STA 518 Homework 3

Anto

2023-10-03

# Let’s load some packages to start.

# Load necessary libraries and the gapminder dataset  
library(gapminder)  
library(skimr)  
library(ggplot2)  
library(ggthemes)  
library(RColorBrewer)  
library(forcats)  
library(scales)  
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

# Load the gapminder dataset using data(“gapminder”).

data("gapminder")

# Load the tidyverse, skimr, and ggthemes packages, and explore the gapminder dataset using glimpse() and ?gapminder to learn about the variables.

glimpse(gapminder)

## Rows: 1,704  
## Columns: 6  
## $ country <fct> "Afghanistan", "Afghanistan", "Afghanistan", "Afghanistan", …  
## $ continent <fct> Asia, Asia, Asia, Asia, Asia, Asia, Asia, Asia, Asia, Asia, …  
## $ year <int> 1952, 1957, 1962, 1967, 1972, 1977, 1982, 1987, 1992, 1997, …  
## $ lifeExp <dbl> 28.801, 30.332, 31.997, 34.020, 36.088, 38.438, 39.854, 40.8…  
## $ pop <int> 8425333, 9240934, 10267083, 11537966, 13079460, 14880372, 12…  
## $ gdpPercap <dbl> 779.4453, 820.8530, 853.1007, 836.1971, 739.9811, 786.1134, …

?gapminder

## starting httpd help server ... done

Quantitative variables:

year: The year of data. lifeExp: Life expectancy at birth. pop: Population. gdpPercap: GDP per capita.

Categorical variables:

country: Country name. continent: Continent to which the country belongs.

# Use skim() from the skimr package to further explore the data set and any missing data patterns. How many missing values are there for this data set?

# Use skim() to check for missing values  
skim(gapminder)

Data summary

|  |  |
| --- | --- |
| Name | gapminder |
| Number of rows | 1704 |
| Number of columns | 6 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| factor | 2 |
| numeric | 4 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: factor**

| skim\_variable | n\_missing | complete\_rate | ordered | n\_unique | top\_counts |
| --- | --- | --- | --- | --- | --- |
| country | 0 | 1 | FALSE | 142 | Afg: 12, Alb: 12, Alg: 12, Ang: 12 |
| continent | 0 | 1 | FALSE | 5 | Afr: 624, Asi: 396, Eur: 360, Ame: 300 |

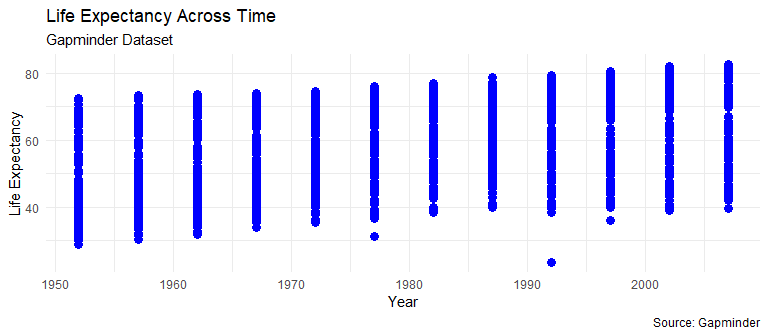
**Variable type: numeric**

| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| year | 0 | 1 | 1979.50 | 17.27 | 1952.00 | 1965.75 | 1979.50 | 1993.25 | 2007.0 | ▇▅▅▅▇ |
| lifeExp | 0 | 1 | 59.47 | 12.92 | 23.60 | 48.20 | 60.71 | 70.85 | 82.6 | ▁▆▇▇▇ |
| pop | 0 | 1 | 29601212.32 | 106157896.74 | 60011.00 | 2793664.00 | 7023595.50 | 19585221.75 | 1318683096.0 | ▇▁▁▁▁ |
| gdpPercap | 0 | 1 | 7215.33 | 9857.45 | 241.17 | 1202.06 | 3531.85 | 9325.46 | 113523.1 | ▇▁▁▁▁ |

There are no missing values.

# Using ggplot2, create a scatterplot showing life expectancy across time, adding descriptive labels of the axes and overall plot. What trend do you notice?

# Scatterplot of life expectancy across time  
ggplot(gapminder, aes(x = year, y = lifeExp)) +  
 geom\_point(color = "blue", size = 3) +  
 labs(x = "Year", y = "Life Expectancy",  
 title = "Life Expectancy Across Time",  
 subtitle = "Gapminder Dataset",  
 caption = "Source: Gapminder") +  
 theme\_minimal()

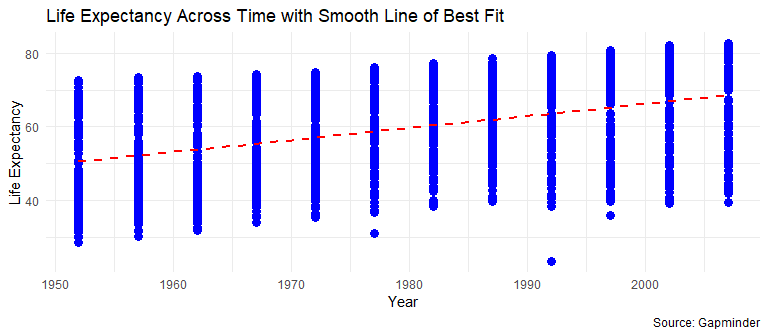


Overall, the trend is globally increasing.

# Recreate the plot of life expectancy across time, this time adding an additional smooth line of best fit through the data using geom\_smooth(). Include the option se = FALSE in geom\_smooth() to suppress the standard error bands around the smooth curves.

ggplot(gapminder, aes(x = year, y = lifeExp)) +  
 geom\_point(color = "blue", size = 3) +  
 geom\_smooth(method = "lm", se = FALSE, color = "red", linetype = "dashed") +  
 labs(x = "Year", y = "Life Expectancy",  
 title = "Life Expectancy Across Time with Smooth Line of Best Fit",  
 caption = "Source: Gapminder") +  
 theme\_minimal()

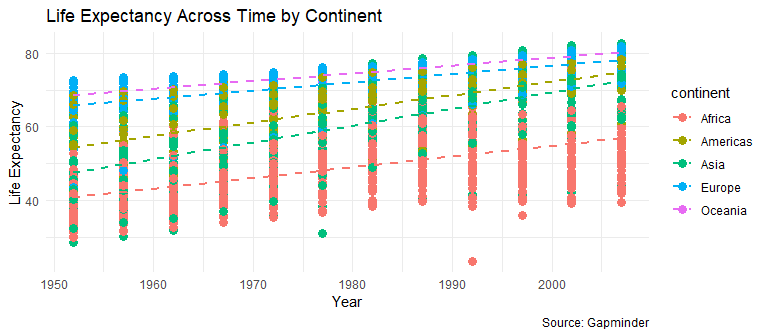
## `geom\_smooth()` using formula = 'y ~ x'



# Color the points based on which continent the points are representing, including smoothed lines through the points using geom\_smooth() so that the lines are still colored by continent as well.

ggplot(gapminder, aes(x = year, y = lifeExp, color = continent)) +  
 geom\_point(size = 3) +  
 geom\_smooth(method = "lm", se = FALSE, aes(group = continent), linetype = "dashed") +  
 labs(x = "Year", y = "Life Expectancy",  
 title = "Life Expectancy Across Time by Continent",  
 caption = "Source: Gapminder") +  
 theme\_minimal()

## `geom\_smooth()` using formula = 'y ~ x'



# Which continent / region has the highest life expectancy on average?.

# Calculate average life expectancy by continent  
average\_life\_expectancy <- gapminder %>%  
 group\_by(continent) %>%  
 summarise(avg\_life\_exp = mean(lifeExp, na.rm = TRUE))  
  
# View the average life expectancy by continent  
print(average\_life\_expectancy)

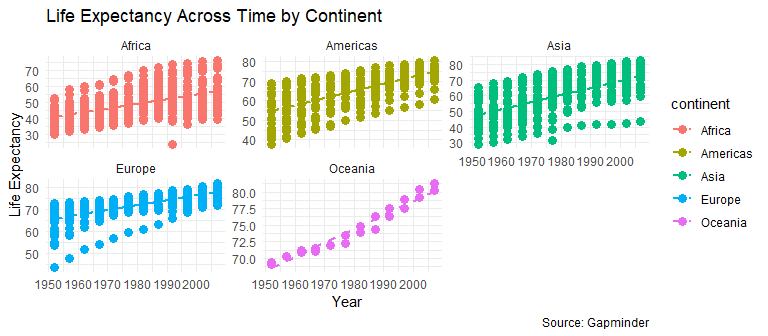
## # A tibble: 5 × 2  
## continent avg\_life\_exp  
## <fct> <dbl>  
## 1 Africa 48.9  
## 2 Americas 64.7  
## 3 Asia 60.1  
## 4 Europe 71.9  
## 5 Oceania 74.3

Oceania has the highest life expectancy on average.

# Extend the plot from the previous part by faceting by the continent associated with each point so that each continent has its own column.

ggplot(gapminder, aes(x = year, y = lifeExp, color = continent)) +  
 geom\_point(size = 3) +  
 geom\_smooth(method = "lm", se = FALSE, aes(group = continent), linetype = "dashed") +  
 labs(x = "Year", y = "Life Expectancy",  
 title = "Life Expectancy Across Time by Continent",  
 caption = "Source: Gapminder") +  
 facet\_wrap(~continent, scales = "free\_y") +  
 theme\_minimal()

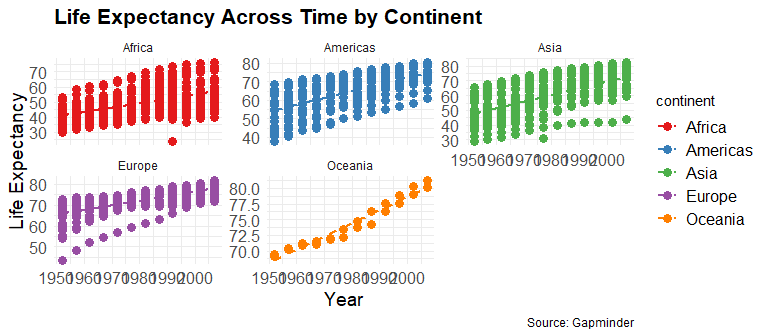
## `geom\_smooth()` using formula = 'y ~ x'



# Modify the colors for each continent to be color-blind friendly using this palette, and separately customize the overall theme of the plot by specifying a complete theme of your choice.

# Define color-blind friendly palette  
color\_palette <- brewer.pal(n = 8, name = "Set1")  
  
# Create the plot with points colored by continent, smoothed lines, and faceted by continent  
ggplot(gapminder, aes(x = year, y = lifeExp, color = continent)) +  
 geom\_point(size = 3) +  
 geom\_smooth(method = "lm", se = FALSE, aes(group = continent), linetype = "dashed") +  
 labs(x = "Year", y = "Life Expectancy",  
 title = "Life Expectancy Across Time by Continent",  
 caption = "Source: Gapminder") +  
 facet\_wrap(~continent, scales = "free\_y") +  
 scale\_color\_manual(values = color\_palette) + # Set color palette  
 theme\_minimal() + # You can customize the theme further if desired  
 theme(  
 plot.title = element\_text(size = 16, face = "bold"),  
 axis.title.x = element\_text(size = 14),  
 axis.title.y = element\_text(size = 14),  
 axis.text.x = element\_text(size = 12),  
 axis.text.y = element\_text(size = 12),  
 legend.position = "right",  
 legend.text = element\_text(size = 12)  
 )

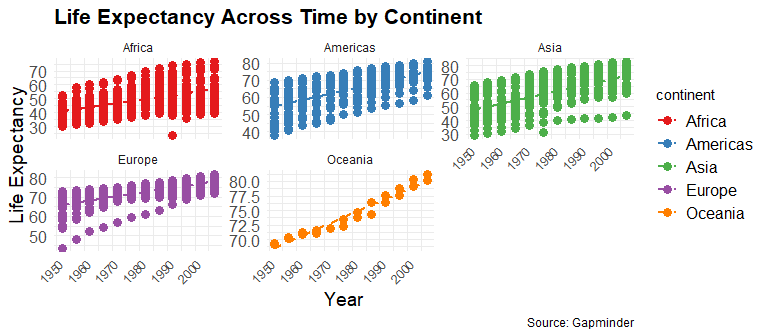
## `geom\_smooth()` using formula = 'y ~ x'



# Rotate the labels on the x-axis 45 degrees by adding a theme() layer with the appropriate option.

ggplot(gapminder, aes(x = year, y = lifeExp, color = continent)) +  
 geom\_point(size = 3) +  
 geom\_smooth(method = "lm", se = FALSE, aes(group = continent), linetype = "dashed") +  
 labs(x = "Year", y = "Life Expectancy",  
 title = "Life Expectancy Across Time by Continent",  
 caption = "Source: Gapminder") +  
 facet\_wrap(~continent, scales = "free\_y") +  
 scale\_color\_manual(values = color\_palette) + # Set color palette  
 theme\_minimal() + # You can customize the theme further if desired  
 theme(  
 axis.text.x = element\_text(angle = 45, hjust = 1), # Rotate x-axis labels by 45 degrees  
 plot.title = element\_text(size = 16, face = "bold"),  
 axis.title.x = element\_text(size = 14),  
 axis.title.y = element\_text(size = 14),  
 axis.text.y = element\_text(size = 12),  
 legend.position = "right",  
 legend.text = element\_text(size = 12)  
 )

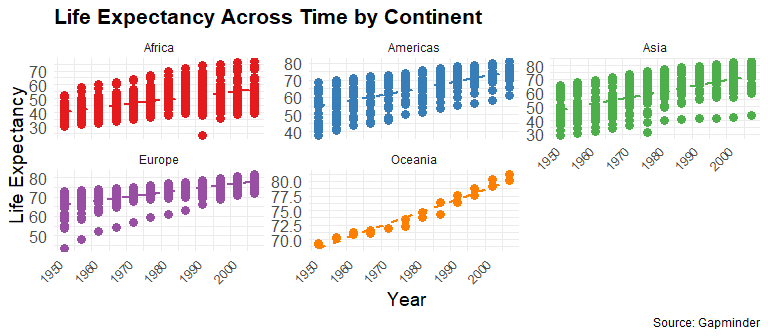
## `geom\_smooth()` using formula = 'y ~ x'



# Specify an argument in the theme() function to suppress the legend.

ggplot(gapminder, aes(x = year, y = lifeExp, color = continent)) +  
 geom\_point(size = 3) +  
 geom\_smooth(method = "lm", se = FALSE, aes(group = continent), linetype = "dashed") +  
 labs(x = "Year", y = "Life Expectancy",  
 title = "Life Expectancy Across Time by Continent",  
 caption = "Source: Gapminder") +  
 facet\_wrap(~continent, scales = "free\_y") +  
 scale\_color\_manual(values = color\_palette) + # Set color palette  
 theme\_minimal() + # You can customize the theme further if desired  
 theme(  
 axis.text.x = element\_text(angle = 45, hjust = 1), # Rotate x-axis labels by 45 degrees  
 legend.position = "none", # Suppress the legend  
 plot.title = element\_text(size = 16, face = "bold"),  
 axis.title.x = element\_text(size = 14),  
 axis.title.y = element\_text(size = 14),  
 axis.text.y = element\_text(size = 12)  
 )

## `geom\_smooth()` using formula = 'y ~ x'

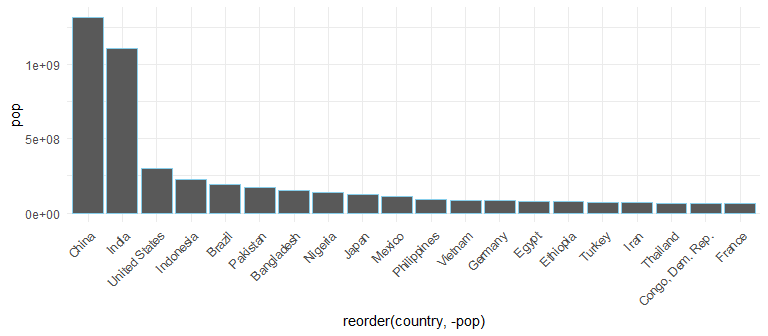


# Use the following code to create a data set called gapminder2007 which contains the top 20 most populated countries in 2007.

# Create the gapminder2007 dataset  
gapminder2007 <- gapminder %>% dplyr::filter(year == 2007) %>% slice\_max(pop, n = 20)

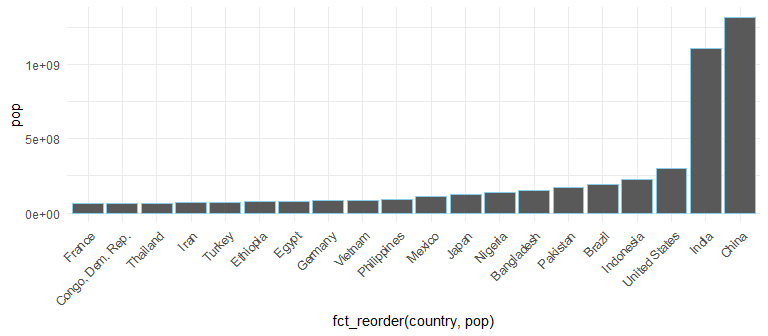
# First, create a bar plot displaying the population of each country using the gapminder2007 dataset using geom\_col(). Note that the country names will be overlapping, but we will fix this later.

ggplot(gapminder2007, aes(x = reorder(country, -pop), y = pop)) +  
 geom\_col(color = "skyblue") +  
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))



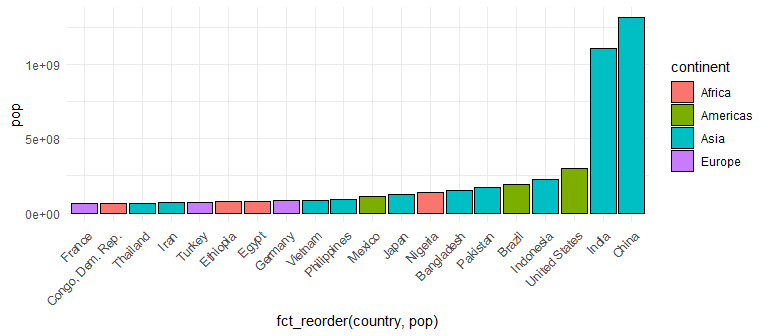
# In a new code chunk, modify the plot in the previous part so that the bars are sorted based on height using the fct\_reorder(country, pop) for the x aesthetic.

ggplot(gapminder2007, aes(x = fct\_reorder(country, pop), y = pop)) +  
 geom\_col(color = "skyblue") +  
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))



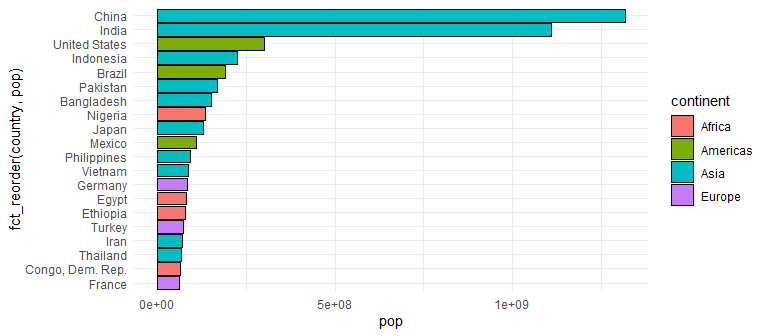
# In another new code chunk, modify the plot so that the color inside of the bars displays which continent each bar represents as well using the fill aesthetic, and change the outline of all bars in the plot to be black by manually setting the color aesthetic.

ggplot(gapminder2007, aes(x = fct\_reorder(country, pop), y = pop, fill = continent)) +  
 geom\_col(color = "black") +  
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))



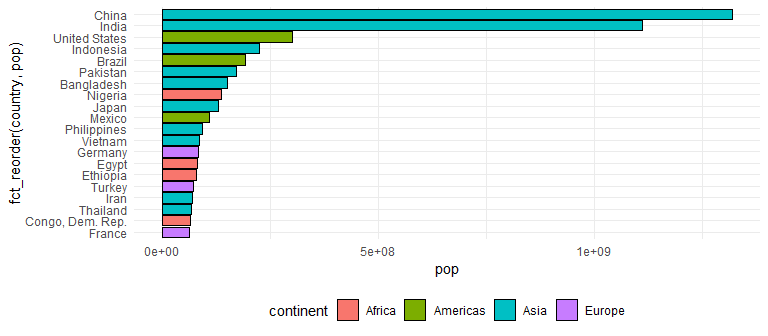
# Use the coord\_flip() function to make the barchart a horizontal bar chart rather than a vertical one to fix the issue of the country names overlapping.

ggplot(gapminder2007, aes(x = fct\_reorder(country, pop), y = pop, fill = continent)) +  
 geom\_col(color = "black") +  
 theme\_minimal() +  
 theme(axis.text.y = element\_text(angle = 0, hjust = 1)) +  
 coord\_flip()



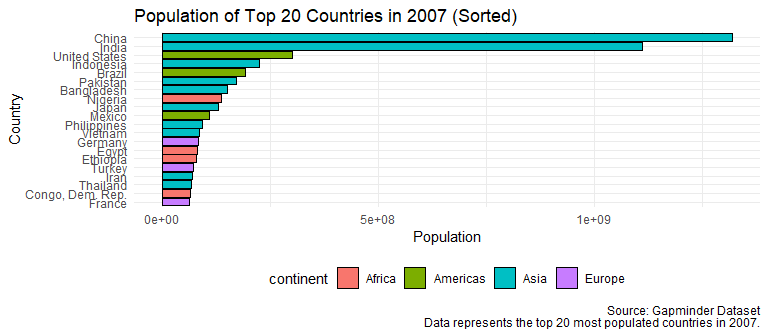
# Move the legend to below the plot (to the “bottom”).

ggplot(gapminder2007, aes(x = fct\_reorder(country, pop), y = pop, fill = continent)) +  
 geom\_col(color = "black") +  
 theme\_minimal() +  
 theme(axis.text.y = element\_text(angle = 0, hjust = 1)) +  
 coord\_flip() +  
 theme(legend.position = "bottom") # Move the legend to the bottom



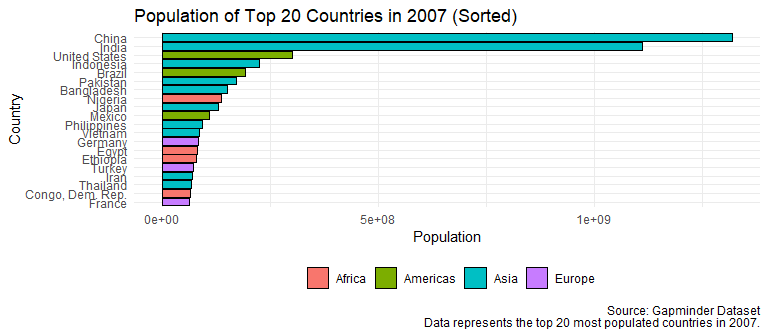
# Add descriptive labels for the axes, title, and a caption below the plot.

ggplot(gapminder2007, aes(x = fct\_reorder(country, pop), y = pop, fill = continent)) +  
 geom\_col(color = "black") +  
 labs(x = "Country", y = "Population",   
 title = "Population of Top 20 Countries in 2007 (Sorted)",  
 caption = "Source: Gapminder Dataset\nData represents the top 20 most populated countries in 2007.") +  
 theme\_minimal() +  
 theme(axis.text.y = element\_text(angle = 0, hjust = 1)) +  
 coord\_flip() +  
 theme(legend.position = "bottom") # Move the legend to the bottom



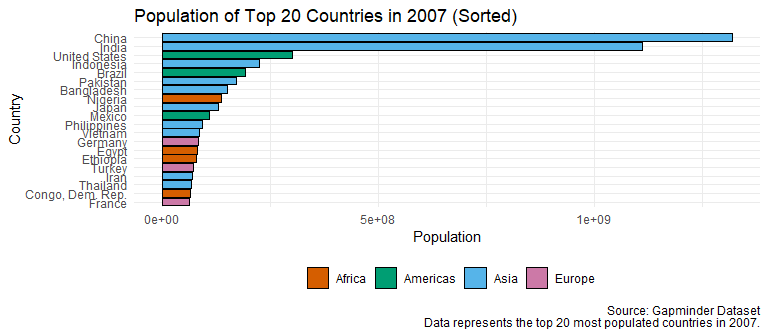
# Add the option legend.title = element\_blank() to the theme() function to remove the legend title.

ggplot(gapminder2007, aes(x = fct\_reorder(country, pop), y = pop, fill = continent)) +  
 geom\_col(color = "black") +  
 labs(x = "Country", y = "Population",   
 title = "Population of Top 20 Countries in 2007 (Sorted)",  
 caption = "Source: Gapminder Dataset\nData represents the top 20 most populated countries in 2007.") +  
 theme\_minimal() +  
 theme(axis.text.y = element\_text(angle = 0, hjust = 1),  
 legend.position = "bottom",  
 legend.title = element\_blank()) + # Remove the legend title  
 coord\_flip()



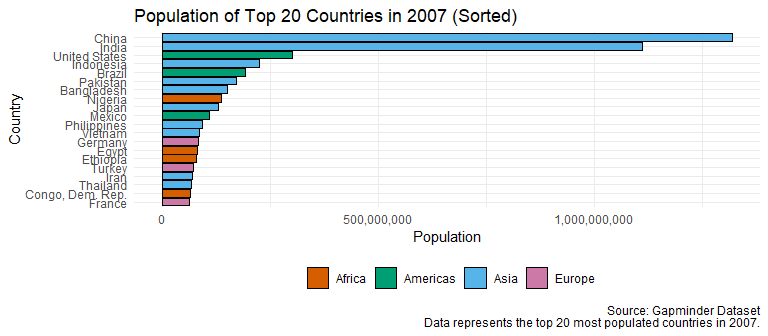
# Use color-blind friendly colors by adding a scale\_fill\_manual() layer using the code below:

# Create a color-blind friendly palette  
color\_palette <- c("#D55E00", "#009E73", "#56B4E9", "#CC79A7")  
  
# Create a horizontal bar plot with bars sorted based on height, colored by continent, and black outline  
ggplot(gapminder2007, aes(x = fct\_reorder(country, pop), y = pop, fill = continent)) +  
 geom\_col(color = "black") +  
 labs(x = "Country", y = "Population",   
 title = "Population of Top 20 Countries in 2007 (Sorted)",  
 caption = "Source: Gapminder Dataset\nData represents the top 20 most populated countries in 2007.") +  
 scale\_fill\_manual(values = color\_palette) + # Set color palette for continents  
 theme\_minimal() +  
 theme(axis.text.y = element\_text(angle = 0, hjust = 1),  
 legend.position = "bottom",  
 legend.title = element\_blank()) + # Remove the legend title  
 coord\_flip()



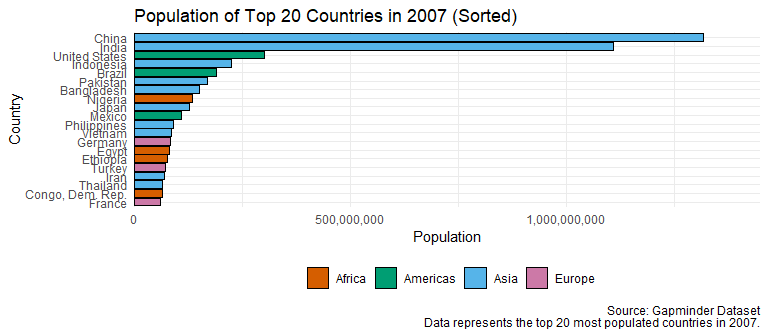
# Display commas in the population numbers rather than scientific notation by adding a scale\_y\_continuous() layer using the code below. Note that the scales package will need to be installed to do this.

color\_palette <- c("#D55E00", "#009E73", "#56B4E9", "#CC79A7")  
  
# Create a horizontal bar plot with bars sorted based on height, colored by continent, and black outline  
ggplot(gapminder2007, aes(x = fct\_reorder(country, pop), y = pop, fill = continent)) +  
 geom\_col(color = "black") +  
 labs(x = "Country", y = "Population",   
 title = "Population of Top 20 Countries in 2007 (Sorted)",  
 caption = "Source: Gapminder Dataset\nData represents the top 20 most populated countries in 2007.") +  
 scale\_fill\_manual(values = color\_palette) + # Set color palette for continents  
 theme\_minimal() +  
 theme(axis.text.y = element\_text(angle = 0, hjust = 1),  
 legend.position = "bottom",  
 legend.title = element\_blank()) + # Remove the legend title  
 scale\_y\_continuous(labels = scales::comma) + # Display commas in population numbers  
 coord\_flip()



# Remove the excess space in between the bars and the axis by specifying the expand argument inside of the scale\_y\_continuous() layer using expand = expansion(mult = c(0, .1)).

ggplot(gapminder2007, aes(x = fct\_reorder(country, pop), y = pop, fill = continent)) +  
 geom\_col(color = "black") +  
 labs(x = "Country", y = "Population",   
 title = "Population of Top 20 Countries in 2007 (Sorted)",  
 caption = "Source: Gapminder Dataset\nData represents the top 20 most populated countries in 2007.") +  
 scale\_fill\_manual(values = color\_palette) + # Set color palette for continents  
 theme\_minimal() +  
 theme(axis.text.y = element\_text(angle = 0, hjust = 1),  
 legend.position = "bottom",  
 legend.title = element\_blank()) + # Remove the legend title  
 scale\_y\_continuous(labels = scales::comma, expand = expansion(mult = c(0, .1))) + # Remove excess space  
 coord\_flip()



# Adding custom theme to a plot.

# Create a color-blind friendly palette  
color\_palette <- c("#D55E00", "#009E73", "#56B4E9", "#CC79A7")  
  
# Create a custom theme from the ggthemes package  
custom\_theme <- theme\_economist() +  
 theme(legend.position = "bottom") # Ensure the legend stays at the bottom  
  
# Create a horizontal bar plot with bars sorted based on height, colored by continent, and black outline  
ggplot(gapminder2007, aes(x = fct\_reorder(country, pop), y = pop, fill = continent)) +  
 geom\_col(color = "black") +  
 labs(x = "Country", y = "Population",   
 title = "Population of Top 20 Countries in 2007 (Sorted)",  
 caption = "Source: Gapminder Dataset\nData represents the top 20 most populated countries in 2007.") +  
 scale\_fill\_manual(values = color\_palette) + # Set color palette for continents  
 theme\_minimal() +  
 theme(axis.text.y = element\_text(angle = 0, hjust = 1),  
 legend.position = "bottom",  
 legend.title = element\_blank()) + # Remove the legend title  
 scale\_y\_continuous(labels = scales::comma, expand = expansion(mult = c(0, .1))) + # Remove excess space  
 coord\_flip() +  
 custom\_theme # Apply the custom theme

