Understanding and Visualizing COVID data

Contents

Section 1: Introduction

Covid-19 has had a significant impact on the world and people's lives for the past 3 years. According to the World Health Organization, there have been over 765 million cases and over 6.9 million deaths due to the virus. I have seen various Covid statistics such as daily trends on news outlets, but I was curious to learn even more about it and get a sense of the impact that it's had. This project is an exercise to look at the big picture and to try to understand and visualize how various countries around the world and particularly the United States were affected and also to observe any correlations between information in the dataset.

Section 2: Data

This project is based on data available via "Our World In Data" which is an open source and open access effort to provide data for many problems(Mathieu et al. 2020).

The Covid data here is a time series data set that has lots of information including cases, deaths, people vaccinated, people hospitalized, population density,median age, and more. The data columns that I chose seemed interesting and there were enough data points in these columns. Some of the columns appeared to have relatively low amounts of recorded data points which discouraged me from using them. For this project the total cases, total deaths, total cases per million, total deaths per million, population density, and median age were chosen.

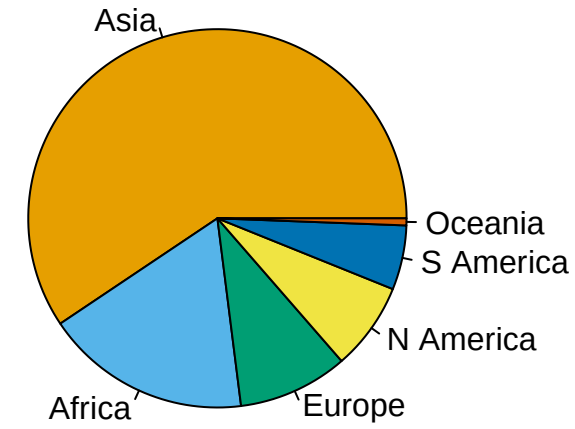
```
# Read data - do not print the column names
covidDf <- read_csv("owid-covid-data.csv", show_col_types=FALSE)
# Keeping only the columns needed
df = select(covidDf, continent, location, total_cases, total_deaths,
                     total_cases_per_million, total_deaths_per_million,
                     population_density, median_age)

# Cleaning up the code. drop_na() deleted every line and so cannot be used
# Replacing all NA columns with a 0 in some of the columns
df$total_cases <-
    ifelse(is.na(df$total_cases), 0, df$total_cases)
df$total_deaths <-
    ifelse(is.na(df$total_deaths), 0, df$total_deaths)
df$total_cases_per_million <-
    ifelse(is.na(df$total_cases_per_million), 0,
           df$total_cases_per_million)
df$total_deaths_per_million <-
```

```
    ifelse(is.na(df$total_deaths_per_million), 0,
          df$total_deaths_per_million)

# Some rows do not have continent information - filter it out
df <- filter(df, !is.na(continent))
```

Section 3. The Big Picture

The percentage of world population in every continent is shown in the pie chart below. Given that Covid transmission depends on proximity, it will be interesting to see if the distribution of cases and deaths match the population levels.

```
continent <- c("Asia", "Africa", "Europe", "N America", "S America", "Oceania")
population <- c(59.4, 17.6, 9.4, 7.5, 5.5, 0.6)
world_df <- data.frame(continent, population)

continent_colors <- c("#E69F00", "#56B4E9", "#009E73", "#F0E442", "#0072B2", "#D55E00")

pie(world_df$population, labels = world_df$continent, main = "Population by Continent", col = continent
```



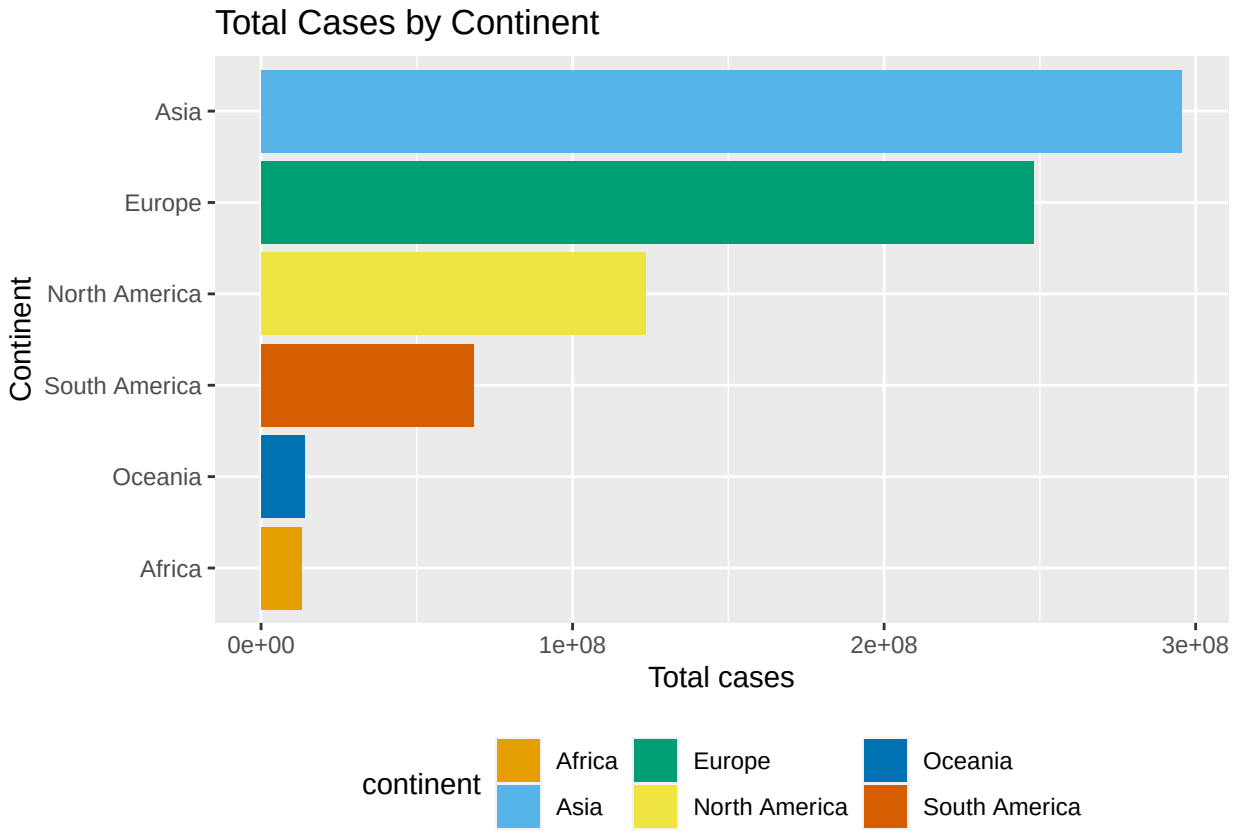
Total Cases in various Continents

Asia recorded the highest number of cases which might partially due to it containing nearly 60% of the world's population. A much less populous continent like Oceania has much fewer cases. However, Africa, which has more people than Oceania(Africa has about 17% of the world's population whereas Oceania has about 0.6% of the world's population) had fewer cases than Oceania. This shows that the amount of Covid cases in an area depended on more than that area's population.

```
# Get cases by country and continent
# Since this is a time series data, we need the max value
cases_country <- df %>%
  group_by(continent, location) %>%
  summarize(final_cases = max(total_cases))

# Get total cases by continent
cases_continent <- cases_country %>%
```

```
  group_by(continent) %>%
  summarize(final_cases = sum(final_cases))

# Get deaths by country and continent
# Since this is a time series data, we need the max value
deaths_country <- df %>%
  group_by(continent, location) %>%
  summarize(final_deaths = max(total_deaths))

# Get total deaths by continent
deaths_continent <- deaths_country %>%
  group_by(continent) %>%
  summarize(final_deaths = sum(final_deaths))

# Combine the tables for final data by continent
final_by_continent = full_join(cases_continent,
                               deaths_continent, by=c("continent") )
```

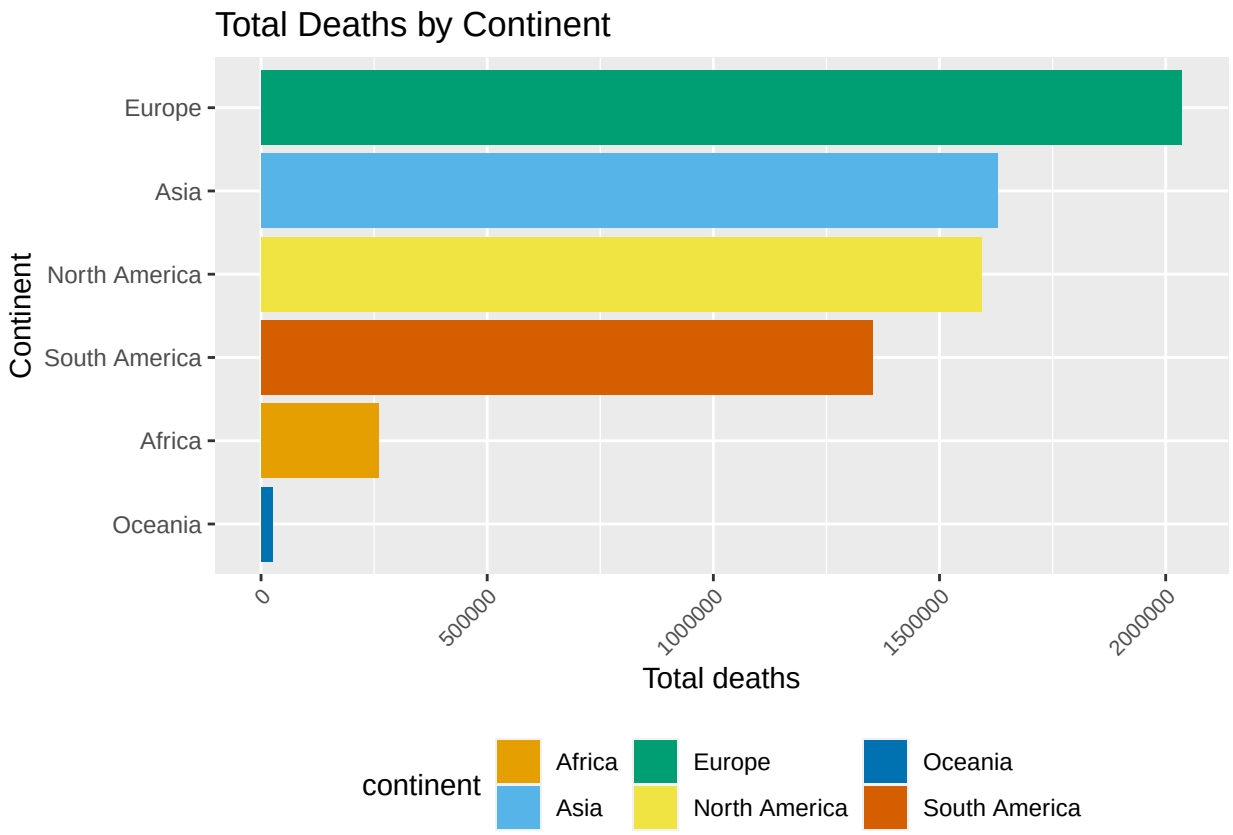
```
continent_colors <- c("#E69F00", "#56B4E9", "#009E73", "#F0E442", "#0072B2", "#D55E00")

# create a bar plot for final_cases
plot1 <- ggplot(final_by_continent, aes(x=final_cases,y=reorder(continent,final_cases),fill = continent)
  geom_col() +
  scale_fill_manual(values = continent_colors) +
  labs(title = "Total Cases by Continent", x = "Total cases", y = "Continent") +
  theme(legend.position="bottom")
plot1
```



Total Deaths in various Continents

Europe led the world in number of deaths with around 2 million deaths. Europe has one sixth of the population of Asia and yet it recorded almost half a million more deaths compared to Asia. Asia had only slightly more deaths than North America despite having a significantly larger population. Despite being the second most populated continent after Asia, Africa had the second lowest number of deaths. This suggests that many factors other than population contributed to the death counts. Some of these factors might be masking and vaccination policies, government shutdown policies, the immunity of the population, and the availability and quality of healthcare.

```
# create a bar plot for final_deaths
plot2 <- ggplot(final_by_continent, aes(x=final_deaths, y=reorder(continent,final_deaths),fill = contin
  geom_col() +
  scale_fill_manual(values = continent_colors) +
  labs(title = "Total Deaths by Continent", x = "Total deaths", y = "Continent") +
  theme(legend.position="bottom")+
  theme(axis.text.x=element_text(angle=45, hjust=1, size=8))
plot2
```

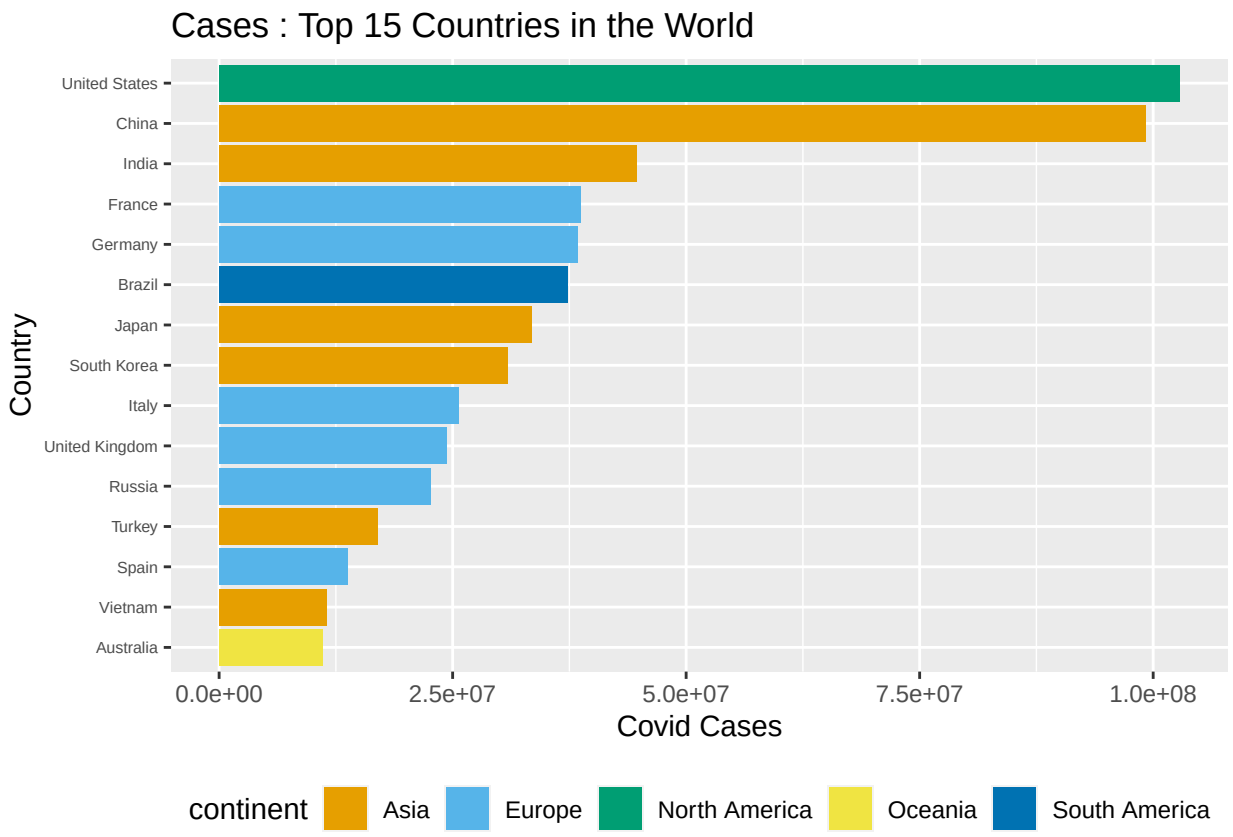


Top 15 countries with Highest Cases Worldwide

United States topped the world with the largest recorded number of Covid cases even though the population of the country is far lower compared to many of the other countries. The rest of the top 15 were mainly in Asia and Europe.

```
# Top 15 countries for Cases in a World
top_15_cases_world<-cases_country %>%
  group_by(location) %>%
  arrange(desc(final_cases)) %>%
```

```
  head(15)

plot2a <- ggplot(top_15_cases_world, aes(x=final_cases, y=reorder(location, final_cases), fill=continen
  geom_col()+
  labs(x="Covid Cases", y="Country", title="Cases : Top 15 Countries in the World") +
  scale_fill_manual(values = continent_colors) +
  theme(axis.text.y = element_text(size = 6)) +
  theme(legend.position="bottom")
plot2a
```

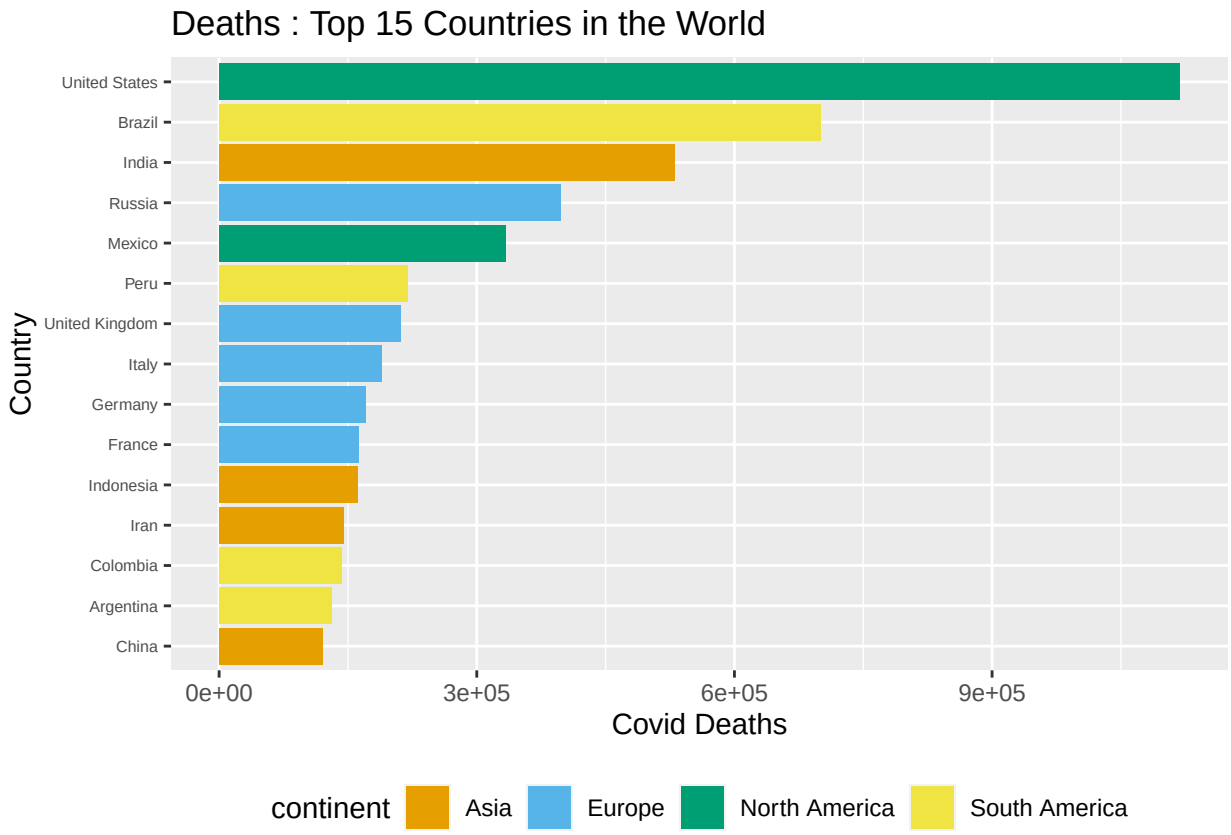


Top 15 countries with Highest Deaths Worldwide

United States has the largest number of Covid deaths of any country. It is interesting that India which is currently the most populous country and has nearly four times the population of the United States had less than half of its death rate. Brazil was another country that was affected disproportionately relative to population compared to other countries during this pandemic.

```
top_15_deaths_world<-deaths_country %>%
  group_by(location) %>%
  arrange(desc(final_deaths)) %>%
  head(15)

plot2b <- ggplot(top_15_deaths_world, aes(x=final_deaths, y=reorder(location, final_deaths), fill=conti
  geom_col()+
  labs(x="Covid Deaths", y="Country", title="Deaths : Top 15 Countries in the World") +
  scale_fill_manual(values = continent_colors) +
  theme(axis.text.y = element_text(size = 6)) +
```

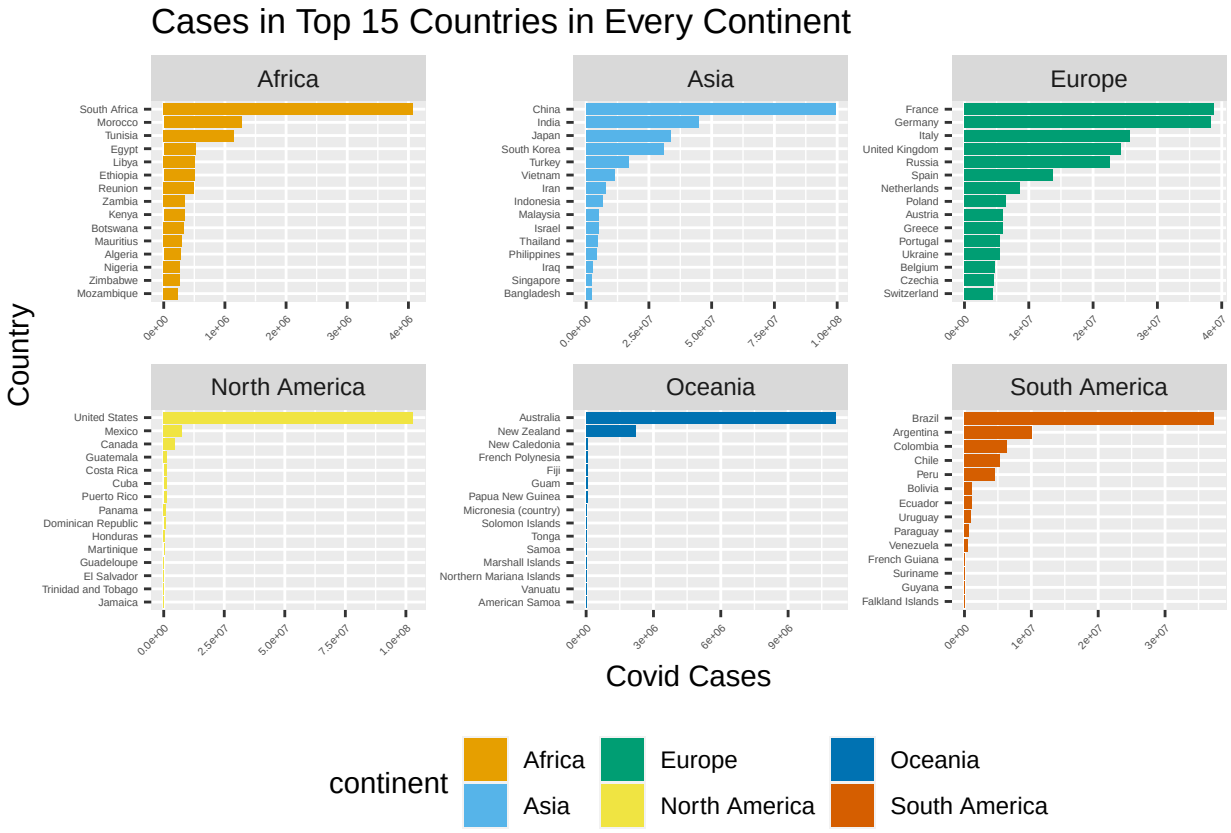
```
  theme(legend.position="bottom")
#  theme(axis.text.x=element_text(angle=45, hjust=1, size=4))
plot2b
```



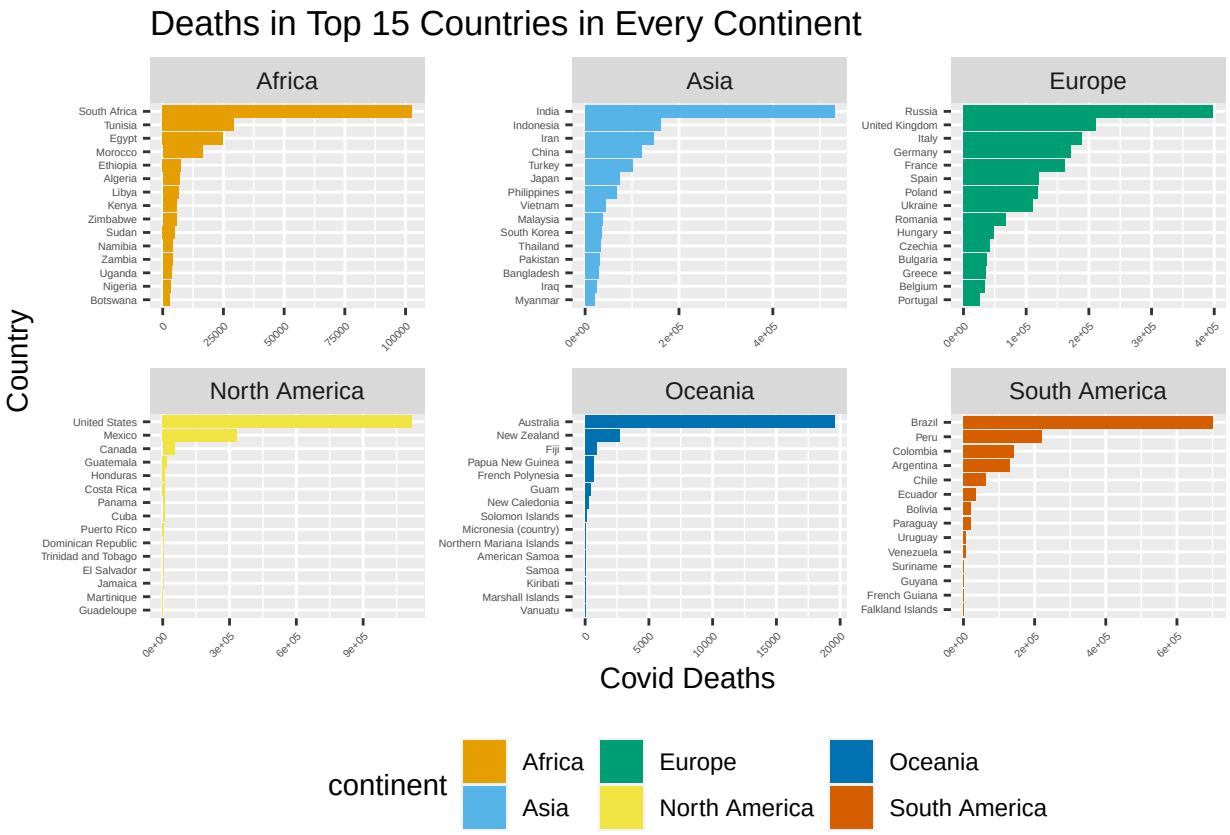
Top 15 Countries in every Continent

Zooming down into Continents, the top 15 countries for cases and deaths are visualized below.

```
# Top 15 countries for Cases and Deaths in each continent
cases_country<-cases_country %>%
  arrange(desc(final_cases)) %>%
  slice_head(n=15)

plot3 <- ggplot(cases_country, aes(x=final_cases, y=reorder(location,final_cases), fill=continent)) +
  geom_col()+
  facet_wrap(~continent, scales="free")+
  labs(x="Covid Cases", y="Country", title="Cases in Top 15 Countries in Every Continent") +
  scale_fill_manual(values = continent_colors) +
  theme(axis.text.y = element_text(size = 4)) +
  theme(legend.position="bottom")+
  theme(axis.text.x=element_text(angle=45, hjust=1, size=4))
plot3
```




```
deaths_country<-deaths_country %>%
  arrange(desc(final_deaths)) %>%
  slice_head(n=15)

plot4 <- ggplot(deaths_country, aes(x=final_deaths, y=reorder(location,final_deaths), fill=continent)) +
  geom_col()+
  facet_wrap(~continent, scales="free")+
  labs(x="Covid Deaths", y="Country", title="Deaths in Top 15 Countries in Every Continent") +
  scale_fill_manual(values = continent_colors) +
  theme(axis.text.y = element_text(size = 4)) +
  theme(legend.position="bottom")+
  theme(axis.text.x=element_text(angle=45, hjust=1, size=4))
plot4
```



Section 4. Any Correlation between Data?

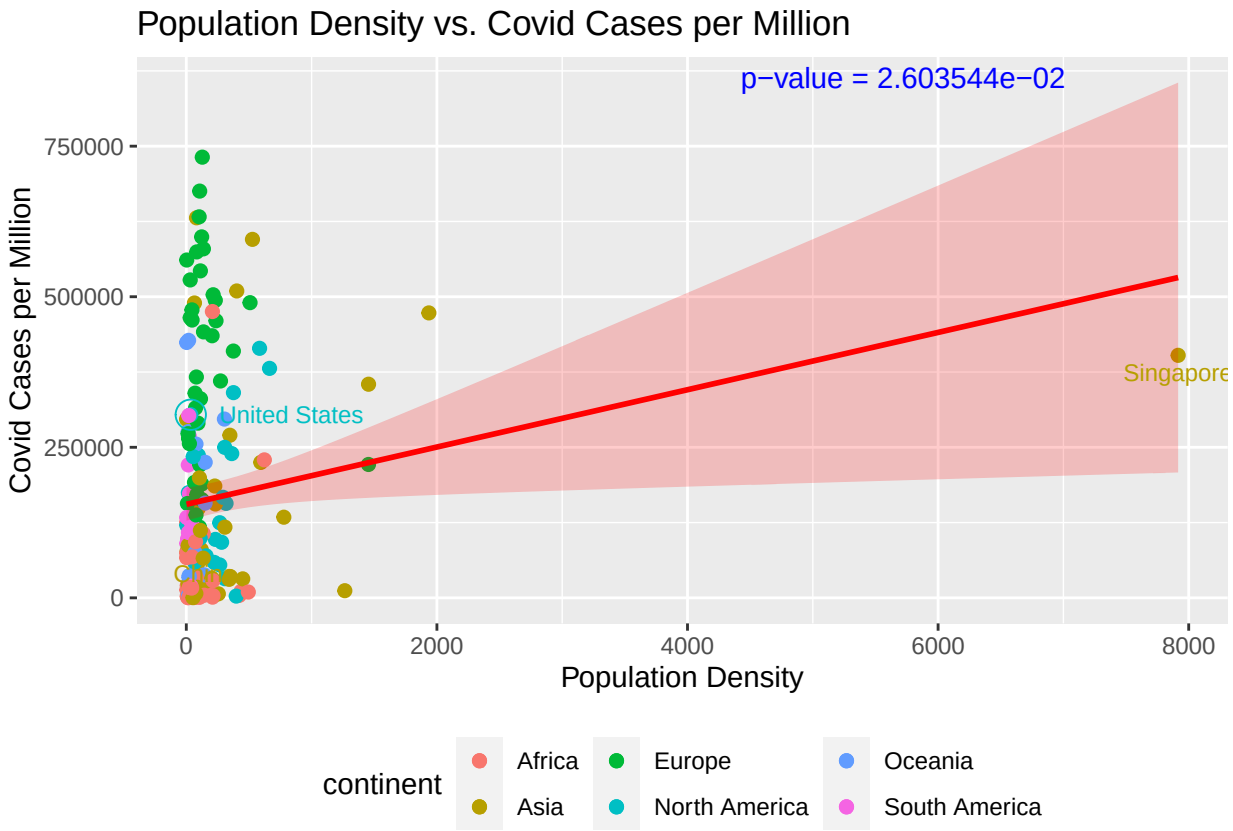
Correlation of Population Density to Cases per Million

Covid is spread through droplets and social distancing is recommended to prevent the spread of the virus. So it is interesting to see how well population density correlates to the total number of cases per million. The p-value of about 0.026 and the upward sloping trendline suggests that the population density and the number of Covid cases per million have a positive correlation. One interesting data point here is Singapore. Despite a very high population density, it still managed to do well in containing the spread of the virus.

```
# Data from all countries
country_data <- df %>%
  group_by(location) %>%
  slice(which.max(total_cases)) %>%
  drop_na()

# Remove data if total_cases is 0
country_data <- country_data %>% filter(total_cases != 0)
us_data <- filter(country_data, location == "United States")
singapore <- filter(country_data, location == "Singapore")
china <- filter(country_data, location == "China")
brazil <- filter(country_data, location == "Brazil")
india <- filter(country_data, location == "India")
france <- filter(country_data, location == "France")
russia <- filter(country_data, location == "Russia")

# Correlation of total_cases to population_density
df_case_pdensity <- country_data %>% filter(population_density != 0)
```

```
model1 <- lm(total_cases_per_million ~ population_density, data=df_case_pdensity)
p_value1 <- summary(model1)$coefficients[2,4]
plot5 <- ggplot(df_case_pdensity, aes(x=population_density, y=total_cases_per_million, color=continent))
       geom_point(size=2) +
       geom_smooth(method = "lm", se = TRUE, color = "red", fill = "red", alpha = 0.2) +
       geom_point(data = us_data, aes(color = "North America"), size = 5, shape=1) +
       geom_text(data = us_data, aes(label = location), hjust = -0.2, vjust = 0.5, size = 3) +
       geom_text(data = singapore, aes(label = location), hjust = 0.5, vjust = 1.5, size = 3) +
       geom_text(data = china, aes(label = location), hjust = 0.5, vjust = 1.5, size = 3) +
       guides(color = guide_legend(override.aes = list(size = 2))) +
       theme(legend.position="bottom") +
       labs(x="Population Density", y="Covid Cases per Million", title="Population Density vs. Covid Cas
plot5 <- plot5 + annotate("text", x = Inf, y = Inf, hjust = 1.5, vjust = 1.5,
                          label = paste0("p-value = ", format(p_value1, scientific = TRUE)),
                         size = 4, color = "blue")
plot5
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

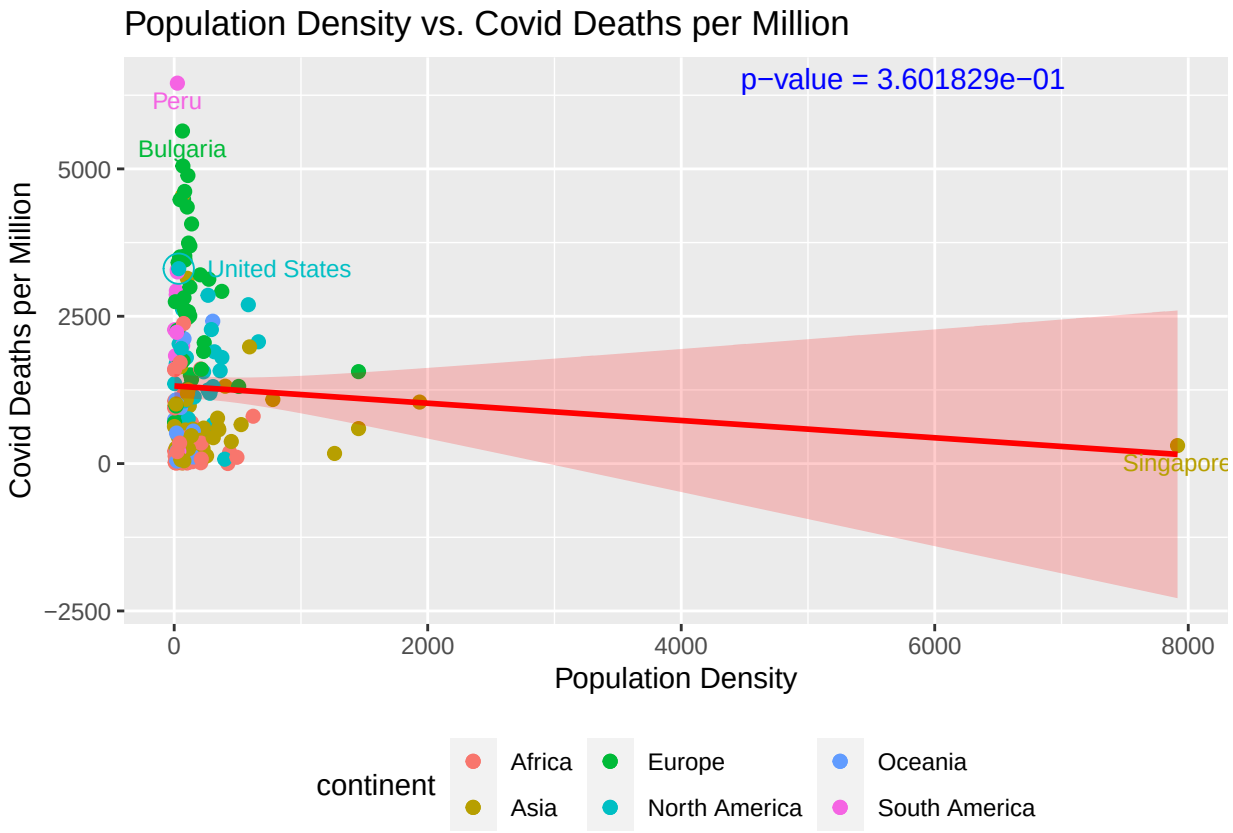


Correlation of Population Density to Deaths per Million

The p-value of 0.36 and the slightly downward sloping trendline indicates that unlike with cases, there isn't such a strong correlation or a positive correlation between population density and deaths per million. One theory for this would be that although lesser population density makes people less likely to get Covid, it might be more difficult to find treatment centers since they might be more spaced out. This would make you more likely to die if you actually do get Covid.

```r
# Data from all countries
country_data <- df %>%
  group_by(location) %>%
  slice(which.max(total_cases)) %>%
  drop_na()

# Remove data if total_cases is 0
country_data <- country_data %>% filter(total_cases != 0)
us_data <- filter(country_data, location == "United States")
singapore <- filter(country_data, location == "Singapore")
peru <- filter(country_data, location == "Peru")
bulgaria <- filter(country_data, location == "Bulgaria")

# Correlation of total_cases to population_density
df_case_pdensity <- country_data %>% filter(population_density != 0)
model1a <- lm(total_deaths_per_million ~ population_density, data=df_case_pdensity)
p_value1a <- summary(model1a)$coefficients[2,4]
plot5 <- ggplot(df_case_pdensity, aes(x=population_density, y=total_deaths_per_million, color=continent)
       geom_point(size=2) +
       geom_smooth(method = "lm", se = TRUE, color = "red", fill = "red", alpha = 0.2) +
       geom_point(data = us_data, aes(color = "North America"), size = 5, shape=1) +
       geom_text(data = us_data, aes(label = location), hjust = -0.2, vjust = 0.5, size = 3) +
       geom_text(data = singapore, aes(label = location), hjust = 0.5, vjust = 1.5, size = 3) +
       geom_text(data = peru, aes(label = location), hjust = 0.5, vjust = 1.5, size = 3) +
  geom_text(data = bulgaria, aes(label = location), hjust = 0.5, vjust = 1.5, size =
3) +
       guides(color = guide_legend(override.aes = list(size = 2))) +
       theme(legend.position="bottom") +
       labs(x="Population Density", y="Covid Deaths per Million", title="Population Density vs. Covid De
plot5 <- plot5 + annotate("text", x = Inf, y = Inf, hjust = 1.5, vjust = 1.5,
                          label = paste0("p-value = ", format(p_value1a, scientific = TRUE)),
                         size = 4, color = "blue")
plot5
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

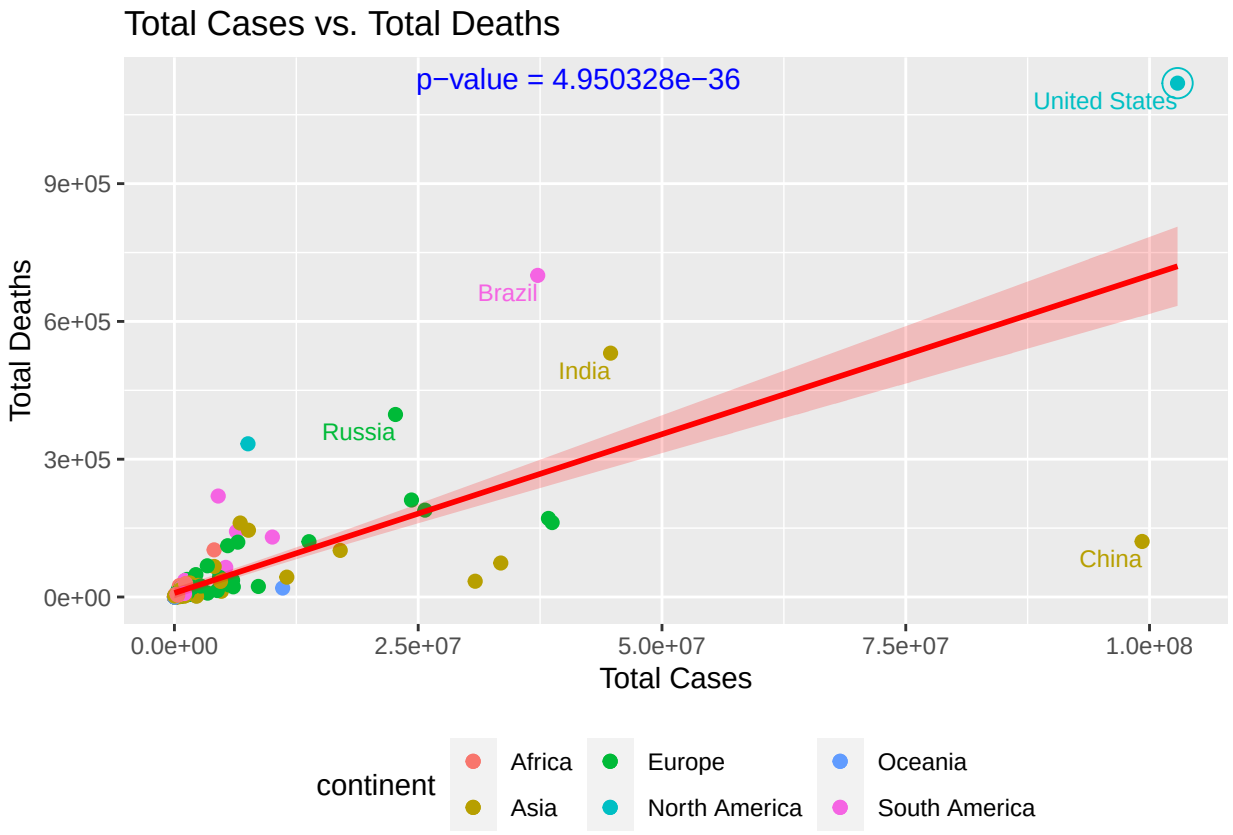


Correlation of Total Cases to Total Deaths

With a very small p-value and an upward sloping trend line, the total cases and deaths clearly have a positive correlation in the graph below. The interesting outliers are United States where the death rate is high compared to the number of cases and China where the death rate is very low to the number of cases. Some other interesting data points are labelled in the figure below.

```
# Correlation of total_cases to total_deaths
model2 <- lm(total_deaths ~ total_cases, data=country_data)
p_value2 <- summary(model2)$coefficients[2,4]
plot6 <- ggplot(country_data, aes(x=total_cases, y=total_deaths, color=continent))+
       geom_point(size=2) +
       geom_smooth(method = "lm", se = TRUE, color = "red", fill = "red", alpha = 0.2) +
       geom_point(data = us_data, aes(color = "North America"), size = 5, shape=1) +
       geom_text(data = us_data, aes(label = location), hjust = 1, vjust = 1.5, size = 3) +
       geom_text(data = china, aes(label = location), hjust = 1, vjust = 1.5, size = 3) +
       geom_text(data = brazil, aes(label = location), hjust = 1, vjust = 1.5, size = 3) +
       geom_text(data = india, aes(label = location), hjust = 1, vjust = 1.5, size = 3) +
       geom_text(data = russia, aes(label = location), hjust = 1, vjust = 1.5, size = 3) +
       guides(color = guide_legend(override.aes = list(size = 2))) +
       theme(legend.position="bottom") +
       labs(x="Total Cases", y="Total Deaths", title="Total Cases vs. Total Deaths")
plot6 <- plot6 + annotate("text", x = Inf, y = Inf, hjust = 2.5, vjust = 1.5,
                          label = paste0("p-value = ", format(p_value2, scientific = TRUE)),
                          size = 4, color = "blue")
plot6
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

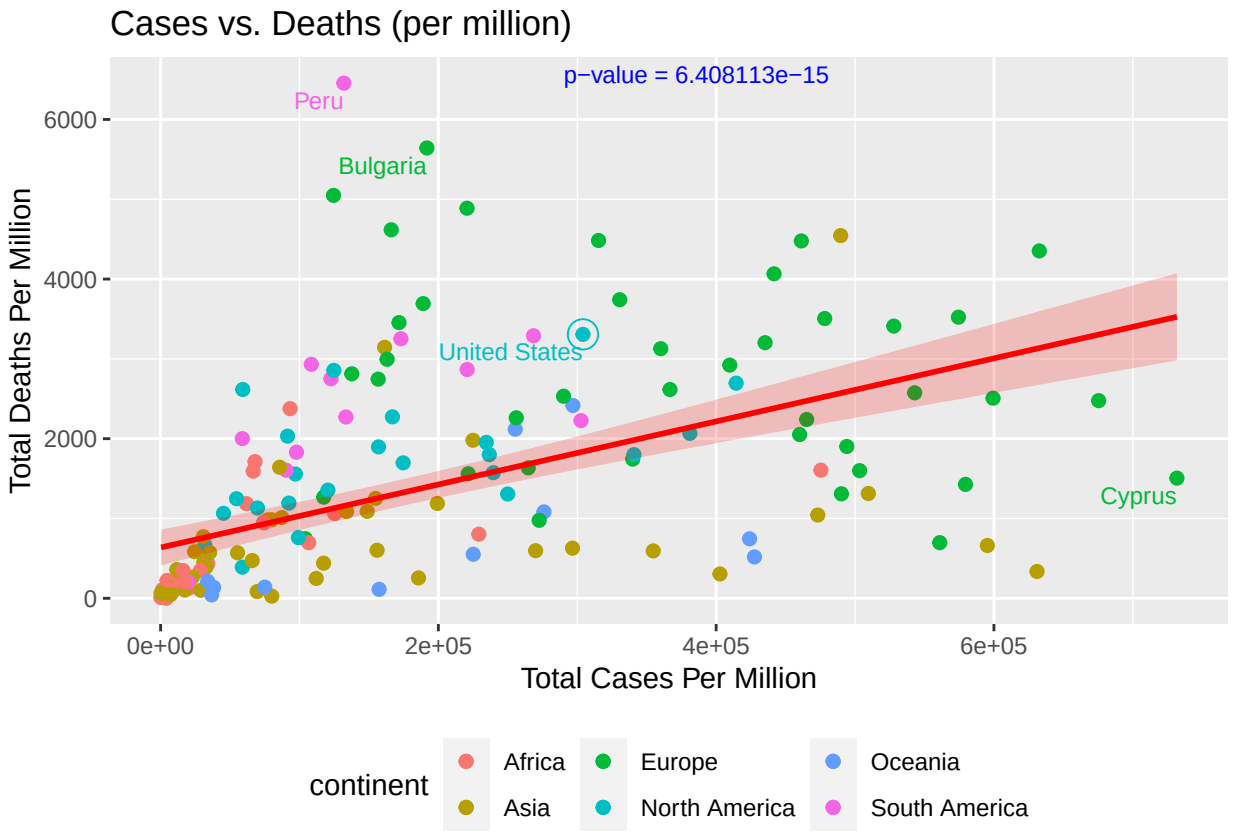


Correlation of Cases per Million to Deaths per Million

Cases per million and deaths per million also show a clear positive correlation. Some interesting data along with the data from United states is highlighted below.

```
bulgaria <- filter(country_data, location == "Bulgaria")
peru <- filter(country_data, location == "Peru")
cyprus <- filter(country_data, location == "Cyprus")
model3 <- lm(total_deaths_per_million ~ total_cases_per_million, data=country_data)
p_value3 <- summary(model3)$coefficients[2,4]
plot6a <- ggplot(country_data, aes(x=total_cases_per_million, y=total_deaths_per_million, color=continer
       geom_point(size=2) +
       geom_smooth(method = "lm", se = TRUE, color = "red", fill = "red", alpha = 0.2) +
       geom_point(data = us_data, aes(color = "North America"), size = 5, shape=1) +
       geom_text(data = us_data, aes(label = location), hjust = 1, vjust = 1.5, size = 3) +
       geom_text(data = cyprus, aes(label = location), hjust = 1, vjust = 1.5, size = 3) +
       geom_text(data = bulgaria, aes(label = location), hjust = 1, vjust = 1.5, size = 3) +
       geom_text(data = peru, aes(label = location), hjust = 1, vjust = 1.5, size = 3) +
       guides(color = guide_legend(override.aes = list(size = 2))) +
       theme(legend.position="bottom") +
       labs(x="Total Cases Per Million", y="Total Deaths Per Million", title="Cases vs. Deaths (per mil
plot6a <- plot6a + annotate("text", x = Inf, y = Inf, hjust = 2.5, vjust = 1.5,
                            label = paste0("p-value = ", format(p_value3, scientific = TRUE)),
                            size = 3, color = "blue")
plot6a
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

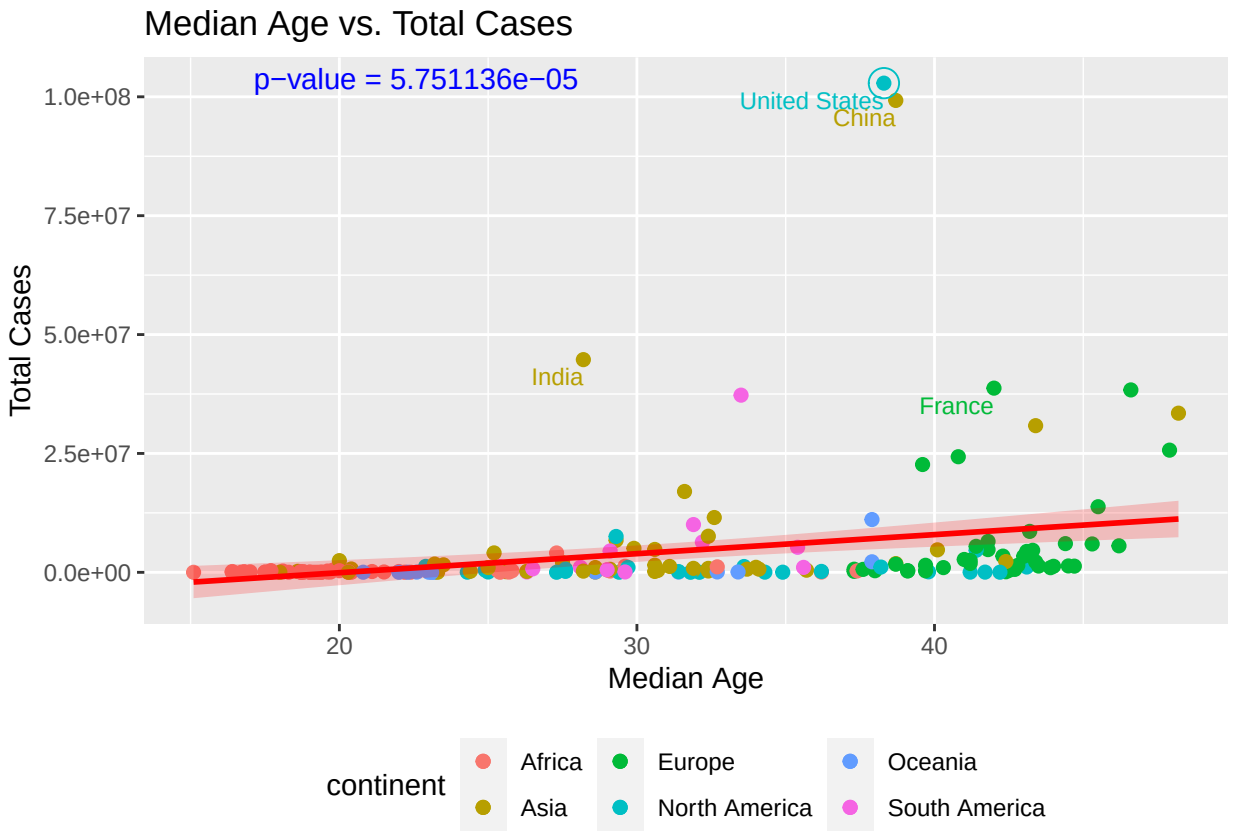


Correlation of Cases to Median Age

With a p-value of approximately 0.000057, the data suggests that the total cases recorded is correlated to the median age of the population in the country. The interesting outliers seems to be the Unites States and China where they population was disproportionately affected compared to the median age of the population.

```
# Correlation of total_cases to median_age
model4 <- lm(total_cases ~ median_age, data=country_data)
p_value4 <- summary(model4)$coefficients[2,4]
plot7 <- ggplot(country_data, aes(x=median_age, y=total_cases, color=continent))+
       geom_point(size=2) +
       geom_smooth(method = "lm", se = TRUE, color = "red", fill = "red", alpha = 0.2) +
       geom_point(data = us_data, aes(color = "North America"), size = 5, shape=1) +
       geom_text(data = us_data, aes(label = location), hjust = 1, vjust = 1.5, size = 3) +
       geom_text(data = china, aes(label = location), hjust = 1, vjust = 1.5, size = 3) +
       geom_text(data = india, aes(label = location), hjust = 1, vjust = 1.5, size = 3) +
       geom_text(data = france, aes(label = location), hjust = 1, vjust = 1.5, size = 3) +
       guides(color = guide_legend(override.aes = list(size = 2))) +
       theme(legend.position="bottom") +
       labs(x="Median Age", y="Total Cases", title="Median Age vs. Total Cases")
plot7 <- plot7 + annotate("text", x = Inf, y = Inf, hjust = 3, vjust = 1.5,
                          label = paste0("p-value = ", format(p_value4, scientific = TRUE)),
                          size = 4, color = "blue")
plot7
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

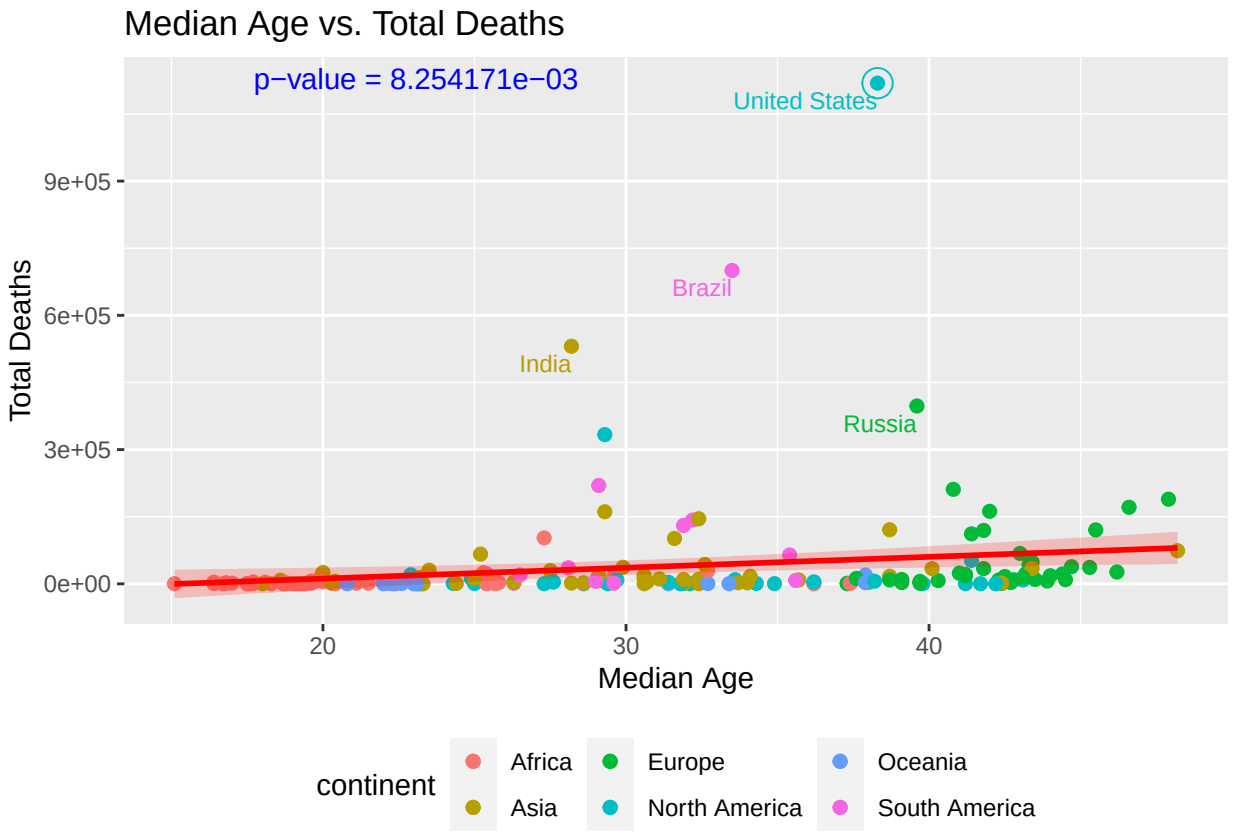


Correlation of Deaths to Median Age

It is known that Covid tends to have a more significant impact on older people, so I was curious to see if there was evidence of this in the data. With a p-value of 0.0082, this data suggests that there is a correlation between median age and total deaths. Again the United States is an interesting outlier here. Other interesting countries like Brazil and India are also noted.

```
# Correlation of total_deaths to median_age
model5 <- lm(total_deaths ~ median_age, data=country_data)
p_value5 <- summary(model5)$coefficients[2,4]
plot8 <- ggplot(country_data, aes(x=median_age, y=total_deaths, color=continent))+
       geom_point(size=2) +
       geom_smooth(method = "lm", se = TRUE, color = "red", fill = "red", alpha = 0.2) +
       geom_point(data = us_data, aes(color = "North America"), size = 5, shape=1) +
       geom_text(data = us_data, aes(label = location), hjust = 1, vjust = 1.5, size = 3) +
       geom_text(data = brazil, aes(label = location), hjust = 1, vjust = 1.5, size = 3) +
       geom_text(data = india, aes(label = location), hjust = 1, vjust = 1.5, size = 3) +
       geom_text(data = russia, aes(label = location), hjust = 1, vjust = 1.5, size = 3) +
       guides(color = guide_legend(override.aes = list(size = 2))) +
       theme(legend.position="bottom") +
       labs(x="Median Age", y="Total Deaths", title="Median Age vs. Total Deaths")
plot8 <- plot8 + annotate("text", x = Inf, y = Inf, hjust = 3, vjust = 1.5,
                          label = paste0("p-value = ", format(p_value5, scientific = TRUE)),
                          size = 4, color = "blue")
plot8
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Section 5. Conclusion

An attempt has been made here to visualize the impact that Covid has had in various parts of the world. The data we collected shows that there is a correlation between things such as cases, deaths, population density, and median age. However, it should be noted that there are confounding variables such as vaccination and masking policies, shutdowns, and healthcare quality that weren't a part of this data set. Furthermore, the collected data is based on the idea that the information about deaths and cases in these countries are accurate. It is possible that some countries did a worse job about collecting Covid data than others.

References

Mathieu, Edouard, Hannah Ritchie, Lucas Rod'e9s-Guirao, Cameron Appel, Charlie Giattino, Joe Hasell, Bobbie Macdonald, et al. 2020. "Coronavirus Pandemic (COVID-19)." *Our World in Data*. https://ourworldindata.org/coronavirus.