Take-home Exercise 1 (Part 2)

Quynh Trang Do

2025-05-08

# Load packages and import data

## Load packages

pacman::p\_load(ggrepel, patchwork,   
 ggthemes, hrbrthemes,  
 tidyverse, ggiraph, plotly,   
 patchwork, DT, dplyr, scales, forcats, dplyr, grid)

## Import data

df <- read\_csv("respopagesex2024.csv")

# Original submission

The original visualization was created by my classmate Andre Ong Jia Kang. Please refer to [this link](https://isss608-ay2025-aprilterm.netlify.app/take-home_ex/take-home_ex01/take-home_ex01) for the original submission.

## Data cleaning and preparation

I follow the same data cleaning and preparation procedure as the original submission to ensure the same output.

### Adding “Region” column

df <- df %>%  
 mutate(  
 Region = case\_when(  
 # Central Region (Core + Rest)  
 PA %in% c(  
 "Downtown Core", "Outram", "Sentosa", "Rochor", "Orchard", "Newton",  
 "River Valley", "Bukit Timah", "Holland Road", "Tanglin", "Novena",  
 "Thomson", "Marina East", "Marina South", "Museum", "Singapore River",  
 "Bishan", "Bukit Merah", "Geylang", "Kallang", "Marine Parade",  
 "Queenstown", "Southern Islands", "Toa Payoh", "Straits View"  
 ) ~ "Central",   
  
 # North Region  
 PA %in% c(  
 "Central Water Catchment", "Lim Chu Kang", "Mandai", "Sembawang",  
 "Simpang", "Sungei Kadut", "Woodlands", "Yishun"  
 ) ~ "North",   
  
 # North-East Region  
 PA %in% c(  
 "Ang Mo Kio", "Hougang", "North-Eastern Islands", "Punggol",  
 "Seletar", "Sengkang", "Serangoon"  
 ) ~ "North-East",   
  
 # East Region  
 PA %in% c("Bedok", "Changi", "Changi Bay", "Paya Lebar", "Pasir Ris", "Tampines"  
 ) ~ "East",   
   
 # West Region  
 PA %in% c(  
 "Bukit Batok", "Bukit Panjang", "Boon Lay", "Pioneer", "Choa Chu Kang",  
 "Clementi", "Jurong East", "Jurong West", "Tengah", "Tuas",  
 "Western Islands", "Western Water Catchment", "Benoi", "Ghim Moh",  
 "Gul", "Pandan Gardens", "Jurong Island", "Kent Ridge", "Nanyang",  
 "Teban Gardens", "Toh Tuck", "Tuas South", "West Coast"  
 ) ~ "West",   
  
 # Catch-all for any PAs that didn’t match  
 TRUE ~ NA\_character\_  
 )  
 )

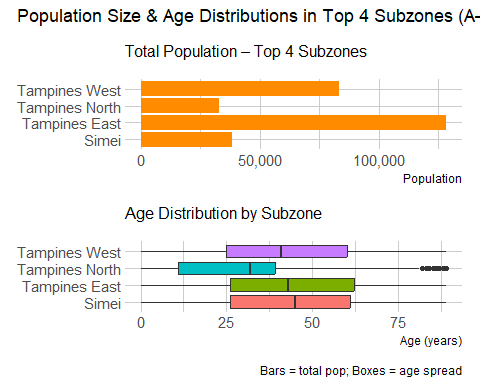
### Removing all rows with population sum equal to 0

df2 <- df %>%  
 group\_by(PA) %>%  
 filter(sum(Pop, na.rm = TRUE) > 0) %>%  
 ungroup()

## Visualization

This visualization demonstrates the top 4 sub-zones with highest population in Tampines planning area and their corresponding box plots of age distribution.

## Plot



## Original code

# 1) Find & alphabetize your top 4 subzones  
top4\_sz <- df2 %>%  
 filter(PA == "Tampines") %>%  
 group\_by(SZ) %>%   
 summarise(TotalPop = sum(Pop), .groups = "drop") %>%   
 slice\_max(TotalPop, n = 4) %>%   
 pull(SZ)  
  
alpha\_levels <- sort(top4\_sz)  
  
# 2) Bar‐chart data (with ordered factor)  
bar\_data <- df2 %>%   
 filter(SZ %in% top4\_sz) %>%   
 group\_by(SZ) %>%   
 summarise(TotalPop = sum(Pop), .groups = "drop") %>%  
 mutate(SZ = factor(SZ, levels = alpha\_levels))  
  
p\_bar <- ggplot(bar\_data, aes(x = SZ, y = TotalPop)) +  
 geom\_col(fill = "darkorange") +  
 coord\_flip() +  
 scale\_y\_continuous(labels = comma) +  
 labs(subtitle = "Total Population – Top 4 Subzones", x = NULL, y = "Population") +  
 theme\_ipsum(plot\_margin = margin(8, 8, 8, 8))  
  
# 3) Boxplot data (with same ordered factor)  
box\_data <- df2 %>%   
 filter(SZ %in% top4\_sz) %>%   
 mutate(  
 Age = as.numeric(Age),  
 SZ = factor(SZ, levels = alpha\_levels)  
 ) %>%   
 select(SZ, Age, Pop) %>%   
 uncount(weights = Pop)  
  
p\_box <- ggplot(box\_data, aes(x = SZ, y = Age, fill = SZ)) +  
 geom\_boxplot(alpha = 1.5, show.legend = FALSE) +  
 coord\_flip() +  
 labs(subtitle = "Age Distribution by Subzone", x = NULL, y = "Age (years)") +  
 theme\_ipsum(plot\_margin = margin(8, 8, 8, 8))  
  
# 4) Combine  
(p\_bar / p\_box) +  
 plot\_annotation(  
 title = "Population Size & Age Distributions in Top 4 Subzones (A–Z)",  
 caption = "Bars = total pop; Boxes = age spread"  
 )

# Comments

## Good design practices

1. **Suitable choice of visualizations to represent quantity and distribution**

* The use of a horizontal bar chart effectively conveys total population size—an intuitive and direct way to compare the number of residents across subzones. Paired with this is a boxplot showing the age distribution within the same subzones, which illustrates not just the median age but also the range, concentration, and presence of outliers in the population. By combining these two perspectives—how many people live in each area and how their ages are spread—the chart enables a more comprehensive analysis than using either visualization alone.

1. **Alignment of axis components and use of a horizontal layout**

* Both the bar chart and the boxplot share the same categorical y-axis representing subzones, and the subzones appear in the same order across both visualizations (e.g., Tampines, Simei, Tampines North, etc.). This consistent alignment allows viewers to easily scan across charts and make immediate, accurate comparisons between total population and age distribution for each subzone without reorienting their focus. Additionally, the horizontal layout improves legibility, especially for subzone names, which are quite long and would be harder to read if rotated vertically.

1. **Logical visual hierarchy**

* The chart’s visual hierarchy reinforces a clear top-down logic that aligns with how viewers naturally process information. By placing the bar chart at the top, it first draws attention to the total population size of each subzone. This is the broad, high-level metric. Beneath it, the boxplot introduces a deeper layer of detail by showing the age distribution, which adds nuance and context to the population figures. This sequence helps direct the viewers to move from general to specific insights: starting with “how many people live here” and then exploring “what is the demographic structure of that population.”

## Areas for improvement

1. **Sorting sub-zones in a logical order**

* The chart displays the top four subzones with the highest population in Tampines, but their arrangement does not follow a consistent or meaningful order—such as descending population size or alphabetical sequence. For instance, Tampines North is positioned above Simei, even though it has a lower population. This weakens the visual impact and does not fully leverage the bar chart’s strength in showing ranked comparisons at a glance. Reordering the bars by population size would improve readability, enhance pattern recognition, and allow viewers to instantly grasp which subzones are more or less populous.

1. **Aligning color coding and adding legend**

* In the current design, the boxplot uses distinct colors to represent each subzone, but the bar chart displays all bars in the same color. Applying the same color scheme to the bar chart would help visually link the two components, improving overall cohesion and visual tracking. Furthermore, the chart lacks a legend to explain the color assignments, which can create confusion and reduce interpretability, especially if more subzones are added. Including a clear, consistent legend would make the chart more intuitive, accessible, and self-explanatory.

1. **Adjusting title and adding data labels**

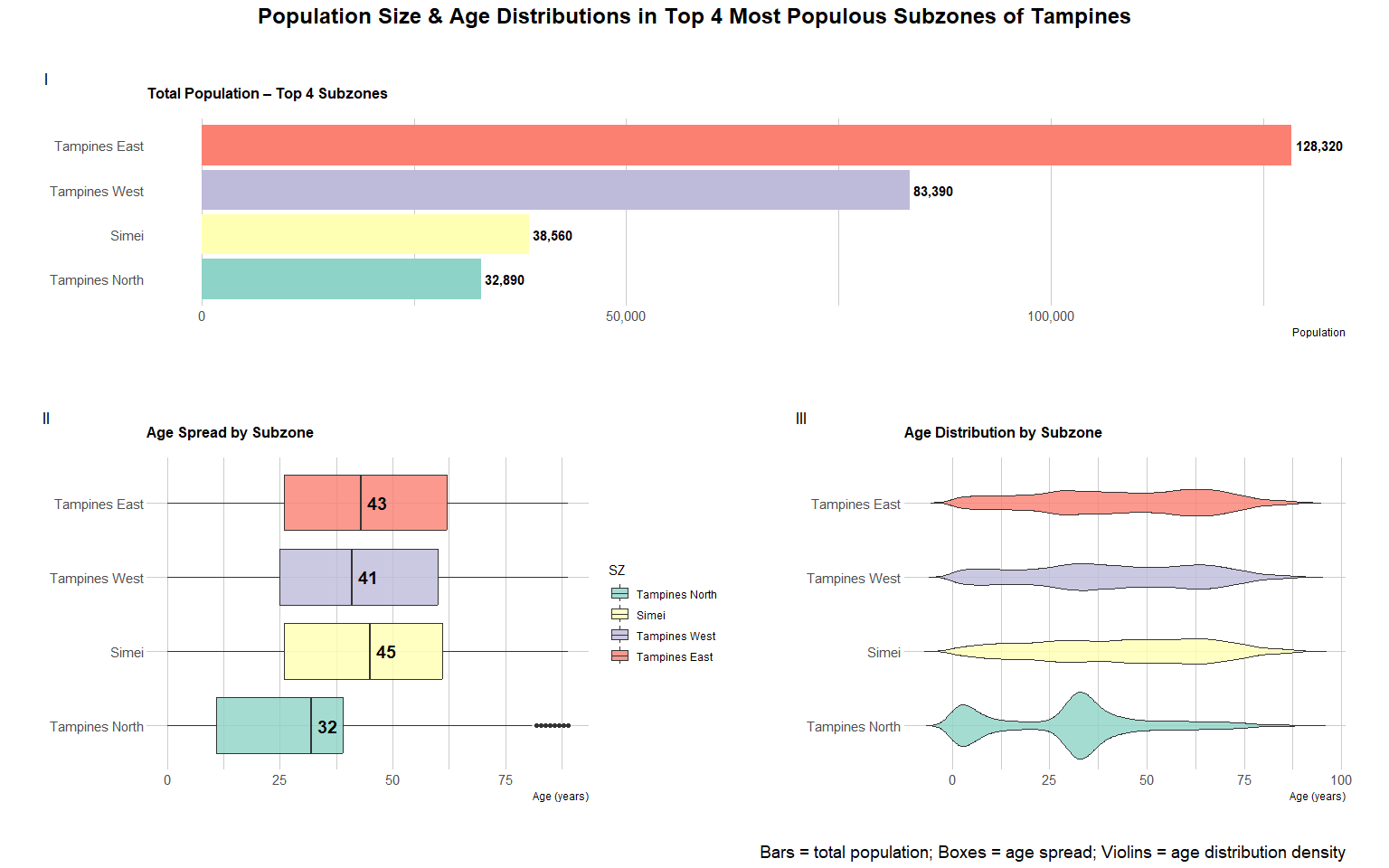
* The original title, “Population Size & Age Distributions in Top 4 Subzones (A–Z),” lacks clarity and precision in a few key ways. First, it does not specify the basis for selecting the top 4 subzones, leaving viewers uncertain about whether the ranking is based on population, age diversity, or another metric. Furthermore, the title omits any geographical context, such as the fact that all subzones belong to the Tampines planning area, which is essential for interpretation.
* The original chart doesn’t include any data label. Adding labels such as total population figures in the bar chart and median age values in the boxplot can help enhance interpretability by providing precise values at a glance, eliminating the need for visual estimation and making the chart more informative and accessible.

1. **Reveal distribution patterns more clearly with Violin Plots**

* While the boxplot effectively displays the median, interquartile range, and outliers, it does not capture the distribution shape or skewness of the data. Adding a violin plot would address this limitation by showing the full density of the age distribution for each subzone. This allows viewers to better understand how age groups are concentrated—whether the population is skewed toward younger or older ages, or if there are multiple peaks. A violin plot provides richer insight into the underlying demographic patterns, making it a more informative alternative for visualizing population spread.

# Makeover

## Plot



## Code

# 1. Get top 4 sub-zones by population  
top4\_sz <- df2 %>%  
 filter(PA == "Tampines") %>%  
 group\_by(SZ) %>%   
 summarise(TotalPop = sum(Pop), .groups = "drop") %>%  
 slice\_max(TotalPop, n = 4)  
  
ordered\_levels <- top4\_sz %>%  
 arrange(desc(TotalPop)) %>%  
 pull(SZ) %>%  
 rev()  
  
# Color palette (match across plots)  
sz\_colors <- RColorBrewer::brewer.pal(n = 4, name = "Set3")  
names(sz\_colors) <- ordered\_levels  
  
# 2. Bar chart  
bar\_data <- df2 %>%   
 filter(SZ %in% ordered\_levels) %>%  
 group\_by(SZ) %>%   
 summarise(TotalPop = sum(Pop), .groups = "drop") %>%  
 mutate(SZ = factor(SZ, levels = ordered\_levels))  
  
p\_bar <- ggplot(bar\_data, aes(x = SZ, y = TotalPop, fill = SZ)) +  
 geom\_col(show.legend = FALSE) +  
 geom\_text(aes(label = comma(TotalPop)),   
 hjust = -0.1,   
 size = 4,   
 fontface = "bold") +  
 coord\_flip() +  
 scale\_fill\_manual(values = sz\_colors, name = "Subzone") +  
 scale\_y\_continuous(labels = comma) +  
 labs(subtitle = "Total Population – Top 4 Subzones", x = NULL, y = "Population") +  
 theme\_ipsum(base\_family="sans") +  
 theme(plot.subtitle = element\_text(face = "bold"),   
 panel.grid.major.y = element\_blank())  
  
# 3. Boxplot data  
box\_data <- df2 %>%   
 filter(SZ %in% ordered\_levels) %>%   
 mutate(  
 Age = as.numeric(Age),  
 SZ = factor(SZ, levels = ordered\_levels)  
 ) %>%   
 select(SZ, Age, Pop) %>%   
 uncount(weights = Pop)  
  
p\_box <- ggplot(box\_data, aes(x = SZ, y = Age, fill = SZ)) +  
 geom\_boxplot(show.legend = TRUE, alpha = 0.8) +  
 stat\_summary(fun = median, geom = "text", aes(label = round(..y.., 1)),   
 size = 5, fontface = "bold", color = "black", hjust = -0.3) +  
 coord\_flip() +  
 scale\_fill\_manual(values = sz\_colors) +  
 labs(subtitle = "Age Spread by Subzone", x = NULL, y = "Age (years)") +  
 theme\_ipsum(base\_family="sans") +  
 theme(plot.subtitle = element\_text(face = "bold"))  
  
# 4. Violin plot  
p\_violin <- ggplot(box\_data, aes(x = SZ, y = Age, fill = SZ)) +  
 geom\_violin(trim = FALSE, show.legend = FALSE, alpha = 0.8) +  
 coord\_flip() +  
 scale\_fill\_manual(values = sz\_colors) +  
 labs(subtitle = "Age Distribution by Subzone", x = NULL, y = "Age (years)") +  
 theme\_ipsum(base\_family="sans") +  
 theme(plot.subtitle = element\_text(face = "bold"))  
  
# 5. Combine all three charts   
(p\_bar / (p\_box | p\_violin)) +  
 plot\_layout(heights = c(0.6, 1)) +  
 plot\_annotation(  
 title = "Population Size & Age Distributions in Top 4 Most Populous Subzones of Tampines",  
 caption = "Bars = total population; Boxes = age spread; Violins = age distribution density",  
 theme = theme(  
 plot.title = element\_text(hjust = 0.5, size = 18, face = "bold"),  
 plot.caption = element\_text(size = 14)  
 ),  
 tag\_levels = "I"  
 )

## Key changes made

* **Changed the title** from a vague and potentially misleading one to  
  “Population Size & Age Distributions in Top 4 Most Populous Subzones of Tampines.” This explicitly states the selection criteria (population-based) and adds geographic context.
* **Subzones are now sorted by** **descending total population**, which makes comparisons more intuitive.
* Added **population labels** to the bar chart for exact values. Added **median age labels** to the boxplot to highlight central tendency, improving interpretability.
* Assigned **consistent subzone colors across all three charts** (bar, box, and violin). A shared legend placed centrally on the right helps reinforce this visual mapping.
* Included a third panel: a **violin plot** beneath the boxplot to reveal the **density and skewness** of age distributions.