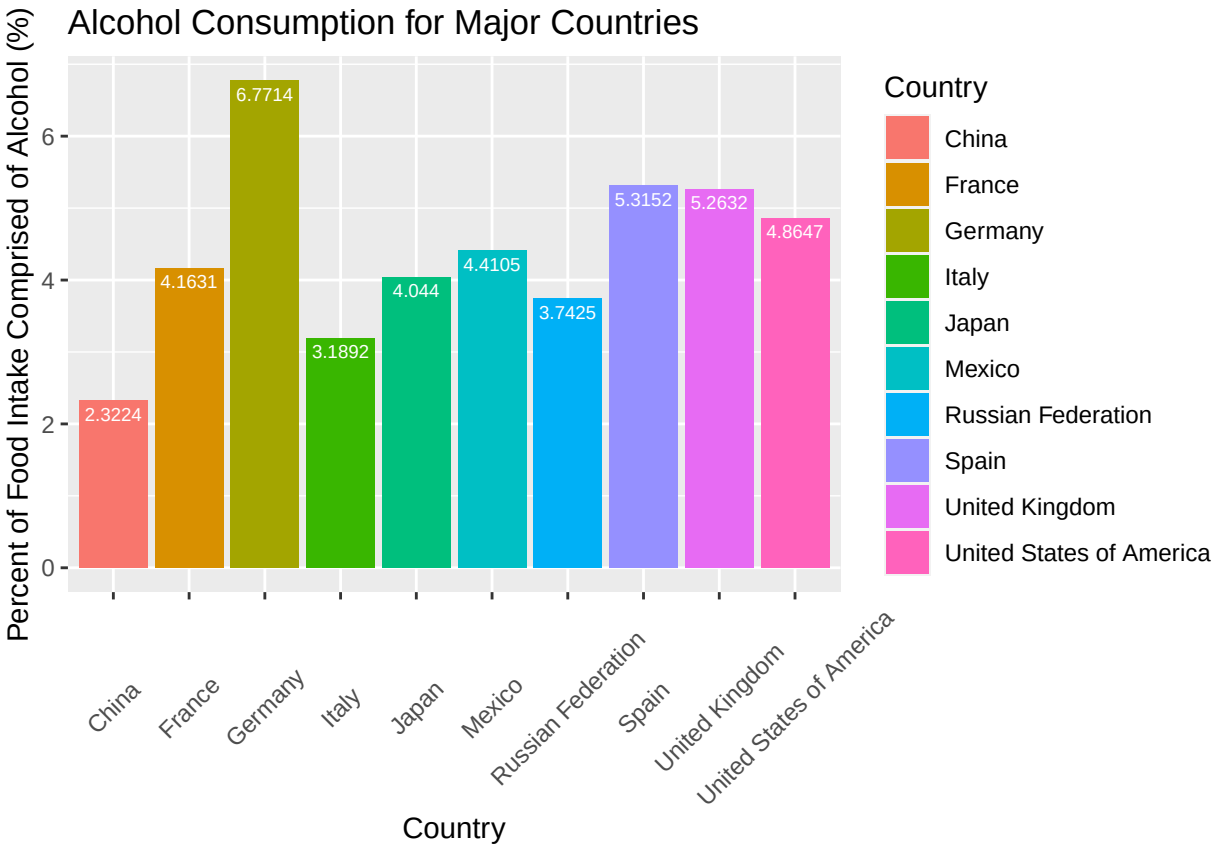
combines numerous other data sets may produce inconsistencies in the variables.

Graph 1-Numerical Summary

Numerical summary bar graph of "percent of diet that is alcohol"

```
majorcountrydata<-coviddata[c(162,127,101,142,160,52,57,75,77,31),]
#majorcountrydata
ggplot(majorcountrydata,aes(x=Country,y=Alcoholic.Beverages,
fill=Country))+geom_bar(stat="identity")+stat_summary(
geom="text",aes(label=..y..),vjust=1.5,color="white",size=2.5)+labs(
title="Alcohol Consumption for Major Countries ",
x="Country",y="Percent of Food Intake Comprised of Alcohol (%)")+theme(
axis.text.x=element_text(angle=45,vjust=0.5))
```

```
## No summary function supplied, defaulting to `mean_se()`
```



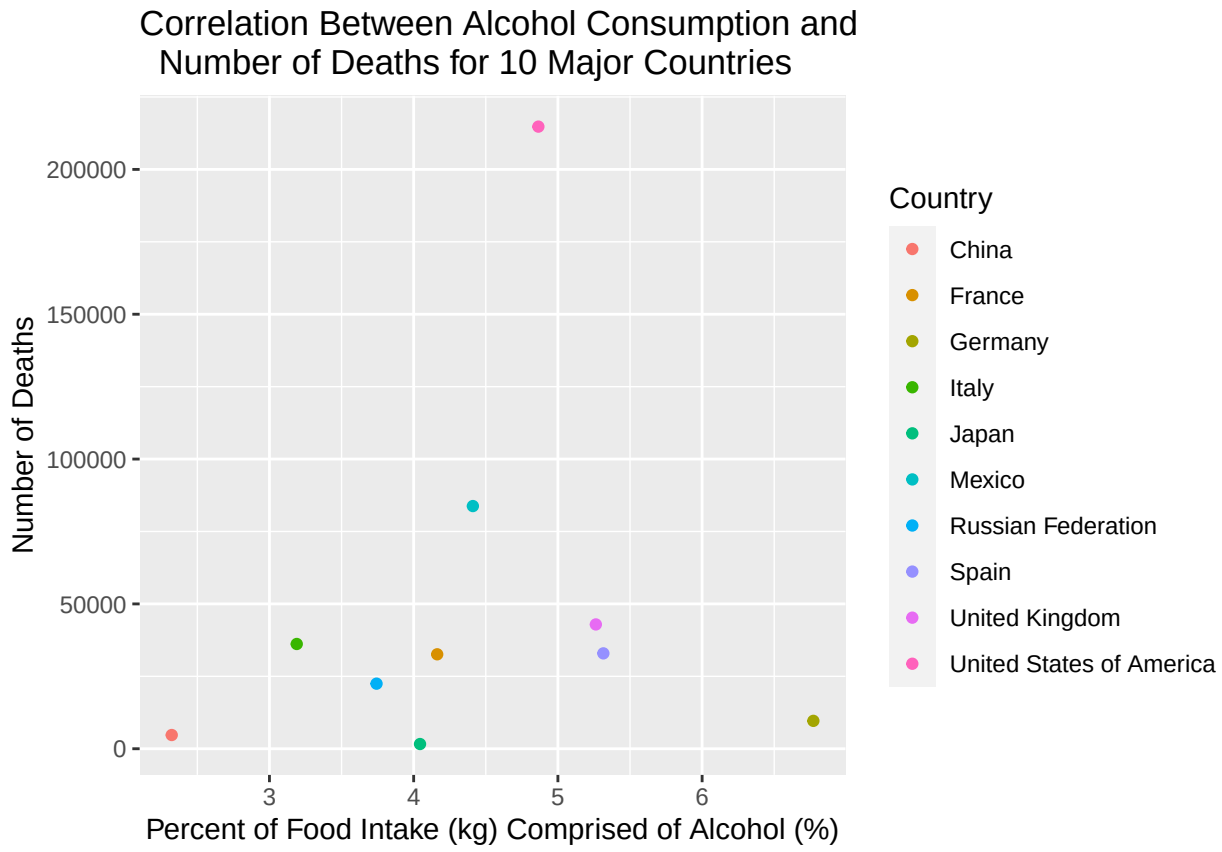
```
#Reference 2: Guardian article provided overview of major countries affected
#by Covid-19
#Reference 6: Professor Martinet helped me adjust the x-axis labels.
```

Graph 2-Graphical Summary

Graphical representation scatter plot of correlation between alcoholic consumption and number of deaths for each of 10 major world nations

```
#Note: "Deaths" variable in data set is a percent itself,
#so dividing by 100 gave the proportion of deaths for each country
#which was then multiplied by "Population" to get "Number of Deaths" on y-axis.

ggplot(majorcountrydata,aes(x=Alcoholic.Beverages,y=((Deaths/100)*Population),
color=Country))+geom_point()+labs(
  title="Correlation Between Alcohol Consumption and
  Number of Deaths for 10 Major Countries"
,x="Percent of Food Intake (kg) Comprised of Alcohol (%)",
y="Number of Deaths")
```



Conclusions from Numerical and Graphical Representations

Graph 1: Numerical summary bar graph of "percent of diet that is alcohol"

This numerical representation displays 10 major world nations that have been heavily impacted by Covid-19 on the x-axis, and respective percentages of their food intake that is comprised of alcohol. The data for choosing the 10 countries can be attributed to (Reference 2). Germany, Spain, the United Kingdom, and the United States had the highest percentages of alcoholic consumption, as displayed by the tallest bars. While the purpose of this first numerical summary is to provide an empirical percentage for each nation's alcoholic consumption, further conclusions will be made using Graph 2 which builds upon this data and measures correlation with the number of deaths for each country. It can also be observed that larger and wealthier nations have a higher alcoholic intake, as displayed by the higher percentages for the United States (4.86%) and the United Kingdom (5.26%).

Graph 2: Graphical representation scatter plot

This graph provides a graphical representation to be used in conjunction with Graph 1. It shows the number of deaths per country, calculated by dividing the Deaths (a percentage) variable by 100 and multiplying that proportion by the Population (counts) variable. Thus, the y-axis displays the death toll for each of the 10 nations as counts. A correlation is made with the x-axis, which displays the percentage of food intake (kilograms) attributed to alcohol. It is evident that the countries from Graph 1 with higher percentages of alcoholic consumption are correlated with a greater number of deaths, since a generally positive trend is seen in both variables from the left to right portions of the graph (rough linear increase). Specifically, the UK, US, Italy, and Mexico align with the trend that an increase in alcoholic intake is correlated with an increase in the number of deaths. Germany and Japan are exceptions in that they show high levels of alcoholic consumption (6.8% and 4% respectively) but have a low death count (approx. 9,600 and 1,600 deaths). This is due to the fact that the y-axis's use of counts to display deaths does not take into account the smaller size of some of the countries measured (which explains the exception for Germany) nor does the data displayed account for other variables including severity of quarantine and self-isolation measures (which may explain the low death count for Japan, where measures were strictly enforced). While the correlation measured is positive, it is not a strong correlation. Thus, the relationship solely between alcoholic consumption and death for each country is not a robust one, as death count is affected by hundreds of other variables that may involve the other food groups in the data, wealth of each nation, or public safety measures taken against the spread of the virus. Future hypothesis testing can study whether there exists a significant relationship between the alcoholic consumption percentages and death, versus a more significant relationship between the percentage of another food group (fish, vegetables, fruit) and death count. Hypothesis testing can be performed for each of the food groups and compared to analyze which, if any, have the greatest impact on death related to Covid-19.

References

1.https://www.kaggle.com/mariaren/covid19-healthy-diet-dataset?select=Food_Supply_Quantity_kg_Data.csv
2.https://www.theguardian.com/world/2020/oct/16/coronavirus-world-map-which-countries-have-the-most-covid-cases-and-deaths
3.https://www.fao.org/faostat/en/#home
4.https://www.prb.org/

5.https://coronavirus.jhu.edu/map.html
6. Professor Martinet's Office Hours