# DSC 520 PROJECT MILESTONE 2 - COVID 19 STATISTICAL ANALYSIS

ANBUSELVAN MAHALINGAM

November 05, 2023

#Milestone 2

#How to import and clean my data step 1: I will be using the read.table fo reading and loading the data into a data frame. step 2: validate the data bt displaying the data and its structure step 3: Create proper column names and convert the data into numerical for supporting machine learning model. step 4: Objective of this abstract is to come up with a statistical analysis predicting the race/ethnicity with their gender category who got registered with more number of covid deaths so that this category people can be given with extra care and studies to handle any future covid like situations with minimal deaths.

#What does the final data set look like? Final Data set available in the same document last pages. I will look for further expansion to add more significant variables supporting the prediction.

#Questions for future steps. we will be using the projected Male/Female percentage data for 2027 and assuming each race and ethnicity follow the same population for each sub groups. Projection estimates calculated using the mid-term population ration of 49.5 percent male and 50.47 percent female.

#What information is not self-evident? Here Gender population with-in each race and ethnicity sub group is not self contained.

#What are different ways you could look at this data? The data set is small and it can be looked in different way by visualization the relation between race/ethnicity and COVID deaths, COVID cases registered..

#How do you plan to slice and dice the data? Gender analysis can be further performed on this data and see, Comparison between male vs. female percentages regarding cases and deaths across races. comparative metrics can be further developed as death rate per 100,000 for each race/ethnicity.

#How could you summarize your data to answer key questions? Creating a Random Forest model for the given data , by setting the threshold and categorize each race and ethnicity had highest covid death and categorize them by Loa and high Category based on death count and number of affected COVID cases.

#What types of plots and tables will help you to illustrate the findings to your questions? Below visualization can be used to for exploring more information from the data set. COVID death and cases comparison by stacked bar Chart for Percentages. COVID death and cases comparison by stacked bar Chart for counts. Bar chart for COVID deaths count and Cases count.

#Do you plan on incorporating any machine learning techniques to answer your research questions? Explain. Yes, by using random forest model, we will be getting high/Low score on COVID deaths for the specific ethnicity/race along with the gender contribution. we will be using the “MeanDecreaseAccuracy” and “MeanDecreaseGini” metrics to measure how much each variable contributes to the prediction accuracy and purity of the model.

#Questions for future steps. For the better high/Low score on COVID deaths for the specific ethnicity/race. do we need to use only populations count instead of percentage? I will be exploring more options and performing further data set preparation and evaluate the random forest model again.

## Abstract

Today, everyone in the world should feel happy and enjoy their life and every current movement because we have passed through the critical covid-19 pandemic and the deadly illness. Corona virus disease (COVID-19) is an infectious disease caused by the SARS-CoV-2 virus. Anyone can get sick with COVID-19 and become seriously ill or die at any age. The objective of this abstract is to perform statistical analysis on available covid 19 data and produce insight on what are the categorical people group who got affected more and what are the categorical people who dies due to this deadly virus infection. Reference WHO, World Health organization, COIV-19 reference, URL:https://www.who.int/health-topics/coronavirus#tab=tab_1

## Introduction

we know COIVD-19 is not the first or the final disease we have seen, it is possible in the future we may get similar virus spread, we need have better prevention plan ad control any similar kind of virus spread in the future.The Objective of this research paper is to come up with statistical analysis and the current insight from covid-19 data, based on the research results we can form a prevention plan .

## Research questions

1. What are the most, impacted gender by the COVID 19 and registered high Number of death.
2. What are the most, impacted race by the COVID 19 and registered high Number of death.
3. What are the most, impacted gender by the COVID 19 and registered high Number of infected cases.
4. What are the most, impacted race by the COVID 19 and registered high Number of infected cases.
5. what are the correlated variables, and the population distribution by each state to the COVID 19 death.
6. what if the COVID like fever hits again, what are the group or category of people we need to provide extra care to reduce the impact. any predictive model needed to find the most impact full group or category of people who needs special attention and care.

## Approach

1. We will be getting the COVID 19 related data from trusted parties.Government, medical departments and Non-profit organizations provides data about COVID 19 death.
2. Perform exploratory data analysis & visualizations to get a clean data,identify significant variables and perform data wrangling process for getting clean data.
3. Perform single and multiple linear regression model analysis and identify a model with most significant variables relating to COVID 19 death count.
4. Evaluate the model performance and produce visualization supporting the created models.
5. If time permits use the prediction model and come up with the possible population by each category Age, race, Sex for each State and use it against model and project the future potential death in case if it happens again.

## How your approach addresses (fully or partially) the problem.

My approach will partially by identifying the significant variables such as sex, age, race related to COVID 19 death. for the prevention we need to predict the needed medical facility,and need to come up with the solution plan .the model we create in this abstract will only give insight about COVID death and solution need to be planed for prevention.

## Data (Minimum of 3 Datasets - but no requirement on number of fields or rows)

CDC Centers for Disease Control and Prevention - Provisional COVID-19 Deaths by Sex and Age Link: https://data.cdc.gov/NCHS/Provisional-COVID-19-Deaths-by-Sex-and-Age/9bhg-hcku

KFF, Population Distribution by Age - Link: https://www.kff.org/other/state-indicator/distribution-by-age/?currentTimeframe=0&sortModel=%7B%22colId%22:%22Location%22,%22sort%22:%22asc%22%7D

KFF, Population Distribution by Sex https://www.kff.org/other/state-indicator/distribution-by-sex/?currentTimeframe=0&sortModel=%7B%22colId%22:%22Location%22,%22sort%22:%22asc%22%7D

## Required Packages

ggplot2, dplyr, Metrics, reshape. Based on implementation need we will consider adding further packages to support data wrangling and visualization.

## Plots and Table Needs

1. Data Distribution Plots - To visualize the distribution of numeric variables using histograms, density plots, or box plots.
2. Categorical Variable Counts - To create bar plots or pie charts to display the distribution of categorical variables.
3. Correlation Matrix and Heatmap - To Correlation Matrix and Heatmap
4. Scatterplots - Create scatterplots to explore relationships between pairs of numeric variables.
5. Outlier Detection Plots - Use box plots, scatterplots, or other visualization techniques to identify outliers in your data.

## Questions for future steps

1. Even today, people are getting infected by COVID and few state has COID death registered, do we need to consider real-time data? or go with the latest data from CDC?
2. do we need to consider adding any other significant variable for this prediction model, as of now we now we are scoping the ,prediction model with variables age, race. and sex distribution across the state.
3. It is possible some of COVID cases and COVID related death were never reported, or by mistake other virus and general sick and illness death for senior citizens might have included as COVID death, any further analysis needs to be performed on this to get better dataset or trust CDC data and proceed with that?

```
#To read data from CDC Centers for Disease Control and Prevention - Provisional COVID-19 Deaths by Sex
#Link: https://data.cdc.gov/NCHS/Provisional-COVID-19-Deaths-by-Sex-and-Age/9bhg-hcku

# COID 19 deaths_by_race_ethnicity__all_age_groups

# Load necessary libraries
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
# Load the covid death data set by race/ethnicity
covid_deaths_race_ethnicity <- read.table("deaths_by_race_ethnicity__all_age_groups.csv", sep = ",", qu

# Load the covid infected case data set by race/ethnicity
covid_cases_race_ethnicity <- read.table("cases_by_race_ethnicity__all_age_groups.csv", sep = ",", quot
```

```
# original data frame & dataset - EDA
covid_deaths_race_ethnicity
```

```
##                                                                                 V1
## 1                                    Deaths by Race/Ethnicity - All Age Groups
## 2  Date generated: Sun Oct 22 2023 13:30:59 GMT-0400 (Eastern Daylight Time)
## 3                                                                  Race/Ethnicity
## 4                                                                Hispanic/Latino
## 5                                American Indian / Alaska Native Non-Hispanic
## 6                                                              Asian Non-Hispanic
## 7                                                              Black Non-Hispanic
## 8                         Native Hawaiian / Other Pacific Islander Non-Hispanic
## 9                                                              White Non-Hispanic
## 10                                                   Multiple/Other Non-Hispanic
##                    V2              V3                         V4
## 1
## 2
## 3  Percent of deaths Count of deaths Percent of US population
## 4                16.4          141073                    18.45
## 5                   1            8962                     0.74
## 6                 3.2           27644                     5.76
## 7                13.1          112166                    12.54
## 8                 0.2            1968                    0.182
## 9                63.9          548781                    60.11
## 10                2.1           18204                     2.22
```

```
# original data frame & dataset - EDA
covid_cases_race_ethnicity
```

```
##                                                                                 V1
## 1                                    Cases by Race/Ethnicity - All Age Groups
## 2  Date generated: Sun Oct 22 2023 13:30:56 GMT-0400 (Eastern Daylight Time)
## 3                                                                  Race/Ethnicity
## 4                                                                Hispanic/Latino
## 5                                American Indian / Alaska Native Non-Hispanic
## 6                                                              Asian Non-Hispanic
## 7                                                              Black Non-Hispanic
## 8                         Native Hawaiian / Other Pacific Islander Non-Hispanic
```

```
## 9                                                    White Non-Hispanic
## 10                                                 Multiple/Other Non-Hispanic
##                  V2             V3                        V4
## 1
## 2
## 3  Percent of cases Count of cases Percent of US population
## 4                24       16480377                    18.45
## 5                 1         719324                     0.74
## 6               4.4        3005495                     5.76
## 7              12.6        8674984                    12.54
## 8               0.3         190128                    0.182
## 9              53.8       36875802                    60.11
## 10              3.8        2624035                     2.22
```

```
# Remove the first 3 rows
covid_deaths_race_ethnicity <- covid_deaths_race_ethnicity[-(1:3),]
covid_cases_race_ethnicity <- covid_cases_race_ethnicity[-(1:3),]

# Reset the row names to maintain consistency
rownames(covid_deaths_race_ethnicity) <- NULL
rownames(covid_cases_race_ethnicity) <- NULL


# Set new column names here the first row was the column name
colnames(covid_deaths_race_ethnicity) <- c("Race/Ethnicity", "Percent of deaths", " Count of deaths", "
colnames(covid_cases_race_ethnicity) <- c("Race/Ethnicity", "Percent of cases", "   Count of cases", "Pe
```

```
#Cleaned dataframe - COVID deaths by race & ethnicity
covid_deaths_race_ethnicity
```

```
##                                      Race/Ethnicity Percent of deaths
## 1                                   Hispanic/Latino              16.4
## 2          American Indian / Alaska Native Non-Hispanic                1
## 3                                 Asian Non-Hispanic               3.2
## 4                                 Black Non-Hispanic              13.1
## 5 Native Hawaiian / Other Pacific Islander Non-Hispanic               0.2
## 6                                 White Non-Hispanic              63.9
## 7                           Multiple/Other Non-Hispanic               2.1
##   \tCount of deaths Percent of US population
## 1            141073                    18.45
## 2              8962                     0.74
## 3             27644                     5.76
## 4            112166                    12.54
## 5              1968                    0.182
## 6            548781                    60.11
## 7             18204                     2.22
```

```
#cleaned dataframe - COVID infected cases by race & ethnicity
covid_cases_race_ethnicity
```

```
##                                      Race/Ethnicity Percent of cases
## 1                                   Hispanic/Latino               24
```

```
## 2          American Indian / Alaska Native Non-Hispanic               1
## 3                                   Asian Non-Hispanic             4.4
## 4                                   Black Non-Hispanic            12.6
## 5 Native Hawaiian / Other Pacific Islander Non-Hispanic             0.3
## 6                                   White Non-Hispanic            53.8
## 7                           Multiple/Other Non-Hispanic             3.8
##   \tCount of cases Percent of US population
## 1         16480377                    18.45
## 2           719324                     0.74
## 3          3005495                     5.76
## 4          8674984                    12.54
## 5           190128                    0.182
## 6         36875802                    60.11
## 7          2624035                     2.22
```

```
# To merge the both the dataframe into combined one dataframe

# Full outer join to merge both the dataframe based on thei Race and ethnicity
merged_Race_Ethnicity_df <- merge(covid_deaths_race_ethnicity, covid_cases_race_ethnicity, by = "Race/E

# drop the duplicate column Percent of US population.x from the merged df
merged_Race_Ethnicity_df$'Percent of US population.x' <- NULL  # Drop the column by assigning NULL

# Rename 'Percent of US population.y' to 'Percent of US population'
names(merged_Race_Ethnicity_df)[names(merged_Race_Ethnicity_df) == "Percent of US population.y"] <- "Pe

# To print the merged data frames
print(merged_Race_Ethnicity_df)
```

```
##                                    Race/Ethnicity Percent of deaths
## 1          American Indian / Alaska Native Non-Hispanic                1
## 2                                   Asian Non-Hispanic              3.2
## 3                                   Black Non-Hispanic             13.1
## 4                                      Hispanic/Latino             16.4
## 5                           Multiple/Other Non-Hispanic              2.1
## 6 Native Hawaiian / Other Pacific Islander Non-Hispanic              0.2
## 7                                   White Non-Hispanic             63.9
##   \tCount of deaths Percent of cases \tCount of cases Percent of US population
## 1              8962                1           719324                     0.74
## 2             27644              4.4          3005495                     5.76
## 3            112166             12.6          8674984                    12.54
## 4            141073               24         16480377                    18.45
## 5             18204              3.8          2624035                     2.22
## 6              1968              0.3           190128                    0.182
## 7            548781             53.8         36875802                    60.11
```

To add gender distribution to the race and ethnicity distribution From www.statista.com, 2027 projection for USA male/female distribution 2027: Projection estimates calculated using the mid-term population ration of 49.5 percent male and 50.47 percent female Refer: https://www.statista.com/statistics/737923/us-population-by-gender/#:~:text=Projection%20estimates%20calculated%20using%20the,US%20Census%20data%20for%2020

```
# we will use the same gender distribution projection and also the assumption o the gender distribution
```

```r
# Add male column 49.50% and female 50.47% to all race and ethnicity

# Assuming you have vectors with actual percentages for each
male_percentages <- c(49.50, 49.50, 49.50, 49.50, 49.50, 49.50, 49.50) # vector data for male percentag
female_percentages <- c(50.47, 50.47, 50.47, 50.47, 50.47, 50.47, 50.47) # vector data for female perce

# Add the male and female percentage data to the dataframe
merged_Race_Ethnicity_df$MalePercentage <- male_percentages
merged_Race_Ethnicity_df$FemalePercentage <- female_percentages

# View the dataframe
print(merged_Race_Ethnicity_df)
```

```
##                                       Race/Ethnicity Percent of deaths
## 1          American Indian / Alaska Native Non-Hispanic                1
## 2                                  Asian Non-Hispanic              3.2
## 3                                  Black Non-Hispanic             13.1
## 4                                     Hispanic/Latino             16.4
## 5                            Multiple/Other Non-Hispanic              2.1
## 6 Native Hawaiian / Other Pacific Islander Non-Hispanic              0.2
## 7                                  White Non-Hispanic             63.9
##   \tCount of deaths Percent of cases \tCount of cases Percent of US population
## 1              8962                1           719324                     0.74
## 2             27644              4.4          3005495                     5.76
## 3            112166             12.6          8674984                    12.54
## 4            141073               24         16480377                    18.45
## 5             18204              3.8          2624035                     2.22
## 6              1968              0.3           190128                    0.182
## 7            548781             53.8         36875802                    60.11
##   MalePercentage FemalePercentage
## 1           49.5            50.47
## 2           49.5            50.47
## 3           49.5            50.47
## 4           49.5            50.47
## 5           49.5            50.47
## 6           49.5            50.47
## 7           49.5            50.47
```

```r
# Clean up column names by removing leading/trailing whitespace and replacing spaces with underscores
names(merged_Race_Ethnicity_df) <- gsub("^\\s+|\\s+$", "", names(merged_Race_Ethnicity_df)) # remove le
names(merged_Race_Ethnicity_df) <- gsub("\\s+", "_", names(merged_Race_Ethnicity_df)) # replace spaces
str(merged_Race_Ethnicity_df)
```

```
## 'data.frame':    7 obs. of  8 variables:
##  $ Race/Ethnicity          : chr  "American Indian / Alaska Native Non-Hispanic" "Asian Non-Hispanic
##  $ Percent_of_deaths       : chr  "1" "3.2" "13.1" "16.4" ...
##  $ Count_of_deaths         : chr  "8962" "27644" "112166" "141073" ...
##  $ Percent_of_cases        : chr  "1" "4.4" "12.6" "24" ...
##  $ Count_of_cases          : chr  "719324" "3005495" "8674984" "16480377" ...
##  $ Percent_of_US_population: chr  "0.74" "5.76" "12.54" "18.45" ...
##  $ MalePercentage          : num  49.5 49.5 49.5 49.5 49.5 49.5 49.5
##  $ FemalePercentage        : num  50.5 50.5 50.5 50.5 50.5 ...
```

```r
# Convert the 'Count_of_deaths' and 'Count_of_cases' to numeric after ensuring they're clean
merged_Race_Ethnicity_df$Count_of_deaths <- as.numeric(as.character(merged_Race_Ethnicity_df$Count_of_de
merged_Race_Ethnicity_df$Count_of_cases <- as.numeric(as.character(merged_Race_Ethnicity_df$Count_of_cas

# Since we have not seen the 'Female_Percentage' column before in your description,
# ensure that it is in the data frame and also numeric, the same way you convert other numeric columns.

# Handle possible NA values after conversion
merged_Race_Ethnicity_df <- na.omit(merged_Race_Ethnicity_df) # Or some other appropriate NA handling t
```

FINAL DATASET

```r
# Rename 'Race/Ethnicity' to 'Race_Ethnicity'
names(merged_Race_Ethnicity_df)[names(merged_Race_Ethnicity_df) == "Race/Ethnicity"] <- "Race_Ethnicity

# Rename 'MalePercentage' to 'Male_Percentage'
names(merged_Race_Ethnicity_df)[names(merged_Race_Ethnicity_df) == "MalePercentage"] <- "Male_Percentag

# Rename 'FemalePercentage' to 'Female_Percentage'
names(merged_Race_Ethnicity_df)[names(merged_Race_Ethnicity_df) == "FemalePercentage"] <- "Female_Perce


#names(merged_Race_Ethnicity_df)[names(merged_Race_Ethnicity_df) == "Percent of deaths"] <- "Percent_of_
#names(merged_Race_Ethnicity_df)[names(merged_Race_Ethnicity_df) == "Count of deaths"] <- "Count_of_dea
#names(merged_Race_Ethnicity_df)[names(merged_Race_Ethnicity_df) == "Percent of cases"] <- "Percent_of_
#names(merged_Race_Ethnicity_df)[names(merged_Race_Ethnicity_df) == "Count of cases"] <- "Count_of_case
#names(merged_Race_Ethnicity_df)[names(merged_Race_Ethnicity_df) == "Percent of US population"] <- "Per



#print(merged_Race_Ethnicity_df)
merged_Race_Ethnicity_df$Percent_of_deaths <- as.numeric(as.character(merged_Race_Ethnicity_df$Percent_
merged_Race_Ethnicity_df$Percent_of_cases <- as.numeric(as.character(merged_Race_Ethnicity_df$Percent_o
merged_Race_Ethnicity_df$Percent_of_US_population <- as.numeric(as.character(merged_Race_Ethnicity_df$P
# Convert factors to numeric if they are encoded as characters
merged_Race_Ethnicity_df$Count_of_deaths <- as.numeric(merged_Race_Ethnicity_df$Count_of_deaths)
merged_Race_Ethnicity_df$Count_of_cases <- as.numeric(merged_Race_Ethnicity_df$Count_of_cases)
```

```r
str(merged_Race_Ethnicity_df)
```

```
## 'data.frame':    7 obs. of  8 variables:
##  $ Race_Ethnicity          : chr  "American Indian / Alaska Native Non-Hispanic" "Asian Non-Hispanic
##  $ Percent_of_deaths       : num  1 3.2 13.1 16.4 2.1 0.2 63.9
##  $ Count_of_deaths         : num  8962 27644 112166 141073 18204 ...
##  $ Percent_of_cases        : num  1 4.4 12.6 24 3.8 0.3 53.8
##  $ Count_of_cases          : num  719324 3005495 8674984 16480377 2624035 ...
##  $ Percent_of_US_population: num  0.74 5.76 12.54 18.45 2.22 ...
##  $ Male_Percentage         : num  49.5 49.5 49.5 49.5 49.5 49.5 49.5
##  $ Female_Percentage       : num  50.5 50.5 50.5 50.5 50.5 ...
```

```r
# Current USA population 339996563
# Step 1 - calculate the total population by each race/ethnicity
# step 2 - calculate the male and female gender population by their percentage
#from each race/ethnicity category.

total_us_population <- 339996563

#Step 1
#Calculate the total population for each race and ethnicity based on their percentage
merged_Race_Ethnicity_df$Total_Population_Count <-
  total_us_population * merged_Race_Ethnicity_df$Percent_of_US_population / 100


# Now calculate the male and female populations again with the new estimated totals
merged_Race_Ethnicity_df$Male_Population_Count <-
  merged_Race_Ethnicity_df$Total_Population_Count * merged_Race_Ethnicity_df$Male_Percentage / 100
merged_Race_Ethnicity_df$Female_Population_Count <-
  merged_Race_Ethnicity_df$Total_Population_Count * merged_Race_Ethnicity_df$Female_Percentage / 100

merged_Race_Ethnicity_df
```

```
##                                       Race_Ethnicity Percent_of_deaths
## 1          American Indian / Alaska Native Non-Hispanic               1.0
## 2                                  Asian Non-Hispanic               3.2
## 3                                  Black Non-Hispanic              13.1
## 4                                     Hispanic/Latino              16.4
## 5                            Multiple/Other Non-Hispanic               2.1
## 6 Native Hawaiian / Other Pacific Islander Non-Hispanic               0.2
## 7                                  White Non-Hispanic              63.9
##   Count_of_deaths Percent_of_cases Count_of_cases Percent_of_US_population
## 1            8962              1.0         719324                    0.740
## 2           27644              4.4        3005495                    5.760
## 3          112166             12.6        8674984                   12.540
## 4          141073             24.0       16480377                   18.450
## 5           18204              3.8        2624035                    2.220
## 6            1968              0.3         190128                    0.182
## 7          548781             53.8       36875802                   60.110
##   Male_Percentage Female_Percentage Total_Population_Count
## 1            49.5             50.47               2515974.6
## 2            49.5             50.47              19583802.0
## 3            49.5             50.47              42635569.0
## 4            49.5             50.47              62729365.9
## 5            49.5             50.47               7547923.7
## 6            49.5             50.47                618793.7
## 7            49.5             50.47             204371934.0
##   Male_Population_Count Female_Population_Count
## 1             1245407.4               1269812.4
## 2             9693982.0               9883944.9
## 3            21104606.7              21518171.7
## 4            31051036.1              31659511.0
## 5             3736222.2               3809437.1
## 6              306302.9                312305.2
## 7           101164107.3             103146515.1
```

```r
# Random Forest model for handling both categorical and numerical values
# By using random forest model we will be getting high/Low score for the specific ethnicity/race along

# Convert the 'Race_Ethnicity' into a factor
merged_Race_Ethnicity_df$Race_Ethnicity <- as.factor(merged_Race_Ethnicity_df$Race_Ethnicity)

if (!require("foreign")) {
  install.packages("randomForest")
}
```

```
## Loading required package: foreign
```

```r
library(randomForest)
```

```
## Warning: package 'randomForest' was built under R version 4.3.2
```

```
## randomForest 4.7-1.1
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
##
## Attaching package: 'randomForest'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```r
# interested in high vs. low death counts based on a threshold
median_deaths <- median(merged_Race_Ethnicity_df$Count_of_deaths)
merged_Race_Ethnicity_df$High_Death_Count <- as.factor(ifelse(merged_Race_Ethnicity_df$Count_of_deaths

# Fit Random Forest model
rf_model <- randomForest(High_Death_Count ~ Race_Ethnicity + Count_of_deaths +
                           Count_of_cases + Total_Population_Count + Male_Population_Count +
                           Female_Population_Count, data = merged_Race_Ethnicity_df,
                         ntree = 500, importance = TRUE)

# View model importance
importance(rf_model)
```

```
##                             High      Low MeanDecreaseAccuracy MeanDecreaseGini
## Race_Ethnicity          0.000000 0.000000             0.000000        0.2685714
## Count_of_deaths         5.426786 5.301958             5.430808        0.6342857
## Count_of_cases          5.042901 4.897830             5.057862        0.5497143
## Total_Population_Count  5.463342 5.246466             5.455016        0.4262857
## Male_Population_Count   5.752237 5.494484             5.663766        0.5097143
## Female_Population_Count 6.534103 6.169337             6.480583        0.6548571
```

```
#FINAL DATA

# From the Results: Male_Population_Count: is the most important feature based
#on the MeanDecreaseAccuracy metric and has the highest MeanDecreaseGini score,
#indicating it plays a significant role in the model's predictions.
merged_Race_Ethnicity_df
```

```
##                                       Race_Ethnicity Percent_of_deaths
## 1          American Indian / Alaska Native Non-Hispanic               1.0
## 2                                  Asian Non-Hispanic               3.2
## 3                                  Black Non-Hispanic              13.1
## 4                                     Hispanic/Latino              16.4
## 5                            Multiple/Other Non-Hispanic               2.1
## 6 Native Hawaiian / Other Pacific Islander Non-Hispanic               0.2
## 7                                  White Non-Hispanic              63.9
##   Count_of_deaths Percent_of_cases Count_of_cases Percent_of_US_population
## 1            8962              1.0         719324                    0.740
## 2           27644              4.4        3005495                    5.760
## 3          112166             12.6        8674984                   12.540
## 4          141073             24.0       16480377                   18.450
## 5           18204              3.8        2624035                    2.220
## 6            1968              0.3         190128                    0.182
## 7          548781             53.8       36875802                   60.110
##   Male_Percentage Female_Percentage Total_Population_Count
## 1            49.5             50.47               2515974.6
## 2            49.5             50.47              19583802.0
## 3            49.5             50.47              42635569.0
## 4            49.5             50.47              62729365.9
## 5            49.5             50.47               7547923.7
## 6            49.5             50.47                618793.7
## 7            49.5             50.47             204371934.0
##   Male_Population_Count Female_Population_Count High_Death_Count
## 1             1245407.4               1269812.4              Low
## 2             9693982.0               9883944.9              Low
## 3            21104606.7              21518171.7             High
## 4            31051036.1              31659511.0             High
## 5             3736222.2               3809437.1              Low
## 6              306302.9                312305.2              Low
## 7           101164107.3             103146515.1             High
```

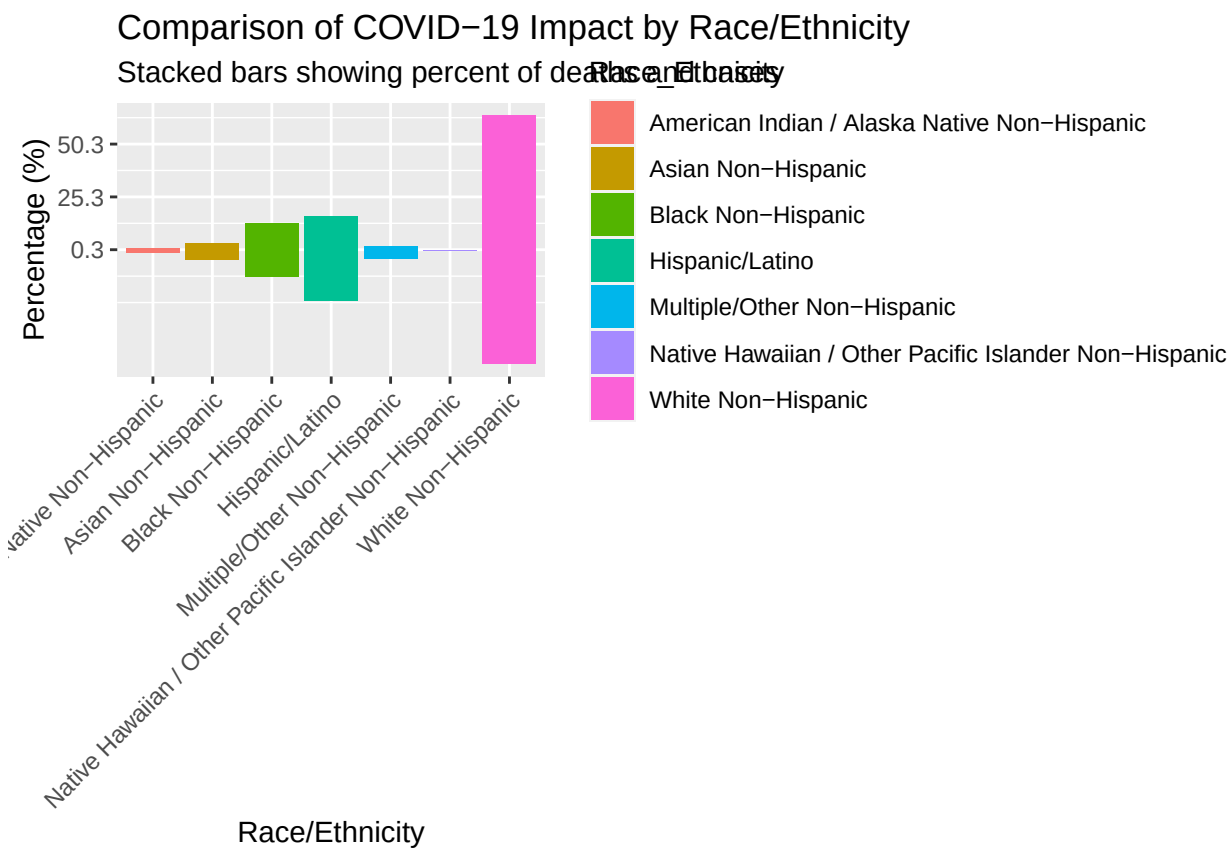
```
library("ggplot2")
```

```
##
## Attaching package: 'ggplot2'

## The following object is masked from 'package:randomForest':
##
##     margin
```

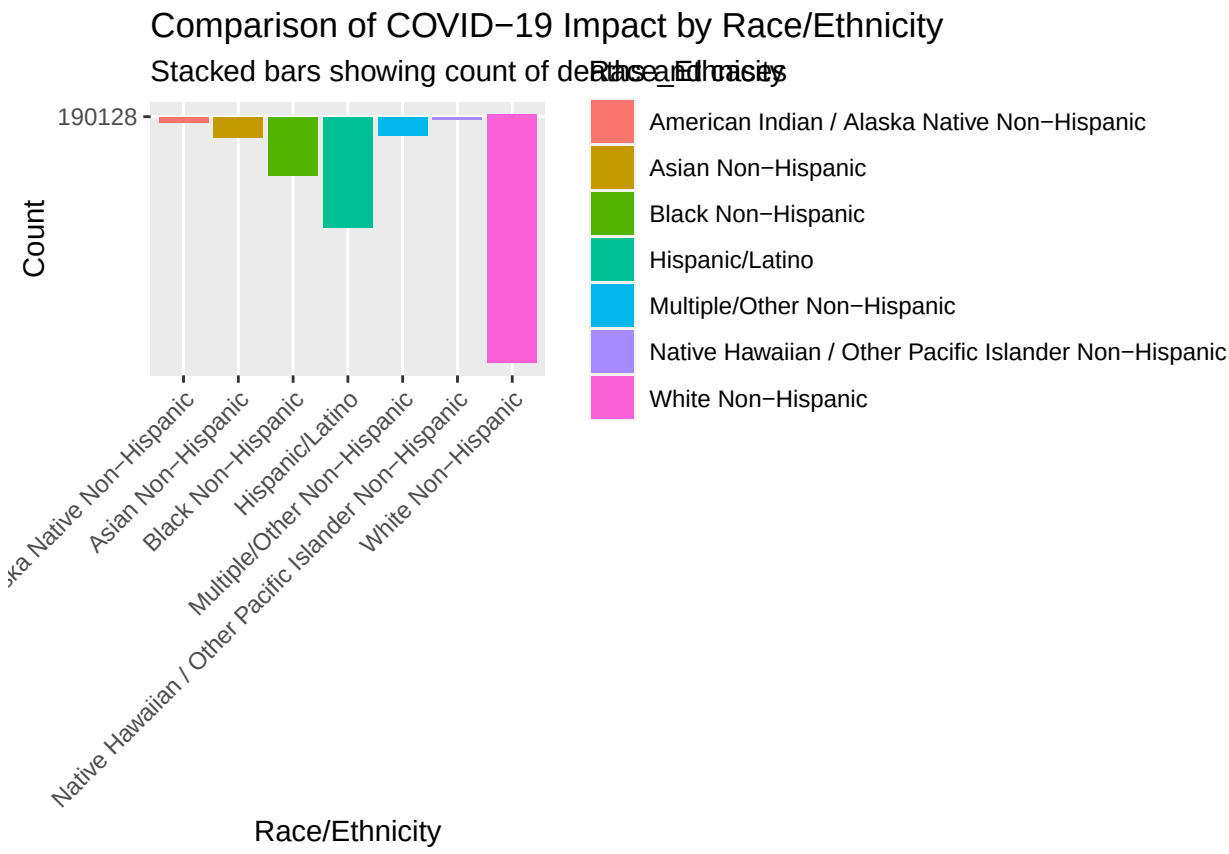
```
# Stacked Bar Chart for Percentages comparison
ggplot(merged_Race_Ethnicity_df, aes(fill=Race_Ethnicity, y=Percent_of_deaths, x=Race_Ethnicity)) +
  geom_bar(position="stack", stat="identity") +
```

```
  geom_bar(aes(y=-Percent_of_cases), position="stack", stat="identity") +
  scale_y_continuous(labels=abs, breaks=abs(seq(min(merged_Race_Ethnicity_df$Percent_of_cases), max(merg
  labs(title="Comparison of COVID-19 Impact by Race/Ethnicity",
       subtitle="Stacked bars showing percent of deaths and cases",
       y="Percentage (%)",
       x="Race/Ethnicity") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



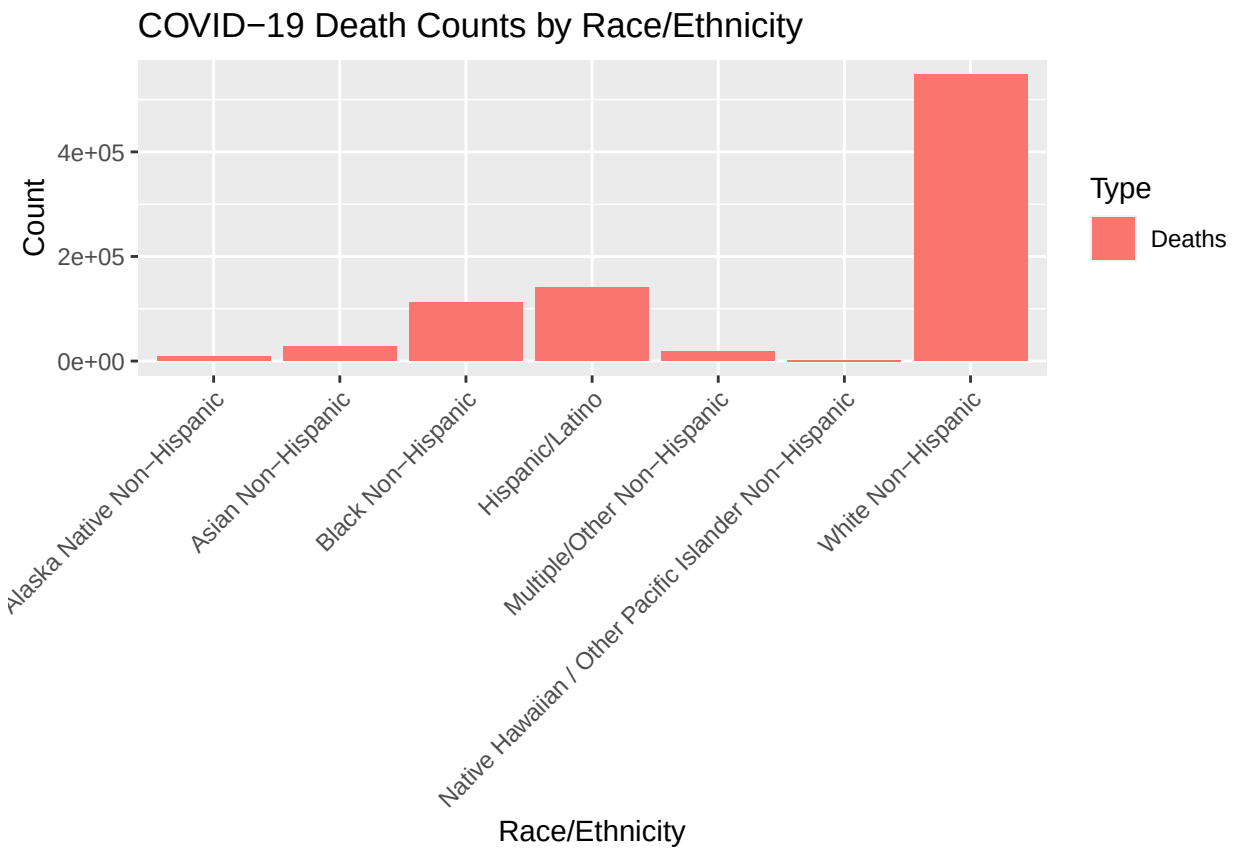

```
  #theme_minimal()
```

```
# Stacked Bar Chart for count comparison
ggplot(merged_Race_Ethnicity_df, aes(fill=Race_Ethnicity, y=Count_of_deaths, x=Race_Ethnicity)) +
  geom_bar(position="stack", stat="identity") +
  geom_bar(aes(y=-Count_of_cases), position="stack", stat="identity") +
  scale_y_continuous(labels=abs, breaks=abs(seq(min(merged_Race_Ethnicity_df$Count_of_cases), max(merge
  labs(title="Comparison of COVID-19 Impact by Race/Ethnicity",
       subtitle="Stacked bars showing count of deaths and cases",
       y="Count",
       x="Race/Ethnicity") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



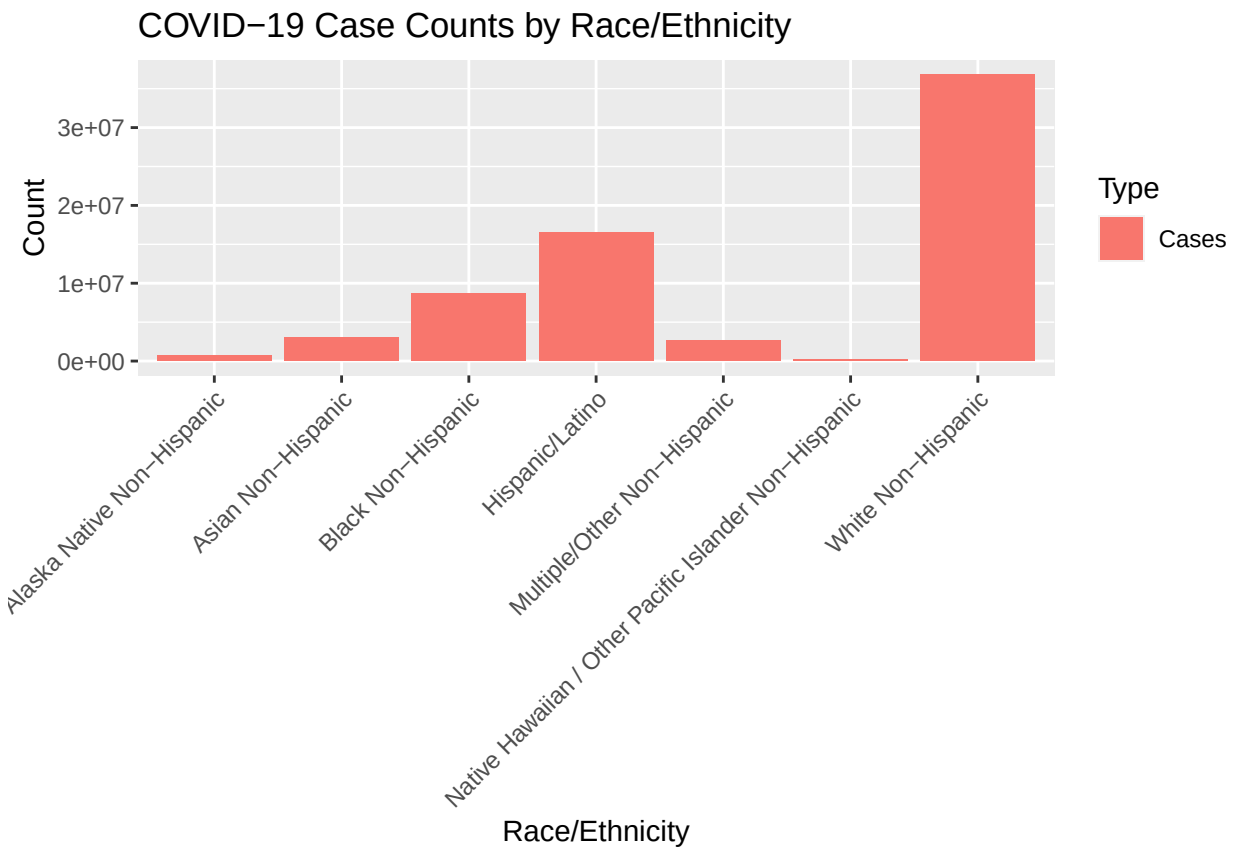

```
  #theme_minimal()
```

```
# Bar Chart for COVID deaths Count Per race/ethnicity
ggplot(merged_Race_Ethnicity_df) +
  geom_bar(aes(x=Race_Ethnicity, y=Count_of_deaths, fill="Deaths"), position="dodge", stat="identity") +
  labs(title="COVID-19 Death Counts by Race/Ethnicity",
       y="Count",
       x="Race/Ethnicity",
       fill="Type") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```
  #theme_minimal()
```

```
#  Bar Chart for COVID cases Counts Per race/ethnicity
ggplot(merged_Race_Ethnicity_df) +
  #geom_bar(aes(x=Race_Ethnicity, y=Count_of_deaths, fill="Deaths"), position="dodge", stat="identity") +
  geom_bar(aes(x=Race_Ethnicity, y=Count_of_cases, fill="Cases"), position="dodge", stat="identity") +
  labs(title="COVID-19 Case Counts by Race/Ethnicity",
       y="Count",
       x="Race/Ethnicity",
       fill="Type") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```
#theme_minimal()
```

## References

KFF data reference for population distribution by sex, age, race and ethnicity Population Distribution by Age: https://www.kff.org/other/state-indicator/distribution-by-age/?currentTimeframe=0&sortModel=%7B%22colId%22:%22Location%22,%22sort%22:%22asc%22%7D Population Distribution by Sex: https://www.kff.org/other/state-indicator/distribution-by-sex/?currentTimeframe=0&sortModel=%7B%22colId%22:%22Location%22,%22sort%22:%22asc%22%7D

CDC Cumulative COVID cases and death data, WHO reference, Link: https://data.cdc.gov/NCHS/Provisional-COVID-19-Deaths-by-Sex-and-Age/9bhg-hcku Covid 19 cases and death count, www.worldometers.