Take-Home\_Ex01 part 2

Teo Wee Siang Roy

2025-05-08

# Reviewed Submission

This review is based on the submission by **Sandra Jacob**, available at:  
🔗 <https://sandrajacob-isss608.netlify.app/takehome_exercises/takehome_ex01/takehomeex01>

# Load packages

pacman::p\_load(tidyverse, haven,  
 ggrepel, ggthemes,  
 ggridges, ggdist,colorspace,ggdist,  
 patchwork, scales, ggplot2,   
 matrixStats, treemapify, showtext)

# Load dataset and data processing

respopagesex2024 <- read\_csv("TH01\_2/data/respopagesex2024.csv")  
show\_col\_types = FALSE  
  
knitr::kable(head(respopagesex2024))

| PA | SZ | Age | Sex | Pop | Time |
| --- | --- | --- | --- | --- | --- |
| Ang Mo Kio | Ang Mo Kio Town Centre | 0 | Males | 10 | 2024 |
| Ang Mo Kio | Ang Mo Kio Town Centre | 0 | Females | 10 | 2024 |
| Ang Mo Kio | Ang Mo Kio Town Centre | 1 | Males | 10 | 2024 |
| Ang Mo Kio | Ang Mo Kio Town Centre | 1 | Females | 10 | 2024 |
| Ang Mo Kio | Ang Mo Kio Town Centre | 2 | Males | 10 | 2024 |
| Ang Mo Kio | Ang Mo Kio Town Centre | 2 | Females | 10 | 2024 |

respopagesex2024[duplicated(respopagesex2024),]

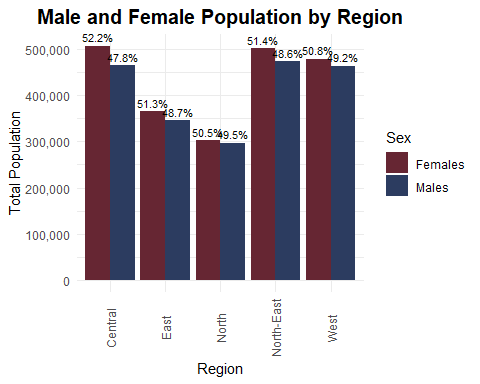
# A tibble: 0 × 6  
# ℹ 6 variables: PA <chr>, SZ <chr>, Age <chr>, Sex <chr>, Pop <dbl>,  
# Time <dbl>

planning\_areas <- read\_csv("TH01\_2/data/Planning\_areas\_of\_Singapore\_2.csv")

respop\_with\_region <- respopagesex2024 %>%  
 left\_join(planning\_areas, by = c("PA" = "Name (English)"))  
  
respop\_cleaned <- respop\_with\_region %>%  
 select(-`Time`, -`Malay`, -`Chinese`, -`Pinyin`, -`Tamil`, -`Area (km2)`, -`Population`, -`Density (/km2)`, -`Subzones`)

# Gender Ratio Analysis by Region (1st plot)

gender\_summary\_long <- respop\_cleaned %>%  
 group\_by(Region, Sex) %>%  
 summarise(Total\_Pop = sum(Pop), .groups = "drop")  
  
#Calculate total population per region  
gender\_summary\_long <- gender\_summary\_long %>%  
 group\_by(Region) %>%  
 mutate(  
 Region\_Total = sum(Total\_Pop),  
 Percent = (Total\_Pop / Region\_Total) \* 100  
 )  
  
#Plot side-by-side bars  
ggplot(gender\_summary\_long, aes(x = Region, y = Total\_Pop, fill = Sex)) +  
 geom\_col(position = "dodge") +  
 geom\_text(  
 aes(label = paste0(round(Percent, 1), "%")),  
 position = position\_dodge(width = 0.9),  
 vjust = -0.5,  
 size = 3, family = "nunito"  
 ) +  
 labs(  
 title = "Male and Female Population by Region",  
 x = "Region",  
 y = "Total Population"  
 ) +  
 scale\_y\_continuous(labels = scales::comma) +  
 theme\_minimal(base\_family = "nunito") + theme(axis.text.x = element\_text(angle = 90, vjust = 0.5),plot.title = element\_text(hjust = 0.5, size = 15, face = "bold", family = "nunito")) +  
 scale\_fill\_manual(values = c("Females" = "#662633", "Males" = "#2C3C60"))



## ✅ Three Good Design Principles Applied

1. **Clear and Descriptive Title with Proper Axis Labels**

The chart’s title — *“Male and Female Population by Region”* — succinctly conveys the purpose of the visualisation. Coupled with clearly labeled axes (*“Region”* and *“Total Population”*), this design adheres to fundamental **graphical integrity** and supports rapid comprehension. It aligns with best practices for enhancing **cognitive accessibility** in data graphics.

1. **Strategic Use of Color for Categorical Differentiation**

The use of high-contrast, intuitive colors (maroon for Females and navy for Males) effectively encodes the **categorical variable (Sex)**. This leverages **pre-attentive visual processing**, enabling users to instantly distinguish between gender groups. The visual encoding is both **aesthetically pleasing and functionally efficient**, which is crucial in comparative visualisations.

1. **Inclusion of Percentage Labels to Add Analytical Depth**

Each bar is annotated with percentage labels that represent the proportion of each gender within its respective region. This additional layer of information transforms a basic bar chart into a more **analytically enriched visual**, allowing users to compare both absolute values and relative gender compositions simultaneously. This reflects the principle of **“deriving insight” rather than merely displaying raw data”** — a key goal in visual analytics.

## ❌ Three Areas for Further Improvement

1. **Label Overlap and Legibility Concerns**

While the inclusion of percentage labels is commendable, their placement **above the bars** results in visual congestion—especially in regions like *North-East* and *West*, where the height differences between segments are minimal. This diminishes readability and increases cognitive load. To resolve this:

* Place labels **inside the bars** using position\_stack(vjust = 0.5)
* Use a **horizontal layout** to increase label spacing
* Reduce font size or apply ggrepel for smart label adjustment

1. **Lack of Alignment Between Data and Message**

Although percentage labels are presented, the **bar height still encodes absolute values**. This can lead viewers to focus on population differences rather than proportional differences. A **100% stacked bar chart** would better align with the intent to highlight **gender balance within each region**.

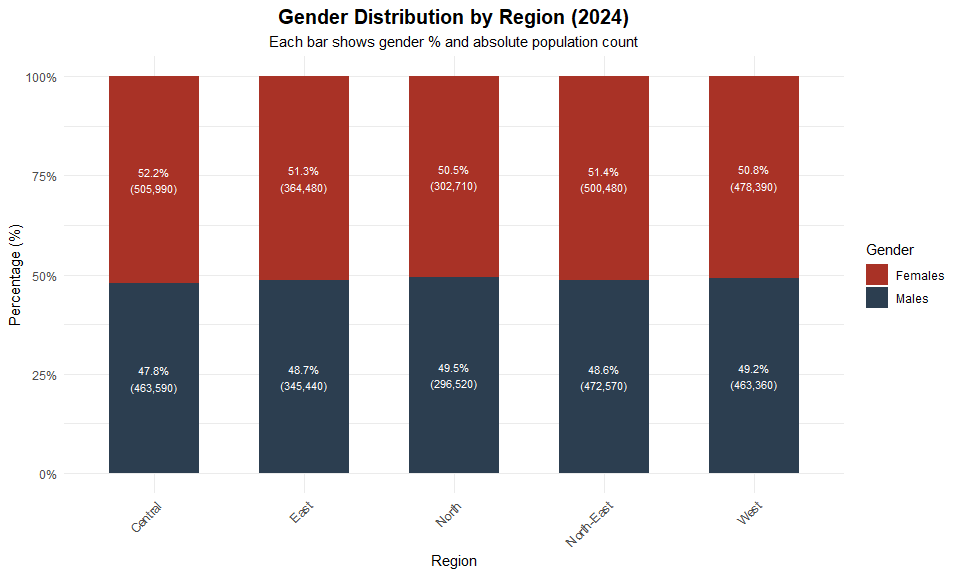
1. **Suboptimal Chart Type for Comparative Ratios**

Grouped bar charts are ideal for comparing totals but are less effective for showing **composition within a group**. A **normalized horizontal stacked bar chart** would:

* More clearly show **gender proportion**
* Eliminate vertical label collisions
* Follow a natural **left-to-right reading flow** that supports comparison

## 🔧 Improvement 1: What Was Improved

gender\_summary\_long <- respop\_cleaned %>%  
 group\_by(Region, Sex) %>%  
 summarise(Total\_Pop = sum(Pop), .groups = "drop") %>%  
 group\_by(Region) %>%  
 mutate(  
 Region\_Total = sum(Total\_Pop),  
 Percent = (Total\_Pop / Region\_Total) \* 100,  
 Label = paste0(round(Percent, 1), "%\n(", scales::comma(Total\_Pop), ")")  
 )  
  
# Plot: stacked bar with % + population inside bars  
ggplot(gender\_summary\_long, aes(x = Region, y = Percent, fill = Sex)) +  
 geom\_col(position = "stack", width = 0.6) +  
  
 # Percentage + population inside each segment  
 geom\_text(  
 aes(label = Label),  
 position = position\_stack(vjust = 0.5),  
 color = "white",  
 size = 3, family = "nunito"  
 ) +  
  
 labs(  
 title = "Gender Distribution by Region (2024)",  
 subtitle = "Each bar shows gender % and absolute population count",  
 x = "Region",  
 y = "Percentage (%)",  
 fill = "Gender"  
 ) +  
 scale\_y\_continuous(labels = scales::percent\_format(scale = 1)) +  
 scale\_fill\_manual(values = c("Females" = "#A93226", "Males" = "#2C3E50")) +  
 theme\_minimal(base\_family = "nunito") +  
 theme(  
 axis.text.x = element\_text(angle = 45, hjust = 1),  
 plot.title = element\_text(hjust = 0.5, face = "bold", size = 15),  
 plot.subtitle = element\_text(hjust = 0.5, size = 11)  
 )



This revised visualisation enhance both analytical clarity and graphical integrity:

### 1. From Absolute Counts to Proportional Comparison

* **Original**: Displayed only raw population counts for each gender, which made it difficult to compare the gender distribution within each region.
* **Improved**: Transformed into a **100% stacked bar chart** that normalises each region to 100%, making it much easier to compare **gender ratios** across regions.

### 2. Embedded Dual-Level Information in Labels

* **Original**: Included either percentage or total population as labels, but not both.
* **Improved**: Combined both **percentage and absolute count** in each label (e.g., 48.7%\n(123,456)), allowing the viewer to simultaneously interpret **relative proportion and actual magnitude**.

### 3. Improved Label Positioning and Readability

* **Original**: Percentage labels were placed outside the bars, leading to overlapping and reduced clarity, especially for regions with similar male/female proportions.
* **Improved**: Labels are now **placed inside the bar segments** using position\_stack(vjust = 0.5), increasing legibility and reducing visual clutter.

### 4. Chart Type Aligns with Analytical Purpose

* **Original**: Used a grouped bar chart, which is more appropriate for comparing absolute values across categories.
* **Improved**: Switched to a **stacked bar layout**, which better supports the analytical objective of comparing **gender balance within each region**.

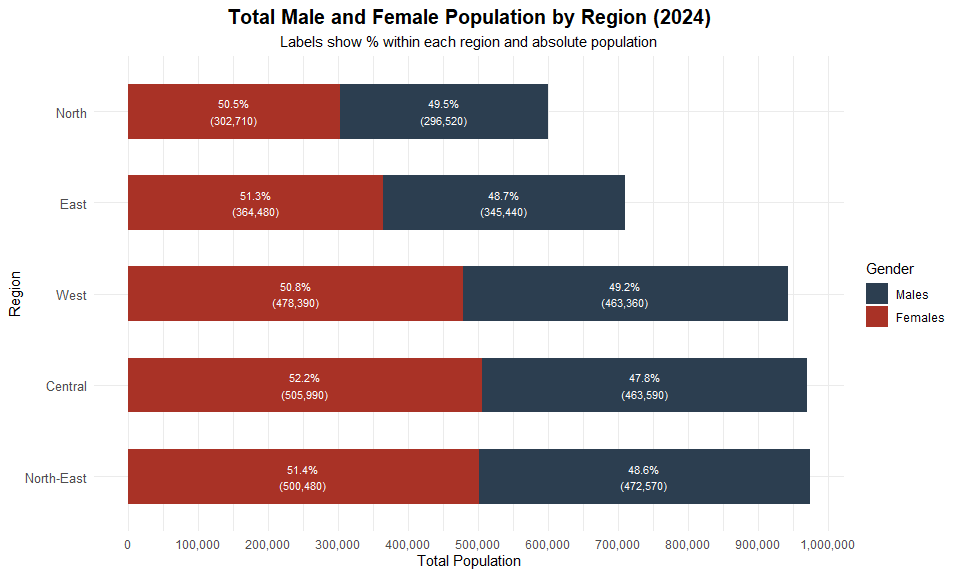
### Improvement outlines

These improvements follow the visual design principles outlined in *Lesson 2: Designing Graphs to Enlighten*, particularly:

* Using the **right chart to match the message**
* Enhancing **interpretability** through derived metrics
* Leveraging **pre-attentive features** like color and label placement

## 🔧 Improvement Summary 2: What Was Further Improved

gender\_summary\_long <- respop\_cleaned %>%  
 group\_by(Region, Sex) %>%  
 summarise(Total\_Pop = sum(Pop), .groups = "drop") %>%  
 group\_by(Region) %>%  
 mutate(  
 Region\_Total = sum(Total\_Pop),  
 Region\_Percent = round((Total\_Pop / Region\_Total) \* 100, 1),  
 Label = paste0(Region\_Percent, "%\n(", scales::comma(Total\_Pop), ")")  
 ) %>%  
 ungroup() %>%  
 mutate(  
 Region = fct\_reorder(Region, Region\_Total, .desc = TRUE), # sort by Region\_Total  
 Sex = factor(Sex, levels = c("Males", "Females")) # ensure stacking order  
 ) %>%  
 arrange(Region, Sex) # ensure correct stacking order  
  
# Plot  
ggplot(gender\_summary\_long, aes(y = Region, x = Total\_Pop, fill = Sex)) +  
 geom\_col(position = "stack", width = 0.6) +  
 geom\_text(  
 aes(label = Label),  
 position = position\_stack(vjust = 0.5),  
 color = "white",  
 size = 3,  
 family = "nunito"  
 ) +  
 labs(  
 title = "Total Male and Female Population by Region (2024)",  
 subtitle = "Labels show % within each region and absolute population",  
 x = "Total Population",  
 y = "Region",  
 fill = "Gender"  
 ) +  
 scale\_x\_continuous(  
 labels = scales::label\_comma(),  
 breaks = seq(0, 1000000, 100000)  
 ) +  
 scale\_fill\_manual(values = c("Females" = "#A93226", "Males" = "#2C3E50")) +  
 theme\_minimal(base\_family = "nunito") +  
 theme(  
 axis.text.y = element\_text(size = 10),  
 plot.title = element\_text(hjust = 0.5, face = "bold", size = 15),  
 plot.subtitle = element\_text(hjust = 0.5, size = 11)  
 )



Building on the earlier stacked percentage chart, this updated visualisation improves the clarity, comparability, and contextual storytelling of the gender distribution across regions in 2024.

### 1. Switched from Percentage to Actual Population on the X-Axis

* **Previous**: X-axis showed gender **percentage share**, which helped with proportion comparison but hid the actual population scale.
* **Improved**: Uses **raw population values** on the x-axis (with comma formatting and 100k breaks), enabling clear comparison of **both gender ratio and total population** per region.

### 2. Horizontal Layout Enhances Label Readability and Comparison

* **Previous**: Vertical bars made category names harder to read, and labels could overlap in crowded segments.
* **Improved**: **Horizontal stacked bars** make region names easier to scan and reduce label collision, improving overall **legibility** and alignment with **natural left-to-right reading flow**.

### 3. Regions Sorted by Total Population

* **Previous**: Regions were presented in arbitrary or default order, which reduced analytical clarity.
* **Improved**: Regions are now **sorted in descending order of total population**, highlighting population scale differences and making the chart easier to interpret at a glance.

### 4. Maintained Rich, Dual-Layer Labels

* Each bar segment still includes a **concise label** that combines the percentage share and absolute count (e.g., 50.5% (302,710)), enabling multi-level insight in a compact, readable form.

### Improvement outlines

These refinements not only retain the strengths of the previous version but enhance the visual’s ability to communicate both **structure (proportions)** and **magnitude (totals)** — an important principle covered in *Lesson 2: Designing Graphs to Enlighten*.