EDA and Report

Ean Anciso Lance He Regi Acierto

2025-12-04

1 Visualizations

Table 1 is a frequency table showing the top 10 countries ranked in descending order by the number of times it appears on in our raw dataset. This means that the countries with the most amount of large populated cities, large population being defined by wikipedia, will show up. Many of the countries in this dataset only appear once, but there are a a few that show multiple times and some that show even more.

Table 1: Countries Ranked by amount of Entries

Table 1: Top 10 Countries by Number of Cities

Country	entries	total_pop	total_city_pop	total_urban_pop
China	18	184151927	264544408	216167475
India	9	127865581	69054182	139999000
United States	9	70248568	22409673	84186000
Japan	4	59073817	20149562	64394000
Brazil	3	34422126	21274580	41006000
Indonesia	3	57666467	15700316	47515000
Egypt	2	32833059	15486760	25008000
Mexico	2	22757212	10595565	27329000
Pakistan	2	36579020	28742135	32555000
Russia	2	19907753	18801911	22777000

Table 2: Urban Area Statistics (Top 6 Countries)

Country	Pop_Mean	Pop_SD	Area_Mean	Area_SD
Brazil	13668667	8927824	2319	1209
China	12009304	6963228	2050	1166
India	15555444	8891563	1095	636
Indonesia	15838333	15520526	1648	1657
Japan	16098500	15380354	3865	3219
United States	9354000	5448314	6367	2461

Table 3: City Proper Statistics (Top 6 Countries)

Country	Pop_Mean	Pop_SD	Area_Mean	Area_SD
Brazil	7091527	4899796	1024	619
China	14696912	6647019	16387	19780
India	7672687	4372810	595	369
Indonesia	5233439	4268480	389	253
Japan	5037390	5671442	771	948
United States	2489964	2638432	668	488

Figure 1: Urban Population by Area

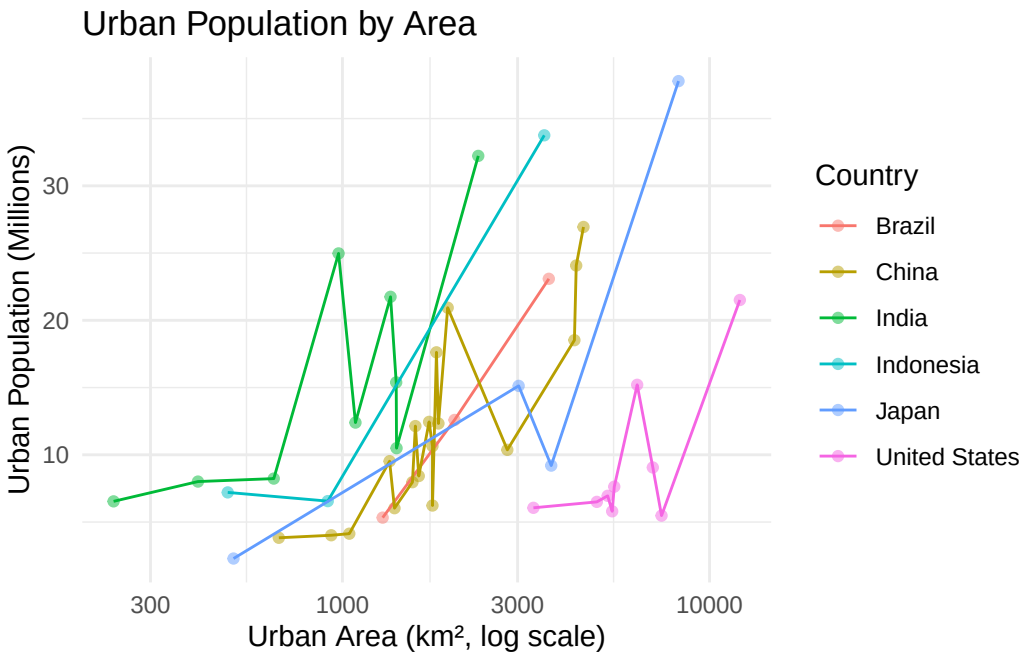


Figure 2: City Proper Population by Area

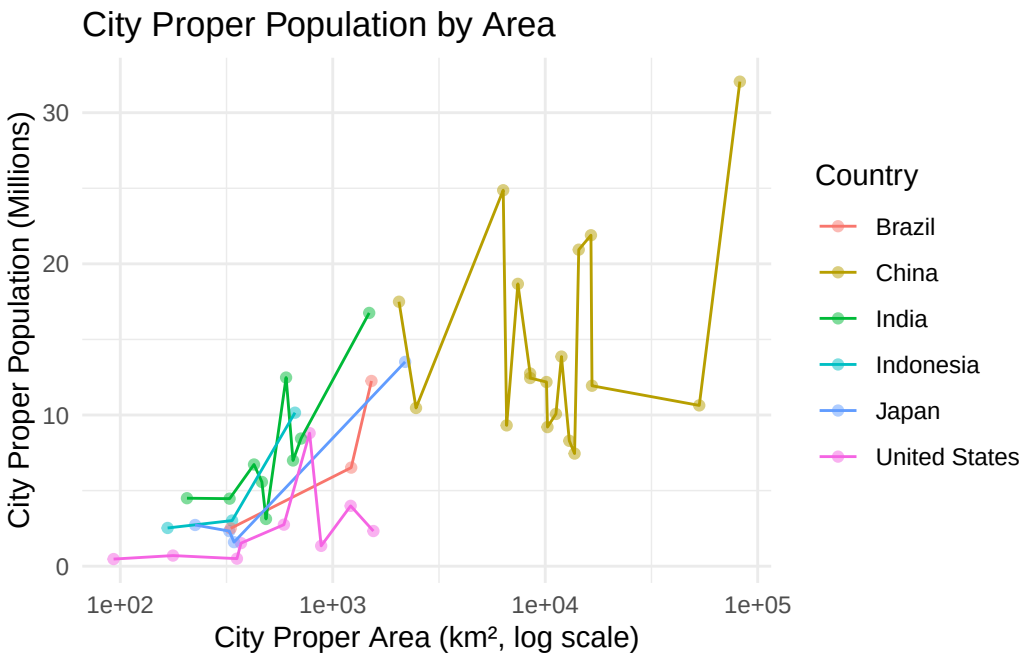


Figure 3: City Proper, Area vs Density for Top 6 Countries with amount of large cities

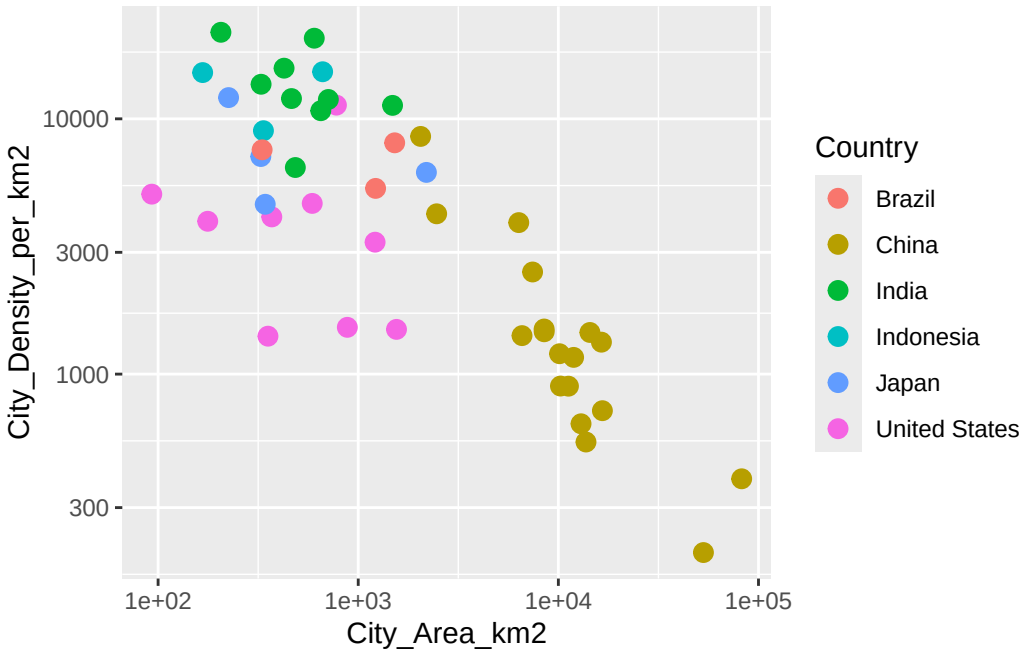
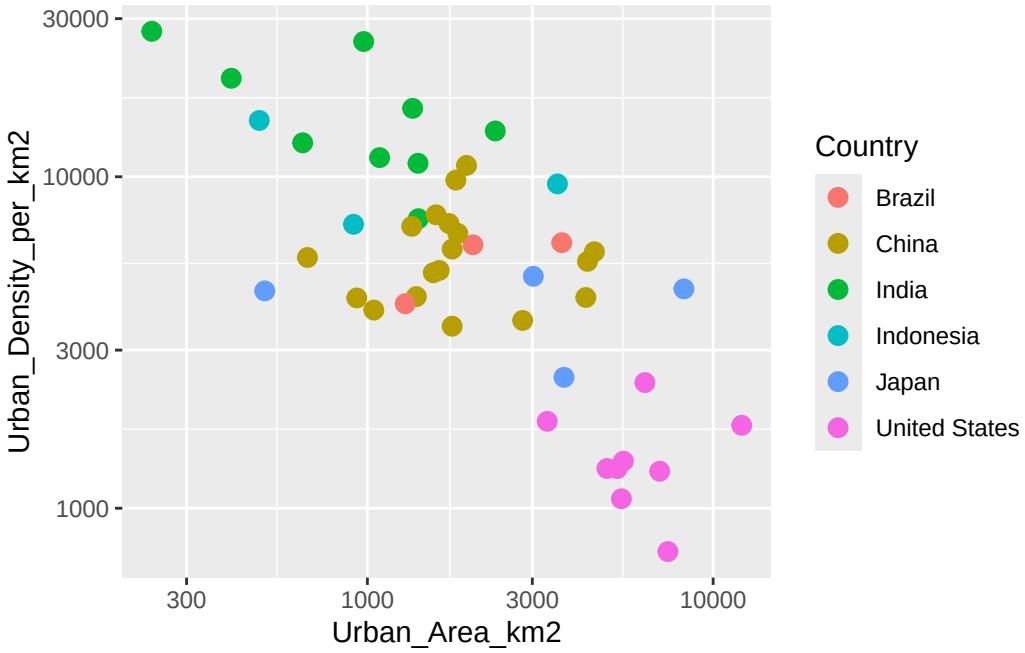


Figure 4: Urban Area, Area vs Density for Top 6 Countries with amount of large cities



Code Appendix

```
# 1. Load Packages
library(rvest)
library(dplyr)
library(stringr)
library(ggplot2)
library(knitr)
library(tidyverse)

# 2. Scrape the table
url <- "https://en.wikipedia.org/wiki/List_of_largest_cities"
webpage <- read_html(url)
tables <- html_table(webpage, fill = TRUE)

# Find the correct table containing "Jakarta"
table_index <- which(sapply(tables, function(t) "Jakarta" %in% t[[1]]))
city_table <- tables[[table_index]]

# 3. Fix Column Names
colnames(city_table) <- c(
  "City", "Country", "UN_Estimate_Pop", "Definition",
  "City_Pop", "City_Area_km2", "City_Density_per_km2",      # City Proper
  "Urban_Pop", "Urban_Area_km2", "Urban_Density_per_km2",   # Urban Area
  "Metro_Pop", "Metro_Area_km2", "Metro_Density_per_km2"    # Metropolitan
)
```

```r
# 4. Data Wrangling Pipeline
clean_data <- city_table %>%
  # Remove Garbage Header Row
  slice(-1) %>%

  select(
    City, Country, UN_Estimate_Pop,
    City_Pop, City_Area_km2, City_Density_per_km2,
    Urban_Pop, Urban_Area_km2, Urban_Density_per_km2
  ) %>%

  # Clean Text (Remove Footnotes and Commas)
  mutate(across(everything(), ~str_remove_all(., "\\[.*?\\]"))) %>%
  mutate(across(everything(), ~str_remove_all(., ","))) %>%

  # Handle Missing Values
  mutate(across(everything(), ~na_if(str_trim(.), ""))) %>%
  mutate(across(everything(), ~na_if(str_trim(.), "-"))) %>%
  mutate(across(everything(), ~na_if(str_trim(.), "-"))) %>%
  mutate(across(everything(), ~na_if(str_trim(.), "-"))) %>%

  # Filter Rows with complete City/Urban data
  filter(
    !is.na(City_Pop), !is.na(City_Area_km2), !is.na(City_Density_per_km2),
    !is.na(Urban_Pop), !is.na(Urban_Area_km2), !is.na(Urban_Density_per_km2)
  ) %>%

  # Convert Columns to Numeric
  mutate(across(3:9, as.numeric)) %>%

  # Clean Country Names
  mutate(Country = str_trim(Country))

# 5. Filter for Target Countries
target_countries <- c("United States", "China", "India", "Japan", "Brazil", "Indonesia")
country_data <- clean_data %>%
  filter(Country %in% target_countries)
firstTable <- clean_data %>%
  group_by(Country) %>%
  summarise(
    entries = n(),
    total_pop = sum(UN_Estimate_Pop, na.rm = TRUE),
    total_city_pop = sum(City_Pop, na.rm = TRUE),
    total_urban_pop = sum(Urban_Pop, na.rm = TRUE)
  ) %>%
  arrange(desc(entries))
```

```r
# Print table using kable for PDF formatting
kable(head(firstTable, 10), caption = "Top 10 Countries by Number of Cities")
# Table A: Urban Area Statistics
urban_area_stats <- country_data %>%
  group_by(Country) %>%
  summarise(
    Pop_Mean = mean(Urban_Pop, na.rm = TRUE),
    Pop_SD   = sd(Urban_Pop, na.rm = TRUE),
    Area_Mean = mean(Urban_Area_km2, na.rm = TRUE),
    Area_SD   = sd(Urban_Area_km2, na.rm = TRUE)
  )

kable(urban_area_stats, digits = 0)
# Table B: City Proper Statistics
city_proper_stats <- country_data %>%
  group_by(Country) %>%
  summarise(
    Pop_Mean = mean(City_Pop, na.rm = TRUE),
    Pop_SD   = sd(City_Pop, na.rm = TRUE),
    Area_Mean = mean(City_Area_km2, na.rm = TRUE),
    Area_SD   = sd(City_Area_km2, na.rm = TRUE)
  )

kable(city_proper_stats, digits = 0)
# Plot 1: Urban Population by Area
# X = Urban Area, Y = Urban Population
urban_data <- country_data # Uses the filtered data from setup

urban_data %>%
  ggplot(
    aes(
      x = Urban_Area_km2,
      y = Urban_Pop / 1000000,
      color = Country,
      group = Country
    )
  ) +
  geom_point(alpha = 0.5) +
  geom_line() +
  scale_x_log10() +
  labs(
    title = "Urban Population by Area",
    x = "Urban Area (km², log scale)",
    y = "Urban Population (Millions)"
  ) +
  theme_minimal()
# Plot 2: City Proper Population by Area
```

```r
# X = City Area, Y = City Population
city_proper_data <- country_data # Uses the filtered data from setup

city_proper_data %>%
  ggplot(
    aes(
      x = City_Area_km2,
      y = City_Pop / 1000000,
      color = Country,
      group = Country
    )
  ) +
  geom_point(alpha = 0.5) +
  geom_line() +
  scale_x_log10() +
  labs(
    title = "City Proper Population by Area",
    x = "City Proper Area (km², log scale)",
    y = "City Proper Population (Millions)"
  ) +
  theme_minimal()
Top_entries <- clean_data %>% # Get target countries(Top 6)
  filter(Country %in% c("China", "India", "United States", "Japan", "Brazil", "Indonesia"))

ggplot(
  Top_entries,
  aes(
    City_Area_km2,
    City_Density_per_km2,
    color = Country
  ),
) +
  geom_point(size = 3) +
  scale_x_log10() +
  scale_y_log10()

Top_entries <- clean_data %>% # Get target countries(Top 6)
  filter(Country %in% c("China", "India", "United States", "Japan", "Brazil", "Indonesia"))

ggplot(
  Top_entries,
  aes(
    Urban_Area_km2,
    Urban_Density_per_km2,
    color = Country
  ),
) +
```

```
  geom_point(size = 3) +
  scale_x_log10() +
  scale_y_log10()
```