Ndanyuzwe_Duncan

Duncan

2025-10-09

NDANYUZWE SEMUGESHI 22217

knitr::opts_chunk$set(message = FALSE, warning = FALSE)

```
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
```

```
library(ggplot2)
library(tidyr)
library(Kendall)
library(corrplot)
```

```
## corrplot 0.95 loaded
```

##2. Data Loading

```
population <- read.csv("~/Downloads/world_population.csv")
co2 <- read.csv("~/Downloads/CO2_emission.csv")
```

Verify the data

```
##   Rank CCA3 Country.Territory          Capital Continent X2022.Population
## 1   36  AFG       Afghanistan            Kabul      Asia         41128771
## 2  138  ALB           Albania           Tirana    Europe          2842321
## 3   34  DZA           Algeria          Algiers    Africa         44903225
## 4  213  ASM    American Samoa        Pago Pago   Oceania            44273
## 5  203  AND           Andorra Andorra la Vella    Europe            79824
##   X2020.Population X2015.Population X2010.Population X2000.Population
## 1         38972230         33753499         28189672         19542982
## 2          2866849          2882481          2913399          3182021
## 3         43451666         39543154         35856344         30774621
## 4            46189            51368            54849            58230
## 5            77700            71746            71519            66097
##   X1990.Population X1980.Population X1970.Population Area..km..
## 1         10694796         12486631         10752971     652230
## 2          3295066          2941651          2324731      28748
## 3         25518074         18739378         13795915    2381741
## 4            47818            32886            27075        199
## 5            53569            35611            19860        468
##   Density..per.km.. Growth.Rate World.Population.Percentage
## 1           63.0587      1.0257                        0.52
## 2           98.8702      0.9957                        0.04
## 3           18.8531      1.0164                        0.56
## 4          222.4774      0.9831                        0.00
## 5          170.5641      1.0100                        0.00
```

```
## 'data.frame':    234 obs. of  17 variables:
##  $ Rank                       : int  36 138 34 213 203 42 224 201 33 140 ...
##  $ CCA3                       : chr  "AFG" "ALB" "DZA" "ASM" ...
##  $ Country.Territory          : chr  "Afghanistan" "Albania" "Algeria" "American Samoa" ...
##  $ Capital                    : chr  "Kabul" "Tirana" "Algiers" "Pago Pago" ...
##  $ Continent                  : chr  "Asia" "Europe" "Africa" "Oceania" ...
##  $ X2022.Population           : int  41128771 2842321 44903225 44273 79824 35588987 15857 93763 4551
##  $ X2020.Population           : int  38972230 2866849 43451666 46189 77700 33428485 15585 92664 4503
##  $ X2015.Population           : int  33753499 2882481 39543154 51368 71746 28127721 14525 89941 4325
##  $ X2010.Population           : int  28189672 2913399 35856344 54849 71519 23364185 13172 85695 4110
##  $ X2000.Population           : int  19542982 3182021 30774621 58230 66097 16394062 11047 75055 3707
##  $ X1990.Population           : int  10694796 3295066 25518074 47818 53569 11828638 8316 63328 32637
##  $ X1980.Population           : int  12486631 2941651 18739378 32886 35611 8330047 6560 64888 280248
##  $ X1970.Population           : int  10752971 2324731 13795915 27075 19860 6029700 6283 64516 238428
##  $ Area..km..                 : int  652230 28748 2381741 199 468 1246700 91 442 2780400 29743 ...
##  $ Density..per.km..          : num  63.1 98.9 18.9 222.5 170.6 ...
##  $ Growth.Rate                : num  1.026 0.996 1.016 0.983 1.01 ...
##  $ World.Population.Percentage: num  0.52 0.04 0.56 0 0 0.45 0 0 0.57 0.03 ...
```

```
##                        Rank                        CCA3
##                           0                           0
##           Country.Territory                     Capital
##                           0                           0
##                   Continent            X2022.Population
##                           0                           0
##            X2020.Population            X2015.Population
##                           0                           0
##            X2010.Population            X2000.Population
##                           0                           0
##            X1990.Population            X1980.Population
##                           0                           0
##            X1970.Population                  Area..km..
##                           0                           0
##           Density..per.km..                 Growth.Rate
##                           0                           0
## World.Population.Percentage
##                           0
```

#3. Data Cleaning

Remove duplicates

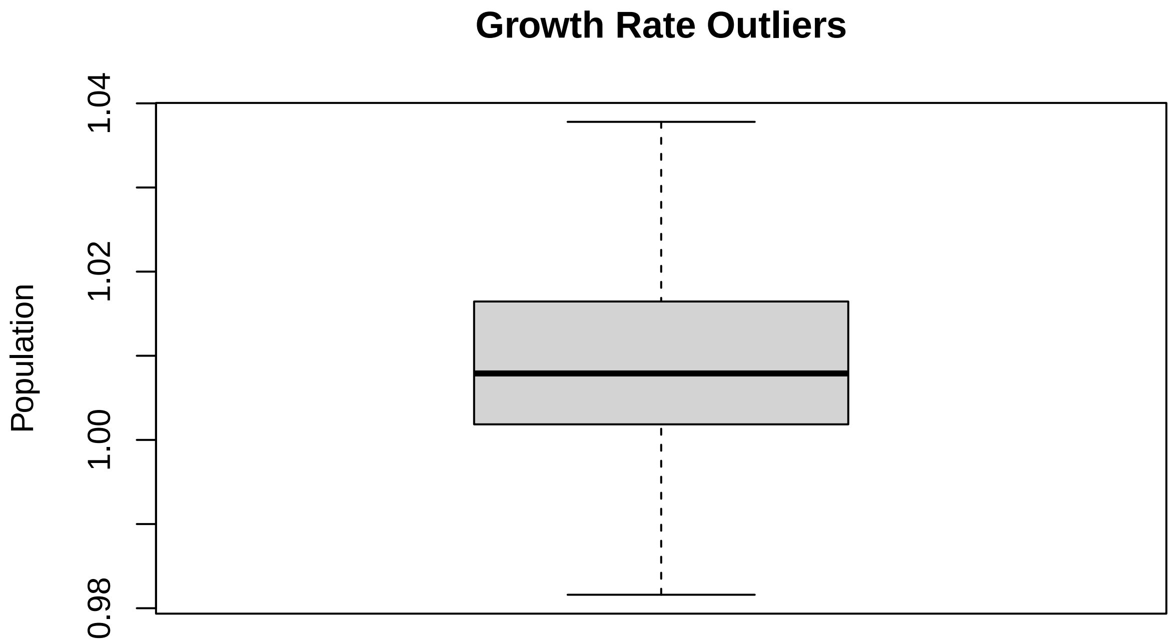
```
population <- population %>% distinct()
```

Detect and remove outliers in Growth Rate

```
detect_outlier <- function(x) {
  Q1 <- quantile(x, 0.25, na.rm = TRUE)
  Q3 <- quantile(x, 0.75, na.rm = TRUE)
  IQR <- Q3 - Q1
  x < (Q1 - 1.5*IQR) | x > (Q3 + 1.5*IQR)
}

remove_outlier <- function(df, cols) {
  for(col in cols){
    df <- df[!detect_outlier(df[[col]]), ]
  }
  df
}

population <- remove_outlier(population, c("Growth.Rate"))
boxplot(population$Growth.Rate, main = "Growth Rate Outliers", ylab = "Population")
```



# 4.3 Generating new Variable by using World Population Dataset

```
population$Growth.Rate <- as.numeric(gsub("%", "", population$Growth.Rate))/100

t <- c(2030-2022)

# prediction of 2030
population$Population_2030 <- population$X2022.Population * exp(population$Growth.Rate * t)


head(population[, c("Country.Territory", "X2022.Population", "Growth.Rate", "Population_2030")])
```

```
##   Country.Territory X2022.Population Growth.Rate Population_2030
## 1       Afghanistan         41128771    0.010257     44645963.54
## 2           Albania          2842321    0.009957      3077990.57
## 3           Algeria         44903225    0.010164     48706944.55
## 4    American Samoa            44273    0.009831        47895.57
## 5           Andorra            79824    0.010100        86541.51
## 6            Angola         35588987    0.010315     38650365.67
```

#4.4 Exploratory Data Analysis #4.4.1 Top 10 Most Populous Countries (2022) ${r} top10populous <- population %>% arrange(desc(X2022.Population)) %>% head(10) top10populous df

```
ggplot(top10populous, aes(x = reorder(Country.Territory, X2022.Population),
                          y = X2022.Population)) +
  geom_bar(stat = "identity", fill = "skyblue", color = "green") +
```

~~~~
  labs(title = "Top 10 Most Populous Countries (2022)",
       x = "Country / Territory",
       y = "Population")
~~~~

#4.4.2 Population Trends (1990–2022) “‘${r} population_long <- population %>% pivot_longer(cols = c(X1990.Population, X2000.Population, X2010.Population, X2015.Population, X2020.Population, X2022.Population), names_to = “Year”, values_to = “Population”) %>% mutate(Year = as.numeric(gsub(“X|\.Population” “ “, Year)), Population = as.numeric(gsub(”,“,” “, Population)))

top10_countries <- top10populous$Country.Territory population_long_top10 <- population_long %>% filter(Country.Territory %in% top10_countries)

ggplot(population_long_top10, aes(x = Year, y = Population, color = Country.Territory)) + geom_line(size = 1.2) + geom_point(size = 2) + labs(title = “Population Trend (1990–2022) for Top 10 Most Populous Countries”, x = “Year”, y = “Population”, color = “Country”) + theme_minimal()

~~~~
#  4.4.3 CO  Emissions Trends (1990-2019)
```${r}
co2_long <- co2 %>%
  pivot_longer(
    cols = starts_with("X"),
    names_to = "Year",
    values_to = "Emission"
  ) %>%
  mutate(
    Year = as.numeric(gsub("X", "", Year)),
    Emission = as.numeric(gsub(",", "", Emission))
  )

co2_top10 <- co2_long %>%
  filter(Country.Name %in% top10_countries, Year >= 1990, Year <= 2019)

ggplot(co2_top10, aes(x = Year, y = Emission, color = Country.Name)) +
  geom_line(size = 1.2) +
  geom_point(size = 2) +
  labs(title = "CO  Emission Trend (1990-2019) for Top 10 Most Populous Countries",
       x = "Year", y = "CO  Emissions", color = "Country") +
  theme_minimal()
~~~~

#4.4 Correlation Analysis names(population) “‘${r} pop_numeric <- population %>% select(Area..km.., Density..per.km.., Growth.Rate, World.Population.Percentage)

cor_matrix <- cor(pop_numeric, use = “complete.obs”) round(cor_matrix, 3)

corrplot(cor_matrix, method = “color”, type = “upper”, tl.col = “black”, tl.srt = 45, addCoef.col = “black”, title = “Correlation Heatmap: Population Metrics”, mar = c(0,0,1,0))

~~~~
#4.5 Merge Population and CO  (2022 & 2019)

``` r
merged_data <- population %>%
  select(Country.Territory, Continent, X2022.Population) %>%
  inner_join(co2 %>% select(Country.Name, Region, X2019),
             by = c("Country.Territory" = "Country.Name")) %>%
  rename(Population2022 = X2022.Population,
~~~~

```
         CO2_2019 = X2019)

head(merged_data)
```

```
##   Country.Territory Continent Population2022                      Region
## 1       Afghanistan      Asia       41128771                  South Asia
## 2           Albania    Europe        2842321       Europe & Central Asia
## 3           Algeria    Africa       44903225  Middle East & North Africa
## 4    American Samoa   Oceania          44273         East Asia & Pacific
## 5           Andorra    Europe          79824       Europe & Central Asia
## 6            Angola    Africa       35588987          Sub-Saharan Africa
##    CO2_2019
## 1 0.1598244
## 2 1.6922483
## 3 3.9776505
## 4        NA
## 5 6.4812174
## 6 0.7921371
```

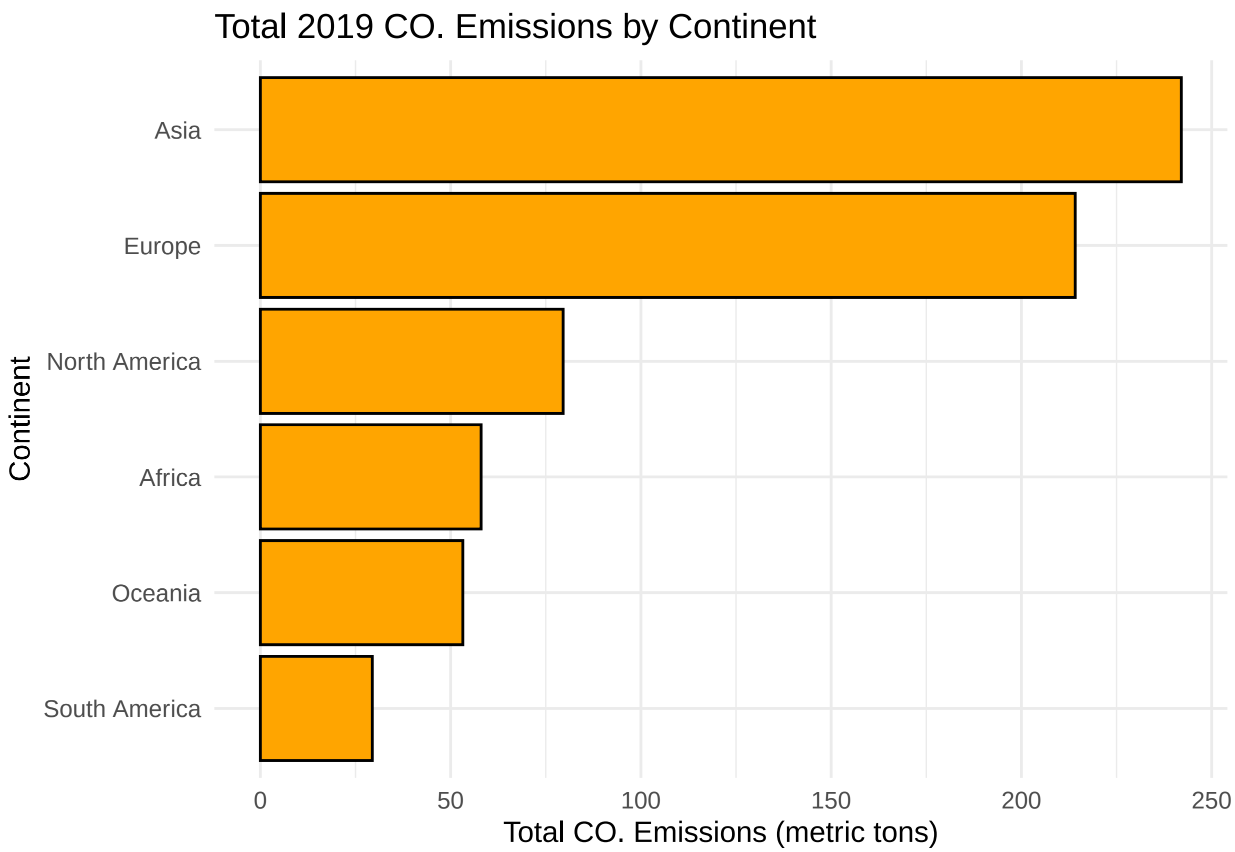
#4.6 CO Emissions by Continent

```r
co2_by_continent <- merged_data %>%
  group_by(Continent) %>%
  summarise(Total_CO2_2019 = sum(CO2_2019, na.rm = TRUE)) %>%
  arrange(desc(Total_CO2_2019))

co2_by_continent
```

```
## # A tibble: 6 x 2
##   Continent     Total_CO2_2019
##   <chr>                  <dbl>
## 1 Asia                   242.
## 2 Europe                 214.
## 3 North America           79.6
## 4 Africa                  58.0
## 5 Oceania                 53.2
## 6 South America           29.4
```

```r
ggplot(co2_by_continent, aes(x = reorder(Continent, Total_CO2_2019), y = Total_CO2_2019)) +
  geom_bar(stat = "identity", fill = "orange", color = "black") +
  coord_flip() +
  labs(title = "Total 2019 CO  Emissions by Continent",
       x = "Continent",
       y = "Total CO  Emissions (metric tons)") +
  theme_minimal()
```



# Key continents

```
cat(" First continent (highest emission):", co2_by_continent$Continent[1], "-", co2_by_continent$Total_(
```

```
##  First continent (highest emission): Asia - 242.0265 metric tons
```

```
cat(" Third continent (CO2 emission):", co2_by_continent$Continent[3], "-", co2_by_continent$Total_CO2_2
```

```
##  Third continent (CO2 emission): North America - 79.56837 metric tons
```

```
cat(" Last continent (lowest emission):", co2_by_continent$Continent[nrow(co2_by_continent)], "-", co2_b
```

```
##  Last continent (lowest emission): South America - 29.41627 metric tons
```