Take-Home Exercise 1: Data Visualisation on Singapore’s Demographics (2024)

Wang Anqi

# **1.1 Overview**

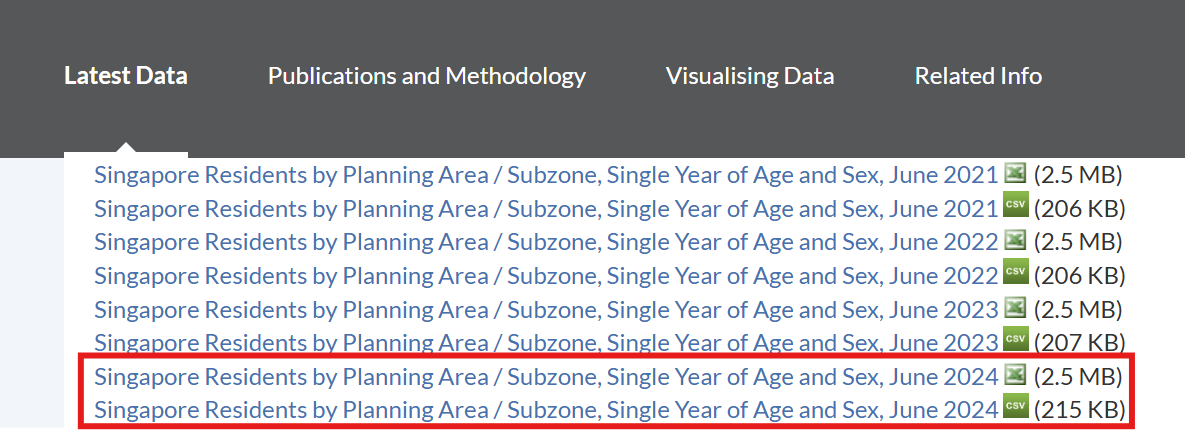
This take-home exercise involves designing insightful data visualisations based on demographic data for Singapore in 2024. The goal is to explore and communicate the **demographic structures and distribution** through effective graphical representations.

## **1.1.1 The data**

All visualisations will be based on the dataset titled

**“Singapore Residents by Planning Area / Subzone, Single Year of Age and Sex, June 2024”**,

provided by the [Department of Statistics, Singapore (DOS)](https://www.singstat.gov.sg/).



# **1.2 Getting started**

## **1.2.1 Installing and loading the required libraries**

In this exercise, beside tidyverse, four R packages will be used. They are:

* tidyverse: a collection of core packages designed for data science, used extensively for data preparation and wrangling.
* ggrepel: an R package provides geoms for ggplot2 to repel overlapping text labels.
* ggthemes: an R package provides some extra themes, geoms, and scales for ‘ggplot2’.
* hrbrthemes: an R package provides typography-centric themes and theme components for ggplot2.
* patchwork: an R package for preparing composite figure created using ggplot2.

Code chunk below will be used to check if these packages have been installed and also will load them onto working R environment.

pacman::p\_load(tidyverse, ggrepel, patchwork,   
 ggthemes, hrbrthemes,)

## **1.2.2 Importing data**

For the purpose of this exercise, a data file called *Exam\_data* will be used. It consists of year end examination grades of a cohort of primary 3 students from a local school. It is in csv file format.

The code chunk below imports *exam\_data.csv* into R environment by using [*read\_csv()*](https://readr.tidyverse.org/reference/read_delim.html) function of [**readr**](https://readr.tidyverse.org/) package. **readr** is one of the tidyverse package.

popu\_data <- read\_csv("data/respopagesex2024.csv", show\_col\_types = FALSE)  
  
cat("Total number of records in the dataset:", nrow(popu\_data), "\n")

Total number of records in the dataset: 60424

# **1.3 Data Wrangling**

## 1.3.1 **Variable Selection**

According to the dataset documentation, The meaning of each column in the dataset is as follows:

* **PA**：Planning Area
* **SZ**：Subzone
* **Age**：Single Year of Age
* **Sex**：Sex
* **Pop**：Resident Count
* **Time**：Time / Period

Since all data points were collected in the year 2024 and the Time column does not provide additional analytical value for this task, it will be removed from the dataset to simplify the analysis.

# Remove the 'Time' column from the dataset  
popu\_data1 <- popu\_data %>% select(-Time)

## 1.3.2 **Handling missing values**

Before conducting the analysis, we need to check for missing values. Later on, when we recode the age groups, if there are NAs in the columns, they may affect the recoding process

is.na() allows us to identify missing values in our dataset and sum() counts the frequency of such missing values.

# Checks for NA in each column  
colSums(is.na(popu\_data1))

PA SZ Age Sex Pop   
 0 0 0 0 0

The result indicates that **there are no missing values in any of the columns** (PA, SZ, Age, Sex, Pop); each column has 0 missing values, confirming that the dataset is complete.

# Here's what the dataset looks like so far  
popu\_data1

# A tibble: 60,424 × 5  
 PA SZ Age Sex Pop  
 <chr> <chr> <chr> <chr> <dbl>  
 1 Ang Mo Kio Ang Mo Kio Town Centre 0 Males 10  
 2 Ang Mo Kio Ang Mo Kio Town Centre 0 Females 10  
 3 Ang Mo Kio Ang Mo Kio Town Centre 1 Males 10  
 4 Ang Mo Kio Ang Mo Kio Town Centre 1 Females 10  
 5 Ang Mo Kio Ang Mo Kio Town Centre 2 Males 10  
 6 Ang Mo Kio Ang Mo Kio Town Centre 2 Females 10  
 7 Ang Mo Kio Ang Mo Kio Town Centre 3 Males 10  
 8 Ang Mo Kio Ang Mo Kio Town Centre 3 Females 10  
 9 Ang Mo Kio Ang Mo Kio Town Centre 4 Males 30  
10 Ang Mo Kio Ang Mo Kio Town Centre 4 Females 10  
# ℹ 60,414 more rows

## **1.3.3 Data Recoding and Type Conversion**

As shown in the output above, the Age column is currently stored as a character type (<chr>), rather than as an integer (<int>). This prevents numerical operations and proper plotting of age-related trends. Therefore, it is necessary to convert the Age variable to an integer format to enable meaningful quantitative analysis.

But the dataset contains a special age value "90 and over" which cannot be directly converted to a numeric format. To maintain consistency and enable numerical analysis, this value was recoded to 90, representing the lower bound of the age group.

# Step 1: Recode "90 and over" to "90"  
popu\_data1 <- popu\_data1 %>%  
 mutate(Age = ifelse(Age == "90 and over", "90", Age))

# Step 2: Convert Age to integer  
popu\_data1 <- popu\_data1 %>%  
 mutate(  
 Age = trimws(Age),  
 Age = ifelse(Age == "90 and over", "90", Age),  
 Age = as.integer(Age)  
 )

## **1.3.4 Standardizing Labels**

popu\_data1 <- popu\_data1 %>%  
 mutate(  
 Sex = recode(Sex, "Males" = "Male", "Females" = "Female")  
 )

# Here's what the dataset looks like so far  
popu\_data1

# A tibble: 60,424 × 5  
 PA SZ Age Sex Pop  
 <chr> <chr> <int> <chr> <dbl>  
 1 Ang Mo Kio Ang Mo Kio Town Centre 0 Male 10  
 2 Ang Mo Kio Ang Mo Kio Town Centre 0 Female 10  
 3 Ang Mo Kio Ang Mo Kio Town Centre 1 Male 10  
 4 Ang Mo Kio Ang Mo Kio Town Centre 1 Female 10  
 5 Ang Mo Kio Ang Mo Kio Town Centre 2 Male 10  
 6 Ang Mo Kio Ang Mo Kio Town Centre 2 Female 10  
 7 Ang Mo Kio Ang Mo Kio Town Centre 3 Male 10  
 8 Ang Mo Kio Ang Mo Kio Town Centre 3 Female 10  
 9 Ang Mo Kio Ang Mo Kio Town Centre 4 Male 30  
10 Ang Mo Kio Ang Mo Kio Town Centre 4 Female 10  
# ℹ 60,414 more rows

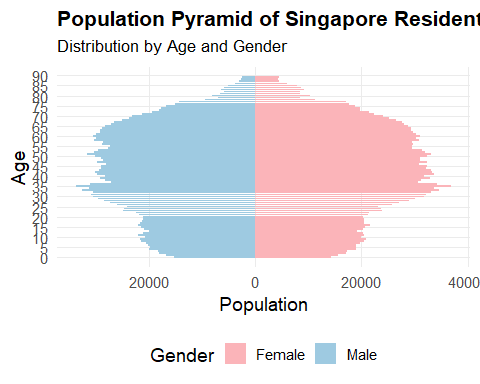
Notes:

1. For June 2024, Planning Areas refer to areas demarcated in the Urban Redevelopment Authority’s Master Plan 2019.
2. Data from 2003 onwards exclude residents who have been away from Singapore for a continuous period of 12 months or longer as at the reference period.
3. The figures have been rounded to the nearest 10.
4. The data may not add up due to rounding.

# **1.4 Exploratory Data Analysis**

## **1.4.1 Population Pyramid by Age and Sex**

## The plot



## The code

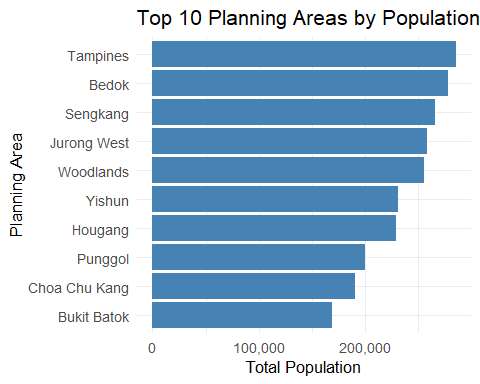
# construct population pyramid data (set male values as negative).  
pyramid\_data <- popu\_data1 %>%  
 mutate(Pop = ifelse(Sex == "Male", -Pop, Pop))  
  
# define color  
gender\_colors <- c("Male" = "#9ecae1", "Female" = "#fbb4b9")  
  
# plot  
ggplot(pyramid\_data, aes(x = Age, y = Pop, fill = Sex)) +  
 geom\_bar(stat = "identity", width = 0.9) +  
 coord\_flip() +  
 scale\_y\_continuous(labels = abs) +  
 scale\_x\_continuous(breaks = seq(0, 90, by = 5)) +  
 scale\_fill\_manual(values = gender\_colors) +  
 labs(  
 title = "Population Pyramid of Singapore Residents (2024)",  
 subtitle = "Distribution by Age and Gender",  
 x = "Age",  
 y = "Population",  
 fill = "Gender"  
 ) +  
 theme\_minimal(base\_size = 14) +  
 theme(  
 plot.title = element\_text(face = "bold", size = 16),  
 plot.subtitle = element\_text(size = 13, margin = margin(b = 10)),  
 legend.position = "bottom",  
 panel.grid.minor = element\_blank()  
 )

The population pyramid shows the age and gender distribution of Singapore residents in 2024. Males are represented on the left and females on the right.

The pyramid has a wide middle, especially between ages 25 to 39, indicating a strong working-age population. The base is relatively narrow, suggesting lower birth rates in recent years. At older ages, especially 75 and above, the female population exceeds that of males, reflecting higher life expectancy for women. Overall, the shape of the pyramid reveals an ageing population with fewer young residents, which may have implications for future labour supply and social support needs.

## **1.4.2 Top 10 Planning Areas by Population**

## The plot



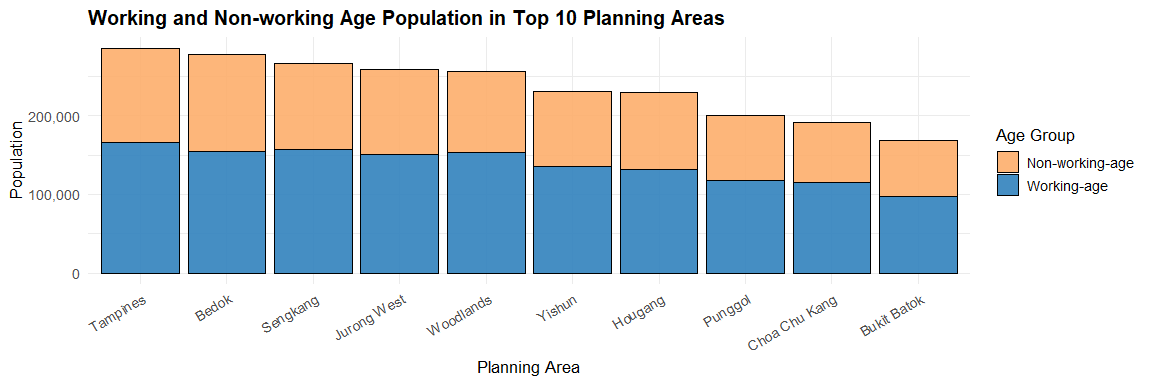
## The code

# Summarise total population by Planning Area (PA)  
top\_pa <- popu\_data1 %>%  
 group\_by(PA) %>%  
 summarise(Total = sum(Pop)) %>%  
 arrange(desc(Total)) %>%  
 slice\_head(n = 10)  
  
# Plot  
library(scales)  
  
ggplot(top\_pa, aes(x = reorder(PA, Total), y = Total)) +  
 geom\_col(fill = "#4682b4") +  
 coord\_flip() +  
 scale\_y\_continuous(labels = comma) +  
 labs(  
 title = "Top 10 Planning Areas by Population (2024)",  
 x = "Planning Area",  
 y = "Total Population"  
 ) +  
 theme\_minimal(base\_size = 13)

The bar chart displays the top 10 planning areas in Singapore by total population in 2024. Tampines ranks first, followed closely by Bedok and Sengkang. These areas are known for their large residential zones and well-developed public amenities. Most of the top-ranked areas are located in the eastern and northern parts of Singapore, reflecting urban decentralisation and population growth in regional centres. The population sizes of these areas are relatively close, each exceeding 150,000 residents.

## **1.4.3 Working and Non-working Age Population in Top 10 Planning Areas**

## The plot



## The code

top10\_pa <- popu\_data1 %>%  
 count(PA, wt = Pop) %>%  
 top\_n(10, n) %>%  
 pull(PA)  
# group and summarize the population by labour force status  
age\_grouped <- popu\_data1 %>%  
 filter(PA %in% top10\_pa) %>%  
 mutate(  
 AgeGroup = case\_when(  
 Age >= 25 & Age <= 64 ~ "Working-age",  
 TRUE ~ "Non-working-age"  
 )  
 ) %>%  
 group\_by(PA, AgeGroup) %>%  
 summarise(Total = sum(Pop), .groups = "drop")  
  
# plot  
ggplot(age\_grouped, aes(x = reorder(PA, -Total), y = Total, fill = AgeGroup)) +  
 geom\_col(color = "black", alpha = 0.9) +  
 scale\_fill\_manual(values = c("Working-age" = "#3182bd", "Non-working-age" = "#fdae6b")) +  
 scale\_y\_continuous(labels = comma) +   
 labs(  
 title = "Working and Non-working Age Population in Top 10 Planning Areas (2024)",  
 x = "Planning Area",  
 y = "Population",  
 fill = "Age Group"  
 ) +  
 theme\_minimal(base\_size = 13) +  
 theme(  
 axis.text.x = element\_text(angle = 30, hjust = 1),  
 plot.title = element\_text(face = "bold", size = 15)  
 )

This stacked bar chart shows the distribution of working-age (25–64) and non-working-age (0–24 and 65+) populations across the top 10 most populous planning areas in Singapore (2024). Working-age individuals consistently make up the majority in every area, with Tampines, Sengkang, and Woodlands having particularly high counts. Interestingly, Bedok and Jurong West, despite their large populations, have a relatively higher share of non-working-age residents—likely due to an older population base. In contrast, newer towns like Punggol and Sengkang have a larger proportion of working-age adults, reflecting younger and growing communities.