Stat184_CourseProject

Zachary Smith

30 June 2025

Table of contents

```
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


Registered S3 method overwritten by 'mosaic':
  method                           from
  fortify.SpatialPolygonsDataFrame ggplot2
```

```
The 'mosaic' package masks several functions from core packages in order to add
additional features.  The original behavior of these functions should not be affected by thi


Attaching package: 'mosaic'


The following object is masked from 'package:Matrix':

    mean


The following object is masked from 'package:ggplot2':

    stat


The following objects are masked from 'package:dplyr':

    count, do, tally


The following objects are masked from 'package:stats':

    binom.test, cor, cor.test, cov, fivenum, IQR, median, prop.test,
    quantile, sd, t.test, var


The following objects are masked from 'package:base':

    max, mean, min, prod, range, sample, sum
```

Introduction

In this project, I will explore the relationship between population, GDP (Gross Domestic Product), abd Olympic success across countries. I will use the following question to guide my project:

1. What is the population distribtion across countries?
2. How do the top five most populatied countries compare in terms of GDP?
3. Is there a correlation between a countries population and the number of gold medals won in the Olympics?

Background Information

I will be using three separate data sets to answer the questions. I wanted to find 3 datasets that related to one another and can find meaningful data points.

Data Summary

My Primary data set contains a dataset with a column of country names and a second column with their respective population counts.

For my secondary data sets, I found a data set with a column of countries and their respective GDP and for the other data set, it has column of country names as well as the amount of gold/silver/bronze models they have each in their own column.

What is the population distribution across countries?

To demonstrate the population distribution across countries, I chose to do the top 10 countries in population so the data wouldn't be cluttered.

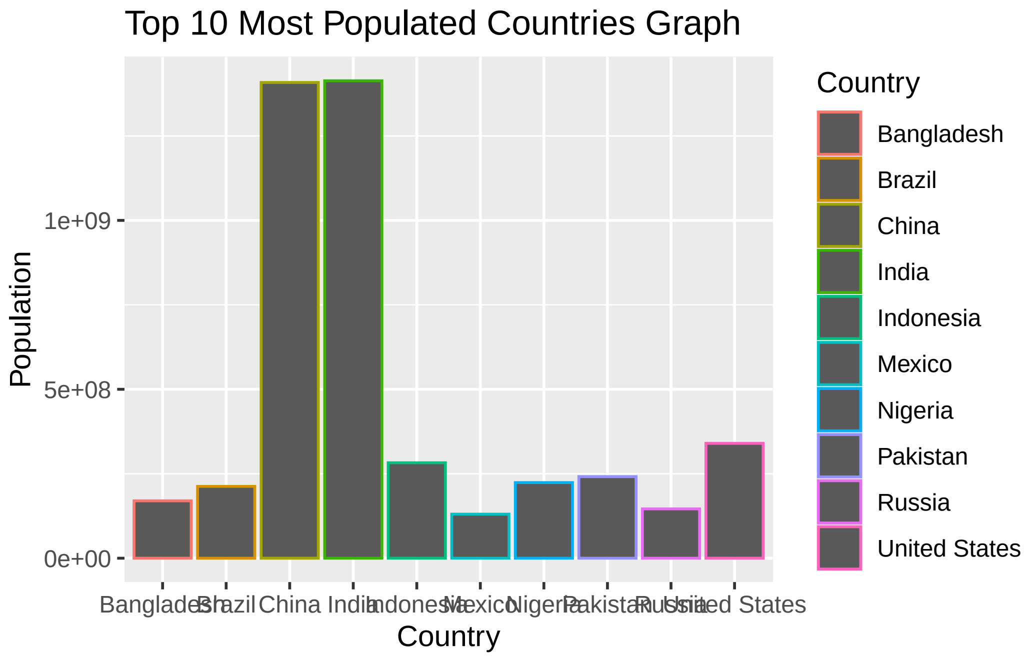

Figure 1: Top 10 Most Populated Countries Graph

As you can see by Figure 1, China and India are close in population but have more than triple the people of the third most populated country. The rest of the countries in Figure 1 are relatively close aside from Mexico and Russia being the lowest by roughly 100 million people compared to the third lowest which would be Bangladesh.

How do the top 5 most populated countries compare in GDP?

The GDP of a country is essentially the health of the countries economy, so the higher the GDP the wealthier the country is. I'm using the GDP data set and the population data set to find if there's any correlation between GDP and population. The Table shows which countries were chosen based on population.

Here's a chart and table of the top 10 most populated countries:

```
Warning: There was 1 warning in `mutate()`.
i In argument: `GDP = as.numeric(gsub(",", "", GDP))`.
Caused by warning:
! NAs introduced by coercion
```

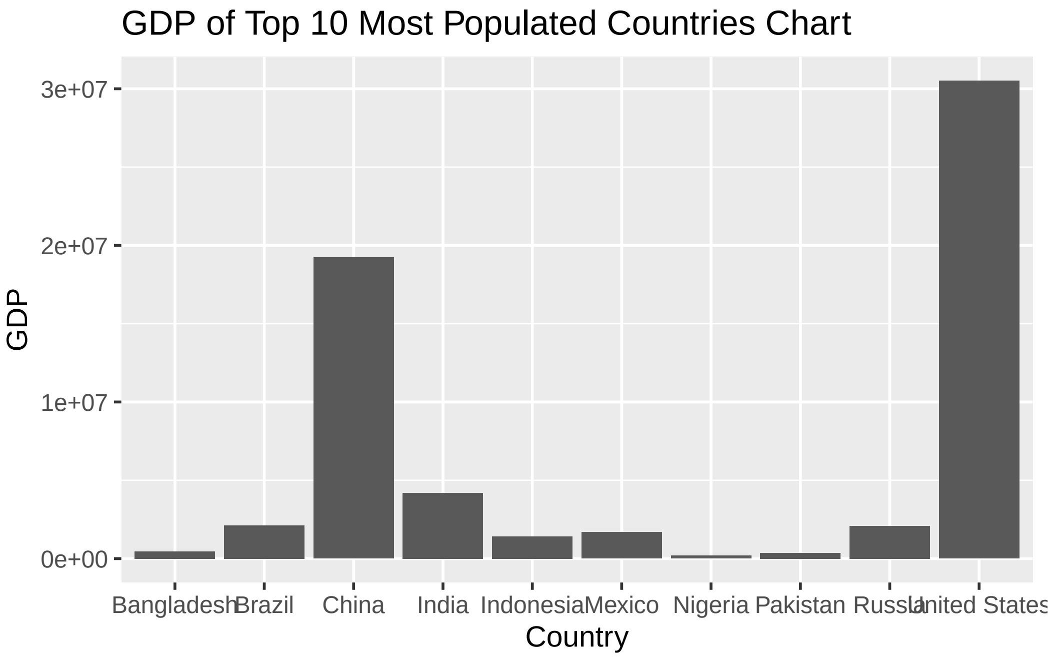

Figure 2: GDP of Top 10 Most Populated Countries Chart

The Table represents adds the population in addition to the GDP.

Table 1: GDP of Top 5 Most Populated Countries Table

Country	GDP	Population
India	4187017	1413324000
China	19231705	1408280000
United States	30507217	340110988
Indonesia	1429743	282477584
Pakistan	373078	241499431
Nigeria	188271	223800000
Brazil	2125958	212583750
Bangladesh	467218	169828911
Russia	2076396	146028325
Mexico	1692640	130417144

Based on both Figure 2 and Table 1, we can see that China and India being the highest in population by more than triple don't have the the largest GDP, but do have the second and third. The largest belonging to the United States.

Is there a correlation between population and the amount of gold medals won?

To find a correlation between population and the amount of gold medals. I needed to join the population and medal data set together. Like the other figures, I used a bar chart, since it was the easiest visual to read in terms of my data. The Table shows which countries were chosen based on population.

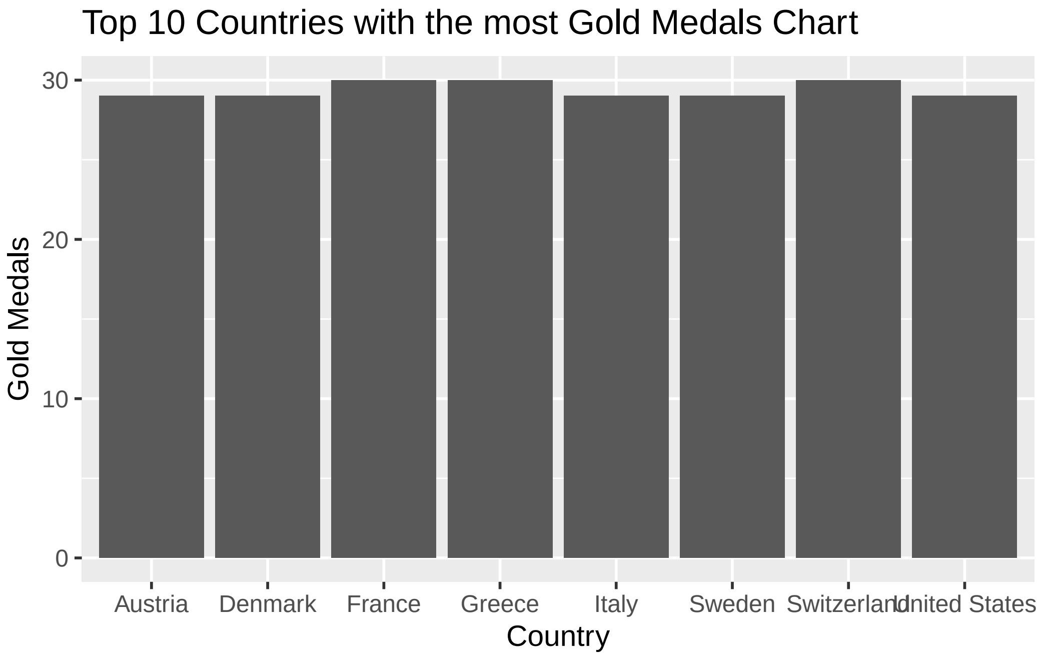

Figure 3: Top 10 Most Populated Countries Gold Medal Count Chart

The Table shows the population and amount of gold medals for each of the top ten countries with the most gold medals.

Table 2: Top 10 Countries with the most Gold Medals Table

Country	Gold Medals	Population
France	30	68633000
Greece	30	10400720
Switzerland	30	9067144
Austria	29	9202428
Denmark	29	6001008
Italy	29	58921111
Sweden	29	10588818
United States	29	340110988

Based on Figure 3 and Table 2, You can see how neither of the top 2 highest countries from Figure 1 appear in the list of the top 10 countries of the most gold medals. In fact, only one country, the United States from the top 5 most populated countries show in the chart. In fact, countries like Austria and Denmark which have the lowest population in Table 2 and Figure 3 are ten times lower in population than India and China. This effectively shows that there is no correlation between population and amount of gold medals.

Conclusion

Working with data was difficult and tidying the data was the hardest task. Showing how to visualize the data sets were also a challenge. After a lot of trial and error, I was able to conclude the research questions and accurately create data visuals that showed that China and India have the highest population by a large margin. United States had the largest GDP by a large margin and didn't even have the largest population count. As for the correlation between population size and gold medals, there is none.

Work Cited

Countries and Population dataset: https://en.wikipedia.org/wiki/List_of_countries_and_dependencies_by_population

Countries and GDP dataset: https://en.wikipedia.org/wiki/List_of_countries_by_GDP_(nominal)

Countries and Gold Medals dataset: https://en.wikipedia.org/wiki/All-time_Olympic_Games_medal_table

Code Appendix

```r
library(rvest)
library(tidyr)
library(dplyr)
library(mosaic)
library(ggplot2)
library(knitr)
URL_pop <- "https://en.wikipedia.org/wiki/List_of_countries_and_dependencies_by_population"
pop_list <- URL_pop %>%
  read_html() %>%
  html_nodes(css = "table") %>%
  html_table(fill = TRUE)

popData <- pop_list[[1]] %>%
  rename(
    Country = "Location"
  ) %>%
  mutate(Population = as.numeric(gsub(",", "", Population)))

clean_popData <- popData %>%
  filter(!grepl("World", Country, ignore.case = TRUE)) %>%
  arrange(desc(Population)) %>%
  head(10)

ggplot(data = clean_popData, aes(x = Country, y = Population)) + geom_bar(stat = "identity")
URL_gdp <- "https://en.wikipedia.org/wiki/List_of_countries_by_GDP_(nominal)"
gdp_list <- URL_gdp %>%
  read_html() %>%
  html_nodes(css = "table") %>%
  html_table(fill = TRUE)
clean_gdpData <- gdp_list[[3]] %>%
  select(1, 2) %>%
  rename(
    Country = 1,
    GDP = 2
  ) %>%
  filter(!Country %in% c("Country/Territory", "World")) %>%
  mutate(GDP = as.numeric(gsub(",", "", GDP)))

pop_gdp_combined <- inner_join(clean_popData, clean_gdpData, by = "Country") %>%
  select(Country, GDP, Population) %>%
  head(10)

ggplot(pop_gdp_combined, aes(x = Country, y = GDP)) +
  geom_bar(stat = "identity") + labs(title = "GDP of Top 10 Most Populated Countries Chart")
kable(pop_gdp_combined, caption = "GDP of Top 5 Most Populated Countries Table")
URL_medal <- "https://en.wikipedia.org/wiki/All-time_Olympic_Games_medal_table"
medal_list <- URL_medal %>%
  read_html() %>%
  html_nodes(css = "table") %>%
  html_table(fill = TRUE)

medalData <- medal_list[[2]] %>%
  select(1, 2) %>%
  rename(
    Country = 1,
    "Gold Medals" = 2
  ) %>%
  mutate(
    Country = gsub("\\[.*?\\]", "", Country),
    `Gold Medals` = as.numeric(gsub("[^0-9]", "", `Gold Medals`))
  ) %>%
  arrange(desc(`Gold Medals`)) %>%
  head(10)

pop_medal_combined <- inner_join(medalData, popData, by = "Country") %>%
  select(Country, "Gold Medals", Population) %>%
  arrange(desc(`Gold Medals`))


ggplot(pop_medal_combined, aes(x = Country, y = `Gold Medals`)) +
  geom_bar(stat = "identity") + labs(title = "Top 10 Countries with the most Gold Medals Cha
kable(pop_medal_combined, caption = "Top 10 Countries with the most Gold Medals Table")
# Population data
URL_pop <- "https://en.wikipedia.org/wiki/List_of_countries_and_dependencies_by_population"
pop_list <- URL_pop %>%
  read_html() %>%
  html_nodes(css = "table") %>%
  html_table(fill = TRUE)

popData <- pop_list[[1]] %>%
  rename(
    Country = "Location"
  ) %>%
  mutate(Population = as.numeric(gsub(",", "", Population)))

clean_popData <- popData %>%
  filter(!grepl("World", Country, ignore.case = TRUE)) %>%
  arrange(desc(Population)) %>%
  head(10)

ggplot(data = clean_popData, aes(x = Country, y = Population)) + geom_point(size = 4)  + aes

# GDP data
URL_gdp <- "https://en.wikipedia.org/wiki/List_of_countries_by_GDP_(nominal)"
gdp_list <- URL_gdp %>%
  read_html() %>%
  html_nodes(css = "table") %>%
  html_table(fill = TRUE)
clean_gdpData <- gdp_list[[3]] %>%
  select(1, 2) %>%
  rename(
    Country = 1,
    GDP = 2
  ) %>%
  filter(!Country %in% c("Country/Territory", "World")) %>%
  mutate(GDP = as.numeric(gsub(",", "", GDP)))

pop_gdp_combined <- inner_join(clean_popData, clean_gdpData, by = "Country") %>%
  select(Country, GDP, Population) %>%
  head(5)

kable(pop_gdp_combined, caption = "GDP of Top 10 Most Populated Countries Table")


ggplot(pop_gdp_combined, aes(x = Country, y = GDP)) +
  geom_bar(stat = "identity") + labs(title = "GDP of Top 10 Most Populated Countries Chart")

# Olympic medals data
# Olympic medals data
URL_medal <- "https://en.wikipedia.org/wiki/All-time_Olympic_Games_medal_table"
medal_list <- URL_medal %>%
  read_html() %>%
  html_nodes(css = "table") %>%
  html_table(fill = TRUE)
```

```
medalData <- medal_list[[2]] %>%
  select(1, 2) %>%
  rename(
    Country = 1,
    "Gold Medals" = 2
  ) %>%
  mutate(
    Country = gsub("\\[.*?\\]", "", Country),
    `Gold Medals` = as.numeric(gsub("[^0-9]", "", `Gold Medals`))
  ) %>%
  arrange(desc(`Gold Medals`)) %>%
  head(10)

pop_medal_combined <- inner_join(medalData, popData, by = "Country") %>%
  select(Country, "Gold Medals", Population) %>%
  arrange(desc(`Gold Medals`))

kable(pop_medal_combined, caption = "Top 10 Most Populated Countries Gold Medal Count Table")

ggplot(pop_medal_combined, aes(x = Country, y = `Gold Medals`)) +
  geom_bar(stat = "identity") + labs(title = "Top 10 Most Populated Countries Gold Medal Cou
```