Exploring 2020 across many variables including Covid-19 By: Suren Bhakta and Johnny Henriquez

2022-10-19

Introduction ~ Suren

2020 was most definitely a step into the unknown for the whole world as Covid-19 ravaged through countries for multiple years. However, 2020 was an especially significant year as many families were destroyed, businesses lost, and lives being brought to a screeching halt. This made 2020 a year to forget, and our research group is no differed. Hence for our project, we want to go back to the year 2020 and look at trends that may have been overlooked at the time. Is GDP a factor into how many cases were present in each continent? Does the development of a country matter? Questions like these and more will be answered in our research. However, to answer these questions we must be able to join the datasets together with some key variable. Luckily, these datasets have a variable that is joinable, “Country.” Although each dataset may not have each country listed under the same variable name, they all share the same value, thus we can rename the variable and join accordingly.

Sources ~ Suren

For our research, we wanted to include many different metrics that may be indicators for Covid-19. Hence, we found data sets that had different variables from one another. For our Population dataset, we used a dataset from “https://www.worldometers.info/world-population/population-by-country/ (https://www.worldometers.info/world-population/population-by-country/)”. Within this dataset, there were 235 entries with a unique observation displaying the country, population, yearly change, net change, and many other factors relating to population. This dataset holds both numerical and categorical variables. For our world GPD dataset, this was found from “https://ourworldindata.org/grapher/gross-domestic-product (https://ourworldindata.org/grapher/gross-domestic-product)” which has 10,457 entries with a unique observation in the dataset displaying the country, GPD, year of observation, and country code. This dataset has exclusively numerical variables besides country. For our Covid-19 dataset, this was found on https://github.com/owid/covid-19-data/blob/master/public/data/README.md (https://github.com/owid/covid-19-data/blob/master/public/data/README.md) which had 226113 observations. The reason for the extreme amount of observations is due to there being an observation for each country for each day. A unique observation from this data set will include iso code, continent, location, date, and many more variables. This dataset holds both numerical and categorical variables.

Expected Trends ~ Suren

After joining the data set, there are trends we expect to see immediately. The first is that for countries with larger populations, there will be a larger amount of Covid cases. Another expected trend is that the death to GDP ratio will be much larger in smaller countries. This is because countries with a smaller GDP indicate that they do not have the infrastructure to fight the virus.

Tidying ~ Johnny

Taking a look at our datasets sourced above, we can see that the data was tidy for our population and gdp datasets. However, the Covid-19 data set was not tidy as we believed there to be variable names for the tests_units variable within a column. For elaboration, tests_units before tidying is a categorical variable which represents how countries were reporting their testing.

```
# "head" used to view top six rows of each dataset to determine if tidying is necessary
head(gross_domestic_product)
```

```
## # A tibble: 6 × 4
##   Entity      Code   Year `GDP (constant 2015 US$)`
##   <chr>       <chr> <dbl>                     <dbl>
## 1 Afghanistan AFG    2002                7228792320
## 2 Afghanistan AFG    2003                7867259392
## 3 Afghanistan AFG    2004                7978511360
## 4 Afghanistan AFG    2005                8874475520
## 5 Afghanistan AFG    2006                9349916672
## 6 Afghanistan AFG    2007               10642666496
```

```
head(pop)
```

```
## # A tibble: 6 × 12
##      no Countr…¹ Popul…² Yearl…³ Net C…⁴ Densi…⁵ Land …⁶ Migra…⁷ Fert.…⁸ Med. …⁹
##   <dbl> <chr>      <dbl> <chr>     <dbl>   <dbl>   <dbl>   <dbl> <chr>   <chr>
## 1     1 China     1.44e9 0.39%    5.54e6     153 9388211 -348399 1.7     38
## 2     2 India     1.38e9 0.99%    1.36e7     464 2973190 -532687 2.2     28
## 3     3 United …  3.31e8 0.59%    1.94e6      36 9147420  954806 1.8     38
## 4     4 Indones…  2.74e8 1.07%    2.90e6     151 1811570  -98955 2.3     30
## 5     5 Pakistan  2.21e8 2.00%    4.33e6     287  770880 -233379 3.6     23
## 6     6 Brazil    2.13e8 0.72%    1.51e6      25 8358140   21200 1.7     33
## # … with 2 more variables: `Urban Pop %` <chr>, `World Share` <chr>, and
## #   abbreviated variable names ¹`Country (or dependency)`, ²`Population 2020`,
## #   ³`Yearly Change`, ⁴`Net Change`, ⁵`Density  (P/Km²)`, ⁶`Land Area (Km²)`,
## #   ⁷`Migrants (net)`, ⁸`Fert. Rate`, ⁹`Med. Age`
```

```
head(owid_covid_data)
```

```
## # A tibble: 6 × 67
##   iso_code continent location date       total…¹ new_c…² new_c…³ total…⁴ new_d…⁵
##   <chr>    <chr>     <chr>    <date>       <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
## 1 AFG      Asia      Afghani… 2020-02-24       5       5  NA          NA      NA
## 2 AFG      Asia      Afghani… 2020-02-25       5       0  NA          NA      NA
## 3 AFG      Asia      Afghani… 2020-02-26       5       0  NA          NA      NA
## 4 AFG      Asia      Afghani… 2020-02-27       5       0  NA          NA      NA
## 5 AFG      Asia      Afghani… 2020-02-28       5       0  NA          NA      NA
## 6 AFG      Asia      Afghani… 2020-02-29       5       0   0.714      NA      NA
## # … with 58 more variables: new_deaths_smoothed <dbl>,
## #   total_cases_per_million <dbl>, new_cases_per_million <dbl>,
## #   new_cases_smoothed_per_million <dbl>, total_deaths_per_million <dbl>,
## #   new_deaths_per_million <dbl>, new_deaths_smoothed_per_million <dbl>,
## #   reproduction_rate <dbl>, icu_patients <dbl>,
## #   icu_patients_per_million <dbl>, hosp_patients <dbl>,
## #   hosp_patients_per_million <dbl>, weekly_icu_admissions <dbl>, …
```

```
# Since the owid_covid_data dataset was not tidy, will use pivot_wider to expand the tes
ts_units column into its own respective columns with values coming from the variable tes
ts_per_case
tidy_owid<-owid_covid_data%>%
  pivot_wider(names_from = tests_units,
              values_from = tests_per_case)
```

Joining/Merging ~ Johnny

We planned to join our various datasets based off country. However, referring to the tidying section above, it is clear to see that not all datasets are named specifically by country. Thus to address this issue, we will be using an assortment of dplyr functions to allow us to join the three datasets together.

Another issue is that our datasets as said above may have an extreme amount of observations due to observations being made each day for each country. Thus before joining, we will filter the data for the data just in the year 2020.

```
# Rename so country is a named variable as well as filter for only values from 2020
GDP<-gross_domestic_product%>%
  rename(GDP= "GDP (constant 2015 US$)")%>%
  rename(Country = "Entity")%>%
  filter(Year=="2020")
nrow(GDP) # Count observations
```

```
## [1] 210
```

```
# 2020 was already the only year for observations so only had to rename such that countr
y is a unique variable name
Population <- pop %>%
  rename(Country = "Country (or dependency)")
nrow(Population) # Count observations
```

```
## [1] 235
```

```
# Once again rename so that country is a unique variable name
# This dataset was a little more tricky to work for only 2020 values since there was an
 observation per country per day.
# So, we filtered for only data found at the end of 2020
Covid<-tidy_owid%>%
  rename(Country="location")%>%
  filter(date=="2020-12-31") # We filter by this data so that we get the total amount of
cases ammased by the end of the year 2020
nrow(Covid) # Count observations
```

```
## [1] 228
```

Now that we have addressed the issue extreme amounts of observations, we now are able to join our datasets. Luckily, the rename function within dplyr allowed us to rename all the columns within the datasets that held the country name with simply the word “Country.” This will now be our key variable as we prepare to join our dataset.

Also before joining, observe how the GDP dataset had 210 observations which had a unique variable of GDP, the Population set having 235 observations which had a unique variable of “Net Change” (measure of pop. change from prev. year to next), and the Covid dataset having 228 observations with a unique variable being “total_cases”. As said above, all countries now share the ID Country. As for ID’s that may be left out, there are countries in the population dataset such as the “Carribean Netherlands” which is not in the other datasets.

```
# Now we can join our data
# First join GDP dataset to population dataset with key variable Country
join1 <- Population %>%
  inner_join(GDP,by="Country")

# Now with our two previously joined datasets, join the Covid dataset with key variable
 country
join2 <- join1 %>%
  inner_join(Covid,by="Country")
nrow(join2)
```

```
## [1] 176
```

As seen with our code above, we joined our three completely different datasets by the key variable “Country.” We opted to use the inner join function to join our data to give us more control of our data. We felt that by using inner join, we would be able to manually select which variables we want to manipulate and which variables we want to remove instead of r joining only by key variable.

After joining our data with inner join, we were left with 176 observations, meaning that 59 rows were dropped. Although at face vaule this does not seem like a problem, this may be an issue later on. Our suspicion is that some of the smaller countries or countries with two word names were dropped from the dataset. This will cause skewing of data that is largely unavoidable.

Wrangling ~ Suren

Now with our joined dataset, it is time to find relations between our specific variables using dplyr functions.

Select

Some variables within our joined dataset serve no purpose now and are extra information. Although we did use inner join earlier, we stand by this decision as we can pick freely from all our variables which were useful. We will use select to only keep variables relevant to our study.

```
# Saving to a new data set
# Using dplyr function select which allows us to select columns by index number rather t
han by name which is far more efficient
new_data <- join2 %>%
  select(2,3,6,7,11,15,17,19,22,40,48,49,63,64,76,74,10,73)
```

filter, group_by, summarize, arrange

Now with our new_data dataset, we can use an assortment to generate results of our data.

A common idea among society is that the Asian continent had the largest amount of cases. This may be because of the virus being said to have originated from China or other ideas stemming from media and society. To see if this is true, we will find the sum of all cases by continents in the joined dataset.

```
new_data%>%
  filter(!is.na(total_cases))%>% # Remove NA values so calculations can be done
  group_by(continent)%>% # Without group_by, we would be given the sum of all cases from
the dataset
  summarize(total_cases_by_continent=sum(total_cases))%>% # Sums total cases
  arrange(desc(total_cases_by_continent)) # Arrange function with descending parameter d
isplays data from greatest to least
```

```
## # A tibble: 6 × 2
##   continent     total_cases_by_continent
##   <chr>                            <dbl>
## 1 North America                 23198329
## 2 Europe                        23059171
## 3 Asia                          20570107
## 4 South America                 13085660
## 5 Africa                         2702570
## 6 Oceania                          31440
```

After the use of the dplyr functions filter, group_by, summarize, and arrange, we found that Asia was neither first nor second, but in fact third in total cases behind Europe and North America with 20570107 people!

We also wanted to observe the mean population by continent along with the mean of total testing to create a ratio to determine of must testing was being performed for each test.

```
new_data%>%
  filter(!is.na(total_tests))%>%
  filter(!is.na(`Population 2020`))%>%
  group_by(continent)%>%
  summarize(mean_pop=mean(`Population 2020`),mean_tests=mean(total_tests), mean_gdp=mean
(GDP))%>%
  mutate(ratio_of_tests_pop=mean_tests/mean_pop)%>%
  arrange(desc(ratio_of_tests_pop))
```

```
## # A tibble: 6 × 5
##   continent      mean_pop mean_tests mean_gdp ratio_of_tests_pop
##   <chr>             <dbl>      <dbl>    <dbl>              <dbl>
## 1 Europe        17786913.   9210436.  4.67e11             0.518
## 2 North America 55652360.  28501224.  2.25e12             0.512
## 3 Oceania        7846834.   3192247.  4.26e11             0.407
## 4 South America 40123511.   5676987.  3.23e11             0.141
## 5 Asia          97527781.  11570517.  5.37e11             0.119
## 6 Africa        46088805.   1363024.  9.57e10             0.0296
```

After using more dplyr functions, we were able to find not only the mean population for each continent, mean gdp for each continent, but also the mean of tests per continent as well. With these two summarized variables, we were able to create a new variable named ratio_of_tests_pop which is a calculated ratio of the mean covid tests performed per mean population. It was found that Europe had the largest ratio of 0.5178 test per person with North America in a close second with a ratio of 0.5121 test per person.

mutate

A common indicator of overall country development is the human development index which is a statistic of life expectancy, education, and per capita income, which is all used to grade human development. Since we do not have a grade for overall country development, we will use this index to create a categorical variable named country development which will be split into four different categories of development.

```
# To create the variable, use the mutate function with nested ifelse statements to pass
 through the human index variable
# If a condition is not met, send to the next condition is met
new_data<-new_data%>%
  mutate(country_development = ifelse(human_development_index<.45, "Under Developed",
         ifelse(human_development_index<.55, "Average Development",
         ifelse(human_development_index<.75, "Slightly Developed",
         ifelse(human_development_index<.85, "Above Avg. Development", "Extremely Develo
ped")))))

new_data%>%
  select(country_development,Country)
```

```
## # A tibble: 176 × 2
##    country_development    Country      
##    <chr>                  <chr>        
##  1 Above Avg. Development China        
##  2 Slightly Developed     India        
##  3 Extremely Developed    United States
##  4 Slightly Developed     Indonesia    
##  5 Slightly Developed     Pakistan     
##  6 Above Avg. Development Brazil       
##  7 Average Development    Nigeria      
##  8 Slightly Developed     Bangladesh   
##  9 Above Avg. Development Russia       
## 10 Above Avg. Development Mexico       
## # … with 166 more rows
```

We can now see the country development for all countries in our joined dataset

One last statistic we want to calculate is the median age per continent.

```
new_data%>%
  filter(`Med. Age` != "N.A.")%>% # Filter NA values
  mutate(median_age=as.numeric(`Med. Age`))%>% # Transform variable into numeric
  group_by(continent)%>%
  summarize(mean_median=mean(median_age))%>% # Find mean per continent
  arrange(desc(mean_median)) # arrange in descending order
```

```
## # A tibble: 6 × 2
##   continent     mean_median
##   <chr>               <dbl>
## 1 Europe               41.9
## 2 North America        32.1
## 3 Asia                 31.3
## 4 South America        30.4
## 5 Oceania              27.5
## 6 Africa               21.2
```

After using an arrangement of dplyr functions, we can see that the continent with the highest median age is the continent of Europe with a average median age of 41.9 years.

summary stats

```
# summary function shows different descriptive stats for each variable
summary(new_data$`Population 2020`)
```

```
##       Min.    1st Qu.     Median       Mean    3rd Qu.       Max. 
##      39242    2400408    8976711   42668533   31533076 1439323776
```

```
summary(new_data$GDP)
```

```
##           Min.        1st Qu.         Median           Mean        3rd Qu. 
##      207699808    10780408320    40607787008   456842613493   247614771200 
##           Max. 
## 19294482071552 
```

```
summary(new_data$total_cases)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's 
##        1     6395    50948   474984   217153 20221642        2 
```

```
summary(as.factor(new_data$`Urban Pop %`))
```

```
##  13%  14%  17%  18%  19%  21%  23%  24%  25%  26%  27%  28%  29%  30%  31%  35% 
##    1    1    1    4    1    2    2    2    1    2    2    1    1    1    4    5 
##  36%  37%  38%  39%  41%  43%  44%  45%  46%  47%  48%  49%  50%  51%  52%  53% 
##    2    1    3    3    1    5    1    2    2    2    1    1    1    1    4    1 
##  54%  55%  56%  57%  58%  59%  60%  61%  62%  63%  64%  65%  66%  67%  68%  69% 
##    1    4    5    6    3    3    1    1    1    4    1    1    1    4    3    4 
##  70%  71%  72%  73%  74%  76%  78%  79%  80%  81%  82%  83%  84%  85%  86%  87% 
##    3    1    1    5    3    4    4    3    4    1    2    3    2    3    4    3 
##  88%  89%  91%  92%  93%  94%  95%  96%  97%  98% N.A. 
##    5    1    1    2    3    1    1    2    2    1    6 
```

```
summary(as.factor(new_data$country_development))
```

```
## Above Avg. Development    Average Development    Extremely Developed 
##                     44                     23                     42 
##     Slightly Developed        Under Developed                   NA's 
##                     55                      5                      7 
```

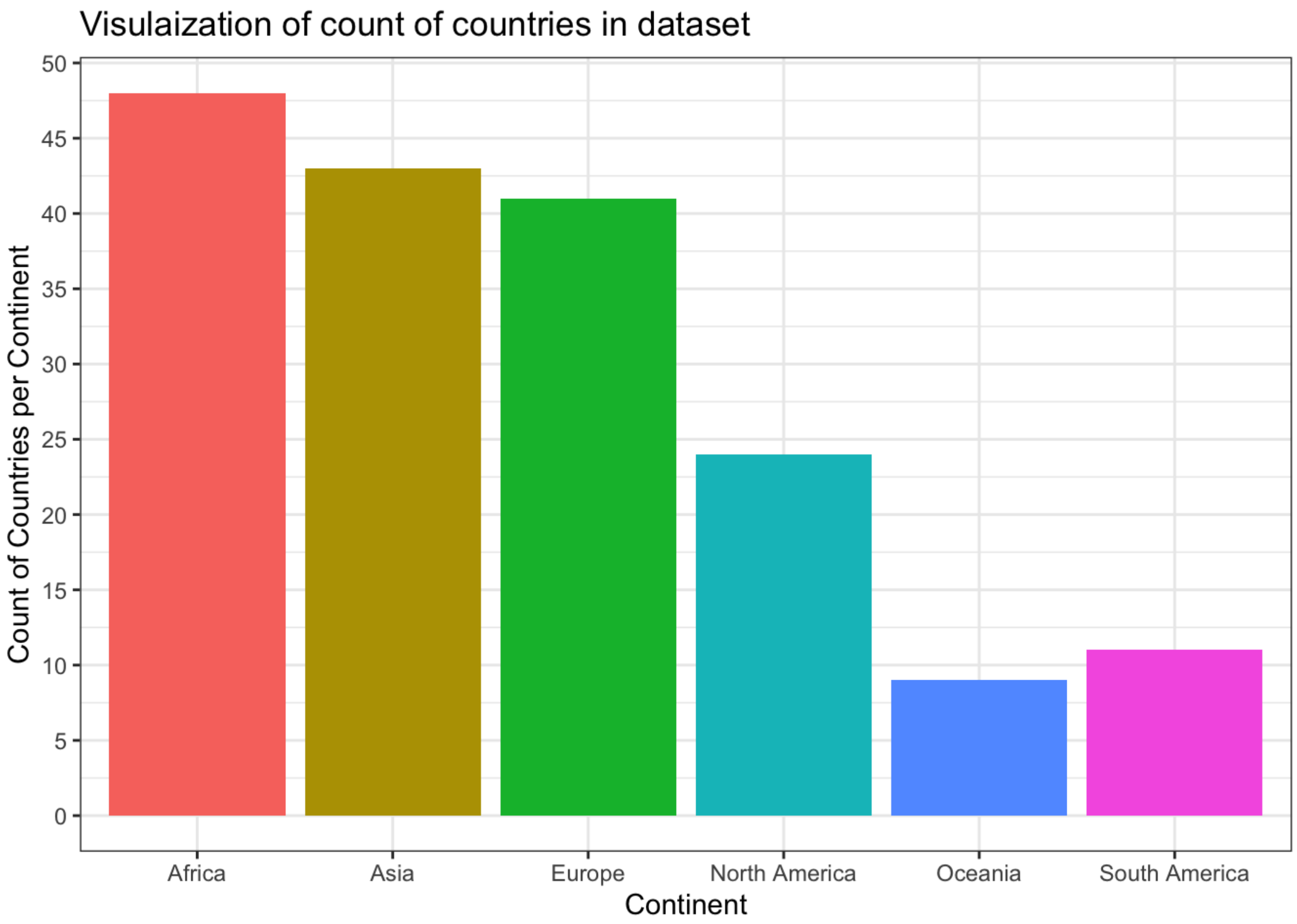
After reviewing all the summary stats for the numeric variables, we were able to see all the min, max, and mean values of the numeric variables and some of those results were very interesting. The results showed that there was a mean population of 42,668,533 people around all countries in 2020, the maximum GDP a country had was $19,294,482,071,552, and the minimum total cases a country had was only 1 case! The categorical variables were very interesting as well, with the highest frequency Urban Population percentage being 57%, and only 5 countries were under developed, while 42 countries are extremely developed. # Visualizing

Visualizing ~ Split equivalently by Suren and Johnny

single variable plot 1 of 2

For our first visualization, we wanted to see which continents were missing countries from our joined dataset. This could be for a number of reasons such as lacking a Key ID or being filtered out in one of the previous chunks.

```
# Use dplyr ggplot
new_data%>%
  group_by(continent)%>% # ensures the countries are summed based off continent
  ggplot(aes(x=continent,fill=continent))+
  geom_bar()+
  labs(x="Continent",y="Count of Countries per Continent",title="Visulaization of count of countries in dataset")+
  scale_y_continuous(breaks = seq(0, 50, 5))+ #Change scale to make counts viewable
  theme_bw()+
  theme(legend.position = "none")
```

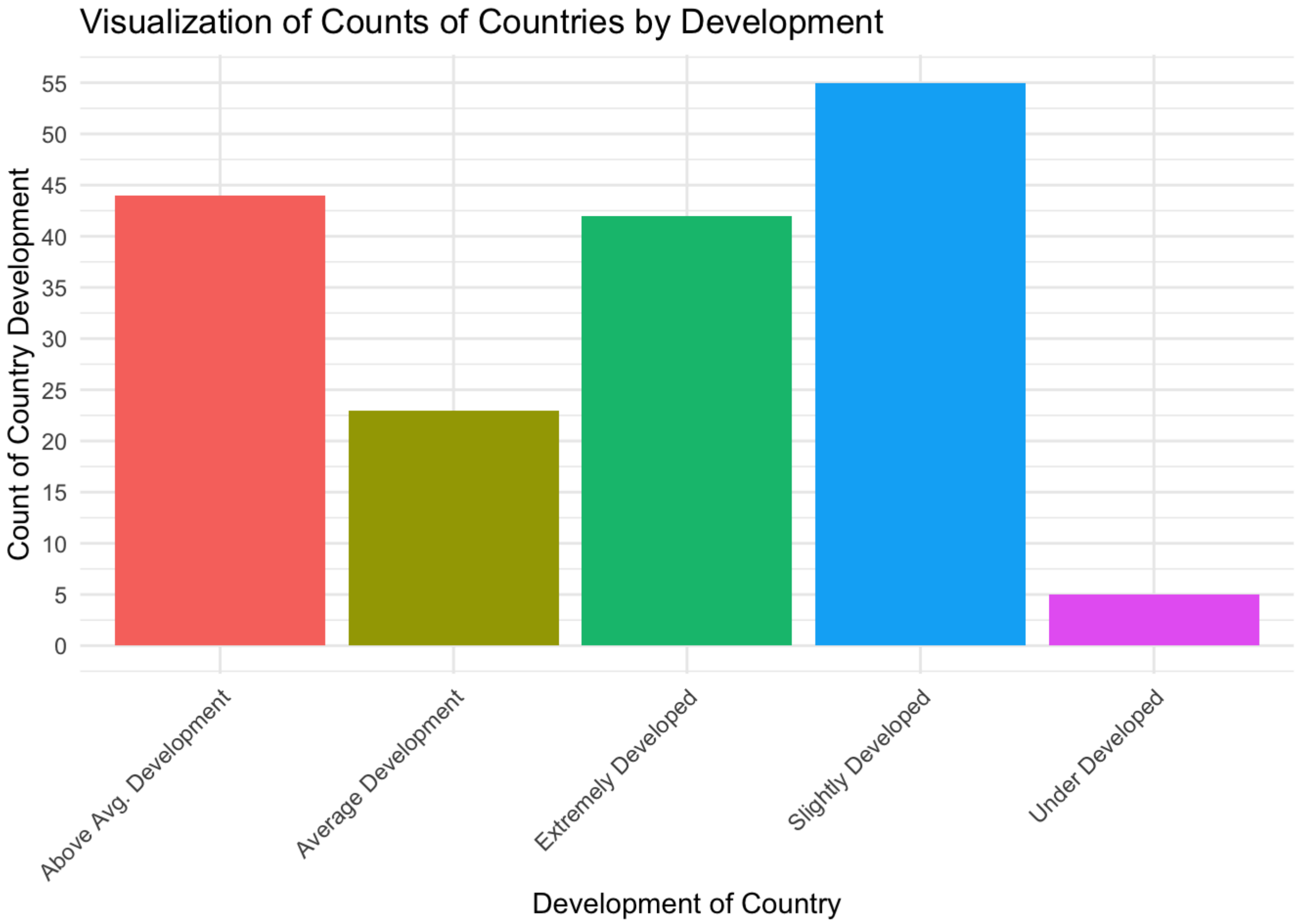


After viewing the visualizations, we are able to see the number of observations per country. However, with deeper research we were able to observe the total countries that existed in each continent in 2020. We can see that Africa is missing six countries from our data set. Asia is missing 5 countries, Europe is missing 3 countries, Oceania is missing 5, and South America is missing one. However, North America has 24 total observation which means there was an extra counted. This could be because U.S. territories were listed as a country in the data set. It is surprising that Oceania, the smallest continent by far, wastied for the most amount of countries missing with Asia.

single variable plot 2 of 2

For this visualization, we wanted to observe the counts of countries for each type of development which was a variable created in the “mutate” section of our report.

```
new_data%>%
  group_by(country_development)%>% # sums observations by type of development
  filter(`country_development` != "N.A.")%>% # do not want ot count NA values
  ggplot(aes(x=`country_development`))+
  geom_bar(aes(fill = country_development))+
  theme_minimal()+
  theme(axis.text.x = element_text(angle = 45, hjust = 1))+ # Make Axis legible
  theme(legend.position = "none")+ #remove legend
  scale_y_continuous(breaks = seq(0, 60, 5))+
  labs(x="Development of Country",y="Count of Country Development",title="Visualization
 of Counts of Countries by Development", colour="Development of Country")
```

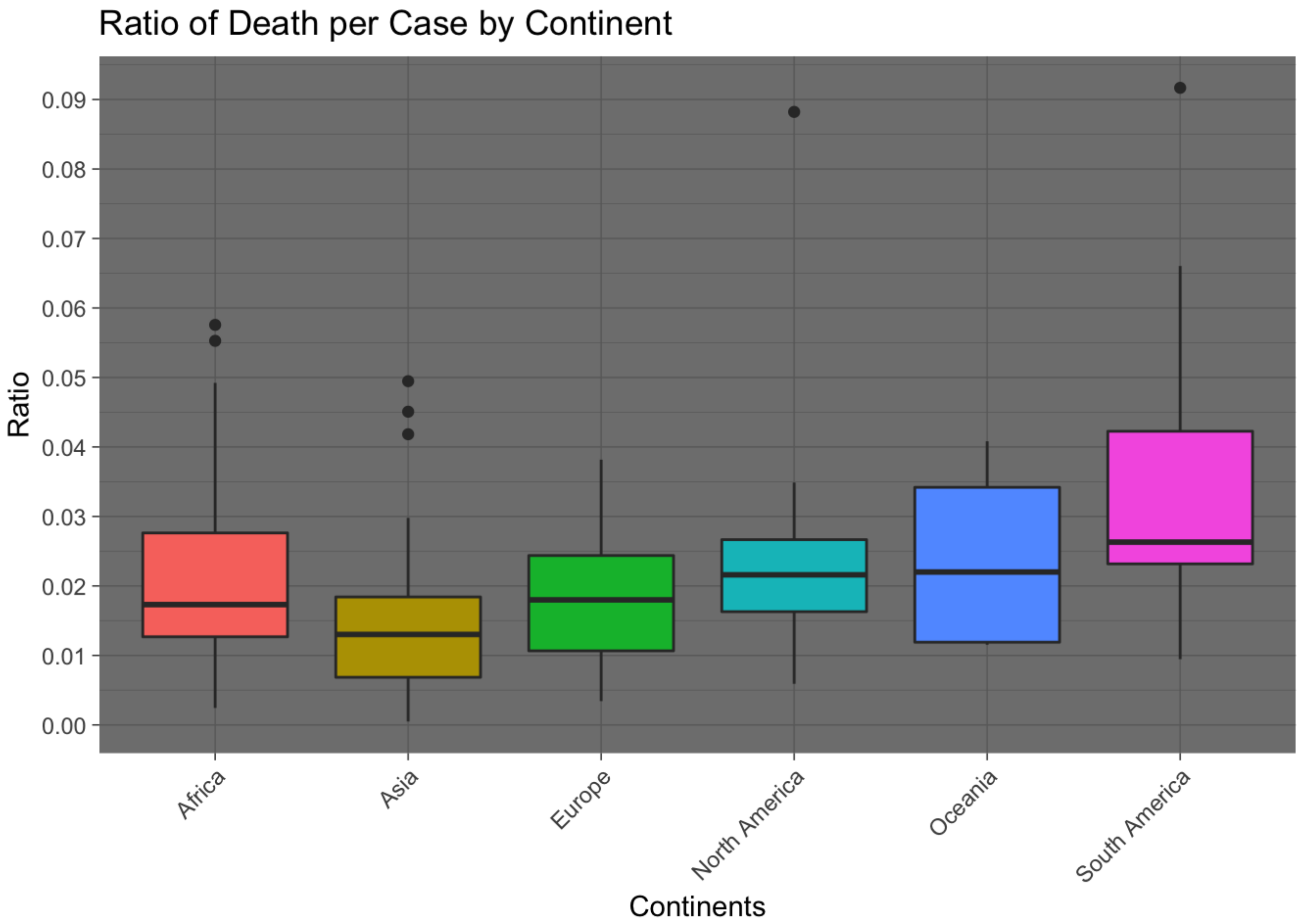


After counting the number of observations for each type of development in our joined dataset, it was found that among all the countries in our joined dataset, the mode of all countries were found to be “Slightly Developed” whereas only 5 countries were found to be “Under Developed” which is an incredible sign for the world. Having an increased development means that there will be better infrastructure to counterract the virus.

two variable plot 1 of 2

For our first bivariate representation, we wanted to have an in depth analysis of which continents were able to treat their covid cases. Since we do not have such a variable, we will create one with the dplyr function mutate. we will call this function death_to_case_ratio which is a measure of deaths per case. This will be colored by continent.

```
new_data%>%
  filter(!is.na(total_cases))%>% #Remove NA values
  filter(!is.na(total_deaths))%>%
  mutate(death_to_case_ratio = total_deaths/total_cases)%>% # Create the ratio variable
  group_by(continent)%>% # Summarize by continent
  ggplot(aes(x=continent, y=death_to_case_ratio , fill=continent))+
  geom_boxplot() +
  theme_dark()+
  theme(axis.text.x = element_text(angle = 45, hjust = 1))+ # angle adjustment for legib
ility
  scale_y_continuous(breaks = seq(0, .1, .01)) + # scale adjustment
  labs(x="Continents",y="Ratio",title="Ratio of Death per Case by Continent")+
  theme(legend.position = "none")
```



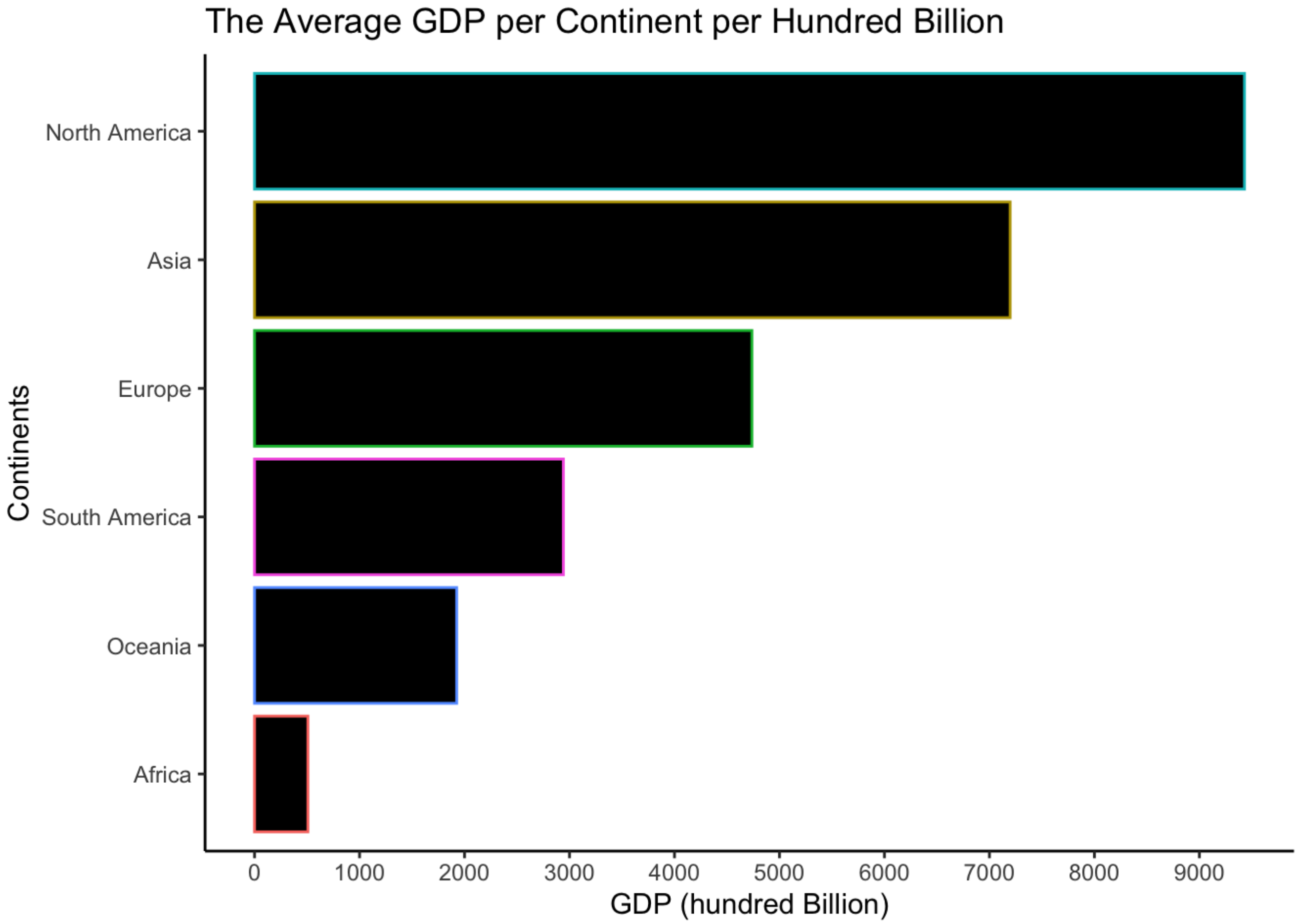
After creating our bivariate representation, there were many obvious conclusions. Overall, South America had the largest death per case ratio which implies the lack of ability to treat cases. Asia on the other hand did a much better job as a whole in treating their covid cases. The remaining three countries had near identical results but Africa had not only the largest distribution among the three, but of all the continents as well. This means that some parts of Africa treated cases moderately well, but some countries could not treat cases as effectively.

two variable plot 2 of 2

For our final bi variate graph, we wanted to explore where mean GDP was the largest among the six continents. However, the variable poses an issue as some values enter the trillions for some countries. Thus to counterract this, we will mutate a variable that will calculate GDP per hundred million.

```
new_data%>%
  group_by(continent)%>%
  mutate(gdp_per_hunmillion = GDP/100000000) %>% # creates variable
  select(gdp_per_hunmillion) %>%
  summarise(gdp_per_hunmillion = mean(gdp_per_hunmillion)) %>% # find mean of gdp
  ggplot(aes(x=gdp_per_hunmillion,y=reorder(continent, +gdp_per_hunmillion),color=contin
ent))+
  geom_bar(stat = "identity", fill = "black") + # Use of stat function
  theme_classic() +
  scale_x_continuous(breaks = seq(0, 10000, 1000)) + # Makes scale more legible
  theme(legend.position = "none") + #removes legend
  labs(x="GDP (hundred Billion)",y="Continents",title="The Average GDP per Continent per
Hundred Billion")
```

```
## Adding missing grouping variables: `continent`
```



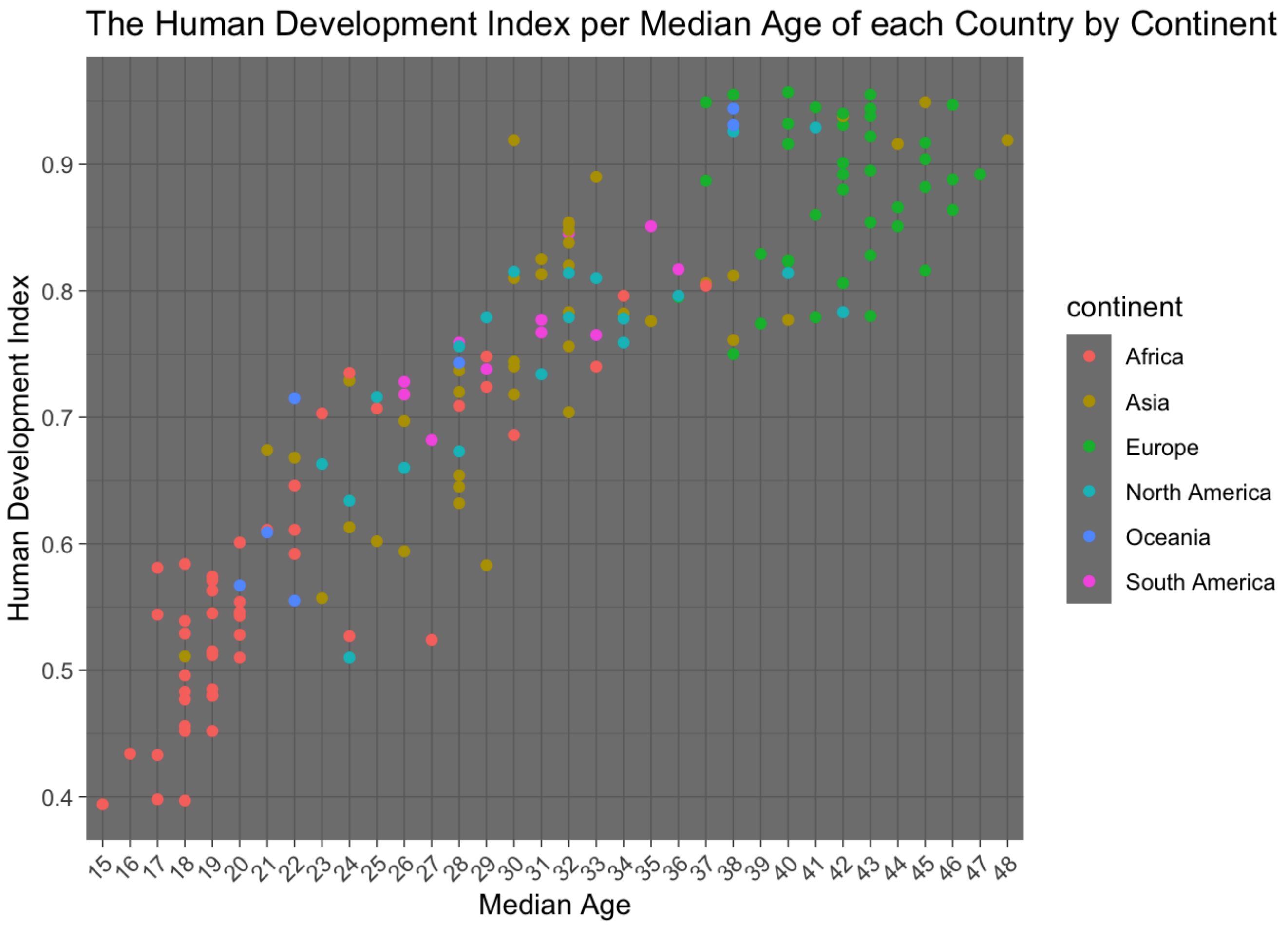
After creating our bivariate graph, it was found that North America had the largest mean GDP. Before creating our bar graph, we had thought that Oceania would have the least mean GDP just because of the lack of countries. However, that was not the case since Africa's mean GDP was more than 3 times smaller than that of Oceania's.

three variable plot 1 of 2

For our first three variable visualization, we wanted to investigate the influence median age had on the variables human development index and by continent. As said earlier, human development index is the measure ofof life expectancy, education, and per capita income in a single metric. With this, the hope is to visualize some sort of connection between the three variables which could be an indicator of covid 19 presence within continents.

```
new_data%>%
  filter(`Med. Age` != "N.A.")%>% # taking out "N.A." values from `Med. Age`
  ggplot(aes(x=as.factor(`Med. Age`),y=human_development_index,color=continent))+
  geom_point() +
  theme_dark()+ # changes theme
  theme(axis.text.x = element_text(angle = 45, hjust = 1))+ #rotates x scale numbers
  scale_y_continuous(breaks = seq(0, 1, .1))+ # rescales the y axis
  labs(x = "Median Age", y = "Human Development Index",
       title = "The Human Development Index per Median Age of each Country by Continent"
) #labels
```

```
## Warning: Removed 4 rows containing missing values (geom_point).
```



After creating a trivariable representation which maps human development index by median age per country by continent, we interestingly found a positive relationship between median age and the human development index. In contextual terms, this means that when median age increases, the development index increases as well. THis makes sense because if a country were to have a higher median age, that means that the country has a much more sustainable health care system which keeps the elderly alive and well. Since this was a trivariable graph, we were able to see where continents ranked among other continents based off the two variables. Based off the graph, one can see that the majority of countries within the African continent had th elowest median ages of all. Although more subtle, at the height of the median age range, that is where the majority of European countries are found.

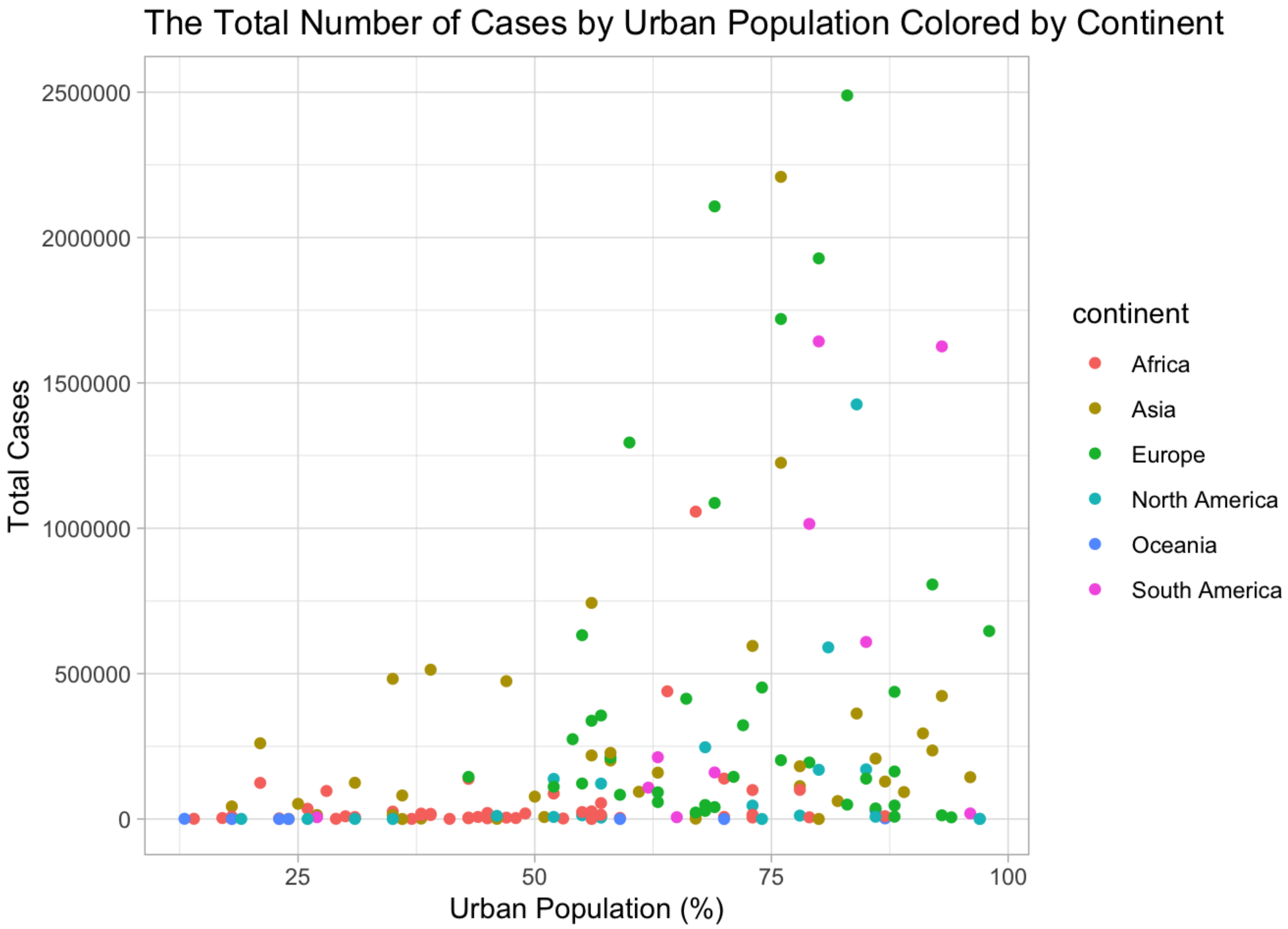
three variable plot 2 of 2

For our final three variable plot, we wanted to investigate if there was a positive trend between urban population and total number of cases. Our prediction is that as the urban percentage increases, as will the total number of cases since that would imply that the population is much more compact into sertain areas. However, sice urban population is a categorical variable, we must do some manipulation to be able to plot the graph accurately.

```
new_data %>%
  mutate(urban_pop = as.numeric(str_remove_all(new_data$`Urban Pop %`, "[%]"))) %>% # Re
moves the percent sign from all observations
  select(continent, total_cases, urban_pop) %>%
  filter(!is.na(total_cases), !is.na(urban_pop)) %>% # remove NA values
  ggplot(aes(x = urban_pop, y = total_cases, color = continent)) +
  geom_point() +
  ylim(0, 2500000) +
  theme_light() +
  labs(x = "Urban Population (%)", y = "Total Cases",
       title = "The Total Number of Cases by Urban Population Colored by Continent")
```

```
## Warning in mask$eval_all_mutate(quo): NAs introduced by coercion
```

```
## Warning: Removed 5 rows containing missing values (geom_point).
```



After creating our final trivariable representation mappingtotal cases for each percentage of urban population, interestngly enough, there was little to no correlation between the two variables. As Urban population did exceed 60%, the number of cases did increase as well. However, there was no constant correlation implying that the two variables were related.