Research Project 1

2023-10-19

Carlos Vazquez, Mayce Hoang, Mindy Li

Unveiling Global Narratives: An Exploratory Analysis on the Development of Countries

Introduction

In a world that is marked by constant evolution, data becomes the key to deciphering the intricate narratives that shape global development, health, and economic prosperity. Our exploratory data analysis gathers insights from three distinct datasets, each offering a unique perspective on the history of global growth. We selected the datasets World GDP, World Population, and World Life Expectancy at Birth. We sourced these datasets from the reputable platform kaggle.com, ensuring the reliability of the data. Our group finds these datasets inherently fascinating for a multitude of reasons. They encapsulate pivotal facets of real-world development, offering insight into the intricate interplay between GDP, population dynamics, and life expectancy. They enable us to examine the unique development trajectories of nations in the socio-economic landscape. Moreover, they have the potential to reveal trends and relationships that can guide future endeavors aimed at enhancing global growth.

Datasets

The World GDP dataset, spanning from 1960 to 2021, gives insights into the macroenvironment of a country. Each unique row represents the country name, the country code, the indicator name, the indicator code, and the corresponding GDP values for each year from 1960 to 2021. These GDP values are distributed across numerical variables such as '1960,' '1970,' '1980,' '1990,' and so on until '2021.' Additionally, the dataset includes categorical variables such as 'Country Name,' 'Country Code,' 'Indicator Name,' and 'Indicator Code.' These variables provide contextual information, allowing for a comprehensive understanding of the economic data for each country over the specified timeframe.

The World Population dataset provides a comprehensive view of population trends for each country and its indicators, spanning from 1970 to 2022. Each unique row in this dataset provides an array of essential information, including a country's population rank, its three-digit country code, the country name and its capital city, the continent to which it belongs, and the population statistics spanning from 1970 to 2022. This dataset also includes valuable insights such as the land area of each country in square kilometers, population density per square kilometer, population growth rates, and the country's contribution to the world's total population. It encompasses numerical variables like 'Rank,' '1970 Population' - '2022 Population,' 'Area (km^2),' 'Density (per km^2),' 'Growth Rate,' and 'World Population Percentage.' Additionally, it incorporates categorical variables, 'CCA3,' 'Country/Territories,' 'Capital,' and 'Continent,' providing essential contextual information for in-depth analysis.

The World Life Expectancy at Birth dataset provides a comprehensive historical record of life expectancy at birth and its indicators on a global scale. Each unique row represents the ISO3 for the country, the country name, the continent name, the name of the hemisphere, the level of Human Development Groups in each country, the UNDP Developing Regions, the Human Development Index Rank for each country in 2021, and the life expectancy at birth from 1990 to 2021. The dataset includes numerical variables 'HDI Rank (2021)' and 'Life Expectancy at Birth from 1990' - 'Life Expectancy at Birth 2021'. Categorical variables in this dataset are 'ISO3,' 'Country,' 'Continent,' 'Hemisphere,' 'Human Development Groups,' and 'UNDP Developing Regions'.

The original datasets can be seen below. There were 195 total observations for the World Life Expectancy at Birth dataset, 266 total observations for the World GDP dataset, and 234 total observations for World Population before tidying and joining.

```r
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
```

```r
library(tidyr)
library(readr)

#Read datasets
Life_Expectancy_at_Birth <- read_csv("~/Downloads/Life Expectancy at Birth.csv")
```

```
## Rows: 195 Columns: 39

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (6): ISO3, Country, Continent, Hemisphere, Human Development Groups, UN...
## dbl (33): HDI Rank (2021), Life Expectancy at Birth (1990), Life Expectancy ...
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```r
GDP_DATA_2 <- read_csv("~/Downloads/GDP DATA 2.csv",
    skip = 2)
```

```
## Rows: 266 Columns: 66
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (4): Country Name, Country Code, Indicator Name, Indicator Code
## num (62): 1960, 1961, 1962, 1963, 1964, 1965, 1966, 1967, 1968, 1969, 1970, ...
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```r
world_population <- read_csv("~/Downloads/world_population.csv")
```

```
## Rows: 234 Columns: 17
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (4): CCA3, Country/Territory, Capital, Continent
## dbl (13): Rank, 2022 Population, 2020 Population, 2015 Population, 2010 Popu...
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

Research Overview

Our key for merging these datasets is the ‘Country’ variable. After tidying the data, there was potential in merging the datasets on ‘Year’ as well, but for the purpose of our exploration, we decided to filter our merged dataset to focus solely on the year 2020. This selective approach enables a holistic cross-analysis of GDP, population, and life expectancy in a particular timeframe, offering a snapshot of global development. Within our exploratory data analysis, we anticipate several potential relationships and trends that could offer valuable insights into the global landscape. We expect a positive correlation between a country’s economic growth, as reflected in GDP, and life expectancy. Generally, higher GDP may indicate better access to healthcare, education, and overall improved living standards, which can contribute to longer life expectancies. We also expect there to be a relationship between a country’s population size and its GDP. Larger populations could result in higher GDP due to a larger labor force and consumer market. However, overpopulation can also strain resources and lead to lower GDP. Moreover, we can explore which regions or continents had higher GDP, longer life expectancies, and different population growth rates in 2020. This could reveal trends in economic development and quality of life on a global scale. With these opportunities for exploration, we find ourselves poised to answer some compelling research questions from our dataset:

1. **What is the correlation between GDP and life expectancy at birth?**
2. **Are there any regional variations in life expectancy and population growth rate trends?**
3. **Is there a relationship between a country’s population size and its population growth rate, and does this relationship vary between the Northern and Southern Hemispheres?**

Tidying

```r
# Tidy life expectancy at birth dataset
tidy_life_expectancy_data <- `Life_Expectancy_at_Birth` %>%
  pivot_longer(
    cols = starts_with("Life Expectancy at Birth"),
    names_to = "Year",
    values_to = "Life_Expectancy_at_Birth"
  )
# Extract just the year from the "Year" column in the same dataset
tidy_life_expectancy_data$Year <- sub("Life Expectancy at Birth \\((\\d{4})\\)", "\\1", tidy_life_expec

# Convert the "Year" column to numeric
tidy_life_expectancy_data$Year <- as.numeric(sub("Life Expectancy at Birth \\((\\d{4})\\)", "\\1", tidy_

print(tidy_life_expectancy_data)
```

```
## # A tibble: 6,240 x 9
##    ISO3  Country     Continent Hemisphere          ‘Human Development Groups‘
##    <chr> <chr>       <chr>     <chr>               <chr>
##  1 AFG   Afghanistan Asia      Northern Hemisphere Low
##  2 AFG   Afghanistan Asia      Northern Hemisphere Low
##  3 AFG   Afghanistan Asia      Northern Hemisphere Low
##  4 AFG   Afghanistan Asia      Northern Hemisphere Low
##  5 AFG   Afghanistan Asia      Northern Hemisphere Low
##  6 AFG   Afghanistan Asia      Northern Hemisphere Low
##  7 AFG   Afghanistan Asia      Northern Hemisphere Low
##  8 AFG   Afghanistan Asia      Northern Hemisphere Low
##  9 AFG   Afghanistan Asia      Northern Hemisphere Low
## 10 AFG   Afghanistan Asia      Northern Hemisphere Low
## # i 6,230 more rows
## # i 4 more variables: ‘UNDP Developing Regions‘ <chr>, ‘HDI Rank (2021)‘ <dbl>,
## #   Year <dbl>, Life_Expectancy_at_Birth <dbl>
```

After tidying the World Life Expectancy at Birth dataset, there are 6,240 total observations. We organized the data by creating a ‘Year’ variable to store the years, which were initially individual columns in the original dataset. Simultaneously, we stored the life expectancy values in a separate variable named ‘Life_Expectancy_at_Birth.’ For convenience, we gave the tidied dataset a more user-friendly name called ‘tidy_life_expectancy_data.’ This numerical outcome aligns with our expectations, as the dataset encompasses 32 unique years and 195 countries, resulting in the calculated total of 6,240 data points.

```
#Tidy GDP dataset
tidy_gdp_data <- GDP_DATA_2 %>%
  pivot_longer(
    cols = -c("Country Name", "Country Code", "Indicator Name", "Indicator Code"),
    names_to = "Year",
    values_to = "GDP_Value")
print(tidy_gdp_data)
```

```
## # A tibble: 16,492 x 6
##    ‘Country Name‘ ‘Country Code‘ ‘Indicator Name‘  ‘Indicator Code‘ Year
##    <chr>          <chr>          <chr>             <chr>            <chr>
##  1 Aruba          ABW            GDP (current US$) NY.GDP.MKTP.CD   1960
##  2 Aruba          ABW            GDP (current US$) NY.GDP.MKTP.CD   1961
##  3 Aruba          ABW            GDP (current US$) NY.GDP.MKTP.CD   1962
##  4 Aruba          ABW            GDP (current US$) NY.GDP.MKTP.CD   1963
##  5 Aruba          ABW            GDP (current US$) NY.GDP.MKTP.CD   1964
##  6 Aruba          ABW            GDP (current US$) NY.GDP.MKTP.CD   1965
##  7 Aruba          ABW            GDP (current US$) NY.GDP.MKTP.CD   1966
##  8 Aruba          ABW            GDP (current US$) NY.GDP.MKTP.CD   1967
##  9 Aruba          ABW            GDP (current US$) NY.GDP.MKTP.CD   1968
## 10 Aruba          ABW            GDP (current US$) NY.GDP.MKTP.CD   1969
## # i 16,482 more rows
## # i 1 more variable: GDP_Value <dbl>
```

After tidying the World GDP dataset, we obtained 16,492 total observations. We organized the data by creating a ‘Year’ variable to store the years, which were initially individual columns in the original dataset. Simultaneously, we stored the GDP values in a separate variable named ‘GDP_Value.’ For convenience, we gave the tidied dataset the more user-friendly name ‘tidy_gdp_data.’ This numerical outcome aligns with our expectations, as the dataset encompasses 62 unique years and 266 countries, resulting in the calculated total of 16,492 data points.

```
# Tidy the world_population dataset and create a "Year" variable
tidy_world_population_data <- world_population %>%
  pivot_longer(
    cols = matches("^\\d{4} Population$"),  # Select columns matching the pattern of a year followed by
    names_to = "Year",
    values_to = "Population" )

tidy_world_population_data$Year <- sub(" (Population)$", "", tidy_world_population_data$Year)
print(tidy_world_population_data)
```

```
## # A tibble: 1,872 x 11
##     Rank CCA3  ‘Country/Territory‘ Capital Continent ‘Area (km²)‘
##    <dbl> <chr> <chr>               <chr>   <chr>            <dbl>
##  1    36 AFG   Afghanistan         Kabul   Asia            652230
##  2    36 AFG   Afghanistan         Kabul   Asia            652230
##  3    36 AFG   Afghanistan         Kabul   Asia            652230
##  4    36 AFG   Afghanistan         Kabul   Asia            652230
##  5    36 AFG   Afghanistan         Kabul   Asia            652230
##  6    36 AFG   Afghanistan         Kabul   Asia            652230
##  7    36 AFG   Afghanistan         Kabul   Asia            652230
##  8    36 AFG   Afghanistan         Kabul   Asia            652230
##  9   138 ALB   Albania             Tirana  Europe           28748
## 10   138 ALB   Albania             Tirana  Europe           28748
## # i 1,862 more rows
## # i 5 more variables: ‘Density (per km²)‘ <dbl>, ‘Growth Rate‘ <dbl>,
## #   ‘World Population Percentage‘ <dbl>, Year <chr>, Population <dbl>
```

After tidying the World Population dataset, there are 1,872 total observations. We organized the data by creating a ‘Year’ variable to store the years, which were initially individual columns in the original dataset. Simultaneously, we stored the population values in a separate variable named ‘Population.’ For convenience, we gave the tidied dataset the more user-friendly name ‘tidy_world_population_data.’ This numerical outcome aligns with our expectations, as the dataset encompasses 8 original year variables and 234 countries, resulting in the calculated total of 1,872 data points.

It is seen that The World Life Expectancy and GDP datasets share a common variable ‘Country Code,’ denoted as ‘ISO3’ in the World Life Expectancy dataset. The Life Expectancy and World Population datasets also share the ‘Continent’ ID. All three datasets now share the variable ‘Year’ after tidying as well as ‘Country’, denoted as ‘Country Name’ in the GDP dataset and ‘Country/Territory’ in the World Population dataset. IDs appearing in the Life Expectancy at Birth dataset that do not appear in the other datasets are ‘Hemisphere,’ ‘Human Development Groups,’ ‘UNDP Developing Regions,’ ‘HDI Rank (2021)’, and ‘Life_Expectancy_at_Birth.’ IDs displayed in the World GDP dataset not found in the other datasets are ‘Indicator Name,’ ‘Indicator Code,’ and’ GDP_Value.’ IDs shown in the World Population dataset not found in the other datasets include ‘Rank,’ ‘Capital,’ ‘Area (km^2), ’Density (per km^2), ’Growth Rate,’ ‘World Population Percentage,’ and ’ Population.’

Filtering

```r
#Filter datasets to show year 2020 only
filtered_tidy_life_expectancy_data <- tidy_life_expectancy_data %>%
  filter(Year == 2020)
print(filtered_tidy_life_expectancy_data)
```

```
## # A tibble: 195 x 9
##    ISO3  Country              Continent Hemisphere        Human Development Gr~1
##    <chr> <chr>                <chr>     <chr>             <chr>
##  1 AFG   Afghanistan          Asia      Northern Hemisph~ Low
##  2 AGO   Angola               Africa    Southern Hemisph~ Medium
##  3 ALB   Albania              Europe    Northern Hemisph~ High
##  4 AND   Andorra              Europe    Northern Hemisph~ Very High
##  5 ARE   United Arab Emirates Asia      Northern Hemisph~ Very High
##  6 ARG   Argentina            America   Southern Hemisph~ Very High
##  7 ARM   Armenia              Asia      Northern Hemisph~ High
##  8 ATG   Antigua and Barbuda  America   Northern Hemisph~ High
##  9 AUS   Australia            Oceania   Southern Hemisph~ Very High
## 10 AUT   Austria              Europe    Northern Hemisph~ Very High
## # i 185 more rows
## # i abbreviated name: 1: ‘Human Development Groups‘
## # i 4 more variables: ‘UNDP Developing Regions‘ <chr>, ‘HDI Rank (2021)‘ <dbl>,
## #   Year <dbl>, Life_Expectancy_at_Birth <dbl>
```

```
filtered_tidy_gdp_data <- tidy_gdp_data %>%
  filter(Year == 2020)
print(filtered_tidy_gdp_data)
```

```
## # A tibble: 266 x 6
##    ‘Country Name‘          ‘Country Code‘ ‘Indicator Name‘ ‘Indicator Code‘ Year
##    <chr>                   <chr>          <chr>            <chr>            <chr>
##  1 Aruba                   ABW            GDP (current US~ NY.GDP.MKTP.CD   2020
##  2 Africa Eastern and So~ AFE            GDP (current US~ NY.GDP.MKTP.CD   2020
##  3 Afghanistan             AFG            GDP (current US~ NY.GDP.MKTP.CD   2020
##  4 Africa Western and Ce~ AFW            GDP (current US~ NY.GDP.MKTP.CD   2020
##  5 Angola                  AGO            GDP (current US~ NY.GDP.MKTP.CD   2020
##  6 Albania                 ALB            GDP (current US~ NY.GDP.MKTP.CD   2020
##  7 Andorra                 AND            GDP (current US~ NY.GDP.MKTP.CD   2020
##  8 Arab World              ARB            GDP (current US~ NY.GDP.MKTP.CD   2020
##  9 United Arab Emirates    ARE            GDP (current US~ NY.GDP.MKTP.CD   2020
## 10 Argentina               ARG            GDP (current US~ NY.GDP.MKTP.CD   2020
## # i 256 more rows
## # i 1 more variable: GDP_Value <dbl>
```

```
filtered_tidy_world_population_data <- tidy_world_population_data %>%
  filter(Year == 2020)
print(filtered_tidy_world_population_data)
```

```
## # A tibble: 234 x 11
##     Rank CCA3  ‘Country/Territory‘ Capital          Continent     ‘Area (km²)‘
##    <dbl> <chr> <chr>               <chr>            <chr>                <dbl>
##  1    36 AFG   Afghanistan         Kabul            Asia                652230
##  2   138 ALB   Albania             Tirana           Europe               28748
##  3    34 DZA   Algeria             Algiers          Africa             2381741
##  4   213 ASM   American Samoa      Pago Pago        Oceania                199
##  5   203 AND   Andorra             Andorra la Vella Europe                 468
##  6    42 AGO   Angola              Luanda           Africa             1246700
##  7   224 AIA   Anguilla            The Valley       North America           91
##  8   201 ATG   Antigua and Barbuda Saint John’s     North America          442
##  9    33 ARG   Argentina           Buenos Aires     South America      2780400
## 10   140 ARM   Armenia             Yerevan          Asia                 29743
## # i 224 more rows
## # i 5 more variables: ‘Density (per km²)‘ <dbl>, ‘Growth Rate‘ <dbl>,
## #   ‘World Population Percentage‘ <dbl>, Year <chr>, Population <dbl>
```

Before joining our datasets, we filtered each dataset to show information from only the year 2020 as this was our focus and gave user-friendly names respectively. The total number of observations for the Life Expectancy dataset was 195, GDP 266, and World Population 234.

Selecting

```r
#Selecting variables and saving into new dataset
GDP <- filtered_tidy_gdp_data %>%
  select(`Country Name`, GDP_Value) %>%
  rename(Country = `Country Name`)

life <- filtered_tidy_life_expectancy_data %>%
  select(Country, Hemisphere, Life_Expectancy_at_Birth, `Human Development Groups`)

pop <- filtered_tidy_world_population_data %>%
  select(`Country/Territory`, Population, Continent, `Growth Rate`) %>%
  rename(Country = `Country/Territory`)
```

We create three new datasets (GDP, life, and pop) by selecting specific columns of interest to retain and standardizing the column name 'Country,' for the purpose of merging these datasets together.

Merging

```r
# Merge datasets
merged_data_2020 <- inner_join(
  GDP, life,
  by = "Country") %>%
  inner_join(
    pop,
    by = "Country")

# Reorder the variables
merged_data_2020 <- merged_data_2020 %>%
  select(Country, Continent, Hemisphere, everything())

print(merged_data_2020)
```

```
## # A tibble: 168 x 8
##    Country              Continent    Hemisphere GDP_Value Life_Expectancy_at_B~1
##    <chr>                <chr>        <chr>          <dbl>                  <dbl>
##  1 Afghanistan          Asia         Northern ~   2.01e10                   62.6
##  2 Angola               Africa       Southern ~   5.36e10                   62.3
##  3 Albania              Europe       Northern ~   1.51e10                   77.0
##  4 Andorra              Europe       Northern ~   2.89e 9                   79.0
##  5 United Arab Emirates Asia         Northern ~   3.59e11                   78.9
##  6 Argentina            South Ameri~ Southern ~   3.90e11                   75.9
##  7 Armenia              Asia         Northern ~   1.26e10                   72.2
##  8 Antigua and Barbuda  North Ameri~ Northern ~   1.37e 9                   78.8
##  9 Australia            Oceania      Southern ~   1.33e12                   84.3
## 10 Austria              Europe       Northern ~   4.33e11                   81.5
## # i 158 more rows
## # i abbreviated name: 1: Life_Expectancy_at_Birth
## # i 3 more variables: ‘Human Development Groups‘ <chr>, Population <dbl>,
## #   ‘Growth Rate‘ <dbl>
```

We first merged the datasets with GDP and life with the common variable 'Country' utilizing the specified columns previously. This resulted in 172 total observations. The reduction in the total number of observations from 266 in the original GDP dataset and 195 in the original life dataset to 172 means 94 observations were dropped. This indicates that some countries present in the GDP dataset did not have corresponding data in the life dataset, vice versa. For example, the GDP dataset has the country 'Africa Eastern and Southern', 'Africa Western and Central', and 'American Samoa' which were not found in the life dataset. These non-matching rows would be excluded from the result. Moreover, there could be variations or discrepancies in how countries are named or represented in the two datasets. These discrepancies can prevent the join operation from identifying matching rows. For instance, the GDP dataset calls the country 'Bahamas, The' whereas in the life dataset it is just 'Bahamas.' If a country was present in both datasets, it would be included which is why we have 172 observations.

After merging the GDP and life datasets, another inner_join was performed with the pop dataset. Like the previous step, the joining is done based on the common 'Country' variable. The subsequent merge with the pop dataset further pruned the data by keeping only the common 'Country' values in all three datasets. This led to the reduction in the total number of observations from 172 to 168, meaning 4 observations were dropped. This indicates that 4 countries present in the GDP and life datasets did not have corresponding data in the pop dataset, or vice versa.

The primary issue to consider is the consistency of country names or representation across the datasets. Variations or discrepancies in how countries are named can lead to unmatched rows and data loss during the merge. Additionally, the dropping of observations indicates that some countries might be excluded from the analysis due to missing data in one or more datasets in our comprehensive analysis, which could lead to some discrepancies.

Following the merging of datasets, the code uses the select function to reorder the columns in the new dataset. The 'Year' variable was omitted because of our initial decision to focus exclusively on data from the year 2020. Since we are interested in understanding the state of global development for that particular year, 'Year' became redundant after filtering. Additionally, several IDs were omitted, including 'Country Code,' 'Indicator Name,' and 'Indicator Code' from the GDP dataset. From the life dataset, the 'ISO3,' 'Continent,' 'UNDP Developing Regions,' and 'HDI Rank (2021)' identifiers were excluded. Lastly, the 'Rank,' 'Country Code,' 'Capital,' 'Area (km^2),' and 'Density (per km^2)' from the pop dataset were dropped in line with the specific column selections made earlier in the process. These variabls were dropped to reduce the dataset's complexity, redundancy, and to maintain focus on the core variables of interest. Each unique row in the new dataset, merged_data_2020, now contain information about a country's name and region, information on GDP, life expectancy, population, as well as additional information such as growth rate and Human Development Groups. The omission and selection of variables in the final merged dataset, merged_data_2020, are the result of a deliberate and thoughtful process aimed at creating a dataset that is relevant to our research questions and conducive to effective data analysis and visualization shown below.

Summary Statistics

```r
# Summary stats (1 of 3 numerical)
# Mean world GDP in $
merged_data_2020 |>
  summarize(mean_gdp = mean(GDP_Value, na.rm = TRUE))
```

```
## # A tibble: 1 x 1
##         mean_gdp
##            <dbl>
## 1 475260159638.
```

The average mean of the world GDP is \$475260159638 in 2020. By knowing the global average GDP, we have a reference point to assess the economic well-being of individual countries.

```r
# Summary stats (2 of 3 numerical)
# Std dev of mean world GDP in $
merged_data_2020 |>
  summarize(sd_gdp = sd(GDP_Value, na.rm = TRUE))
```

```
## # A tibble: 1 x 1
##     sd_gdp
##      <dbl>
## 1 2.07e12
```

The standard deviation of the mean world GDP is 2.067831e+12 in 2020. The standard deviation provides insight into the degree of variability, with some countries having much higher or lower GDP values compared to the global average mean GDP in 2020.

```
# Summary stats (3 of 3 numerical)
# Mean life expectancy at birth in years by continent
merged_data_2020 |>
  group_by(Continent) |>
  summarize(mean_life_expectancy = mean(Life_Expectancy_at_Birth, na.rm = TRUE)) |>
  arrange(desc(mean_life_expectancy))
```

```
## # A tibble: 6 x 2
##   Continent     mean_life_expectancy
##   <chr>                        <dbl>
## 1 Europe                        79.3
## 2 North America                 74.2
## 3 Asia                          73.5
## 4 South America                 73.4
## 5 Oceania                       70.0
## 6 Africa                        63.3
```

The result is a summary that shows the mean life expectancy at birth for each continent, sorted in descending order in 2020. This provides insights into the variations in life expectancy across different continents in the year 2020, with Europe having the highest mean life expectancy of approximately 79.27 years and Africa with the lowest mean life expectancy of approximately 63 years.

```
# Summary stats (1 of 2 categorical)
# Find proportion of continents
merged_data_2020 %>%
  group_by(Continent) %>%
  summarize(n = n() / 168) %>% # Divide count of each continent by total rows
  arrange(desc(n))
```

```
## # A tibble: 6 x 2
##   Continent          n
##   <chr>          <dbl>
## 1 Africa        0.286
## 2 Europe        0.244
## 3 Asia          0.214
## 4 North America 0.113
## 5 Oceania       0.0774
## 6 South America 0.0655
```

The result is a summary that shows the proportion of continents, sorted in descending order in 2020. The continent with the highest number of countries is Africa, containing about 29% of the countries in our dataset. The continent with the lowest number of countries is South America, which contains only 6.5% of the countries.

```
# Summary stats (2 of 2 categorical)
# Find proportion of hemispheres
merged_data_2020 %>%
  group_by(Hemisphere) %>%
  summarize(n = n() / 168) %>%
  arrange(desc(n))
```

```
## # A tibble: 2 x 2
##   Hemisphere              n
##   <chr>               <dbl>
## 1 Northern Hemisphere 0.774
## 2 Southern Hemisphere 0.226
```

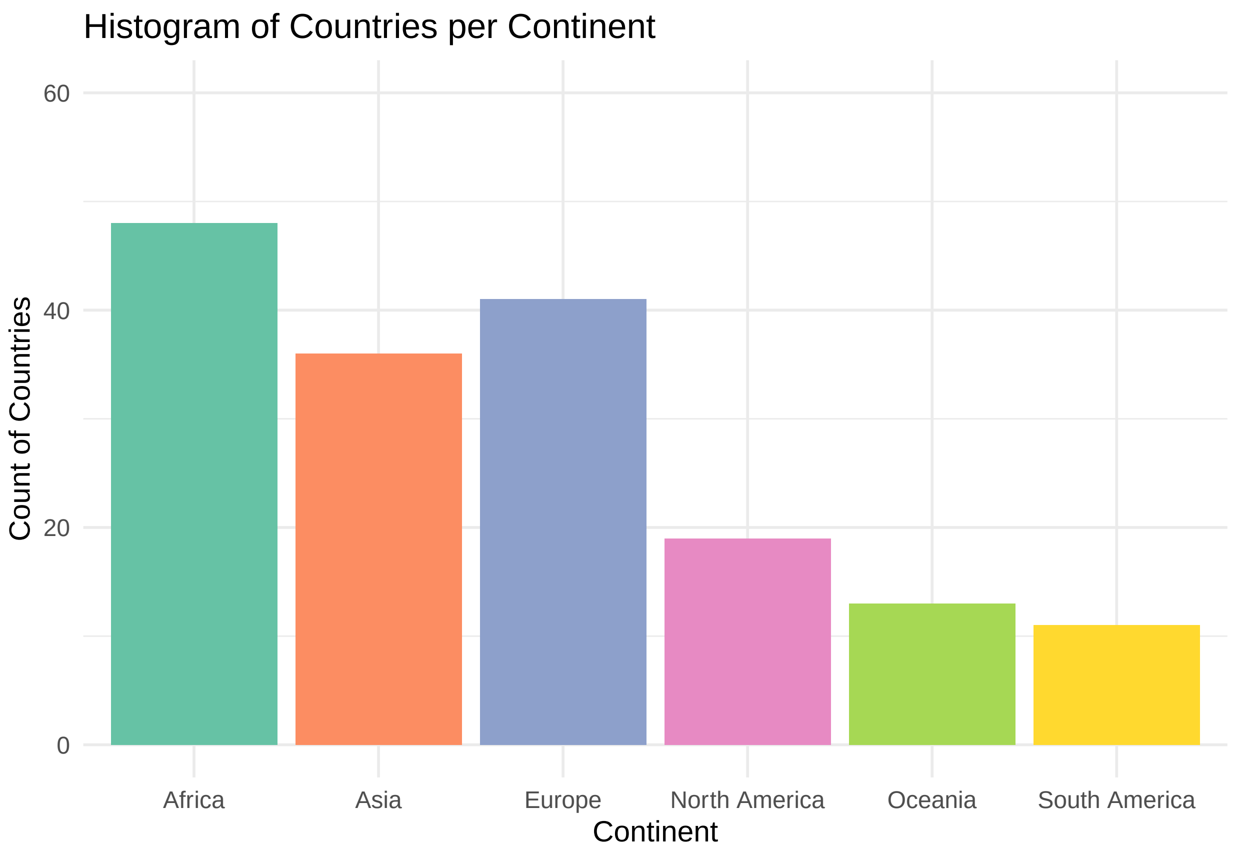
The result is a summary that shows the proportion of hemispheres, sorted in descending order in 2020. The Northern Hemisphere contains a large majority of the countries in our dataset, 77%, and the Southern Hempisphere encompasses the other 23%.

Visualization 1

```
library(ggplot2)
library(RColorBrewer)

# Generate custom colors for each continent
continent_colors <- brewer.pal(nlevels(factor(merged_data_2020$Continent)), "Set2")

# Histogram by hemisphere (plot 1 of 2 of one variable)
merged_data_2020 |>
  ggplot(aes(x = Continent, fill = Continent)) +
  geom_histogram(stat = 'count') +
  scale_fill_manual(values = continent_colors) +  # Use custom colors
  labs(x = 'Continent', y = 'Count of Countries', title = 'Histogram of Countries per Continent') +
  theme_minimal() +
  scale_y_continuous(limits = c(0, 60)) +
  guides(fill = FALSE)
```

```
## Warning in geom_histogram(stat = "count"): Ignoring unknown parameters:
## ‘binwidth‘, ‘bins‘, and ‘pad‘
```

```
## Warning: The ‘<scale>‘ argument of ‘guides()‘ cannot be ‘FALSE‘. Use "none" instead as
## of ggplot2 3.3.4.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```



This is a histogram that depicts the count of countries in the world by continent. The histogram shows Africa and Europe are the continents with the most countries in the world and Oceania and South America are the continents with the least countries in the world. Overall, this histogram allows us to quickly grasp the distribution of countries across continents, highlighting regions with varying degrees of geodiversity.

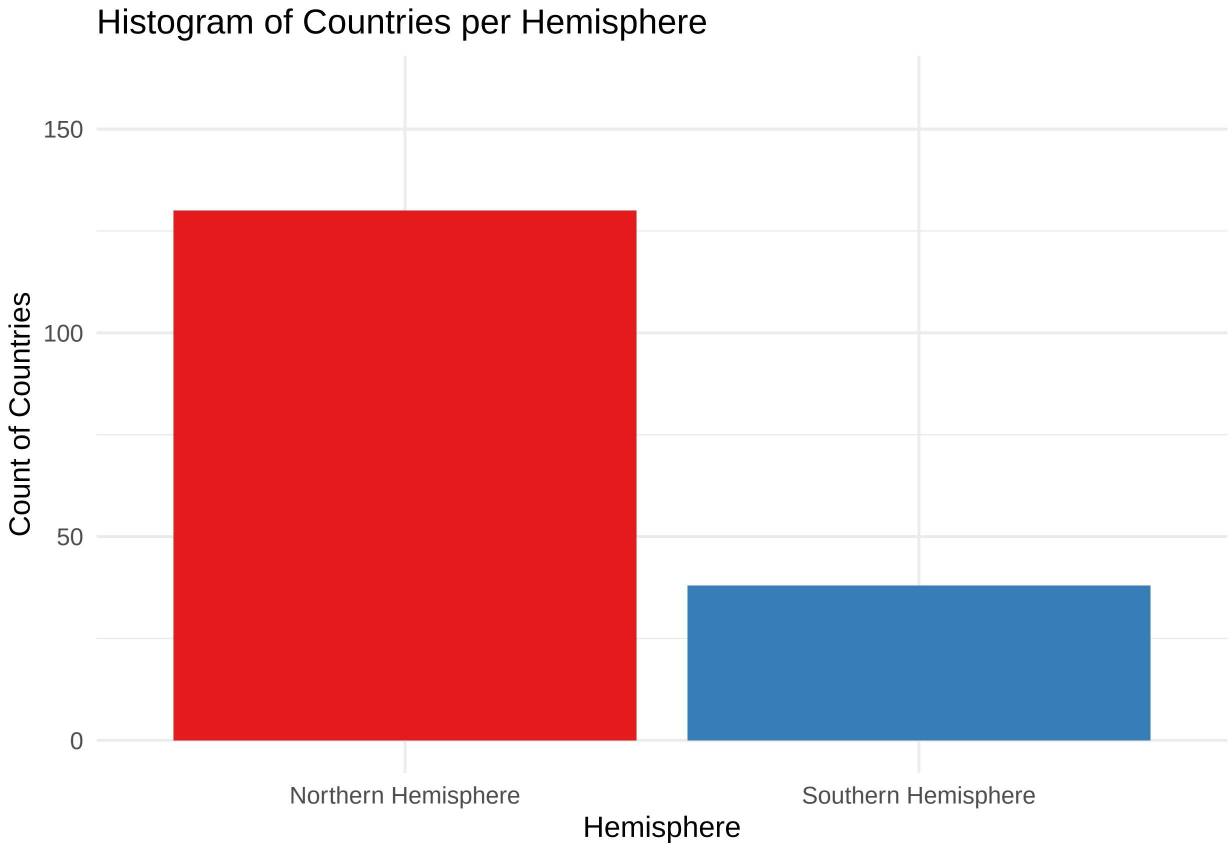
Visualization 2

```
library(RColorBrewer)

# Generate custom colors for each hemisphere
hemisphere_colors <- brewer.pal(nlevels(factor(merged_data_2020$Hemisphere)), "Set1")
```

```
## Warning in brewer.pal(nlevels(factor(merged_data_2020$Hemisphere)), "Set1"): minimal value for n is
```

```
# Histogram by hemisphere (plot 2 of 2 of one variable)
merged_data_2020 |>
  ggplot(aes(x = Hemisphere, fill = Hemisphere)) +
  geom_histogram(stat = 'count') +
  labs(x = 'Hemisphere', y = 'Count of Countries', title = 'Histogram of Countries per Hemisphere') +
  theme_minimal() +
  scale_y_continuous(limits = c(0, 160)) +
  scale_fill_manual(values = hemisphere_colors) +  # Set custom colors
   theme(legend.position = 'none')
```

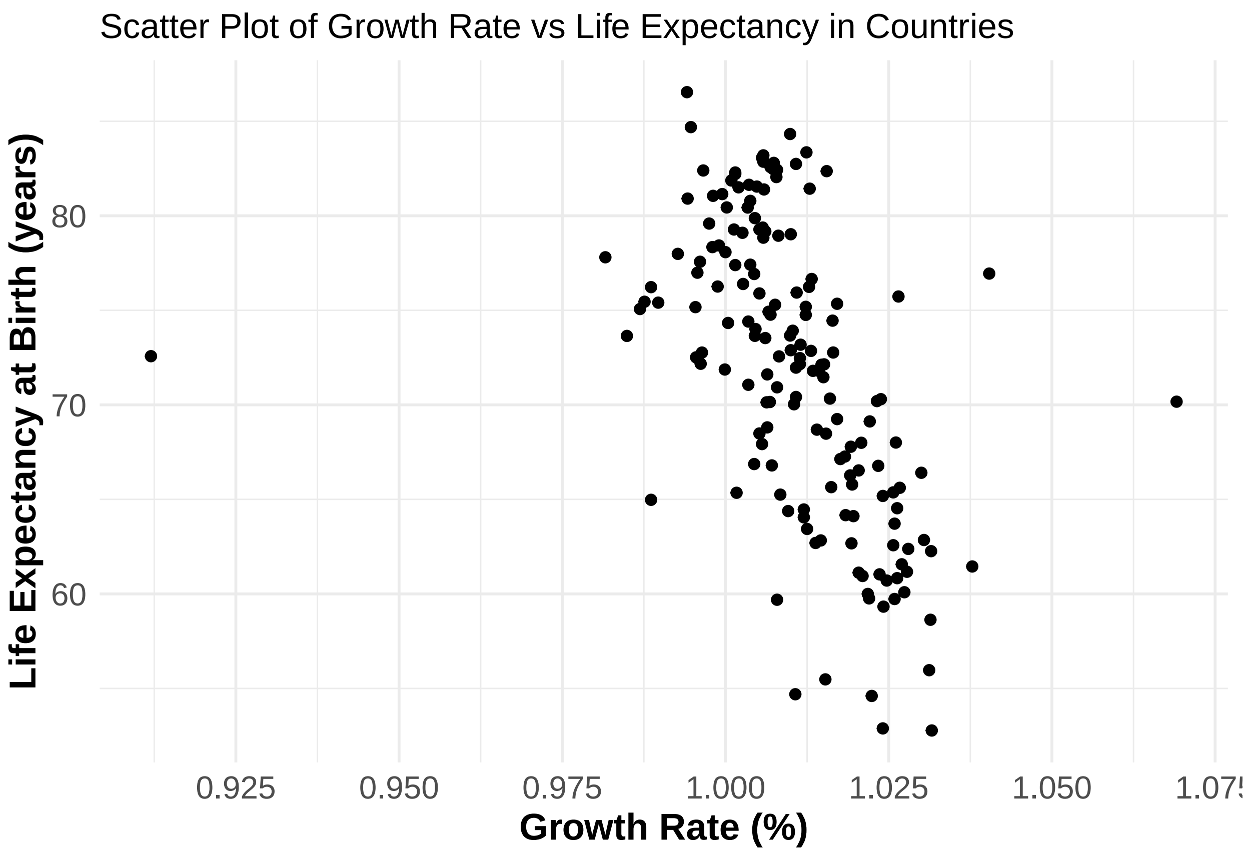
```
## Warning in geom_histogram(stat = "count"): Ignoring unknown parameters:
## ‘binwidth‘, ‘bins‘, and ‘pad‘
```



This is a histogram that depicts the count of countries in the world by hemisphere. The histogram shows that the Northern Hemisphere has almost five times as many countries as the Southern Hemisphere. We can see that most countries in this dataset are within the Northen Hemisphere.

Visualization 3

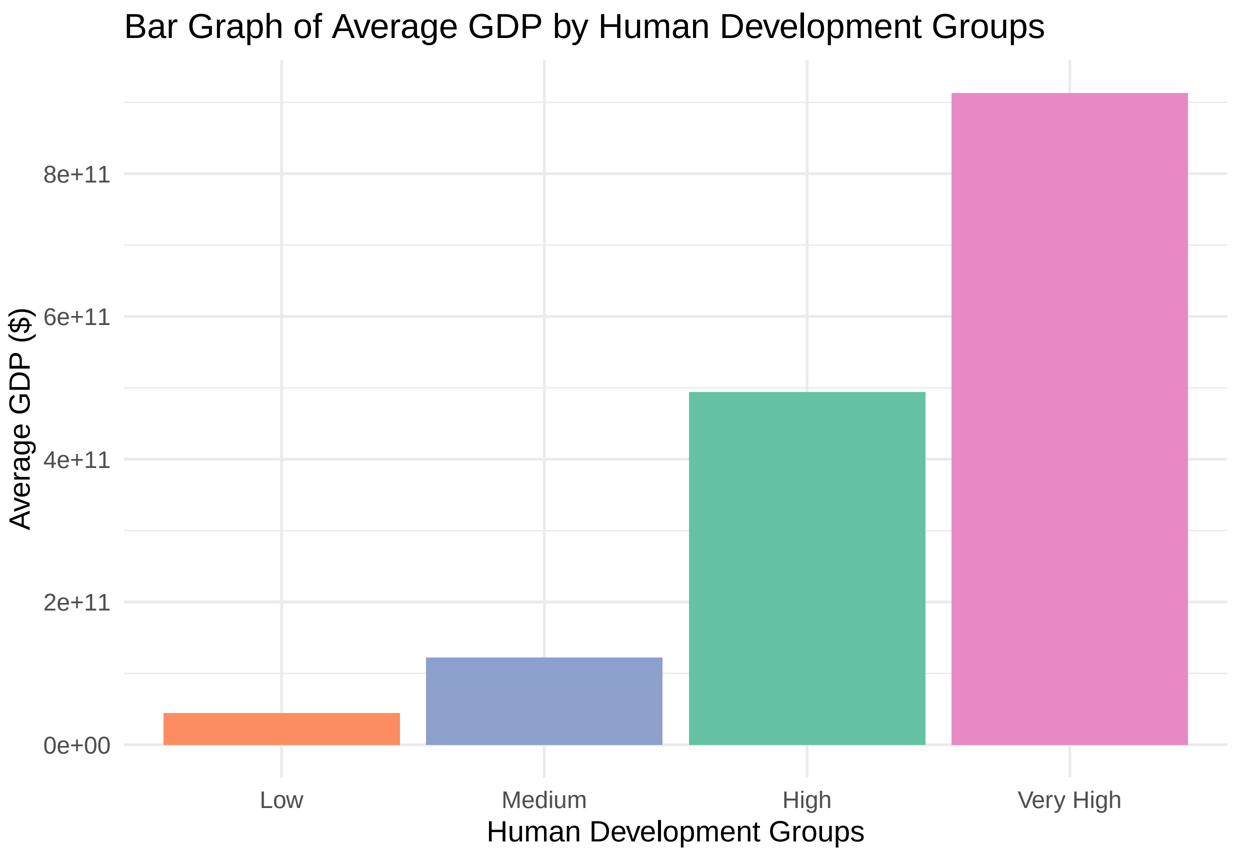
```
# Growth rate vs life expectancy plot (1 of 2 two variable)
merged_data_2020 |>
  ggplot(aes(x = `Growth Rate`, y = Life_Expectancy_at_Birth)) +
  geom_point() +
  labs(x = 'Growth Rate (%)', y = 'Life Expectancy at Birth (years)', title = 'Scatter Plot of Growth Ra
  theme_minimal() +
  scale_x_continuous(n.breaks = 7) +
  theme(axis.text = element_text(size = 12),  # Modify axis text size
        axis.title = element_text(size = 14, face = "bold"))  # Modify axis title size and style
```



This is a scatter plot of population growth rate in percent vs life expectancy in years. Each dot represents a country in the world. In other words, as a country's growth rate increases, its life expectancy tends to decrease. Conversely, countries with lower growth rates tend to have higher life expectancies. There are a few outliers such as Ukraine having the lowest growth rate but a higher life expectancy of about 72 years. However, most points follow the negative trend.

Visualization 4

```
# Human development groups vs GDP plot (2 of 2 two variable)
library(RColorBrewer)

# Define custom colors for each Human Development Group
group_colors <- brewer.pal(4, "Set2")

merged_data_2020 %>%
  filter(!is.na(`Human Development Groups`)) %>%
  group_by(`Human Development Groups`) %>%
  summarize(avg_gdp = mean(GDP_Value, na.rm = TRUE)) %>%
  ggplot(aes(x = factor(`Human Development Groups`, levels = c('Low', 'Medium', 'High', 'Very High')),
  geom_bar(stat = 'identity') +
  labs(x = 'Human Development Groups', y = 'Average GDP ($)', title = 'Bar Graph of Average GDP by Huma
  scale_fill_manual(values = group_colors) +  # Set custom colors
  theme_minimal() +
  scale_y_continuous(n.breaks = 7) +
  theme(legend.position = 'none')
```



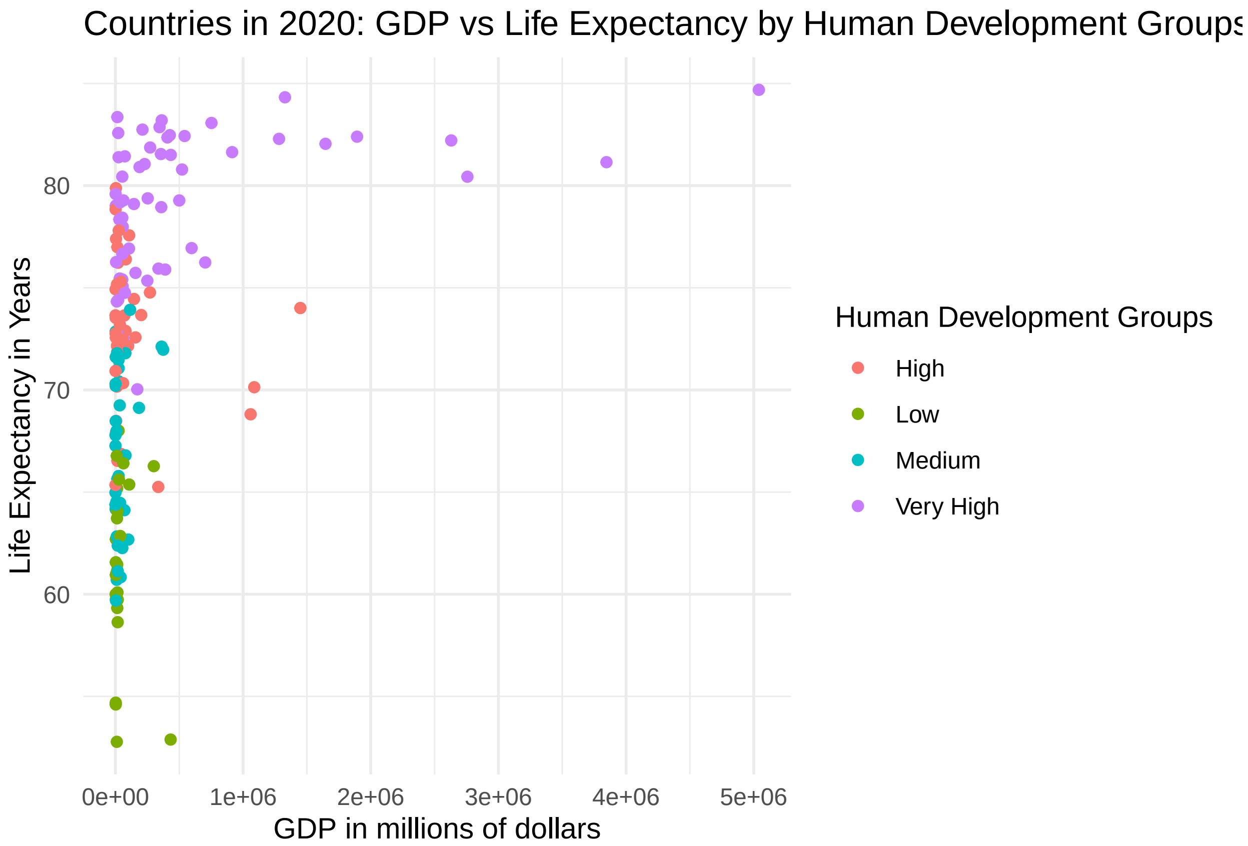
This is a bar graph showing the average GDP in dollars for Human Development Groups. By categorizing countries into Human Development Groups, which are indicative of their overall development and living standards, we can see a clear trend. As countries advance from "Low" to "Medium," "High," and "Very High" Human Development Groups, their average GDP tends to increase. This implies a strong correlation between a country's economic prosperity (GDP) and its level of human development, which encompasses factors like healthcare, education, and living standards. Specifically, countries with higher human development tend to have a higher average GDP.

Visualization 5

```
# Create new variable GPD_zscore

mean_GDP = mean(merged_data_2020$GDP_Value, na.rm = TRUE) # Find the mean value of GDP, ignoring missin
sd_GDP = sd(merged_data_2020$GDP_Value, na.rm = TRUE) # Find the standard deviation of GDP, ignoring mi

merged_data_2020 <- merged_data_2020 %>%
  mutate(GDP_zscore = (GDP_Value - mean_GDP) / sd_GDP) # Create a column that represents the z-score va
```

```
# Create new variable Pop_zscore

mean_pop = mean(merged_data_2020$Population, na.rm = TRUE) # Find the mean value of Population, ignorin
sd_pop = sd(merged_data_2020$Population, na.rm = TRUE) # Find the standard deviation of Population, ign

merged_data_2020 <- merged_data_2020 %>%
  mutate(Pop_zscore = (Population - mean_pop) / sd_pop) # Create a column that represents the z-score v
```

```r
# Plot GDP, Life expectancy, and Human Development Groups

merged_data_2020 %>%
  # Filter out GDP and Population outliers
  filter(abs(GDP_zscore) < 4 & abs(Pop_zscore) < 5) %>%
  # Filter out NA values for Human Development Groups
  filter(!is.na(`Human Development Groups`)) %>%

  # GDP on x-axis, divide to put in millions, life expectancy on y
  ggplot(aes(y = Life_Expectancy_at_Birth, x = GDP_Value / 1000000)) +
  # Color by Human Development Groups
  geom_point(aes(color = `Human Development Groups`)) +
  scale_y_continuous(n.breaks = 5) +
  # Add labels and title
  labs(x = 'GDP in millions of dollars',
       y = 'Life Expectancy in Years',
       title = 'Countries in 2020: GDP vs Life Expectancy by Human Development Groups') +
  theme_minimal()
```

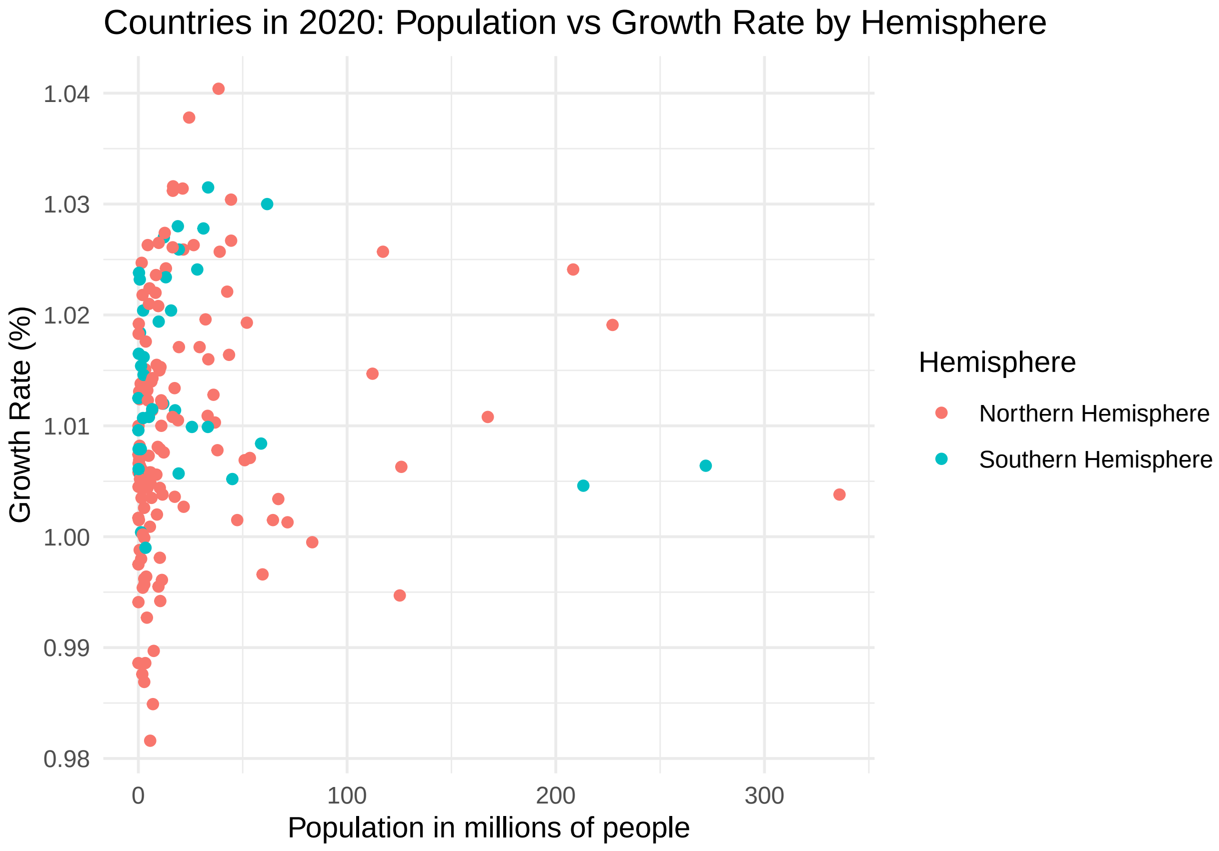


This plot represents GDP vs Life expectancy, and also groups together Human Development Groups. According to the graph, there seems to be some positive relationship between GDP and Life Expectancy. However, this relationship is limited. Once a country reaches a certain GDP threshold, the life expectancy flattens out. It seems that regardless of how high a country's GDP is, life expectancy will not increase past 85 years. It also seems to be the case that Human Development Groups has a high correlation with life expectancy. Countries with high life expectancy are also high in Human Development, and countries that have a lower life expectancy rank lower in Human Development. This relationship underscores that a country's life ex-

pectancy isn't solely determined by economic prosperity but is also influenced by other aspects like education and healthcare, as captured by Human Development Groups.

Visualization 6

```
# Plot Population, Growth Rate, and Hemisphere
merged_data_2020 %>%
  # Filter out NA values for Hemisphere
  filter(!is.na(Hemisphere)) %>%
  # Filter out Growth Rate and Population outliers
  filter(abs(`Growth Rate` - 1) < .05 & abs(Pop_zscore) < 5) %>%

  # Population on x-axis, divide to put in millions, growth rate on y
  ggplot(aes(x = Population / 1000000, y = `Growth Rate`)) +
  # Color by Hemisphere
  geom_point(aes(color = Hemisphere)) +
  scale_y_continuous(n.breaks = 10) +
  # Add labels and title
  labs(x = 'Population in millions of people',
       y = 'Growth Rate (%)',
       title = 'Countries in 2020: Population vs Growth Rate by Hemisphere') +
  theme_minimal()
```



This plot depicts total population of a country and its growth rate, and also groups together countries according to hemisphere. This graph shows little in the way of trends, the vast majority of the points cluster together on the left side, meaning that most countries have a population lower than 50 million. The Northern and

Southern Hemisphere groups have very little difference between them. There are considerably more countries in the Northern Hemisphere, but the spread is about the same for both. The main takeaway from this graph is that as the countries populations increase, the growth rates have less variation. The countries with smaller populations have much greater growth rate variation, above and below 1.

Discussion

Having conducted data analysis and examined the visualizations, we can now provide insights into our research questions:

1. What is the correlation between GDP and life expectancy at birth?
Visualization 5 depicts the correlation between GDP and life expectancy at birth. It is evident that as a country's GDP increases, its life expectancy tends to rise. This suggests a positive correlation between these two variables. However, an interesting observation from this visualization that our group did not expect was that the relationship has diminishing returns. Beyond a certain GDP threshold, which is around 85 years on the life expectancy axis, the increase in GDP does not significantly impact life expectancy. This plateau implies that other factors, such as healthcare and education, become crucial in determining life expectancy. By visualizing this correlation, we have learned that the GDP, can indeed contribute to longer life expectancy, but there are diminishing marginal returns.

2. Are there any regional variations in life expectancy and population growth rate trends?
Visualization 3 demonstrates that as a country's growth rate increases, its life expectancy tends to decrease. This suggests regional variations and trends in population and growth rate, with implications for factors influencing these trends. The negative correlation observed in the scatter plot suggests that regions or countries experiencing rapid population growth may face challenges related to healthcare, access to education, and overall living conditions. Conversely, regions with slower population growth rates may have more stable resources and potentially better access to essential services, leading to higher life expectancies. By visualizing this relationship, we've learned that regional variations in population growth rates and life expectancy are interconnected. However, it should be noted that there are outliers, and other factors can contribute to these variations.

3. Is there a relationship between a country's population size and its population growth rate, and does this relationship vary between the Northern and Southern Hemispheres?
Visualization 6, which displays a scatter plot of a country's total population and its growth rate while grouping countries by hemisphere, allows us to explore the relationship between population size and population growth rate while considering any potential variations between the Northern and Southern Hemispheres. In this visualization, we observe that most countries are clustered on the left side, indicating that they have relatively smaller populations and varying growth rates. The primary takeaway from this visualization is it is difficult to identify a clear relationship between these factors. To gain deeper insights and quantify these relationships, further statistical analysis may be required, such as correlation calculations or regression models. We learned that visualizations can serve as a starting point for more in-depth investigations into the complex factors, and in this case, factors influencing population growth rates in countries with varying population sizes.

Reflection

Overall, this project was very insightful and helped us hone our R skills. We learned the importance of thoroughly exploring and evaluating available datasets to tailor our research questions, leveraging each individual's strengths, and extracting meaningful insights from the visualizations. The most significant challenge we faced as a group was finding three distinct datasets that could be successfully merged into one cohesive dataset. We initially wanted to explore the statistics of streaming services. However, the inability of finding suitable datasets led us to adapt and refine our research questions. We realized that sometimes the availability and quality of data can shape the direction of a project. Being flexible and open to adjusting our research questions was a valuable lesson.

Acknowledgements

These individual contributions were vital in the successful completion of the project:

- Mayce: Responsible for data cleaning, dataset acquisition, reformatting visualizations, and formatting the report.
- Carlos: Contributed to the creation of Visualizations 5 and 6 and provided summary statistics.
- Mindy: Created Visualizations 1-4 and handled summary statistics.