Take-Home\_Ex01\_Part1

Andre Ong Jia Kang

2025-05-01

## Overview

This study is obtained from a local Singapore dataset to reveal how Singapore’s resident population is distributed by age and sex across its various planning areas and subzones as of June 2024. It helps to map age cohorts geographically, examine sex-ratio, key-demographic locations and population density.

I will be taking a look at the dataset to provide some key insights of the following stated below:

1. Regional breakdown for policy makers allocating services such as transport based on geographic distributions.
2. Target professionals providing community healthcare or age-targeted programs for the population.

## Data Cleaning

### Launching R packages

Launching R packages and libraries required.

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

The dataset “Singapore Residents by Planning Area / Subzone, Single Year of Age and Sex, June 2024” was obtained from Department of Statistics, Singapore.

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

Before I analyse the data, I will have a preview of the dataset to find out the number of columns ad character type of it.

glimpse(df)

Rows: 60,424  
Columns: 6  
$ PA <chr> "Ang Mo Kio", "Ang Mo Kio", "Ang Mo Kio", "Ang Mo Kio", "Ang Mo K…  
$ SZ <chr> "Ang Mo Kio Town Centre", "Ang Mo Kio Town Centre", "Ang Mo Kio T…  
$ Age <chr> "0", "0", "1", "1", "2", "2", "3", "3", "4", "4", "5", "5", "6", …  
$ Sex <chr> "Males", "Females", "Males", "Females", "Males", "Females", "Male…  
$ Pop <dbl> 10, 10, 10, 10, 10, 10, 10, 10, 30, 10, 20, 10, 20, 30, 30, 10, 3…  
$ Time <dbl> 2024, 2024, 2024, 2024, 2024, 2024, 2024, 2024, 2024, 2024, 2024,…

**Observations**

* Based on the variable and data type. The dataset consists of 6 columns and 60,424 rows. 4 Columns are Character class and 2 are Number class.
* For Age column, it is in Character class. I will keep it in this state first. In the plots that require transforming of the age data points, I will change it to Number class.
* PA - Planning Area, SZ - Subzone, Age - Single Year of Age, Sex - Male/Female, Pop - Resident Count, Time - Time/Period

Before using the data, I want to check if there is any missing values present.

which(is.na(df))

integer(0)

sum(is.na(df))

[1] 0

Based on the result, there is no missing values or “NA” values. The dataset is ready for analysis.

### Adding a column called “Region” to group the PA into the 5 Singapore Regions

I want to further visualise the PA easier into the 5 Regions in Singapore. Using dplyr’s mutate() + case\_when(), I will map each Planning Area (PA) into one of the five regions North, South, East, West and North-East using the following source website. (https://www.newlaunchesreview.com/regions-of-singapore/)

# 3. Define the mapping and create the new 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", # :contentReference[oaicite:0]{index=0}  
  
 # North Region  
 PA %in% c(  
 "Central Water Catchment", "Lim Chu Kang", "Mandai", "Sembawang",  
 "Simpang", "Sungei Kadut", "Woodlands", "Yishun"  
 ) ~ "North", # :contentReference[oaicite:1]{index=1}  
  
 # North-East Region  
 PA %in% c(  
 "Ang Mo Kio", "Hougang", "North-Eastern Islands", "Punggol",  
 "Seletar", "Sengkang", "Serangoon"  
 ) ~ "North-East", # :contentReference[oaicite:2]{index=2}  
  
 # East Region  
 PA %in% c("Bedok", "Changi", "Changi Bay", "Paya Lebar", "Pasir Ris", "Tampines"  
 ) ~ "East", # :contentReference[oaicite:3]{index=3}  
  
 # 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", # :contentReference[oaicite:4]{index=4}  
  
 # Catch-all for any PAs that didn’t match  
 TRUE ~ NA\_character\_  
 )  
 )  
  
# 4. Inspect  
head(df)

# A tibble: 6 × 7  
 PA SZ Age Sex Pop Time Region   
 <chr> <chr> <chr> <chr> <dbl> <dbl> <chr>   
1 Ang Mo Kio Ang Mo Kio Town Centre 0 Males 10 2024 North-East  
2 Ang Mo Kio Ang Mo Kio Town Centre 0 Females 10 2024 North-East  
3 Ang Mo Kio Ang Mo Kio Town Centre 1 Males 10 2024 North-East  
4 Ang Mo Kio Ang Mo Kio Town Centre 1 Females 10 2024 North-East  
5 Ang Mo Kio Ang Mo Kio Town Centre 2 Males 10 2024 North-East  
6 Ang Mo Kio Ang Mo Kio Town Centre 2 Females 10 2024 North-East

I would like to check the if there are any “NA” values in the Region column.

df %>%  
 summarise(na\_count = sum(is.na(Region)))

# A tibble: 1 × 1  
 na\_count  
 <int>  
1 0

With no “NA” values present, I will proceed with check the Population count on how many “0” values present for the Sum and the respective Planning Areas.

zero\_pop\_pas <- df %>%  
 group\_by(PA) %>%  
 summarise(TotalPop = sum(Pop, na.rm = TRUE), .groups = "drop") %>%  
 filter(TotalPop == 0)  
  
# View the PAs  
print(zero\_pop\_pas)

# A tibble: 13 × 2  
 PA TotalPop  
 <chr> <dbl>  
 1 Boon Lay 0  
 2 Central Water Catchment 0  
 3 Changi Bay 0  
 4 Lim Chu Kang 0  
 5 Marina East 0  
 6 Marina South 0  
 7 North-Eastern Islands 0  
 8 Paya Lebar 0  
 9 Pioneer 0  
10 Simpang 0  
11 Straits View 0  
12 Tuas 0  
13 Western Islands 0

# If you just want a character vector of the names:  
zero\_pa\_names <- zero\_pop\_pas %>% pull(PA)  
print(zero\_pa\_names)

[1] "Boon Lay" "Central Water Catchment"  
 [3] "Changi Bay" "Lim Chu Kang"   
 [5] "Marina East" "Marina South"   
 [7] "North-Eastern Islands" "Paya Lebar"   
 [9] "Pioneer" "Simpang"   
[11] "Straits View" "Tuas"   
[13] "Western Islands"

### Cleaning of dataset by removing Population Sum = 0

There are 13 Planning Areas with Population Sum = 0 is not useful for visualization. I will remove these 13 PA’s from the dataset and call the new dataset as df2.

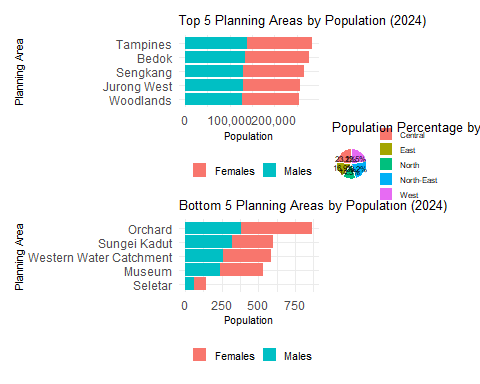
df2 <- df %>%  
 group\_by(PA) %>%  
 filter(sum(Pop, na.rm = TRUE) > 0) %>%  
 ungroup()

With this, I will start the visualization graphs below.

## 1) Top 5 Planning Areas by Total Population

I would want a piechart plot showing the percentage spread of the population in the 5 regions. This will show where majority of the people in the transformed dataset is living at. On top of this, 2 more barchart plots to show the “Top 5” and “Bottom 5” most populated Planning Areas are in Singapore.

# identify top 5 PAs  
top5\_pa <- df2 %>%  
 group\_by(PA) %>%  
 summarise(TotalPop = sum(Pop), .groups = "drop") %>%  
 arrange(desc(TotalPop)) %>%  
 slice\_head(n = 5) %>%  
 pull(PA)  
  
# prepare stacked bars by sex  
p1\_data <- df2 %>%  
 filter(PA %in% top5\_pa) %>%  
 group\_by(PA, Sex) %>%  
 summarise(Pop = sum(Pop), .groups = "drop")  
  
p1 <- ggplot(p1\_data, aes(x = fct\_reorder(PA, Pop), y = Pop, fill = Sex)) +  
 geom\_col() +  
 coord\_flip() +  
 scale\_y\_continuous(labels = comma) +  
 labs(  
 title = "Top 5 Planning Areas by Population (2024)",  
 x = "Planning Area",  
 y = "Population",  
 fill = ""  
 ) +  
 theme\_minimal() +  
 theme(legend.position = "bottom") +  
 theme(  
 plot.margin = margin(5, 5, 5, 5),   
 legend.key.size = unit(0.4, "cm"),   
 legend.text = element\_text(size = 8),   
 plot.title = element\_text(size = 10),  
 axis.title.x = element\_text(size = 8),  
 axis.title.y = element\_text(size = 8)  
 )  
  
## identify bottom 5 PAs  
bottom10\_pa <- df2 %>%  
 group\_by(PA) %>%  
 summarise(TotalPop = sum(Pop), .groups = "drop") %>%  
 arrange(TotalPop) %>%  
 slice\_head(n = 5) %>%  
 pull(PA)  
  
# 2. prepare stacked bars by sex  
p2\_data <- df2 %>%  
 filter(PA %in% bottom10\_pa) %>%  
 group\_by(PA, Sex) %>%  
 summarise(Pop = sum(Pop), .groups = "drop")  
  
# 3. plot  
p2 <- ggplot(p2\_data, aes(x = fct\_reorder(PA, Pop), y = Pop, fill = Sex)) +  
 geom\_col() +  
 coord\_flip() +  
 scale\_y\_continuous(labels = comma) +  
 labs(  
 title = "Bottom 5 Planning Areas by Population (2024)",  
 x = "Planning Area",  
 y = "Population",  
 fill = ""  
 ) +  
 theme\_minimal() +  
 theme(legend.position = "bottom") +  
 theme(  
 plot.margin = margin(5, 5, 5, 5),  
 legend.key.size = unit(0.4, "cm"),  
 legend.text = element\_text(size = 8),  
 plot.title = element\_text(size = 10),  
 axis.title.x = element\_text(size = 8),  
 axis.title.y = element\_text(size = 8)  
 )  
  
# 3a. Pie chart of Region shares  
p3\_data <- df2 %>%   
 group\_by(Region) %>%   
 summarise(TotalPop = sum(Pop, na.rm = TRUE), .groups = "drop") %>%   
 mutate(Percent = TotalPop / sum(TotalPop) \* 100)  
  
# 2) Draw pie chart  
p3 <- ggplot(p3\_data, aes(x = "", y = TotalPop, fill = Region)) +  
 geom\_col(width = 1, color = "white") +  
 coord\_polar(theta = "y") +  
 geom\_text(aes(label = paste0(round(Percent, 1), "%")),  
 position = position\_stack(vjust = 0.5), size = 2) +  
 labs(title = "Population Percentage by Region",  
 x = NULL, y = NULL, fill = NULL) +  
 theme\_void() +  
 theme(  
 plot.title = element\_text(size = 10, margin = margin(b = 5), vjust = -28),  
 legend.position = "right",  
 legend.key.size = unit(0.4, "cm"),  
 legend.text = element\_text(size = 6),  
 legend.title = element\_blank(),  
 plot.margin = margin(t = 5, r = 5, b = 5, l = 5)  
 )  
  
  
combined <- (p1 / p2) | p3 +  
 plot\_layout(widths = c(2, 1.5)) +  
 # --- ADJUSTMENT for overall ---  
 plot\_annotation(theme = theme(  
 plot.margin = margin(10, 10, 10, 10)  
 ))  
  
print(combined)



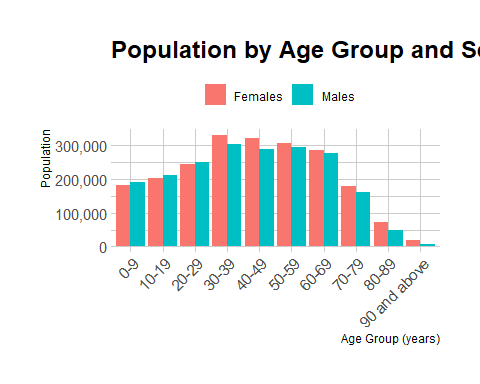
**Discussion:**

* Highest population were from Central, West and North-East Regions based on the Pie Chart.
* The top 5 Planning Areas collectively houses the highest resident counts above 200,000 where Tampines leads the distribution. For the bottom 5, the resident counts are below 1000 with Seletar being the lowest.
* In each, the male/female split is roughly even with no majority of sex in a particular area. Though the bottom 5 areas has a slight female majority for some, it does not accurate reflect the cluster due to its low resident count.
* The differences in bar lengths emphasize varying suburban densities. This can be a good insight for Planners and infrastructure teams to prioritize resource allocation (schools, clinics) where populations are largest.

## 2) Population of male/female across the ages

Based on the initial view, I would like to check the spread of ages across the population in the dataset. To visualize the distribution for male and female, we will plot a bar chart with Age as the X-axis, Population as the Y-axis and bar graphs representing the sex. For the age, I will bin them in 10-year bands for better aesthetic purposes.

df3 <- df2 %>%  
 # ensure Age is numeric  
 mutate(Age = as.numeric(Age)) %>%  
 # now bin into 10-year groups  
 mutate(  
 AgeGroup = cut(  
 Age,  
 breaks = seq(0, 100, by = 10),  
 right = FALSE,  
 labels = paste0(seq(0, 90, by = 10), "-", seq(9, 99, by = 10))  
 )  
 ) %>%  
 group\_by(AgeGroup, Sex) %>%  
 summarise(Pop = sum(Pop), .groups = "drop")  
  
df3 <- df3 %>%  
 mutate(  
 AgeGroup = fct\_explicit\_na(AgeGroup, na\_level = "90 and above")  
 )  
  
p4 <- ggplot(df3, aes(x = AgeGroup, y = Pop, fill = Sex)) +  
 geom\_col(position = "dodge") +  
 scale\_y\_continuous(labels = comma) +  
 labs(  
 title = "Population by Age Group and Sex (2024)",  
 x = "Age Group (years)",  
 y = "Population",  
 fill = ""  
 ) +  
 theme\_ipsum() +  
 theme(  
 axis.text.x = element\_text(angle = 45, hjust = 1),  
 legend.position = "top"  
 )  
  
p4



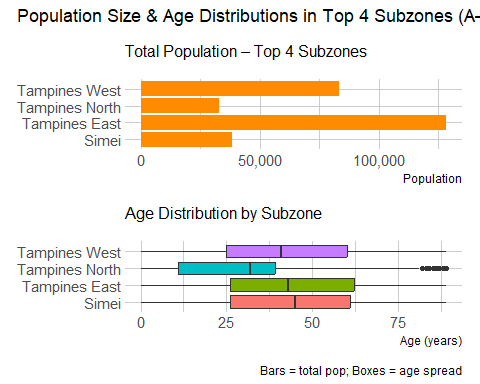
**Discussion:**

* The age distribution has a positive skew with most of the population in the middle age ranges.
* Females do appear slightly higher than males from 30 years old onwards which might reflect women having longer lifespan.However, the disparity isn’t very obvious and it is well balanced distribution.

## 3) Top 4 subzones with highest population with boxplots of the age distribution

I would like to find out the highest 4 subzones in Tampines based on the population graph. To see which subzones contributed to it and what is the age group living there. I will first plot a barchart and also boxplot to visualize these properly.

# 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"  
 )



**Discussion:** - There is a obvious majority of the population in Tampines staying in Tampines West & East.

* Based on the boxplot, the IQR (Lower, Median, Upper Quantile) is between 25 to 60 years old. Only residents at Tampines North is lower between 10 to 35 years old. This could be due to newer house such as BTO at Tampines North areas.
* This trend can inform on transport services focusing in the busiest areas at Tampines West & East as the population is significantly larger as compared to Tampines North & Simei.