Stat-5405 Quiz on Data Visualization

Pooja Raj Lakshmi | Student ID: 33251787

2025-09-03

Introduction

The dataset `international-tourist-arrivals-by-region-of-origin.csv` contains information about international tourist trips by region of departure. These trips include visitors who arrive from abroad and stay overnight.
In this analysis, we aim to:

1. Plot international tourist arrivals by region and year
2. Interpret patterns in the data
3. Discuss the usefulness of applying a logarithmic transformation

Data Importation

```r
getwd()
```

```
[1] "/Users/pokeapokemon/playground_pooja/Stats"
```

Loading the csv and previewing first few records of the data

```r
# Loading CSV
data <- read.csv("international-tourist-arrivals-by-region-of-origin.csv")

# Preview data
head(data)
```

```
          Entity Code Year International.tourist.arrivals.by.region.of.origin
1 Africa (UNWTO)   NA 1995                                           12825738
2 Africa (UNWTO)   NA 1996                                           14147503
3 Africa (UNWTO)   NA 1997                                           13960251
4 Africa (UNWTO)   NA 1998                                           15610950
5 Africa (UNWTO)   NA 1999                                           15517119
6 Africa (UNWTO)   NA 2000                                           16493972
```

Data Exploration

The dataframe contains 196 observations and 4 variables:

- `Entity` (region of origin)
- `Code` (entirely missing values)
- `Year` (1995 onward)
- `International.tourist.arrivals.by-region-of-origin` (counts of international tourist trips with overnight stays).

Each row represents the number of arrivals from a specific region in a given year. The `Code` column contains no useful information and can be safely excluded from the analysis. The dataset is otherwise complete, with no missing values in the key variables.

```r
str(data)
```

```
'data.frame':   196 obs. of  4 variables:
 $ Entity                                            : chr  "Africa (UNWTO)" "Africa (UNWTO)"
 $ Code                                              : logi  NA NA NA NA NA NA ...
 $ Year                                              : int  1995 1996 1997 1998 1999 2000 20
 $ International.tourist.arrivals.by.region.of.origin: int  12825738 14147503 13960251 15610
```

```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
```

```
glimpse(data)
```

```
Rows: 196
Columns: 4
$ Entity                                           <chr> "Africa (UNWTO)", "~
$ Code                                             <lgl> NA, NA, NA, NA, NA,~
$ Year                                             <int> 1995, 1996, 1997, 1~
$ International.tourist.arrivals.by.region.of.origin <int> 12825738, 14147503,~
```

```
# Summarizing data
summary(data)
```

```
    Entity              Code              Year     
 Length:196         Mode:logical   Min.   :1995  
 Class :character   NA's:196       1st Qu.:2002  
 Mode  :character                  Median :2008  
                                   Mean   :2008  
                                   3rd Qu.:2015  
                                   Max.   :2022  
 International.tourist.arrivals.by.region.of.origin
 Min.   :  4775750                                 
 1st Qu.: 19052982                                 
 Median : 37908274                                 
 Mean   :152479716                                 
 3rd Qu.:196376748                                 
 Max.   :878514600                                 
```

```
# Missing data
colSums(is.na(data))
```

```
                                            Entity 
                                                 0 
                                              Code 
                                               196 
                                              Year 
                                                 0 
International.tourist.arrivals.by.region.of.origin 
                                                 0 
```

```
# Getting distinct values of categorical column i.e. Entity
unique(data$Entity)
```

```
[1] "Africa (UNWTO)"                    "Americas (UNWTO)"
[3] "East Asia and the Pacific (UNWTO)" "Europe (UNWTO)"
[5] "Middle East (UNWTO)"               "Other (UNWTO)"
[7] "South Asia (UNWTO)"
```

```
length(unique(data$Entity))
```

```
[1] 7
```

```
dim(data)      # rows and columns
```

```
[1] 196   4
```

```
nrow(data)     # number of rows
```

```
[1] 196
```

```
ncol(data)     # number of columns
```

```
[1] 4
```

Data Preparation

Removing the column Code since its null and doesn't capture any useful information. Also, renaming the columns for the ease of usability.

```
# Keeping the original data intact, creating a cleaned copy

tourism_data <- data %>%
  select(-Code) %>%   # removing Code column
  rename(
    Region = Entity,
    Arrivals = International.tourist.arrivals.by.region.of.origin
  ) %>%
  select(Year, Region, Arrivals)
```

```
# Checking the cleaned structure
glimpse(tourism_data)
```

```
Rows: 196
Columns: 3
$ Year     <int> 1995, 1996, 1997, 1998, 1999, 2000, 2001, 2002, 2003, 2004, 2~
$ Region   <chr> "Africa (UNWTO)", "Africa (UNWTO)", "Africa (UNWTO)", "Africa~
$ Arrivals <int> 12825738, 14147503, 13960251, 15610950, 15517119, 16493972, 1~
```

Data Analysis

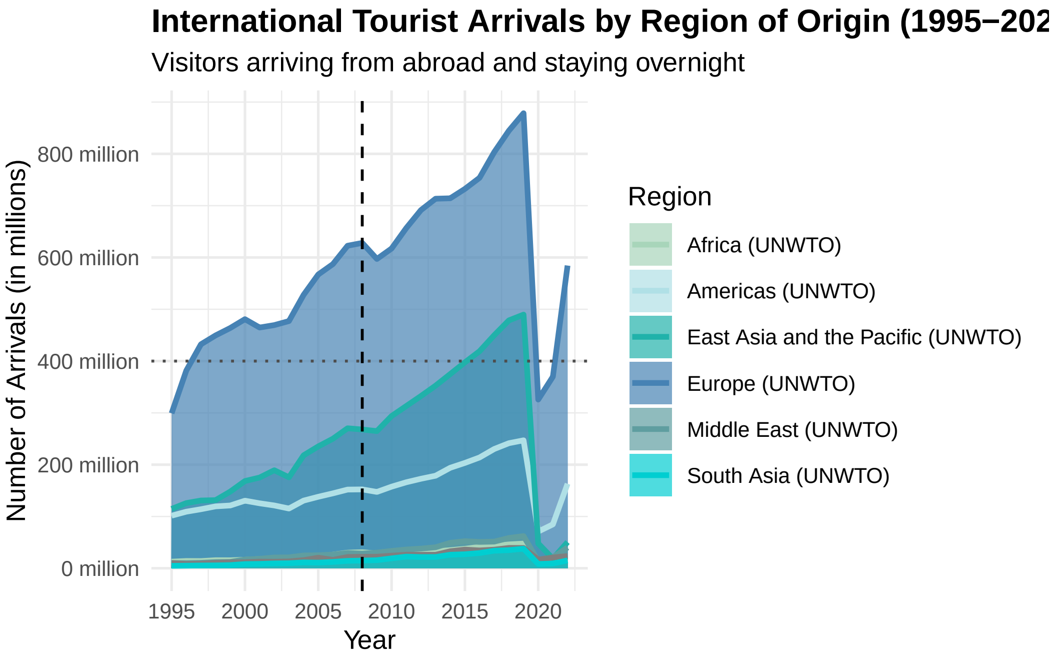
1. Plot the number of International tourist arrivals by region of origin by Year, for each region. Put all six regions on one plot, and use different colors and line types for each region. Annotate the axes appropriately. Use reference lines (using the abline function in R). Add a legend and a title.

As was noticed in data exploration, our dataframe has raw counts as big as 878514600. For a number this big, y-axis will end up in scientific notation (8e+08). Hence to fix the readability, re-scaling the y-axis (Arrivals column).

I have created an overlapping area plot showing international tourist arrivals (in millions) by region of origin for 1995–2022. The x-axis shows Year, the y-axis shows actual arrivals (in millions). Different colors distinguish regions, and a legend/title annotate the chart appropriately.

```
# ---- Libraries ----
library(ggplot2)
library(dplyr)

# ---- Data prep ----
tourism_data <- tourism_data %>% mutate(Arrivals_millions = Arrivals / 1e6)

# ---- Custom palette (sea-green / blue shades) ----
my_colors <- c(
  "Africa (UNWTO)"                   = "#A8D5BA",
  "Americas (UNWTO)"                 = "#B0E0E6",
  "East Asia and the Pacific (UNWTO)"= "#20B2AA",
  "Europe (UNWTO)"                   = "#4682B4",
  "Middle East (UNWTO)"              = "#5F9EA0",
  "South Asia (UNWTO)"               = "#00CED1"
)
```

```r
# ---- Plot ----
ggplot(tourism_data, aes(x = Year, y = Arrivals_millions)) +
  # shaded areas
  geom_area(aes(fill = Region), position = "identity", alpha = 0.7, color = NA) +
  # outlines
  geom_line(aes(color = Region), linewidth = 1) +

  # reference lines (abline equivalents)
  geom_hline(yintercept = 400, linetype = "dotted", color = "grey30") +
  geom_vline(xintercept = 2008, linetype = "dashed", color = "black") +

  # labels
  labs(
    title = "International Tourist Arrivals by Region of Origin (1995-2022)",
    subtitle = "Visitors arriving from abroad and staying overnight",
    x = "Year",
    y = "Number of Arrivals (in millions)",
    fill = "Region",
    color = "Region"
  ) +
  # scaling years and arrivals
  scale_x_continuous(
    breaks = seq(min(tourism_data$Year, na.rm = TRUE),
                 max(tourism_data$Year, na.rm = TRUE), 5)
  ) +
  scale_y_continuous(
    breaks = seq(0, max(tourism_data$Arrivals_millions, na.rm = TRUE), 200),
    labels = function(x) paste0(x, " million")
  ) +
  # custom palette
  scale_fill_manual(values = my_colors) +
  scale_color_manual(values = my_colors) +

  theme_minimal(base_size = 10) +
  theme(
    plot.title = element_text(face = "bold"),
    legend.position = "right",
    legend.box = "vertical"
  )
```



2. Takeaway from graph?

   Europe Dominates Global Tourism, By 2018, Europe contributed **over 600 million trips**, making it the **largest source market** for global tourism.

   East Asia and the Pacific grew from under **100 million** in the mid-1990s to **over 300 million by 2018**, showcasing a promising fastest growth rate among all regions.

   The Americas show **steady growth**, reaching ~**200–250 million by 2018**, but their growth is not as steep as Asia's.

   Africa, South Asia and Middle East regions start from very low counts but show **gradual upward trends**. South Asia in particular, while still small, has seen recent acceleration.

   The sharp drop around **2020** corresponds to the **COVID-19 pandemic**, which halted international travel.

   Smaller dips (like 2008–2009) reflect the **global financial crisis**.

```
# Checking arrivals from all regions in 2010 for validation
tourism_data %>%
  filter(Year == 2010) %>%
  select(Region, Arrivals, Arrivals_millions)
```

```
                             Region  Arrivals Arrivals_millions
1                   Africa (UNWTO)  31104820          31.10482
2                 Americas (UNWTO) 157586380         157.58638
3 East Asia and the Pacific (UNWTO) 293878600         293.87860
4                   Europe (UNWTO) 617303550         617.30355
5              Middle East (UNWTO)  33533844          33.53384
6                    Other (UNWTO)  24320732          24.32073
7               South Asia (UNWTO)  19202424          19.20242
```

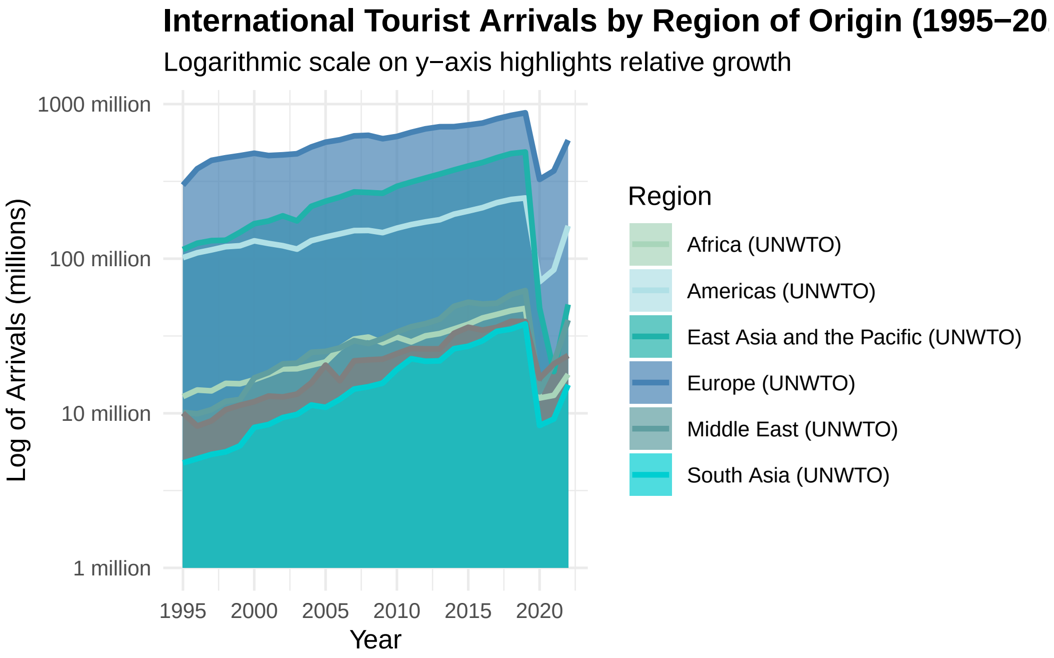
3. Would it be a good idea to use the logarithm of the counts? Why or why not?

   On a log10 scale, 600 million **8.78** whereaslog10 20 million **7.30.** The gap is now **just 1.5 units** although the difference between the two is huge. Which is why instead of growing linearly, log shrinks big numbers and grows very slowly.

   In our dataset, Europe Region is very large whereas Africa is so small that it almost looks flat. So when we will do the log, the scale will compress the gap between big and small values. Hence using logarithmic transformation could be only useful where we want to highlight relative growth rates across regions.

   Hence, the raw counts are more informative for understanding absolute tourism impact, since log-transformed values are less intuitive, but a log scale might be useful if the goal is to compare **growth patterns across regions**.

```r
ggplot(tourism_data, aes(x = Year, y = Arrivals_millions)) +
  geom_area(aes(fill = Region), position = "identity", alpha = 0.7, color = NA) +
  geom_line(aes(color = Region), linewidth = 1) +
  labs(
    title = "International Tourist Arrivals by Region of Origin (1995-2022)",
    subtitle = "Logarithmic scale on y-axis highlights relative growth",
    x = "Year",
    y = "Log of Arrivals (millions)",
    fill = "Region",
    color = "Region"
  ) +
  scale_x_continuous(
    breaks = seq(min(tourism_data$Year, na.rm = TRUE),
                 max(tourism_data$Year, na.rm = TRUE), 5)
  ) +
  scale_y_log10(
    labels = function(x) paste0(x, " million")
  ) +
  scale_fill_manual(values = my_colors) +
  scale_color_manual(values = my_colors) +
  theme_minimal(base_size = 10) +
  theme(
    plot.title = element_text(face = "bold"),
    legend.position = "right",
    legend.box = "vertical"
  )
```


International Tourist Arrivals by Region of Origin (1995−20
Logarithmic scale on y−axis highlights relative growth
Log of Arrivals (millions)
1000 million
100 million
10 million
1 million
Year
Region
Africa (UNWTO)
Americas (UNWTO)
East Asia and the Pacific (UNWTO)
Europe (UNWTO)
Middle East (UNWTO)
South Asia (UNWTO)