Take-home\_Ex01B

2025-05-07

# Take-home Exercise 1 - Phase 2

**Phase 2**: to selecting one submission provided by classmate, critic three good design principles and three areas for further improvement. With reference to the comment, prepare the makeover version of the data visualisation. I am selecting this submission from other classmate, as shown [here](https://isss608-cindy.netlify.app/take-home_exercise/take-home_ex01/take-home_ex01a).

## **The Designing Tool**

The data should be processed by using appropriate **tidyverse** family of packages and the data visualisation must be prepared using **ggplot2** and **its extensions**.

pacman::p\_load(ggrepel, patchwork,   
 ggthemes, hrbrthemes,  
 tidyverse, ggplot2)

## Import data

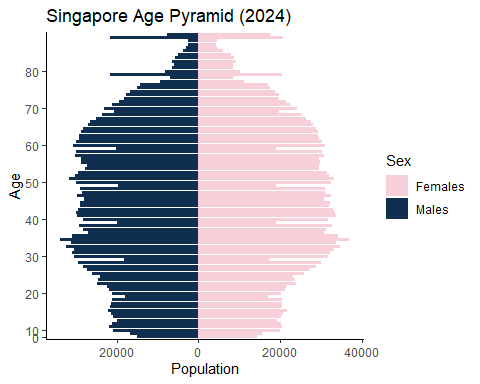
To accomplish the task, *Singapore Residents by Planning Area / Subzone, Single Year of Age and Sex, June 2024* dataset shared by [Department of Statistics, Singapore (DOS)](https://www.singstat.gov.sg/) will be used and we wil load it as follows:

pop\_data <- read\_csv("data/respopagesex2024.csv", col\_names = TRUE)

## Original Data Visualization

### **Population pyramid (Sex vs Age)**

pyramid\_data <- pop\_data %>%  
 group\_by(Age, Sex) %>%  
 summarise(Pop = sum(Pop), .groups = "drop") %>%  
 mutate(Pop = ifelse(Sex == "Males", -Pop, Pop))  
  
ggplot(pyramid\_data, aes(x = Age, y = Pop, fill = Sex)) +  
 geom\_bar(stat = "identity") +  
 coord\_flip() +  
 scale\_y\_continuous(labels = abs) +  
 scale\_fill\_manual(values = c("Males" = "#102E50", "Females" = "#F7CFD8")) +   
 scale\_x\_discrete(breaks = seq(0, 100, by = 10)) +  
 labs(title = "Singapore Age Pyramid (2024)",  
 x = "Age", y = "Population") +  
 theme\_classic()



**Comments**:

**Three good design principles:**

* Different contrast colours are used to differentiate between Males and Females Resident count
* Appropriate use of chart (pyramid chart) instead of points to represent discrete values
* Clear labeling for chart title, legend, x axis title and y axis title

**Three areas for further improvement**

* **Axis Labeling and Scale Consistency**
* Confusing x axis (starts from 0) and y axis (0 starts from the center) labeling. And, the distance between 0 to 10 to 20 on the y-axis is not consistent which signalling there is an error in the data preparation. The y-axis values ranges from 0 mark to 80 mark, whereas the dataset for Age ranges from 0 to 90+. There should be a 90 mark in the y-axis to represent the clearer and more accurate representation of the chart.
* **Graphical Integrity**: The top chart shows 2 wide bars after tapering off of the top pyramid which is not a true representation of the dataset. There should be a data cleaning performed before the visualization to change the data type for Age column from strings to integer and there is 1 value in the Age column: 90\_and\_Over that needs to be recoded to a numeric number. The lack of data preparation has led to the wrong representation of the data in this chart.
* **Group Ages into 5-Year Bins**: The current age pyramid displays age in single-year intervals, resulting in a visually dense and harder-to-read chart. Binning the ages into 5-year groups (e.g., 0–4, 5–9, …, 85–89, 90+) would simplify the structure and highlight broader population trends more effectively. Additionally, including a vertical line to indicate the median age would provide a valuable reference point, making it easier to interpret the overall age distribution and identify demographic imbalance

**Makeover version of the Chart**

library(dplyr)  
library(ggplot2)  
library(scales)  
  
# Data preparation  
pyramid\_data <- pop\_data %>%  
 mutate(Age = ifelse(Age == "90\_and\_Over", "91", Age),  
 Age = as.numeric(Age)) %>%  
 filter(!is.na(Age)) %>%  
 mutate(AgeGroup = cut(Age,  
 breaks = c(0, 5, 10, 15, 20, 25, 30, 35, 40, 45,  
 50, 55, 60, 65, 70, 75, 80, 85, 90, Inf),  
 right = TRUE,  
 include.lowest = TRUE,  
 labels = c("0–5", "6–10", "11–15", "16–20", "21–25",  
 "26–30", "31–35", "36–40", "41–45", "46–50",  
 "51–55", "56–60", "61–65", "66–70", "71–75",  
 "76–80", "81–85", "86–90", "90+"))) %>%  
 group\_by(AgeGroup, Sex) %>%  
 summarise(Pop = sum(Pop, na.rm = TRUE), .groups = "drop") %>%  
 mutate(Pop = ifelse(Sex == "Males", -Pop, Pop),  
 Label = comma(abs(Pop), accuracy = 1),  
 Rank = dense\_rank(desc(AgeGroup)))  
  
median\_group <- pyramid\_data %>%  
 group\_by(AgeGroup) %>%  
 summarise(TotalPop = sum(abs(Pop))) %>%  
 mutate(CumSum = cumsum(TotalPop),  
 MedianFlag = CumSum >= sum(TotalPop) / 2) %>%  
 filter(MedianFlag) %>%  
 slice(1) %>%  
 pull(AgeGroup)  
  
# Plot  
ggplot(pyramid\_data, aes(x = AgeGroup, y = Pop, fill = Sex)) +  
 geom\_bar(stat = "identity", width = 0.9) +  
 geom\_text(aes(label = Label,  
 hjust = case\_when(  
 AgeGroup %in% c("90+", "86–90") & Sex == "Males" ~ 1.1,  
 AgeGroup %in% c("90+", "86–90") & Sex == "Females" ~ -0.1,  
 Sex == "Males" ~ 0, # centered inside left bar  
 Sex == "Females" ~ 1.5 # centered inside right bar  
 )),  
 size = 3, color = "black") + # <- Added the missing plus sign here  
 geom\_vline(xintercept = 0, color = "black") +  
 coord\_flip() +  
 scale\_y\_continuous(labels = NULL, breaks = NULL) + # Remove x-axis tick values  
 scale\_fill\_manual(values = c("Males" = "#7EC8E3", "Females" = "#F7CFD8")) +  
 labs(title = "Singapore Population Age Pyramid (2024)",  
 x = "Age Group (Years)", y = NULL, fill = "Sex") +  
 theme\_minimal(base\_size = 11) +  
 theme(plot.title = element\_text(size = 14, face = "bold"),  
 axis.text.y = element\_text(size = 9),  
 panel.grid.major.y = element\_blank(),  
 legend.position = "right") +  
 annotate("text", x = median\_group, y = 0,  
 label = paste("Median:", median\_group),  
 vjust = -0.8, fontface = "italic", color = "gray40", size = 4)

