Take-home Exercise 1 - Data Visualizations in Singapore Residence by Planning Area

PENG Hsiao Yuan

## 1. Environment Set-up

## 1.1 Background Information and Data Source

The data set is downloaded from “https://www.singstat.gov.sg/”

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

## 1.2 Data Processing

Importing Data and Install Packages

echo = FALSE  
pacman::p\_load(  
 tidyverse, haven, knitr,  
 patchwork, ggthemes, scales,  
 ggridges, ggpubr, gganimate,  
 ggdist, ggtext, ggalt,  
 cowplot, ggnewscale,  
 sf, viridis  
)

Loading and Installing Packages

## 1.3 Read the file and Clean BOM

raw\_lines <- readLines("respopagesex2024/respopagesex2024.csv", encoding = "UTF-8")  
writeLines(raw\_lines, "respopagesex2024/respopagesex2024\_clean.csv", useBytes = TRUE)

## 1.4 Read the Clean File

echo = FALSE  
df <- read\_csv("respopagesex2024/respopagesex2024\_clean.csv", skip\_empty\_rows = TRUE)  
  
df <- df %>%  
 mutate(  
 Age = suppressWarnings(as.numeric(Age)),  
 Pop = suppressWarnings(as.numeric(Pop))  
 )  
  
df\_clean <- df %>%  
 filter(!is.na(Pop) & !is.na(Age))  
show\_col\_types = FALSE

## 2 Population Distributions by Planning Area

A local online media company that publishes daily content on digital platforms is planning to release an article on demographic structures and distribution of Singapore in 2024.

df <- tibble(  
 PA = c("Planning Area"),  
 SZ = c("Subzone"),  
 Age = c("Age"),  
 Sex = c("Sex"),  
 Pop = c("Population"),  
 Time = c("Time")  
)  
  
df %>%  
 knitr::kable(caption = "Column Information")

Column Information

| PA | SZ | Age | Sex | Pop | Time |
| --- | --- | --- | --- | --- | --- |
| Planning Area | Subzone | Age | Sex | Population | Time |

## 2.1 Task Objective

Assuming the role of the graphical editor of the media company, you are tasked to prepare at most three data visualizations for the article.

## 2.2 Data Preparation

df\_summary <- df\_clean %>%  
 group\_by(PA) %>%  
 summarise(  
 Total\_Pop = sum(Pop, na.rm = TRUE),  
 .groups = "drop"  
 ) %>%  
 arrange(desc(Total\_Pop))

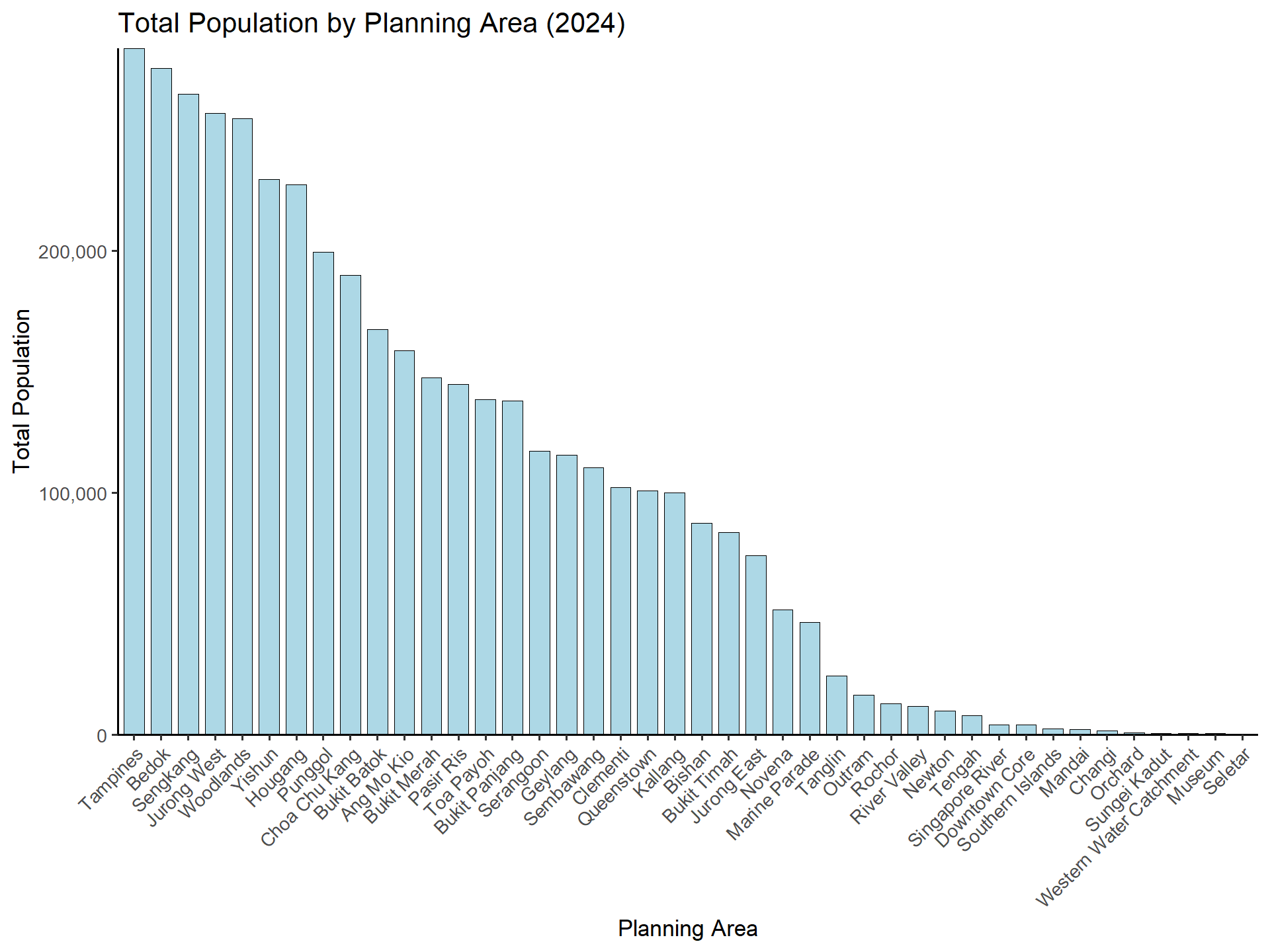
## 2.3 Plot the Graph

The code below plots the distributions of population.

X: Planning area

Y: Total population by area in 2024

bottom10\_pa <- df\_summary %>%  
 slice\_min(Total\_Pop, n = 10) %>%  
 pull(PA)  
  
df\_summary\_filtered <- df\_summary %>%  
 filter(!PA %in% bottom10\_pa)  
  
ggplot(df\_summary\_filtered, aes(x = reorder(PA, -Total\_Pop), y = Total\_Pop)) +  
 geom\_bar(stat = "identity", color = "black", fill = "lightblue", width = 0.75) +  
 scale\_y\_continuous(labels = scales::comma, expand = c(0, 0)) +   
 labs(  
 x = "Planning Area",  
 y = "Total Population",  
 title = "Total Population by Planning Area (2024)"  
 ) +  
 theme\_classic(base\_size = 26) +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))



## 2.4 Analysis

The bar chart above suggests that **Tampines** are the most populated area according to the survey. This outcome also indicates that the outskirt area such as **Tampines**, **Bedok**, **Jurong West** and **Woodlands** are among the most populated residential area. This could be attributed to the higher rent in the central area, convenient public transportation nationwide or greener environment.

As a media company, the marketing department is able to allocate more resources on advertisement on the area which has higher population density. By analyzing population of each planning area and the resources invested in that area, companies are able to evaluate the KPI of the advertisement to further adjust the strategies.

##3.The Generation Strucutre Ratio by Planning Area

library(tidyverse)  
  
df <- read\_csv("respopagesex2024/respopagesex2024\_clean.csv",  
 skip\_empty\_rows = TRUE,  
 show\_col\_types = FALSE) %>%  
 rename(Age\_raw = Age) %>%   
 mutate(  
 Age = suppressWarnings(as.numeric(Age\_raw)),   
 Pop = suppressWarnings(as.numeric(Pop))  
 )  
  
df\_clean <- df %>%  
 filter(!is.na(Pop))

## 3.1 Grouping

df <- tibble(  
 Children = c("0-14"),  
 Adults = c("15-64"),  
 Seniors = c("Above 65"),  
)  
  
df %>%  
 knitr::kable(caption = "Column Information")

Column Information

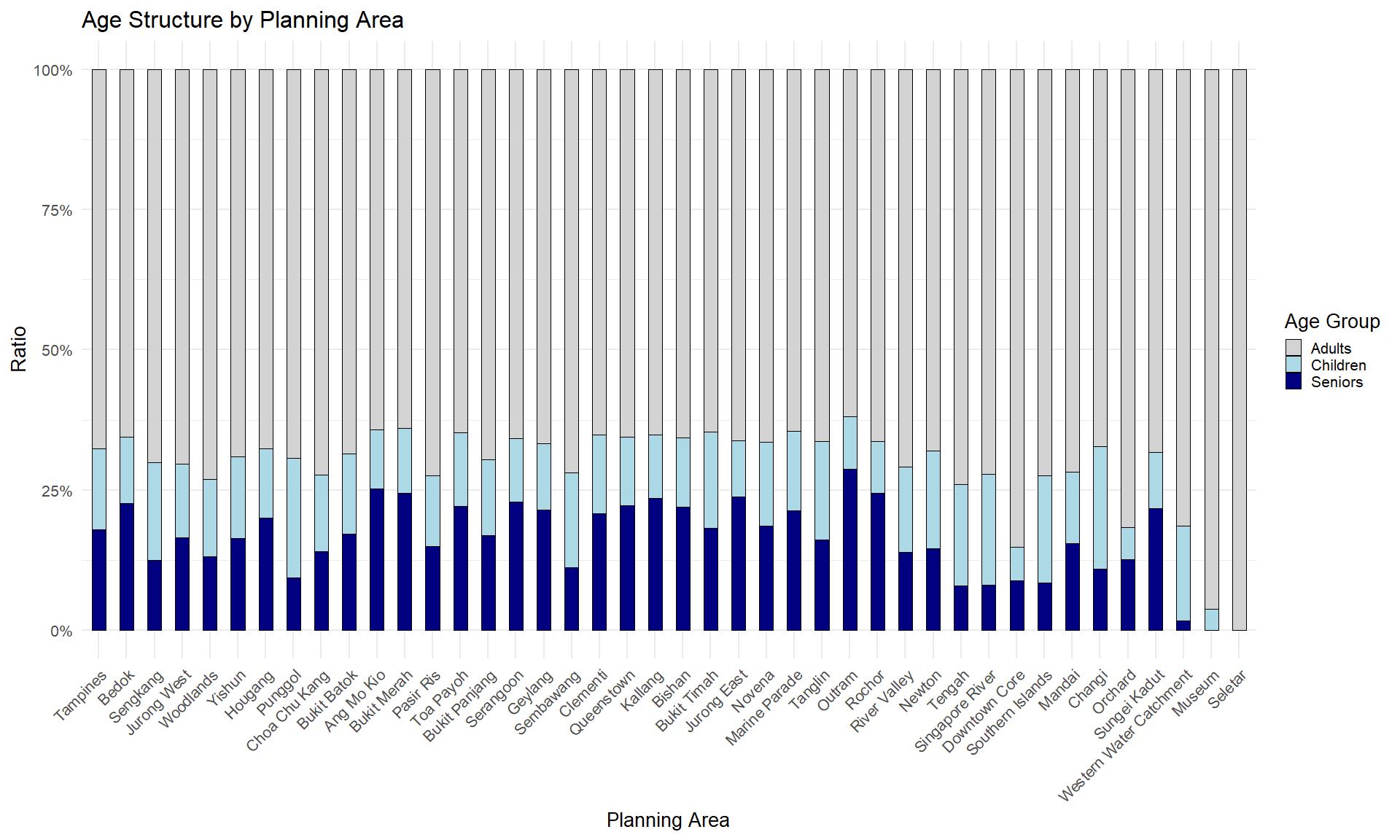
| Children | Adults | Seniors |
| --- | --- | --- |
| 0-14 | 15-64 | Above 65 |

## 3.2 Data Processing and Cleaning

df\_clean <- df\_clean %>%  
 mutate(AgeGroup = case\_when(  
 Age\_raw == "90\_and\_Over" ~ "Seniors",   
 Age >= 0 & Age <= 14 ~ "Children",  
 Age >= 15 & Age <= 64 ~ "Adults",  
 Age >= 65 ~ "Seniors",  
 TRUE ~ NA\_character\_  
 )) %>%  
 filter(!is.na(AgeGroup))   
  
df\_percent <- df\_clean %>%  
 group\_by(PA, AgeGroup) %>%  
 summarise(Pop = sum(Pop, na.rm = TRUE), .groups = "drop") %>%  
 group\_by(PA) %>%  
 mutate(  
 Total\_Pop = sum(Pop),  
 Percent = Pop / Total\_Pop \* 100  
 ) %>%  
 filter(Percent > 0)

## 3.3 Plotting the Graph

ggplot(df\_percent, aes(  
 x = reorder(PA, -Total\_Pop),  
 y = Percent,  
 fill = AgeGroup  
 )) +  
 geom\_bar(stat = "identity",  
 position = "stack",  
 width = 0.5,  
 color = "black") +  
 scale\_y\_continuous(labels = scales::percent\_format(scale = 1)) +  
 scale\_fill\_manual(  
 values = c("Children" = "lightblue", "Adults" = "lightgray", "Seniors" = "navy")  
 ) +  
 labs(  
 title = "Age Structure by Planning Area",  
 x = "Planning Area",  
 y = "Ratio",  
 fill = "Age Group"  
 ) +  
 theme\_minimal(base\_size = 20) +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))



## 3.4 Analysis

By analyzing the population structures of a specific area, media companies are able to advertise the products or the service according to the age distributions. By targeting the majority generation of a specific area, tha advertisements are more likely to reach it’s target customer, further improving the KPI of the expenses.If done properly, the advertisement could enhance the sales of a products significantly.

# 4. Part 2 - Graph Evaluation

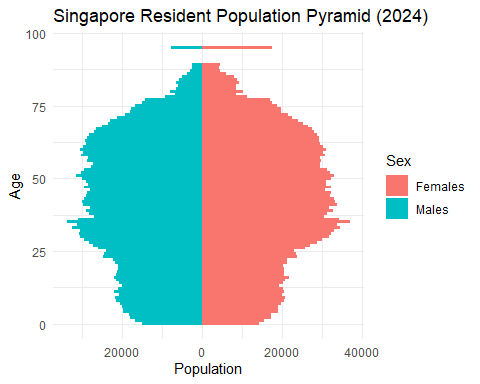
# 4.1 Environment Setup

data <- read\_csv("respopagesex2024/respopagesex2024.csv") %>%  
 mutate(Age = ifelse(Age == "90\_and\_Over", "95", Age),  
 Age = as.numeric(Age)) %>%  
 drop\_na(Age)

# 4.2 Figure 1

# 4.2.1 Plot the Graph

pyramid\_data <- data %>%  
 group\_by(Age, Sex) %>%  
 summarise(Pop = sum(Pop), .groups = "drop") %>%  
 mutate(Pop = if\_else(Sex == "Males", -Pop, Pop))  
  
pyramid\_plot <- ggplot(pyramid\_data, aes(x = Age, y = Pop, fill = Sex)) +  
 geom\_bar(stat = "identity", width = 1) +  
 coord\_flip() +  
 scale\_y\_continuous(labels = abs) +  
 labs(title = "Singapore Resident Population Pyramid (2024)",  
 x = "Age", y = "Population") +  
 theme\_minimal()  
  
pyramid\_plot



# 4.2.2 Comment :

The width of the bars are too thin to distinguish the counterside of the other gender. If the number of the bars can be reduced, the graph will be more clear and easier to compare between genders.

# 4.2.3 Improvement:

library(tidyverse)  
data\_binned <- data %>%  
 mutate(  
 age\_group = cut(  
 Age,  
 breaks = seq(0, 100, by = 5),  
 right = FALSE,  
 labels = paste0(seq(0, 95, by = 5), "-", seq(4, 99, by = 5))  
 ),  
 age\_group = factor(age\_group, levels = paste0(seq(0, 95, by = 5), "-", seq(4, 99, by = 5)))  
 )  
pyramid\_data <- data\_binned %>%  
 group\_by(age\_group, Sex) %>%  
 summarise(Pop = sum(Pop), .groups = "drop") %>%  
 mutate(Pop = if\_else(Sex == "Males", -Pop, Pop))  
ggplot(pyramid\_data, aes(x = age\_group, y = Pop, fill = Sex)) +  
 geom\_col(width = 0.8) +  
 coord\_flip() +  
 scale\_y\_continuous(labels = abs) +  
 labs(  
 title = "Singapore Resident Population Pyramid (2024)",  
 x = "Age Group",  
 y = "Population"  
 ) +  
 theme\_minimal() +  
 theme(  
 axis.text.y = element\_text(size = 8),  
 axis.text.x = element\_text(size = 8)  
 )

