CITIES BY GDP

DATA ANALYSIS BY - JITHIN JAYACHANDRAN

INTRODUCTION

In this project, we aim to perform an exploratory data analysis (EDA) on a dataset containing information about various cities around the world, focusing specifically on their Gross Domestic Product (GDP) and population metrics. The analysis is designed to uncover patterns, relationships, and insights from the data, facilitating a deeper understanding of economic and demographic distributions across different regions.

Objectives:

Data Exploration and Cleaning:
Load and inspect the dataset to understand its structure and content.
Perform data cleaning to ensure consistency and accuracy, including handling missing values and converting data types as necessary.

Descriptive Statistics:
Generate summary statistics to get an overview of key metrics such as GDP and population.
Visualize the distribution of GDP and population among cities to identify any noticeable trends or outliers.

Comparative Analysis:
Identify the top cities by GDP and population to highlight major economic hubs and densely populated areas.
Analyze the average, total, and median GDP and population by country/region to compare economic and demographic characteristics across different areas.

Correlation Analysis:
Investigate the relationship between GDP and population to understand how these two variables are correlated.

Threshold Analysis:
Filter and visualize cities that exceed specific thresholds for GDP and population, providing insights into significant economic and demographic centers.

Methodology:
The analysis will be conducted using R, a powerful statistical programming language. The project will leverage several R packages including dplyr for data manipulation, ggplot2 for data visualization, and tidyr for data tidying. The workflow will consist of data loading, cleaning, exploratory data analysis, and visualization, ensuring a comprehensive examination of the dataset.

By the end of this project, we aim to provide a clear and insightful overview of the economic and population landscape of cities worldwide, supported by detailed visualizations and statistical analysis.

DATASET OVERVIEW

THE DATASET USED IN THIS ANALYSIS CONTAINS INFORMATION ABOUT VARIOUS CITIES AROUND THE WORLD, FOCUSING ON THEIR ECONOMIC AND DEMOGRAPHIC ATTRIBUTES.

Index	Metropolitian Area/City	Country/Region	Estimated_GDP_Billion_USD	Metropolitian_Population
1	A CoruÃ±a metropolitan area	Spain	28.819	11,21,815
2	Aachen	Germany	24.296	2,49,070
3	Aalborg	Denmark	31.855	2,19,487
4	Aarhus	Denmark	54.927	3,52,751
5	Abbotsford, British Columbia	Canada	6.239	2,03,907
6	Aberdeen, Scotland	United Kingdom	23	4,89,840
7	Abidjan	Ivory Coast	27	56,00,000
8	Abilene, TX MSA	United States	9.469	1,79,308
9	Abu Dhabi metropolitan area	United Arab Emirates	151.073	16,60,000
10	Adelaide	Australia	64.461	13,44,368
11	Ahmedabad	India	80	93,00,000
12	Aix-Marseille-Provence Metropolis	France	135.038	19,11,657
13	Akron, OH MSA	United States	44.562	6,97,627
14	Albany-Lebanon, OR MSA	United States	6.108	1,30,467
15	Albany-Schenectady-Troy, NY MSA	United States	80.303	9,04,617
16	Albany, GA MSA	United States	7.312	1,45,786
17	Albuquerque, NM MSA	United States	53.862	9,19,543
18	Alexandria	Egypt	36	59,50,000
19	Alexandria, LA MSA	United States	7.46	1,49,189
20	Alicante metropolitan area	Spain	38.851	19,15,282
21	Allentown-Bethlehem-Easton, PA-NJ MSA	United States	54.323	8,71,229
22	Almaty	Kazakhstan	41.485	22,50,000
23	Altoona, PA MSA	United States	7.238	1,21,032
24	Alxa League	China	4.419	2,62,361
25	Amarillo, TX MSA	United States	17.376	2,71,171
26	Ames, IA MSA	United States	8.104	1,25,767
27	Amiens	France	19.418	3,04,331
28	Amsterdam metropolitan area	Netherlands	237.841	28,90,428
29	AnÃ¡polis	Brazil	6.74	3,34,623
30	Anchorage, AK MSA	United States	31.57	4,00,470
31	Angers Loire MÃ©tropole	France	29.379	3,94,256

COLUMN DETAILS AND DESCRIPTION

THE DATASET CONSISTS OF SEVERAL COLUMNS, EACH PROVIDING SPECIFIC INFORMATION ABOUT THE CITIES. BELOW IS A DETAILED DESCRIPTION OF EACH COLUMN:

Column Name	Description
`Metropolitian_Area_City`	The name of the metropolitan area or city.
`Country_Region`	The name of the country or region where the city is located.
`Estimated_GDP_Billion_USD`	The estimated Gross Domestic Product (GDP) of the city in billions of USD.
`Metropolitian_Population`	The population of the metropolitan area.

METROPOLITIAN_AREA_CITY:
TYPE: CHARACTER
DESCRIPTION: THIS COLUMN CONTAINS THE NAMES OF VARIOUS METROPOLITAN AREAS OR CITIES INCLUDED IN THE DATASET.

COUNTRY_REGION:
TYPE: CHARACTER
DESCRIPTION: THIS COLUMN SPECIFIES THE COUNTRY OR REGION WHERE EACH CITY IS LOCATED.

ESTIMATED_GDP_BILLION_USD:
TYPE: NUMERIC (AFTER CLEANING)
DESCRIPTION: REPRESENTS THE ESTIMATED GDP OF EACH CITY IN BILLIONS OF USD. THIS VALUE PROVIDES AN INDICATION OF THE ECONOMIC SIZE OF THE CITY.

METROPOLITIAN_POPULATION:
TYPE: NUMERIC (AFTER CLEANING)
DESCRIPTION: INDICATES THE POPULATION OF THE METROPOLITAN AREA. THIS VALUE HELPS UNDERSTAND THE SCALE AND DENSITY OF THE POPULATION IN EACH CITY.

DATA LOADING AND DATA HANDLING

```r
# Load necessary libraries
library(dplyr)
library(ggplot2)
library(tidyr)
```

LOADING NECESSARY LIBRARIES:

TO PERFORM DATA MANIPULATION AND VISUALIZATION IN R, WE UTILIZE SEVERAL ESSENTIAL LIBRARIES. HERE ARE THE LIBRARIES LOADED FOR THIS PROJECT:

DPLYR: FOR DATA MANIPULATION.
GGPLOT2: FOR DATA VISUALIZATION.
TIDYR: FOR DATA TIDYING.

```r
# Set Directory
setwd('C:/Users/Jithin Jayachandran/Downloads/R')

# Load Data
df <- read.csv('cities_by_gdp.csv')
```

DATA LOADING:

TO BEGIN OUR ANALYSIS, WE FIRST LOAD THE DATASET AND SET THE WORKING DIRECTORY IN R.

SETTING THE WORKING DIRECTORY: THIS COMMAND SETS THE WORKING DIRECTORY TO THE SPECIFIED PATH WHERE THE DATASET IS STORED.
LOADING THE DATA: THE READ.CSV FUNCTION IS USED TO LOAD THE DATA FROM A CSV FILE INTO A DATAFRAME NAMED DF.

```r
# Convert columns to appropriate data types
df$Estimated_GDP_Billion_USD <- as.numeric(gsub(",", "", df$Estimated_GDP_Billion_USD))
df$Metropolitian_Population <- as.numeric(gsub(",", "", df$Metropolitian_Population))

# Adjust the column names to remove special characters and spaces for easier handling
names(df) <- gsub(" ", "_", names(df))
names(df) <- gsub("/", "", names(df))
```

DATA HANDLING:

AFTER LOADING THE DATA, WE PERFORM SEVERAL DATA HANDLING STEPS TO ENSURE THE DATASET IS CLEAN AND READY FOR ANALYSIS.

DATA CLEANING STEPS:

CONVERT COLUMNS TO APPROPRIATE DATA TYPES:
REMOVE COMMAS FROM NUMERIC FIELDS AND CONVERT THEM TO NUMERIC DATA TYPES.
ADJUST COLUMN NAMES TO REMOVE SPACES AND SPECIAL CHARACTERS FOR EASIER HANDLING.

DISPLAY THE STRUCTURE OF THE DATAFRAME:

**STR(DF): DISPLAYS THE STRUCTURE OF THE DATAFRAME,
SHOWING THE DATA TYPES OF EACH COLUMN AND A PREVIEW OF THE DATA.**

```
str(df)
```

```
> STR(DF)
'DATA.FRAME':       903 OBS. OF  5 VARIABLES:
 $ INDEX                : INT  1 2 3 4 5 6 7 8 9 10 ...
 $ METROPOLITIAN.AREA.CITY  : CHR  "A CORUÑA METROPOLITAN AREA" "AACHEN" "AALBORG" "AARHUS" ...
 $ COUNTRY.REGION           : CHR  "SPAIN" "GERMANY" "DENMARK" "DENMARK" ...
 $ ESTIMATED_GDP_BILLION_USD: NUM  28.82 24.3 31.86 54.93 6.24 ...
 $ METROPOLITIAN_POPULATION : NUM  1121815 249070 219487 352751 203907 ...
```

DISPLAY SUMMARY STATISTICS:

**SUMMARY(DF): GENERATES SUMMARY STATISTICS FOR EACH COLUMN IN THE DATAFRAME,
INCLUDING MINIMUM, MAXIMUM, MEAN, AND QUARTILES FOR NUMERICAL COLUMNS, AND
FREQUENCY COUNTS FOR CATEGORICAL COLUMNS.**

```
summary(df)
```

```
> SUMMARY(DF)
     INDEX      METROPOLITIAN.AREA.CITY COUNTRY.REGION     ESTIMATED_GDP_BILLION_USD METROPOLITIAN_POPULATION
 MIN.   :  1.0   LENGTH:903          LENGTH:903         MIN.   :  0.009           MIN.   :    561
 1ST QU.:226.5   CLASS :CHARACTER    CLASS :CHARACTER   1ST QU.:  10.812          1ST QU.:  212960
 MEDIAN :452.0   MODE  :CHARACTER    MODE  :CHARACTER   MEDIAN :  25.459          MEDIAN :  564024
 MEAN   :452.0                                          MEAN   :  71.718          MEAN   : 2273056
 3RD QU.:677.5                                          3RD QU.:  60.160          3RD QU.: 2418908
 MAX.   :903.0                                          MAX.   :2163.210          MAX.   :40700000
```

CHECK FOR MISSING VALUES:

SUM(IS.NA(DF)): CALCULATES THE TOTAL NUMBER OF MISSING VALUES IN THE DATAFRAME.

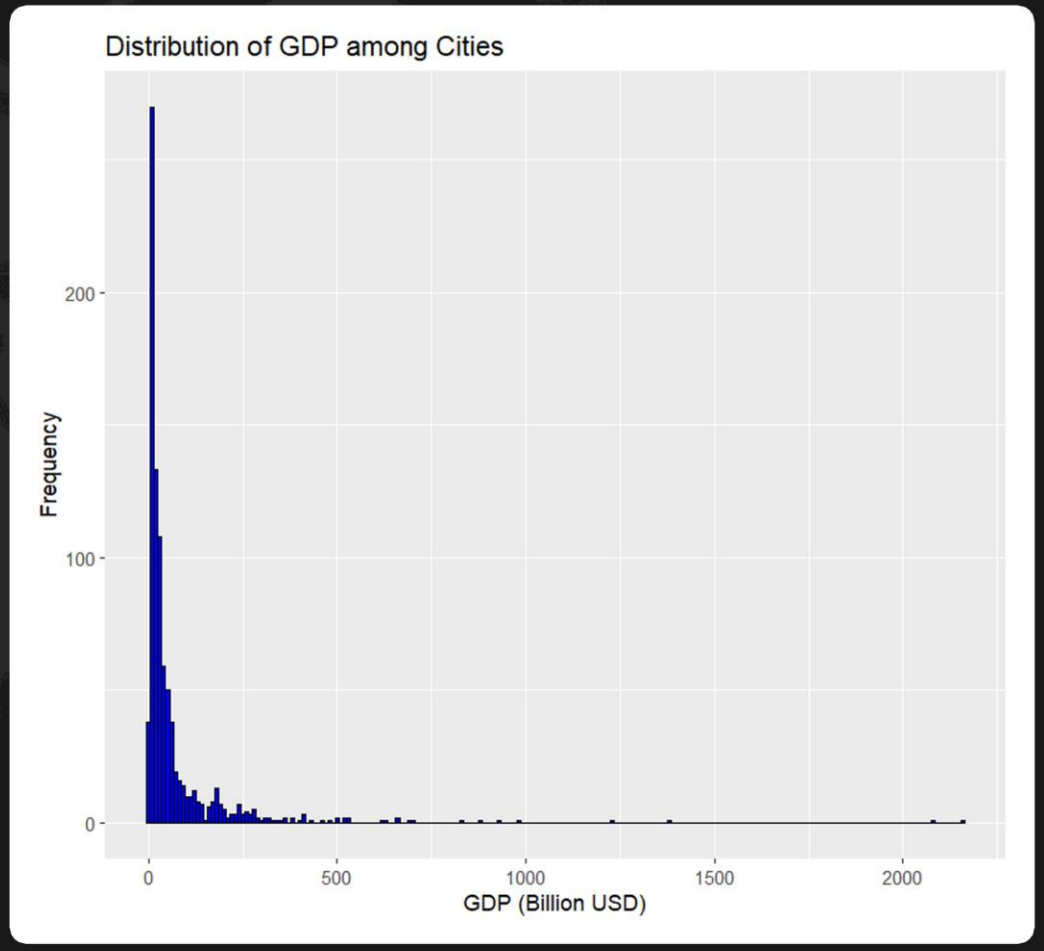
```
sum(is.na(df))
```

```
> SUM(IS.NA(DF))
[1] 0
```

DATA ANALYSIS

1. DISTRIBUTION OF GDP AMONG CITIES

```
# Distribution of GDP among cities
ggplot(df, aes(x=Estimated_GDP_Billion_USD)) +
  geom_histogram(binwidth=10, fill='blue', color='black') +
  labs(title="Distribution of GDP among Cities", x="GDP (Billion USD)", y="Frequency")
```



EXPLANATION:

GGPLOT(DF, AES(X=ESTIMATED_GDP_BILLION_USD)): INITIALIZES THE PLOT WITH DF AS THE DATA SOURCE AND ESTIMATED_GDP_BILLION_USD AS THE VARIABLE FOR THE X-AXIS.

GEOM_HISTOGRAM(BINWIDTH=10, FILL='BLUE', COLOR='BLACK'): CREATES A HISTOGRAM WITH A BIN WIDTH OF 10, BLUE FILL COLOR, AND BLACK BORDER FOR THE BINS.

LABS(TITLE="DISTRIBUTION OF GDP AMONG CITIES", X="GDP (BILLION USD)", Y="FREQUENCY"): ADDS A TITLE AND LABELS FOR THE X AND Y AXES.

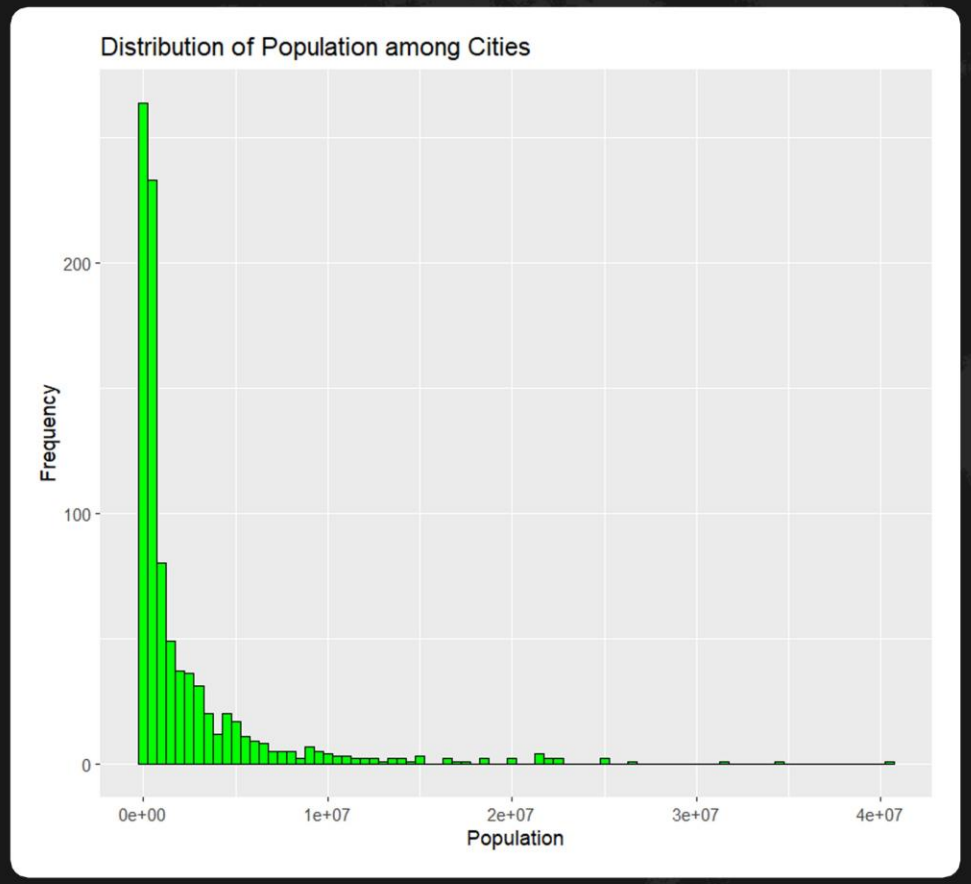
THE HISTOGRAM PROVIDES A VISUAL REPRESENTATION OF HOW GDP IS DISTRIBUTED AMONG DIFFERENT CITIES.

IT HIGHLIGHTS THE FREQUENCY OF CITIES WITHIN SPECIFIC GDP RANGES.

HELPS IDENTIFY ANY OUTLIERS OR CITIES WITH EXCEPTIONALLY HIGH OR LOW GDP VALUES.

2. WHAT IS THE DISTRIBUTION OF POPULATIONS AMONG CITIES?

TO EXPLORE THE DISTRIBUTION OF POPULATIONS AMONG CITIES, WE WILL CREATE A HISTOGRAM. THIS VISUALIZATION WILL HELP US UNDERSTAND HOW POPULATIONS ARE DISTRIBUTED ACROSS DIFFERENT CITIES AND IDENTIFY ANY PATTERNS OR ANOMALIES.




```r
# Distribution of populations among cities
ggplot(df, aes(x=Metropolitian_Population)) +
  geom_histogram(binwidth=500000, fill='green', color='black') +
  labs(title="Distribution of Population among Cities", x="Population", y="Frequency")
```

EXPLANATION:

GGPLOT(DF, AES(X=METROPOLITIAN_POPULATION)): INITIALIZES THE PLOT WITH DF AS THE DATA SOURCE AND METROPOLITIAN_POPULATION AS THE VARIABLE FOR THE X-AXIS.

GEOM_HISTOGRAM(BINWIDTH=500000, FILL='GREEN', COLOR='BLACK'): CREATES A HISTOGRAM WITH A BIN WIDTH OF 500,000, GREEN FILL COLOR, AND BLACK BORDER FOR THE BINS.

LABS(TITLE="DISTRIBUTION OF POPULATION AMONG CITIES", X="POPULATION", Y="FREQUENCY"): ADDS A TITLE AND LABELS FOR THE X AND Y AXES.

THE HISTOGRAM PROVIDES A VISUAL REPRESENTATION OF HOW POPULATION IS DISTRIBUTED AMONG DIFFERENT CITIES.

IT HIGHLIGHTS THE FREQUENCY OF CITIES WITHIN SPECIFIC POPULATION RANGES.

HELPS IDENTIFY ANY OUTLIERS OR CITIES WITH EXCEPTIONALLY HIGH OR LOW POPULATION VALUES.

3. WHICH ARE THE TOP 10 CITIES BY GDP?

TO IDENTIFY THE TOP 10 CITIES BY GDP, WE WILL SORT THE DATA AND CREATE A BAR CHART.THIS VISUALIZATION WILL HELP US UNDERSTAND WHICH CITIES HAVE THE HIGHEST ECONOMIC OUTPUT.

STEPS:

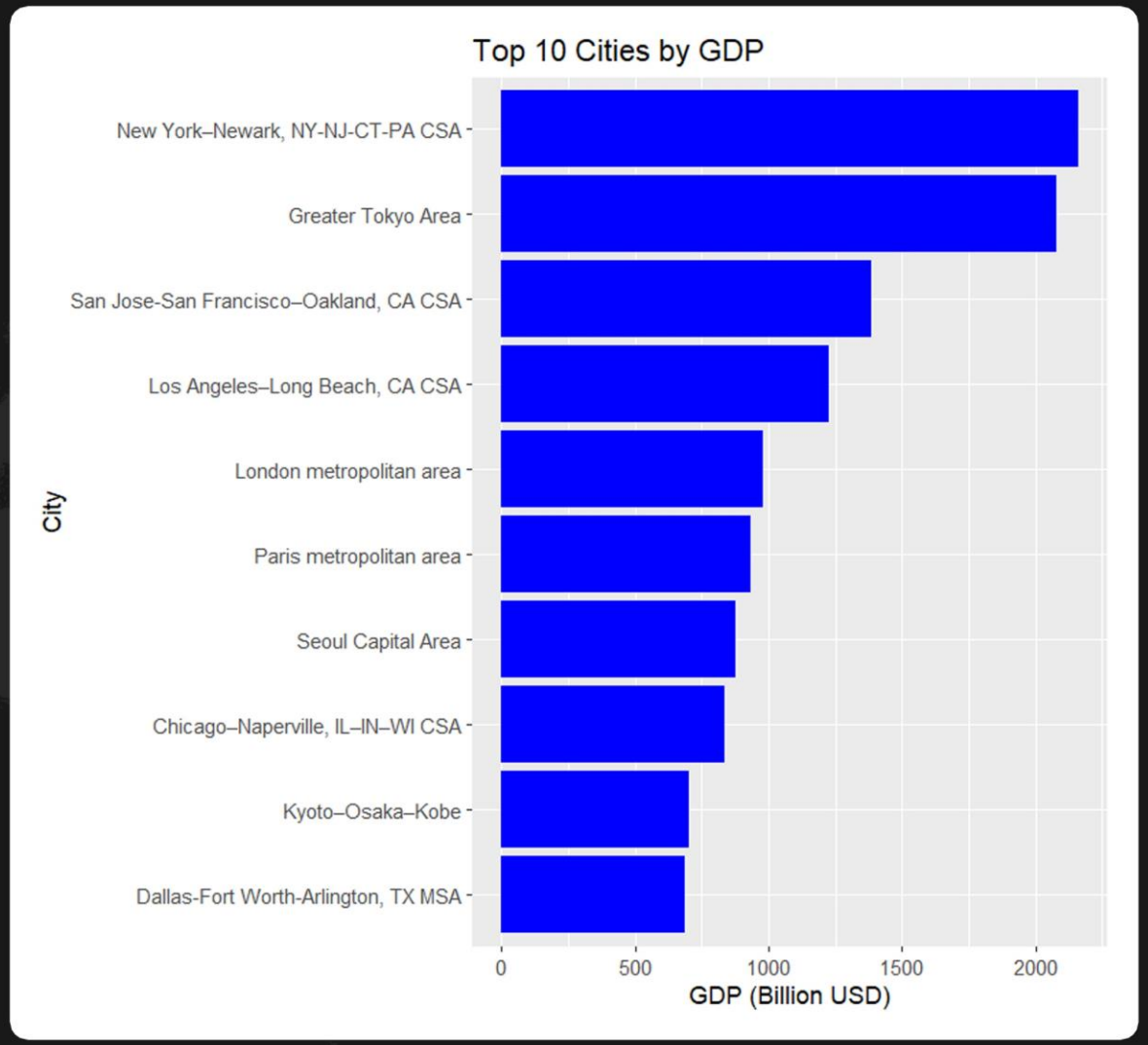
FILTER THE TOP 10 CITIES BY GDP:

USING DPLYR TO ARRANGE THE DATA IN DESCENDING ORDER OF ESTIMATED_GDP_BILLION_USD AND SELECT THE TOP 10 CITIES.

CREATE A BAR CHART:

USING GGPLOT2 TO CREATE A BAR CHART.
SET THE X AESTHETIC TO METROPOLITIAN_AREA_CITY AND Y AESTHETIC TO ESTIMATED_GDP_BILLION_USD.
CUSTOMIZE THE BAR CHART WITH APPROPRIATE COLORS, ORIENTATION, AND LABELS.

```
# Top 10 cities by GDP
top_10_gdp <- df %>% arrange(desc(Estimated_GDP_Billion_USD)) %>% head(10)
ggplot(top_10_gdp, aes(x=reorder(Metropolitian_Area_City, Estimated_GDP_Billion_USD), y=Estimated_GDP_Billion_USD)) +
  geom_bar(stat='identity', fill='blue') +
  coord_flip() +
  labs(title="Top 10 Cities by GDP", x="City", y="GDP (Billion USD)")
```



4. WHICH ARE THE TOP 10 CITIES BY POPULATION?

TO IDENTIFY THE TOP 10 CITIES BY POPULATION, WE WILL SORT THE DATA AND CREATE A BAR CHART. THIS VISUALIZATION WILL HELP US UNDERSTAND WHICH CITIES HAVE THE LARGEST POPULATIONS.

THE BAR CHART DISPLAYS THE TOP 10 CITIES WITH THE LARGEST POPULATIONS.

PROVIDES A CLEAR COMPARISON OF THE POPULATION SIZES OF THESE MAJOR CITIES.

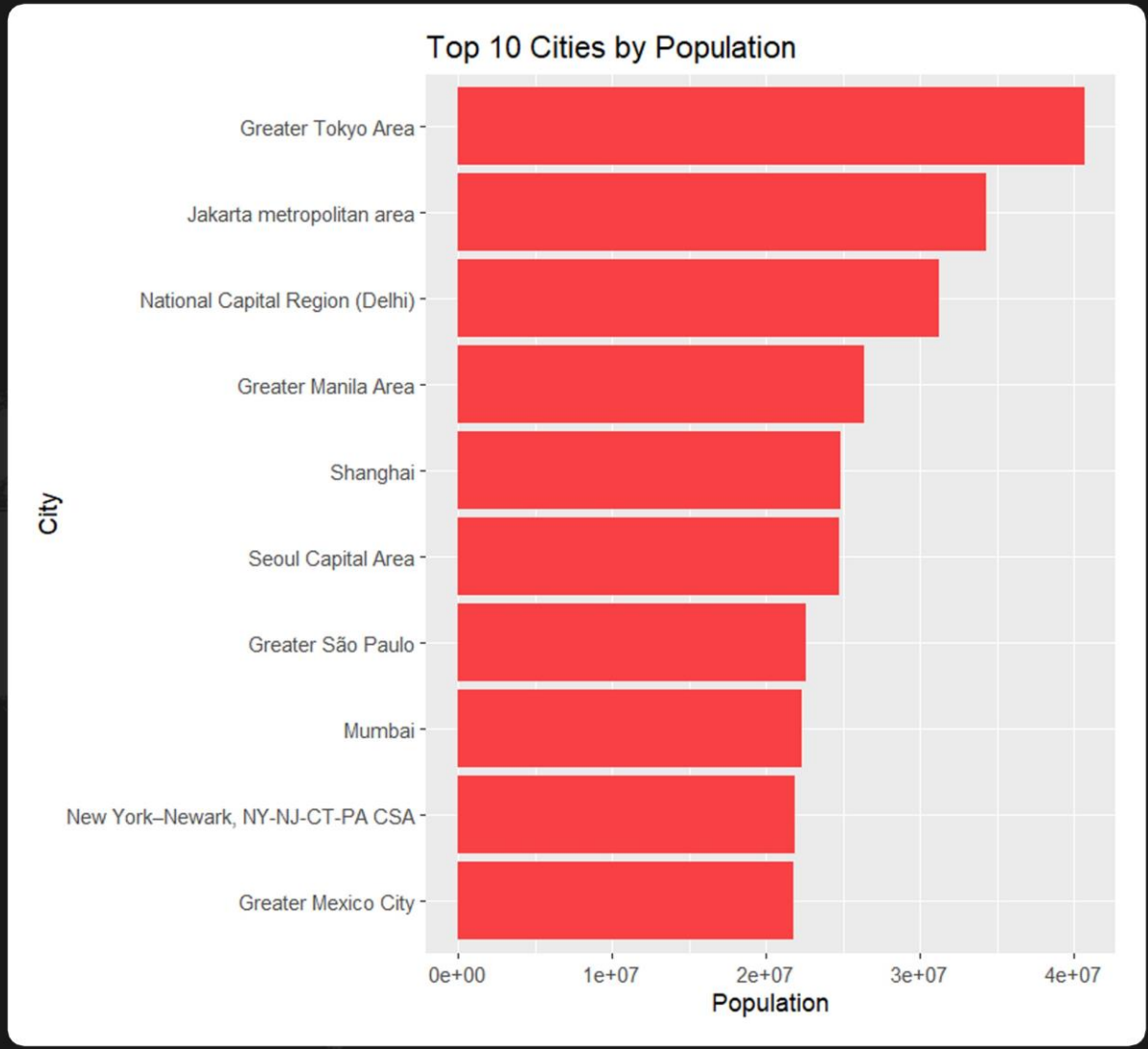
HIGHLIGHTS THE SIGNIFICANT URBAN CENTERS IN TERMS OF POPULATION DENSITY.

```
# Top 10 cities by population
top_10_population <- df %>% arrange(desc(Metropolitian_Population)) %>% head(10)
ggplot(top_10_population, aes(x=reorder(Metropolitian_Area_City, Metropolitian_Population), y=Metropolitian_Population)) +
  geom_bar(stat='identity', fill='green') +
  coord_flip() +
  labs(title="Top 10 Cities by Population", x="City", y="Population")
```

THE BAR CHART DISPLAYS THE TOP 10 CITIES WITH THE LARGEST POPULATIONS.

PROVIDES A CLEAR COMPARISON OF THE POPULATION SIZES OF THESE MAJOR CITIES.

HIGHLIGHTS THE SIGNIFICANT URBAN CENTERS IN TERMS OF POPULATION DENSITY.



5. IS THERE A CORRELATION BETWEEN GDP AND POPULATION AMONG CITIES?

TO EXPLORE THE CORRELATION BETWEEN GDP AND POPULATION, WE WILL CREATE A SCATTER PLOT WITH A REGRESSION LINE. THIS VISUALIZATION WILL HELP US UNDERSTAND THE RELATIONSHIP BETWEEN THE ECONOMIC OUTPUT AND POPULATION SIZE OF CITIES.

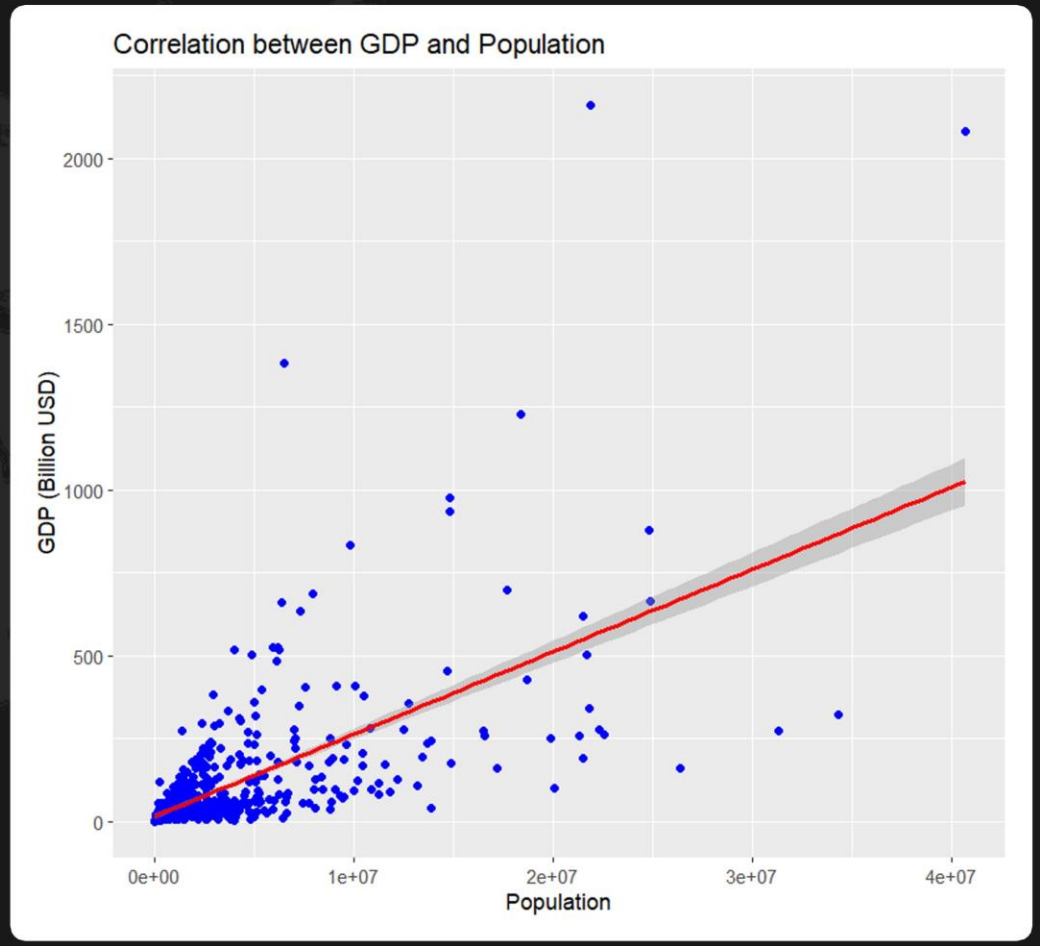
THE SCATTER PLOT WITH THE REGRESSION LINE SHOWS THE RELATIONSHIP BETWEEN GDP AND POPULATION ACROSS CITIES.

THE CORRELATION COEFFICIENT QUANTIFIES THE STRENGTH AND DIRECTION OF THIS RELATIONSHIP.

HELPS DETERMINE IF HIGHER POPULATIONS ARE ASSOCIATED WITH HIGHER GDPS IN CITIES.

```
# Correlation between GDP and population
ggplot(df, aes(x=Metropolitian_Population, y=Estimated_GDP_Billion_USD)) +
  geom_point(color='blue') +
  geom_smooth(method='lm', color='red') +
  labs(title="Correlation between GDP and Population", x="Population", y="GDP (Billion USD)")

# Calculate the correlation coefficient
correlation_coefficient <- cor(df$Metropolitian_Population, df$Estimated_GDP_Billion_USD, use='complete.obs')
correlation_coefficient
```



6. WHAT IS THE AVERAGE GDP BY COUNTRY/REGION?

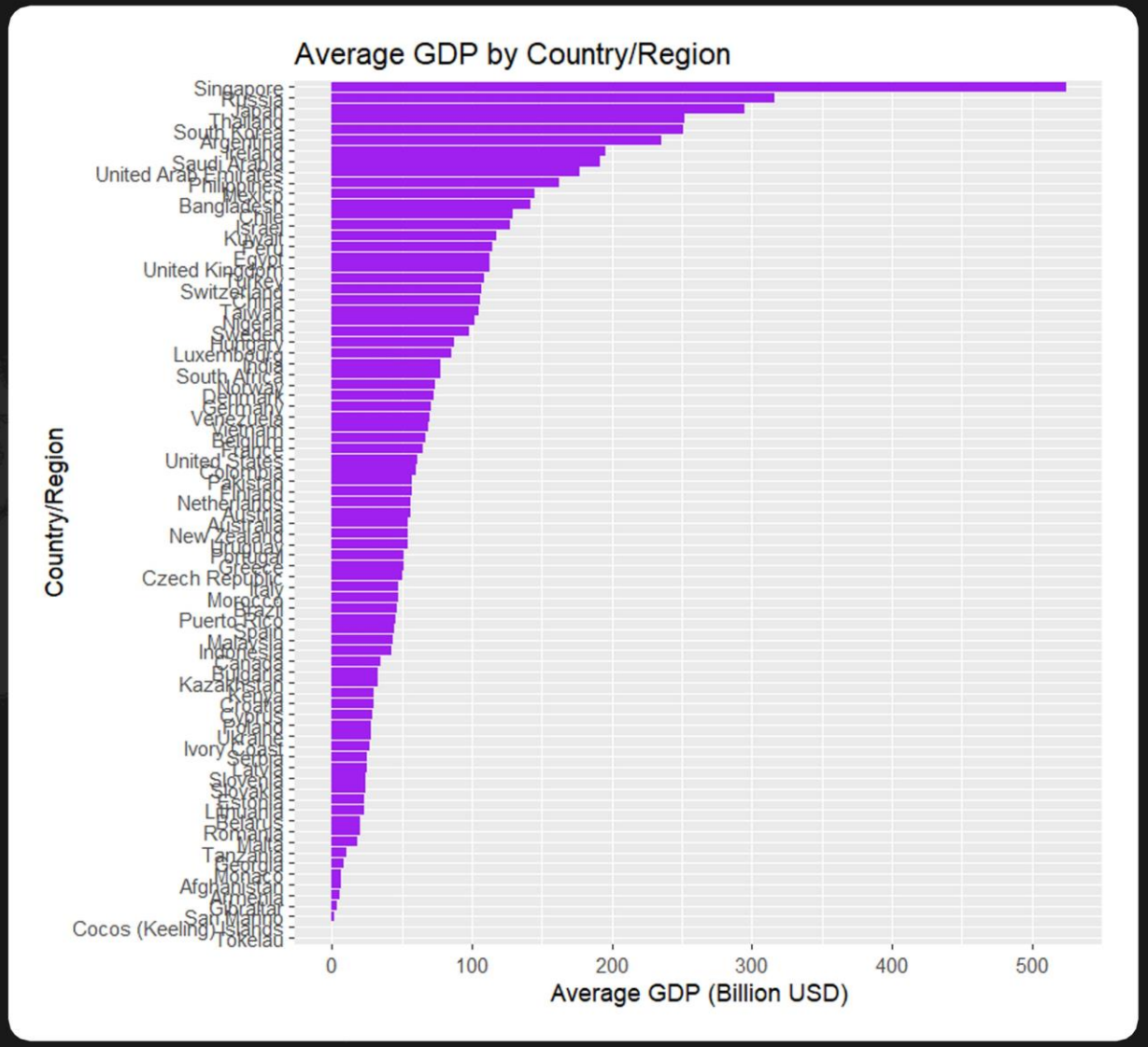
TO ANALYZE THE AVERAGE GDP BY COUNTRY OR REGION, WE WILL GROUP THE DATA BY COUNTRY.REGION AND CALCULATE THE MEAN GDP FOR EACH GROUP. WE WILL VISUALIZE THE RESULTS USING A BAR CHART.

THE BAR CHART DISPLAYS THE AVERAGE GDP FOR EACH COUNTRY OR REGION.

PROVIDES A CLEAR COMPARISON OF THE ECONOMIC PERFORMANCE OF DIFFERENT COUNTRIES/REGIONS.

HIGHLIGHTS THE COUNTRIES/REGIONS WITH THE HIGHEST AND LOWEST AVERAGE GDP.

```r
# Average GDP by country/region
avg_gdp_country <- df %>% group_by(Country.Region) %>% summarize(avg_GDP = mean(Estimated_GDP_Billion_USD, na.rm=TRUE))
ggplot(avg_gdp_country, aes(x=reorder(Country.Region, avg_GDP), y=avg_GDP)) +
  geom_bar(stat='identity', fill='purple') +
  coord_flip() +
  labs(title="Average GDP by Country/Region", x="Country/Region", y="Average GDP (Billion USD)")
```



7. WHAT IS THE AVERAGE POPULATION BY COUNTRY/REGION?

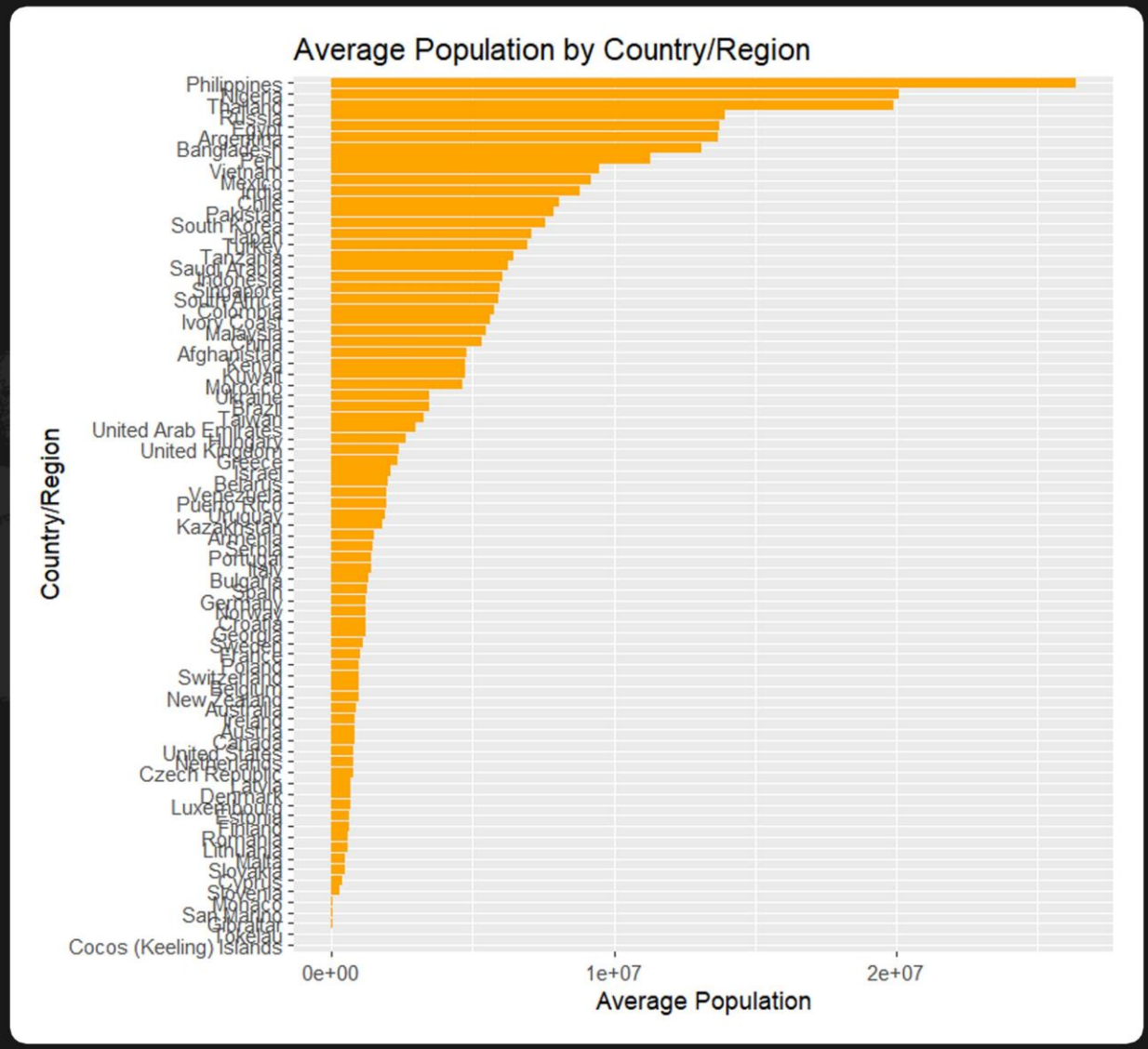
TO ANALYZE THE AVERAGE POPULATION BY COUNTRY OR REGION, WE WILL GROUP THE DATA BY COUNTRY.REGION AND CALCULATE THE MEAN POPULATION FOR EACH GROUP. WE WILL VISUALIZE THE RESULTS USING A BAR CHART.

THE BAR CHART DISPLAYS THE AVERAGE POPULATION FOR EACH COUNTRY OR REGION.

PROVIDES A CLEAR COMPARISON OF THE POPULATION SIZES OF DIFFERENT COUNTRIES/REGIONS.

HIGHLIGHTS THE COUNTRIES/REGIONS WITH THE HIGHEST AND LOWEST AVERAGE POPULATIONS.

```r
# Average population by country/region
avg_pop_country <- df %>% group_by(Country.Region) %>% summarize(avg_Population = mean(Metropolitian_Population, na.rm=TRUE))
ggplot(avg_pop_country, aes(x=reorder(Country.Region, avg_Population), y=avg_Population)) +
  geom_bar(stat='identity', fill='orange') +
  coord_flip() +
  labs(title="Average Population by Country/Region", x="Country/Region", y="Average Population")
```



8. WHAT IS THE TOTAL GDP BY COUNTRY/REGION?

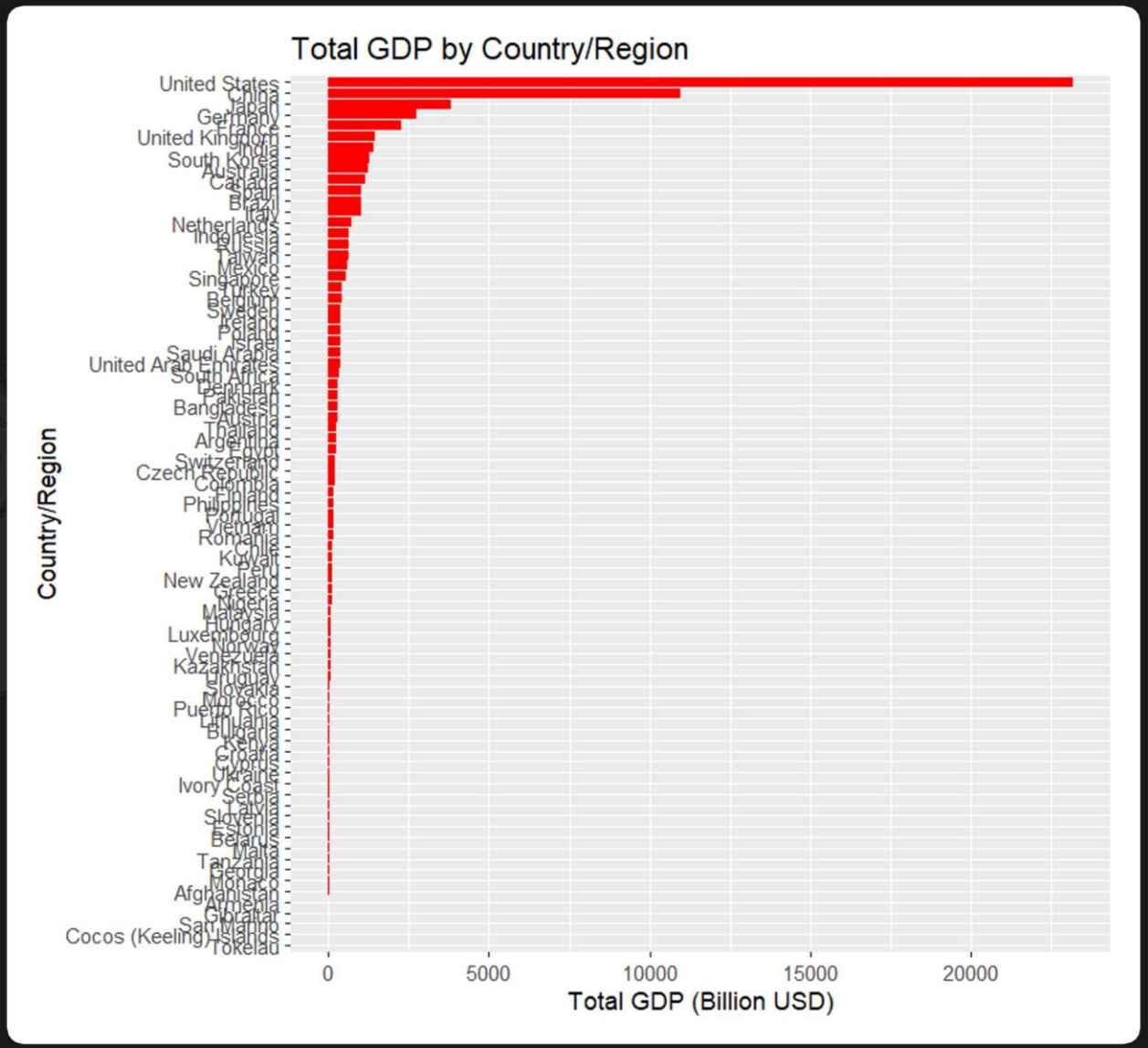
TO ANALYZE THE TOTAL GDP BY COUNTRY OR REGION, WE WILL GROUP THE DATA BY COUNTRY.REGION AND CALCULATE THE SUM OF GDP FOR EACH GROUP. WE WILL VISUALIZE THE RESULTS USING A BAR CHART.

THE BAR CHART DISPLAYS THE TOTAL GDP FOR EACH COUNTRY OR REGION.

PROVIDES A CLEAR COMPARISON OF THE ECONOMIC OUTPUT OF DIFFERENT COUNTRIES/REGIONS.

HIGHLIGHTS THE COUNTRIES/REGIONS WITH THE HIGHEST AND LOWEST TOTAL GDP.

```r
# Total GDP by country/region
total_gdp_country <- df %>% group_by(Country.Region) %>% summarize(total_GDP = sum(Estimated_GDP_Billion_USD, na.rm=TRUE))
ggplot(total_gdp_country, aes(x=reorder(Country.Region, total_GDP), y=total_GDP)) +
  geom_bar(stat='identity', fill='red') +
  coord_flip() +
  labs(title="Total GDP by Country/Region", x="Country/Region", y="Total GDP (Billion USD)")
```



9. WHAT IS THE TOTAL POPULATION BY COUNTRY/REGION?

TO ANALYZE THE TOTAL POPULATION BY COUNTRY OR REGION, WE WILL GROUP THE DATA BY COUNTRY.REGION AND CALCULATE THE SUM OF POPUL ATION FOR EACH GROUP. WE WILL VISUALIZE THE RESULTS USING A BAR CHART.

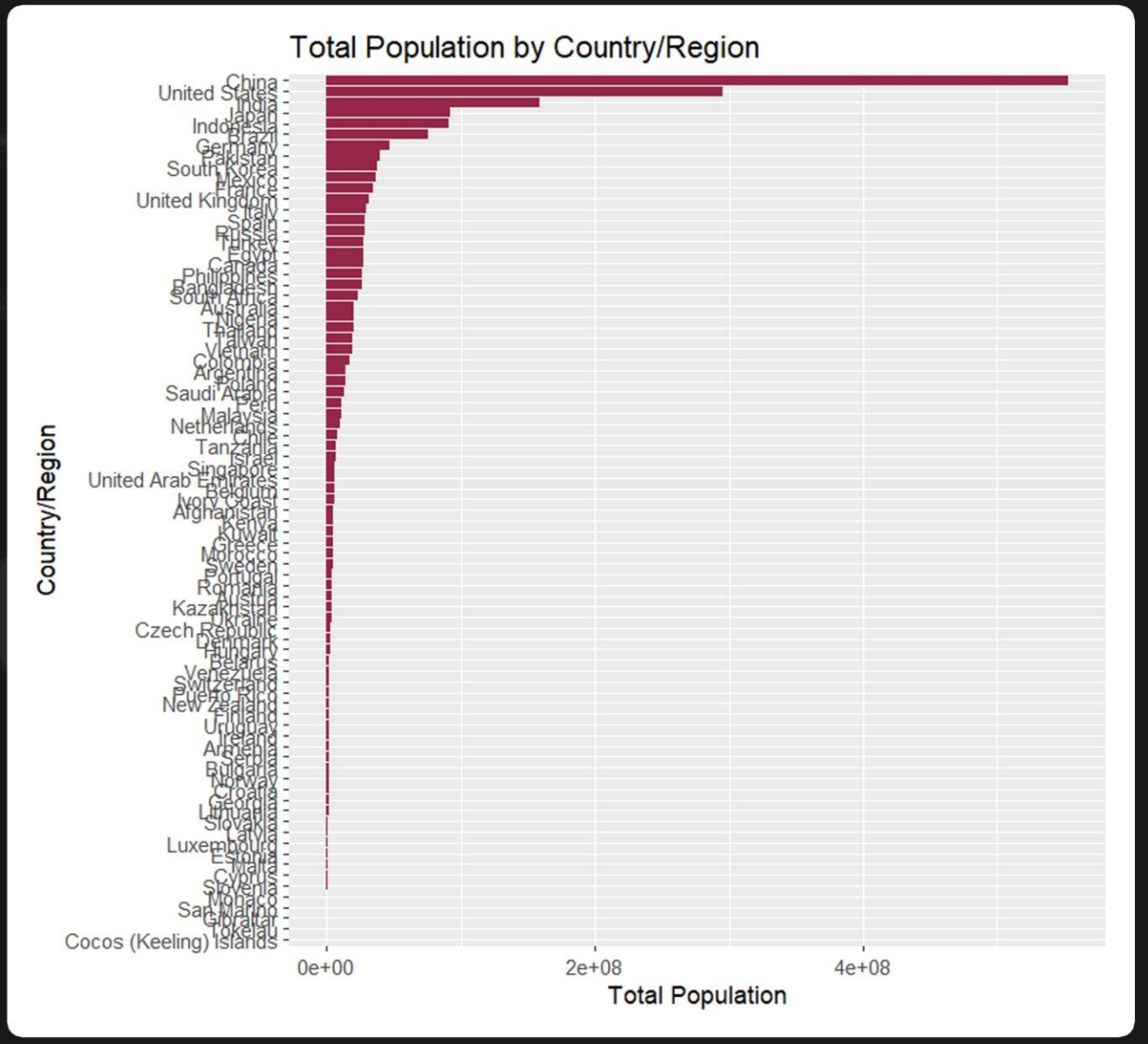
TOTAL_POPULATION_COUNTRY <- DF %>% GROUP_BY(COUNTRY.REGION) %>% SUMMARIZE(TOTAL_POPULATION = SUM(METROPOLITIAN_ POPULATION, NA.RM=TRUE)): GROUPS THE DATAFRAME BY COUNTRY. REGION AND CALCULATES THE TOTAL POPULATION FOR EACH GROUP, EXCLUDING MISSING VALUES.
GGPLOT(TOTAL_POPULATION_COUNTRY, AES(X=REORDER(COUNTRY. REGION, TOTAL_POPULATION), Y=TOTAL_POPULATION)): INITIALIZES THE PLOT WITH THE TOTAL POPULATION DATA GROUPED BY COUNTRY/REGION.
GEOM_BAR(STAT='IDENTITY', FILL='YELLOW'): CREATES A BAR CHART WITH YELLOW BARS REPRESENTING THE TOTAL POPULATION OF EACH COUNTRY/REGION.
COORD_FLIP(): FLIPS THE COORDINATES TO MAKE THE BARS HORIZONTAL FOR BETTER READABILITY.
LABS(TITLE="TOTAL POPULATION BY COUNTRY/REGION", X="COUNTRY /REGION", Y="TOTAL POPULATION"): ADDS A TITLE AND LABELS FOR THE X AND Y AXES.

THE BAR CHART DISPLAYS THE TOTAL POPULATION FOR EACH COUNTRY OR REGION.

PROVIDES A CLEAR COMPARISON OF THE POPULATION SIZES OF DIFFERENT COUNTRIES/REGIONS.

HIGHLIGHTS THE COUNTRIES/REGIONS WITH THE HIGHEST AND LOWEST TOTAL POPULATIONS.

```r
# Total population by country/region
total_population_country <- df %>% group_by(Country.Region) %>% summarize(total_Population = sum(Metropolitian_Population, na.rm=TRUE))
ggplot(total_population_country, aes(x=reorder(Country.Region, total_Population), y=total_Population)) +
  geom_bar(stat='identity', fill='yellow') +
  coord_flip() +
  labs(title="Total Population by Country/Region", x="Country/Region", y="Total Population")
```



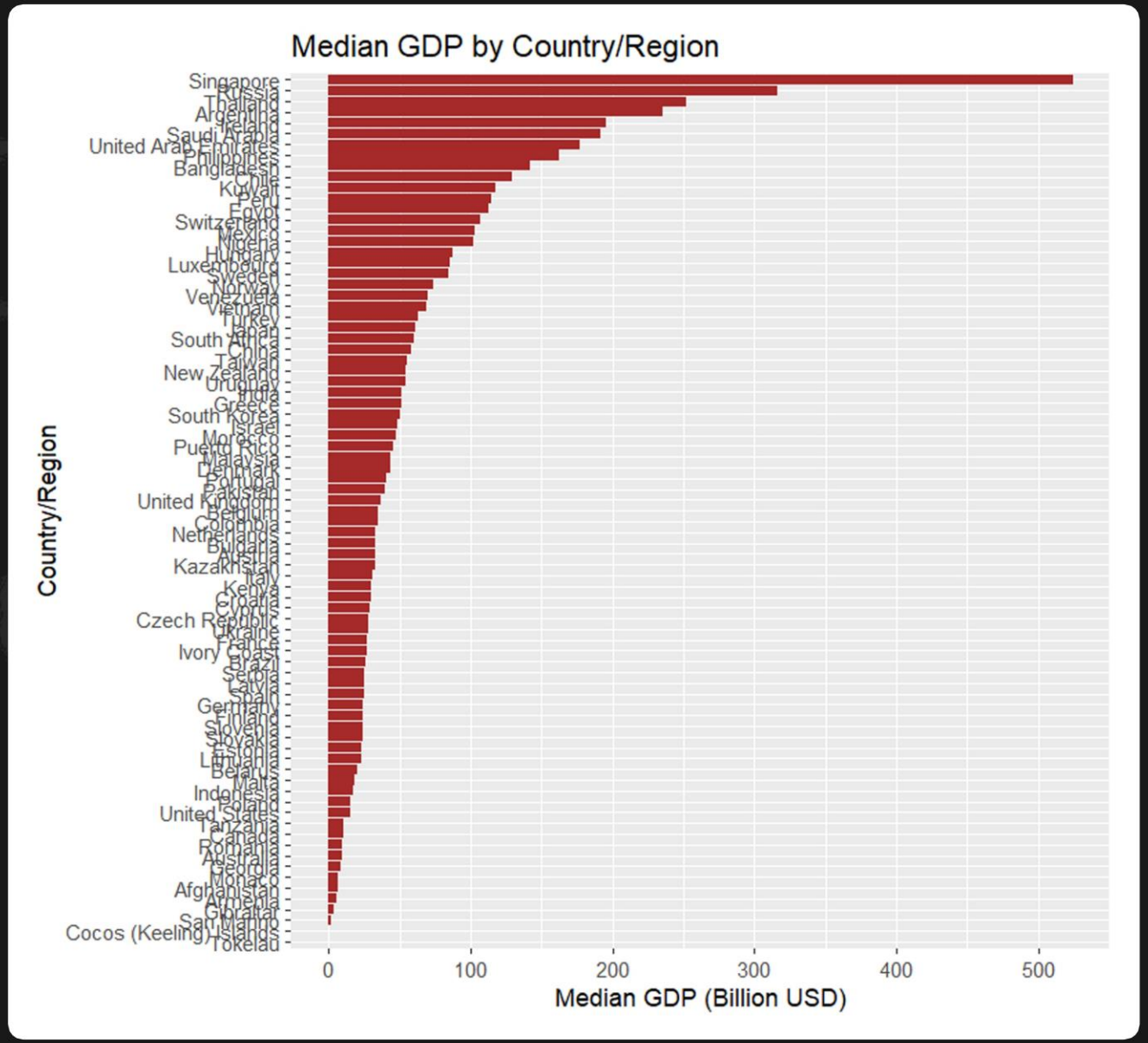
10. WHAT IS THE MEDIAN GDP BY COUNTRY/REGION?

TO ANALYZE THE MEDIAN GDP BY COUNTRY OR REGION, WE WILL GROUP THE DATA BY COUNTRY.REGION AND CALCULATE THE MEDIAN GDP FOR EACH GROUP. WE WILL VISUALIZE THE RESULTS USING A BAR CHART.

MEDIAN_GDP_COUNTRY <- DF %>% GROUP_BY(COUNTRY.REGION) %>% SUMMARIZE(MEDIAN_GDP = MEDIAN(ESTIMATED_GDP_BILLION_USD, NA.RM=TRUE)): GROUPS THE DATAFRAME BY COUNTRY.REGION AND CALCULATES THE MEDIAN GDP FOR EACH GROUP, EXCLUDING MISSING VALUES.
GGPLOT(MEDIAN_GDP_COUNTRY, AES(X=REORDER(COUNTRY.REGION, MEDIAN_GDP), Y=MEDIAN_GDP)): INITIALIZES THE PLOT WITH THE MEDIAN GDP DATA GROUPED BY COUNTRY/REGION.
GEOM_BAR(STAT='IDENTITY', FILL='BROWN'): CREATES A BAR CHART WITH BROWN BARS REPRESENTING THE MEDIAN GDP OF EACH COUNTRY/REGION.
COORD_FLIP(): FLIPS THE COORDINATES TO MAKE THE BARS HORIZONTAL FOR BETTER READABILITY.
LABS(TITLE="MEDIAN GDP BY COUNTRY/REGION", X="COUNTRY/REGION", Y="MEDIAN GDP (BILLION USD)"): ADDS A TITLE AND LABELS FOR THE X AND Y AXES.

THE BAR CHART DISPLAYS THE MEDIAN GDP FOR EACH COUNTRY OR REGION.
PROVIDES A MEASURE OF CENTRAL TENDENCY FOR THE ECONOMIC PERFORMANCE OF DIFFERENT COUNTRIES/REGIONS.
HIGHLIGHTS THE COUNTRIES/REGIONS WITH THE MOST TYPICAL GDP VALUES.

```r
# Median GDP by country/region
median_gdp_country <- df %>% group_by(Country.Region) %>% summarize(median_GDP = median(Estimated_GDP_Billion_USD, na.rm=TRUE)
ggplot(median_gdp_country, aes(x=reorder(Country.Region, median_GDP), y=median_GDP)) +
  geom_bar(stat='identity', fill='brown') +
  coord_flip() +
  labs(title="Median GDP by Country/Regiom", x="Country/Region", y="Median GDP (Billion USD)")
```



11. WHAT IS THE MEDIAN POPULATION BY COUNTRY/REGION?

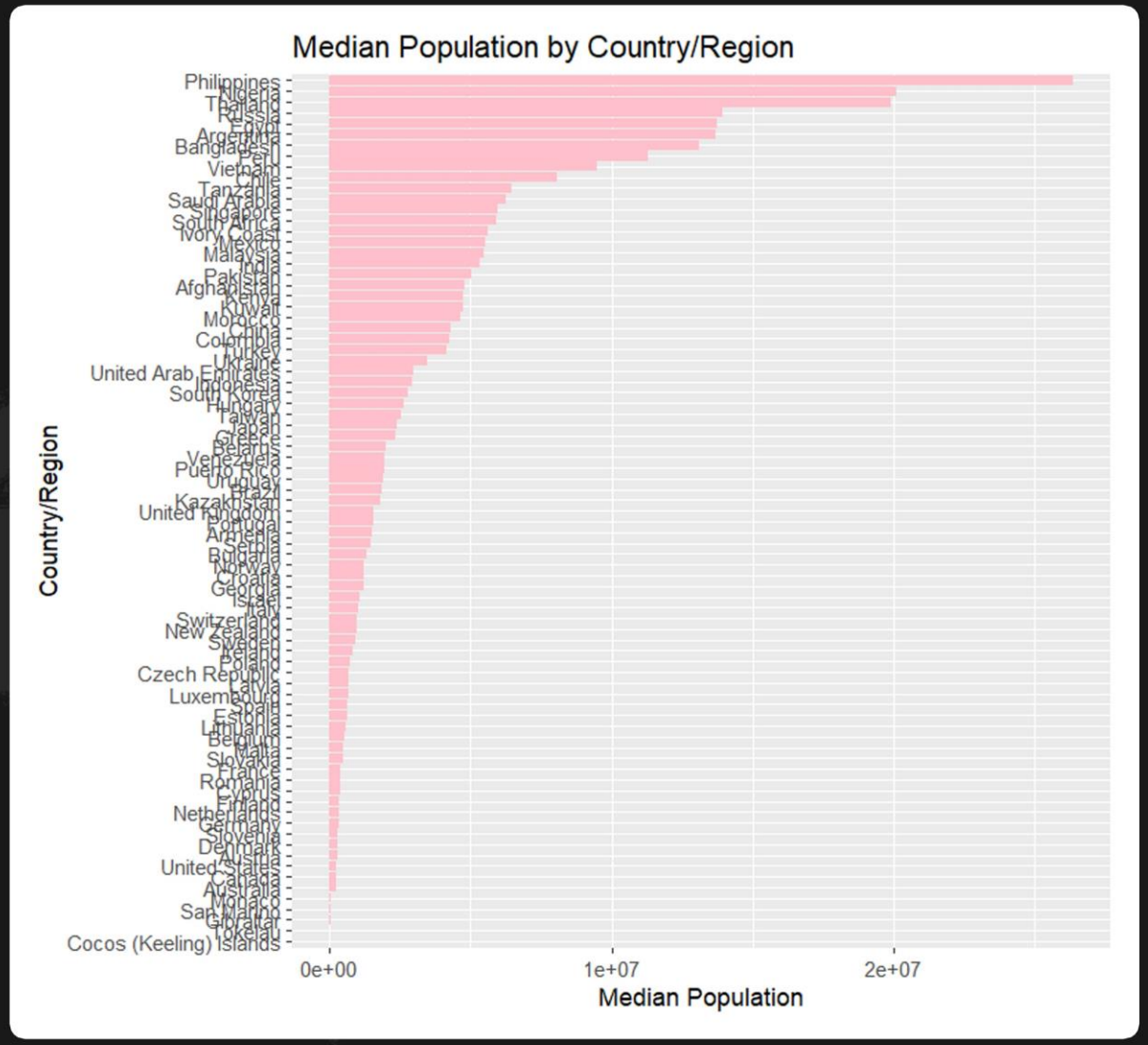
TO ANALYZE THE MEDIAN POPULATION BY COUNTRY OR REGION, WE WILL GROUP THE DATA BY COUNTRY.REGION AND CALCULATE THE MEDIAN POPULATION FOR EACH GROUP. WE WILL VISUALIZE THE RESULTS USING A BAR CHART.

THE BAR CHART DISPLAYS THE MEDIAN POPULATION FOR EACH COUNTRY OR REGION.

PROVIDES A MEASURE OF CENTRAL TENDENCY FOR THE DEMOGRAPHIC DISTRIBUTION OF DIFFERENT COUNTRIES/REGIONS.

HIGHLIGHTS THE COUNTRIES/REGIONS WITH THE MOST TYPICAL POPULATION SIZES.

```r
# Median population by country/region
median_population_country <- df %>% group_by(Country.Region) %>% summarize(median_Population = median(Metropolitian_Population, na.rm=TRUE))
ggplot(median_population_country, aes(x=reorder(Country.Region, median_Population), y=median_Population)) +
  geom_bar(stat='identity', fill='pink') +
  coord_flip() +
  labs(title="Median Population by Country/Region", x="Country/Region", y="Median Population")
```

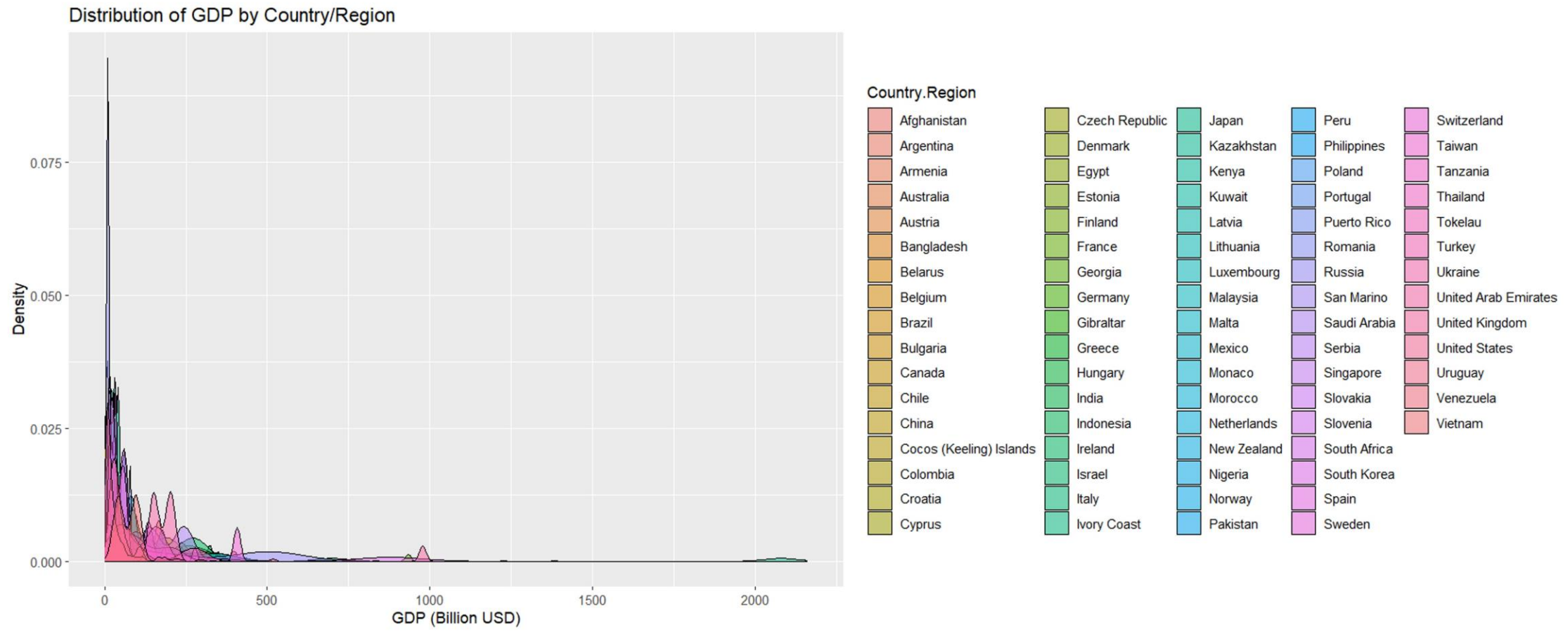


12. WHAT IS THE DISTRIBUTION OF GDP BY COUNTRY/REGION?

TO ANALYZE THE DISTRIBUTION OF GDP BY COUNTRY OR REGION, WE WILL CREATE A DENSITY PLOT TO VISUALIZE THE SPREAD OF GDP VALUES. THIS WILL HELP US UNDERSTAND THE VARIABILITY AND SHAPE OF THE GDP DISTRIBUTION FOR EACH COUNTRY/REGION.

THE DENSITY PLOT ILLUSTRATES THE DISTRIBUTION OF GDP VALUES FOR EACH COUNTRY OR REGION.
HELPS IDENTIFY COUNTRIES/REGIONS WITH DIFFERENT GDP DISTRIBUTIONS AND VARIABILITY.
ENABLES COMPARISON OF THE GDP DISTRIBUTION SHAPES ACROSS COUNTRIES/REGIONS.

```r
# Distribution of GDP by country/region
ggplot(df, aes(x=Estimated_GDP_Billion_USD, fill=Country.Region)) +
  geom_density(alpha=0.5) +
  labs(title="Distribution of GDP by Country/Region", x="GDP (Billion USD)", y="Density")
```

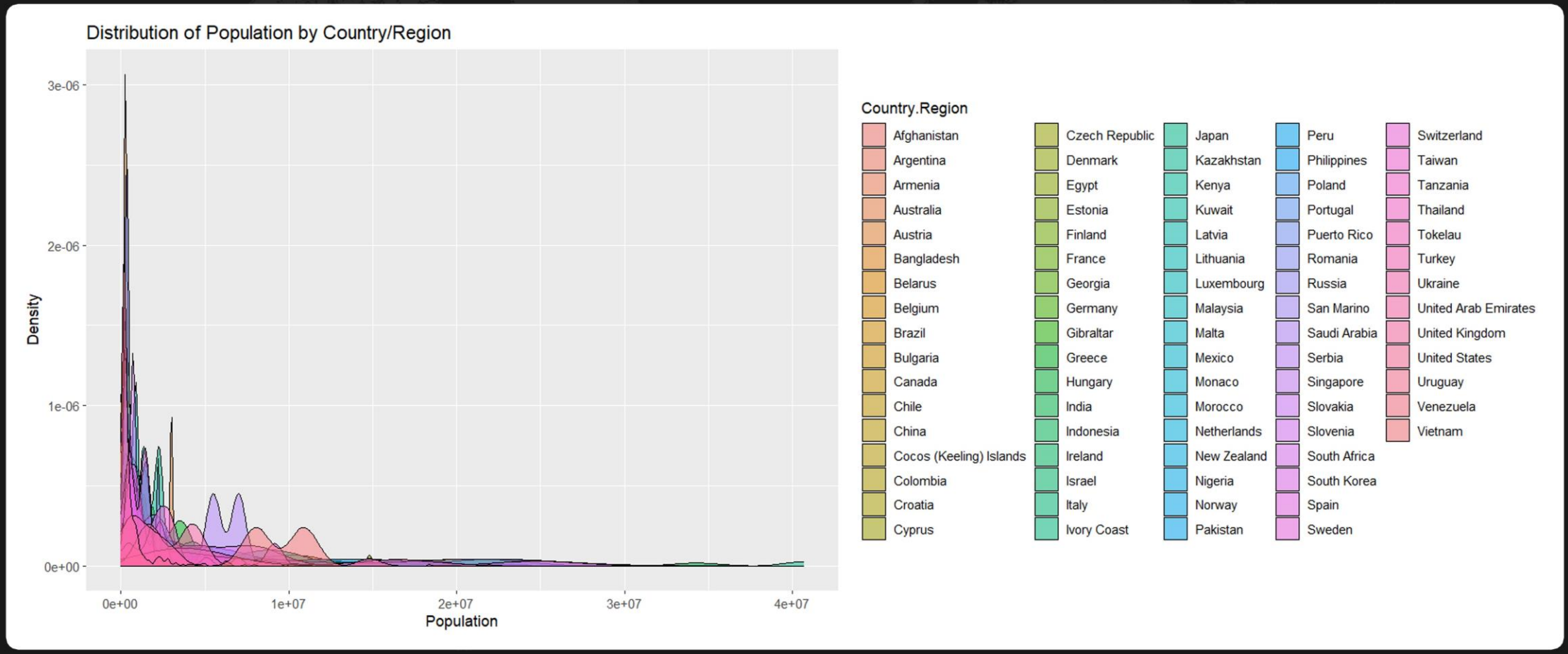


13. WHAT IS THE DISTRIBUTION OF POPULATION BY COUNTRY/REGION?

TO ANALYZE THE DISTRIBUTION OF POPULATION BY COUNTRY OR REGION, WE WILL CREATE A DENSITY PLOT TO VISUALIZE THE SPREAD OF POPULATION VALUES. THIS WILL HELP US UNDERSTAND THE VARIABILITY AND SHAPE OF THE POPULATION DISTRIBUTION FOR EACH COUNTRY/REGION.

THE DENSITY PLOT ILLUSTRATES THE DISTRIBUTION OF POPULATION VALUES FOR EACH COUNTRY OR REGION.
HELPS IDENTIFY COUNTRIES/REGIONS WITH DIFFERENT POPULATION DISTRIBUTIONS AND VARIABILITY.
ENABLES COMPARISON OF THE POPULATION DISTRIBUTION SHAPES ACROSS COUNTRIES/REGIONS.

```r
# Distribution of population by country/region
ggplot(df, aes(x=Metropolitian_Population, fill=Country.Region)) +
  geom_density(alpha=0.5) +
  labs(title="Distribution of Population by Country/Region", x="Population", y="Density")
```



14. WHICH CITIES HAVE A GDP GREATER THAN A SPECIFIC THRESHOLD (E.G., 100 BILLION USD)?

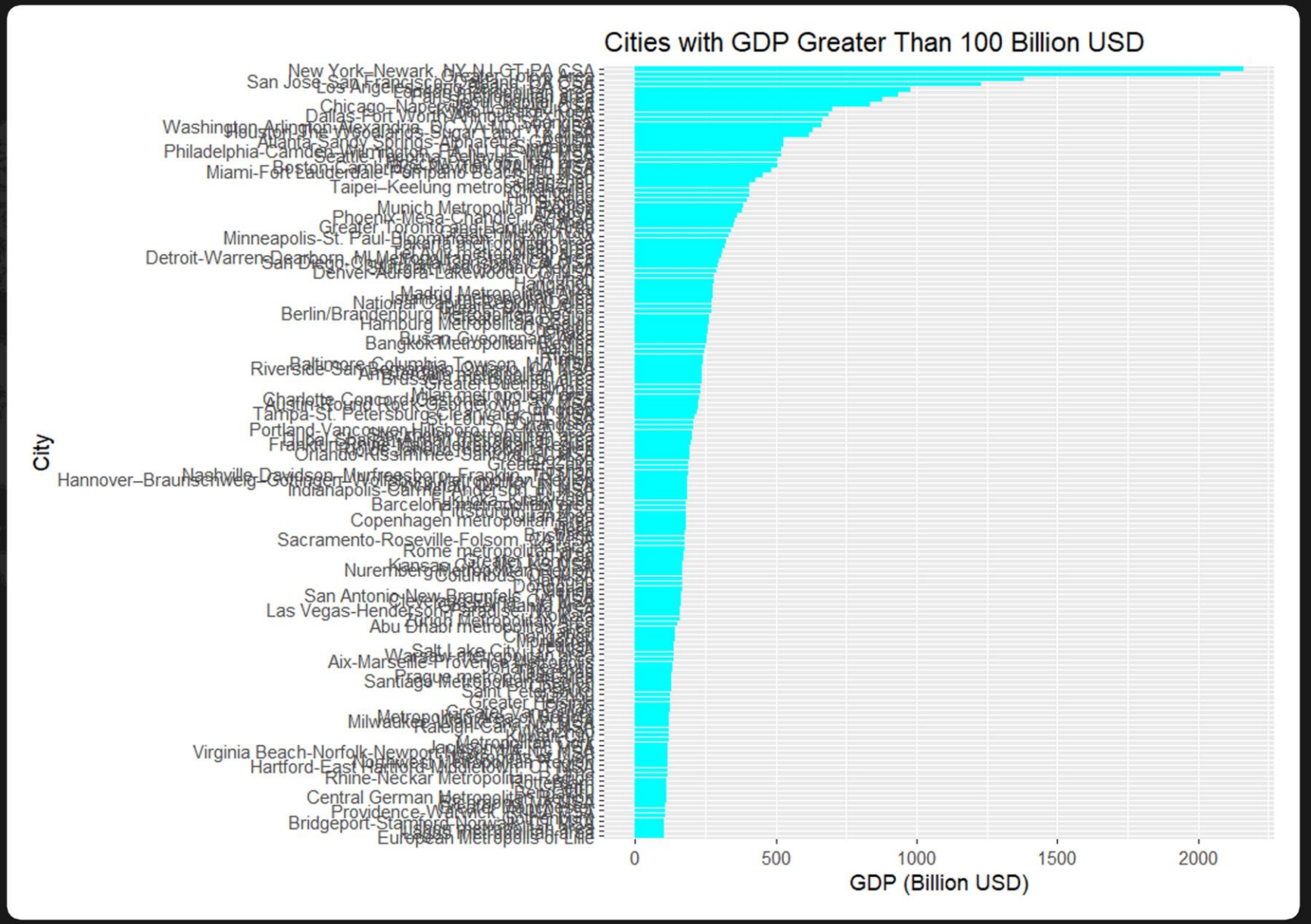
TO IDENTIFY CITIES WITH A GDP GREATER THAN A SPECIFIC THRESHOLD, WE WILL FILTER THE DATAFRAME BASED ON THE THRESHOLD VALUE AND VISUALIZE THE RESULTS USING A BAR CHART.

THE BAR CHART DISPLAYS CITIES WITH A GDP GREATER THAN THE SPECIFIED THRESHOLD.

PROVIDES INSIGHTS INTO THE ECONOMIC STRENGTH OF CITIES SURPASSING THE THRESHOLD.

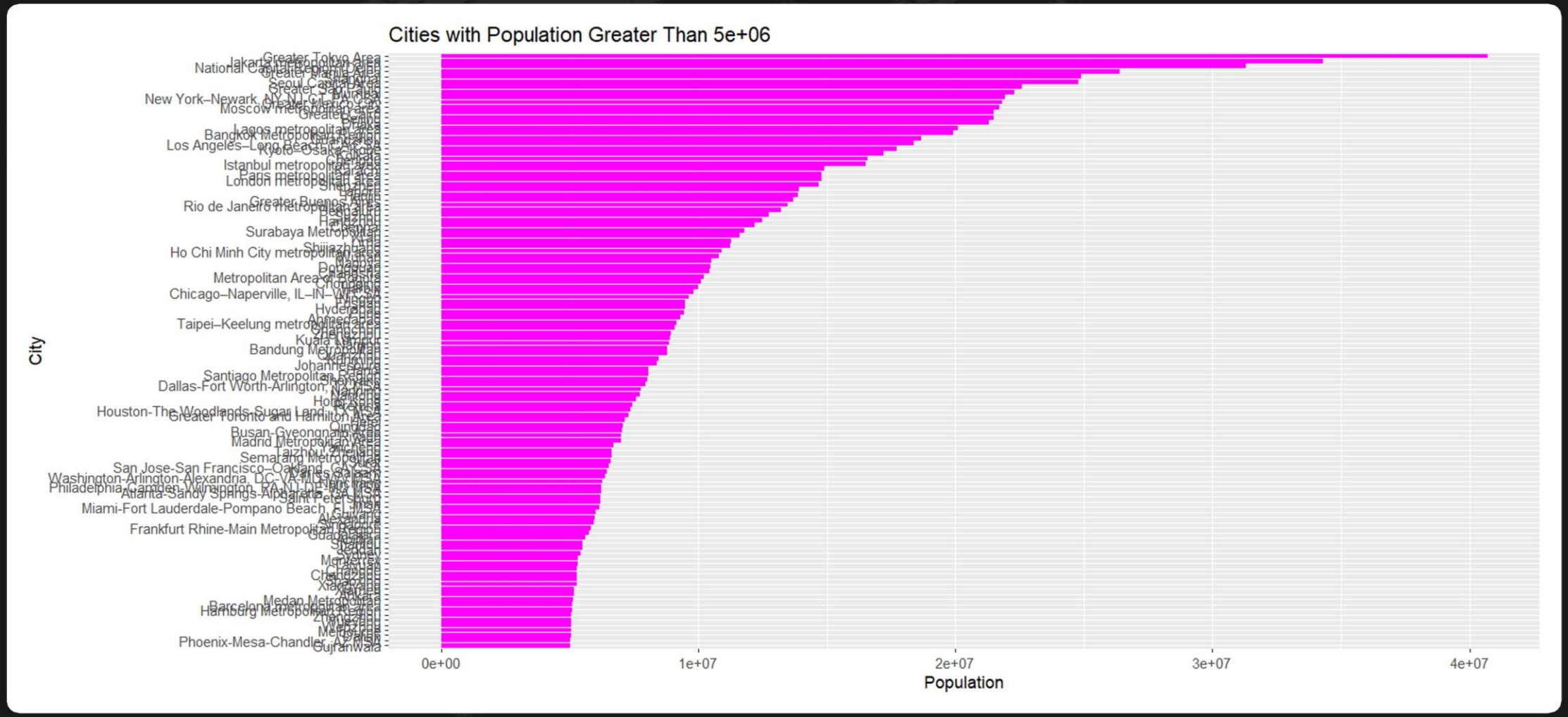
HELPS IDENTIFY KEY ECONOMIC HUBS DRIVING GDP GROWTH.

```
# Cities with GDP greater than a specific threshold
threshold <- 100
high_gdp_cities <- df %>% filter(Estimated_GDP_Billion_USD > threshold)
ggplot(high_gdp_cities, aes(x=reorder(Metropolitan.Area.City, Estimated_GDP_Billion_USD), y=Estimated_GDP_Billion_USD)) +
  geom_bar(stat='identity', fill='cyan') +
  coord_flip() +
  labs(title=paste("Cities with GDP Greater Than", threshold, "Billion USD"), x="City", y="GDP (Billion USD)")
```



15. WHICH CITIES HAVE A POPULATION GREATER THAN A SPECIFIC THRESHOLD (E.G., 5 MILLION)?

TO IDENTIFY CITIES WITH A POPULATION GREATER THAN A SPECIFIC THRESHOLD, WE WILL FILTER THE DATAFRAME BASED ON THE THRESHOLD VALUE AND VISUALIZE THE RESULTS USING A BAR CHART.



CONCLUSION

THE ANALYSIS OF THE CITIES BY GDP DATASET PROVIDES VALUABLE INSIGHTS INTO THE ECONOMIC AND DEMOGRAPHIC CHARACTERISTICS OF METROPOLITAN AREAS ACROSS DIFFERENT REGIONS.

HERE ARE THE KEY FINDINGS:

GDP DISTRIBUTION:
THE DISTRIBUTION OF GDP AMONG CITIES SHOWS SIGNIFICANT VARIATION, WITH A FEW CITIES HAVING EXCEPTIONALLY HIGH GDP VALUES. THIS INDICATES THAT ECONOMIC POWER IS CONCENTRATED IN A SELECT NUMBER OF METROPOLITAN AREAS.
THE TOP 10 CITIES BY GDP INCLUDE GLOBALLY RECOGNIZED ECONOMIC HUBS, EMPHASIZING THEIR ROLE AS MAJOR CENTERS OF ECONOMIC ACTIVITY.

POPULATION DISTRIBUTION:
SIMILAR TO GDP, THE POPULATION DISTRIBUTION AMONG CITIES IS ALSO UNEVEN, WITH SOME CITIES HAVING POPULATIONS IN THE TENS OF MILLIONS WHILE OTHERS ARE MUCH SMALLER.
THE TOP 10 CITIES BY POPULATION ARE PRIMARILY LOCATED IN HIGHLY URBANIZED AND POPULOUS COUNTRIES.

CORRELATION BETWEEN GDP AND POPULATION:
THERE IS A POSITIVE CORRELATION BETWEEN GDP AND POPULATION, INDICATING THAT LARGER CITIES TEND TO HAVE HIGHER ECONOMIC OUTPUT.
THIS RELATIONSHIP UNDERSCORES THE IMPORTANCE OF POPULATION SIZE IN DRIVING ECONOMIC ACTIVITY.

COUNTRY/REGION ANALYSIS:
THE AVERAGE, TOTAL, AND MEDIAN GDP AND POPULATION BY COUNTRY/REGION REVEAL SIGNIFICANT DIFFERENCES. SOME COUNTRIES HAVE CITIES WITH VERY HIGH AVERAGE GDPS, WHILE OTHERS HAVE A LARGER NUMBER OF CITIES WITH LOWER GDPS.
COUNTRIES LIKE THE UNITED STATES AND CHINA HAVE A SUBSTANTIAL NUMBER OF CITIES WITH HIGH GDP AND POPULATION, REFLECTING THEIR LARGE AND DIVERSE URBAN LANDSCAPES.

ECONOMIC CONCENTRATION:
THE ANALYSIS HIGHLIGHTS THE CONCENTRATION OF ECONOMIC POWER IN A FEW MAJOR CITIES. FOR INSTANCE, CITIES WITH GDPS GREATER THAN 100 BILLION USD ARE LIMITED, EMPHASIZING THE DOMINANCE OF CERTAIN METROPOLITAN AREAS IN THE GLOBAL ECONOMY.

THRESHOLD ANALYSIS:
IDENTIFYING CITIES WITH GDP GREATER THAN SPECIFIC THRESHOLDS (E.G., 100 BILLION USD) AND POPULATIONS GREATER THAN CERTAIN THRESHOLDS (E.G., 5 MILLION) HELPS PINPOINT KEY ECONOMIC AND POPULATION CENTERS.

RECOMMENDATIONS:

ECONOMIC DIVERSIFICATION:
POLICYMAKERS SHOULD CONSIDER STRATEGIES TO DIVERSIFY ECONOMIC ACTIVITY BEYOND THE MAJOR METROPOLITAN AREAS TO PROMOTE BALANCED REGIONAL DEVELOPMENT.

URBAN PLANNING:
URBAN PLANNERS SHOULD FOCUS ON SUSTAINABLE GROWTH STRATEGIES TO MANAGE THE POPULATION AND ECONOMIC PRESSURES ON THE LARGEST CITIES.

INVESTMENT IN INFRASTRUCTURE:
INVESTING IN INFRASTRUCTURE IN CITIES WITH HIGH GDP AND POPULATION CAN HELP SUPPORT THEIR CONTINUED GROWTH AND IMPROVE QUALITY OF LIFE FOR RESIDENTS.

SUPPORT FOR EMERGING CITIES:
SUPPORTING EMERGING CITIES WITH POTENTIAL FOR GROWTH CAN HELP REDUCE THE ECONOMIC CONCENTRATION IN A FEW METROPOLITAN AREAS AND PROMOTE MORE EQUITABLE DEVELOPMENT.

THIS ANALYSIS PROVIDES A FOUNDATION FOR FURTHER EXPLORATION INTO THE FACTORS DRIVING ECONOMIC AND POPULATION GROWTH IN METROPOLITAN AREAS, AS WELL AS THE CHALLENGES AND OPPORTUNITIES THESE CITIES FACE IN THE CONTEXT OF GLOBAL ECONOMIC DYNAMICS.


THANK YOU