Take-home Exercise 1

Visual Analytics of Population Structure Across Different Planning Areas in Singapore

Li Yuquan

2025-04-30

# 1 Overview

## 1.1 Setting the scene

A local online media company that publishes daily content on digital platforms is planning to release an article in 2024 focusing on the population structure and distribution of Singapore. The number of resident population is a key indicator in understanding population structure, as it helps analyze aspects such as age composition, gender ratio, regional population density, and developmental disparities. This study aims to explore the differences in Singapore’s population structure and distribution through the organization and visual representation of such data.

## 1.2 Objectives

As a graphical editor of a median company, I conduct exploratory data analysis and visualize the data from the following three aspects:

1. **Age Group Composition across Planning Areas**

* Identify spatial disparities in youth, middle-aged, and elderly population distribution to reveal areas with aging or younger communities.

1. **Gender Distribution and Elderly Gender Imbalance**

* Explore male-female ratios and highlight significant gender differences in the elderly population across planning areas.

1. **Regional Demographics by Age and Gender**

* Compare age and gender structures across five major regions to uncover regional demographic trends and planning implications.

# 2 Data pre-processing & data exploration

## 2.1 Dataset Overview

The dataset used for this analysis is “Singapore Resident Population by Planning Area/Subzone, Single Year of Age and Sex, June 2024,” published by the Singapore Department of Statistics (DOS) on <https://www.singstat.gov.sg>. The data covers all planning areas and subzones across the Singapore, listing the number of resident population (Residents) by age and sex.

## 2.2 Loading R packages

| Library | Purpose |
| --- | --- |
| tidyverse | Include dplyr for data manipulation (mutate, group\_by, summarize), ggplot2 for visualization (histograms, boxplots), and forcats for factor handling and variable ordering |
| ggrepel | an R package provides geoms for ggplot2 to repel overlapping text labels |
| ggthemes | an R package provides some extra themes, geoms, and scales for ‘ggplot2’ |
| hrbrthemes | an R package provides typography-centric themes and theme components for ggplot2 |
| patchwork | an R package for preparing composite figure created using ggplot2 |
| ggdist | an R package for visualisation of distributions and uncertainity. |
| ggridges | an R package for creating density plots for visualisation of continuous distribution. |
| forcats | an R package for working with categorical variables (factors), including tools for reordering and modifying factor levels |
| viridis | an R package providing color scales that are perceptually uniform |
| scales | an R package providing tools to customize axis labels, legends, and formatting in visualizations |

pacman::p\_load(ggrepel, patchwork,   
 ggthemes, hrbrthemes,  
 tidyverse, ggdist, ggridges, forcats, viridis, scales  
)

## 2.3 Importing the data

The dataset used in this study is in CSV format. The dataset is imported using the  read\_csv()  function and is named as “data”.

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

## 2.4 Data exploration

### 2.4.1 Overview of Data Structure and Statistics

Explore the basic information of the dataset by using  summary() ,  glimpse()  and  file.info() . The dataset contains 60,424 observations and 6 variables, describing population counts disaggregated by planning areas, subzones, age groups, sex, and time (year).The size of the dataset is 2.5MB.

| Variable | Description |
| --- | --- |
| PA (Planning Area) | Character variable indicating planning areas |
| SZ (Subzone) | Character variable indicating subzones |
| Age | Character variable representing age groups |
| Sex | Character variable indicating gender |
| Pop | Numeric variable (<dbl>) representing the population count. |
| Time | Numeric variable (<dbl>) indicating the year, with all values are 2024. |

#### summary()

summary(data)

PA SZ Age Sex   
 Length:60424 Length:60424 Length:60424 Length:60424   
 Class :character Class :character Class :character Class :character   
 Mode :character Mode :character Mode :character Mode :character   
   
   
   
 Pop Time   
 Min. : 0.0 Min. :2024   
 1st Qu.: 0.0 1st Qu.:2024   
 Median : 20.0 Median :2024   
 Mean : 69.4 Mean :2024   
 3rd Qu.: 90.0 3rd Qu.:2024   
 Max. :1180.0 Max. :2024

#### glimpse()

glimpse(data)

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,…

#### file size

file.info("data/respopagesex2024.csv")

size isdir mode mtime  
data/respopagesex2024.csv 2521616 FALSE 644 2024-09-23 17:44:00  
 ctime atime uid gid  
data/respopagesex2024.csv 2025-05-02 17:13:57 2025-05-11 19:05:04 501 20  
 uname grname  
data/respopagesex2024.csv liyuquan staff

### 2.4.2 Check missing value

Map all possible null values in the data to NA. Use  is.na  to check the data and output FALSE, indicating that there are no missing values in the dataset.

data[data == ""] <- NA  
any(is.na(data))

[1] FALSE

### 2.4.3 Check for duplication

Using the duplicated function, the output is “# A tibble: 0 × 6”, which indicates that there are no duplicate entries in the data.

data[duplicated(data),]

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

## 2.5 Data pre-processing

### 2.5.1 Convert data type

In the original dataset, the data type of the Age variable is character. Among them, there is data “90\_and\_Over”. We will replace this data with “90” and convert the data type of this variable to numeric.

data1 <- data %>%  
 mutate(  
 Age = if\_else(Age == "90\_and\_Over", "90", as.character(Age)),  
 Age = as.numeric(Age)  
 )

### 2.5.2 Construct new variable

#### 2.5.2.1 Grouped by broad age categories → v1 → Age\_Group

Use mutate() to create a new variable and apply case\_when() to categorize the values of Age into three groups: Youth (0–21 years), Middle (22–64 years), and Elderly (65–90 years).

v1 <- data1 %>%  
 mutate(  
 Age\_Group = case\_when(  
 Age <= 21 ~ "Youth",  
 Age < 65 ~ "Middle",  
 TRUE ~ "Elderly"  
 )  
 )

#### 2.5.2.2 Grouped by fine-grained 5-year age intervals → v2 → Age\_Group

Use mutate() to create a new variable and apply cut() to bin the Age data into 5-year intervals from 0 to 89 years, with an additional group at the end labeled as “90\_and\_over” for ages 90 and above.

v2 <- data1 %>%  
 mutate(  
 Age\_Group = cut(  
 Age,  
 breaks = c(seq(0,   
 90,   
 by = 5),   
 Inf),  
 right = FALSE,  
 labels = c(  
 "00\_to\_04", "05\_to\_09", "10\_to\_14", "15\_to\_19", "20\_to\_24",  
 "25\_to\_29", "30\_to\_34", "35\_to\_39", "40\_to\_44", "45\_to\_49",  
 "50\_to\_54", "55\_to\_59", "60\_to\_64", "65\_to\_69", "70\_to\_74",  
 "75\_to\_79", "80\_to\_84", "85\_to\_89", "90\_and\_over"  
 )  
 )  
 )

#### 2.5.2.3 Group the planning areas by regions → v3\_region → region

Use tribble() to categorize the 55 planning areas (PA) in the original dataset into five regions: West, Central, East, North, and North-East. Name this new variable Region, and use left\_join() to merge the resulting table into the v3\_region dataset.

region\_mapping <- tribble(  
 ~PA, ~Region,  
 # Central  
 "Bishan", "Central",  
 "Bukit Merah", "Central",  
 "Bukit Timah", "Central",  
 "Downtown Core", "Central",  
 "Geylang", "Central",  
 "Kallang", "Central",  
 "Marina East", "Central",  
 "Marina South", "Central",  
 "Marine Parade", "Central",  
 "Museum", "Central",  
 "Newton", "Central",  
 "Novena", "Central",  
 "Orchard", "Central",  
 "Outram", "Central",  
 "Queenstown", "Central",  
 "River Valley", "Central",  
 "Rochor", "Central",  
 "Singapore River", "Central",  
 "Southern Islands", "Central",  
 "Straits View", "Central",  
 "Tanglin", "Central",  
 "Toa Payoh", "Central",  
  
 # East  
 "Bedok", "East",  
 "Changi", "East",  
 "Changi Bay", "East",  
 "Pasir Ris", "East",  
 "Paya Lebar", "East",  
 "Tampines", "East",  
  
 # North  
 "Central Water Catchment", "North",  
 "Lim Chu Kang", "North",  
 "Mandai", "North",  
 "Sembawang", "North",  
 "Simpang", "North",  
 "Sungei Kadut", "North",  
 "Woodlands", "North",  
 "Yishun", "North",  
  
 # North-East  
 "Ang Mo Kio", "North-East",  
 "Hougang", "North-East",  
 "North-Eastern Islands", "North-East",  
 "Punggol", "North-East",  
 "Seletar", "North-East",  
 "Sengkang", "North-East",  
 "Serangoon", "North-East",  
  
 # West  
 "Boon Lay", "West",  
 "Bukit Batok", "West",  
 "Bukit Panjang", "West",  
 "Choa Chu Kang", "West",  
 "Clementi", "West",  
 "Jurong East", "West",  
 "Jurong West", "West",  
 "Pioneer", "West",  
 "Tengah", "West",  
 "Tuas", "West",  
 "Western Islands", "West",  
 "Western Water Catchment","West"  
)  
  
v3\_region <- data1 %>%  
 left\_join(region\_mapping, by = "PA") %>%  
 filter(!is.na(Region))

### 2.5.3 Data prepare for Visualization

#### 2.5.3.1 Data prepare for Visualization 3.1

#### Description

v1\_grouped\_all summarizes the total population counts and corresponding ratios of the Elderly (≥ 65), Youth (≤ 21), and Middle-aged (22–64) groups for each Planning Area (PA). It ensures that only areas with valid computed ratios for all three groups are retained, forming the analytical base for plots p1, p2, and p3.

The top\_bottom\_elderly, top\_bottom\_youth, and top\_bottom\_middle subsets extract the top and bottom 10 Planning Areas based on the proportions of elderly, youth, and middle-aged populations respectively, with the PA factor reordered by each corresponding age group ratio to facilitate ranked comparisons in plots p1, p2, and p3.

v1\_p4 prepares the full age group composition by Planning Area for visualization in plot p4. It calculates the population share (Ratio) of each age group within each PA. Additionally, each PA is sorted by the ratio of middle-aged residents, and the PA variable is re-leveled accordingly to ensure consistent display order.

v1\_p5 dataset calculates the overall proportion of each age group across Singapore, and derives variables (xmin, xmax, label\_x) to position percentage labels for a horizontal stacked visualization.

#### The Code

v1\_grouped\_all <- v1 %>%  
 group\_by(PA) %>%  
 summarise(  
 Elderly = sum(Pop[Age >= 65],   
 na.rm = TRUE),  
 Youth = sum(Pop[Age <= 21],   
 na.rm = TRUE),  
 Middle = sum(Pop[Age > 21 & Age < 65],   
 na.rm = TRUE),  
 Total = sum(Pop,   
 na.rm = TRUE),  
 Elderly\_Rate = Elderly / Total,  
 Youth\_Rate = Youth / Total,  
 Middle\_Rate = Middle / Total,  
 .groups = "drop"  
 ) %>%  
 filter(!is.na(Elderly\_Rate) & !is.na(Youth\_Rate) & !is.na(Middle\_Rate))  
  
# filter top\_bottom for p1 p2 p3  
top\_bottom\_elderly <- bind\_rows(  
 slice\_max(v1\_grouped\_all,   
 Elderly\_Rate,   
 n = 10),  
 slice\_min(v1\_grouped\_all,   
 Elderly\_Rate,   
 n = 10)  
) %>%  
 mutate(PA = fct\_reorder(PA,   
 Elderly\_Rate))  
  
top\_bottom\_youth <- bind\_rows(  
 slice\_max(v1\_grouped\_all,   
 Youth\_Rate,   
 n = 10),  
 slice\_min(v1\_grouped\_all,   
 Youth\_Rate,   
 n = 10)  
) %>%  
 mutate(PA = fct\_reorder(PA,   
 Youth\_Rate))  
  
top\_bottom\_middle <- bind\_rows(  
 slice\_max(v1\_grouped\_all,   
 Middle\_Rate,   
 n = 10),  
 slice\_min(v1\_grouped\_all,   
 Middle\_Rate,   
 n = 10)  
) %>%  
 mutate(PA = fct\_reorder(PA,   
 Middle\_Rate))  
  
# p4  
v1\_p4 <- v1 %>%  
 group\_by(PA, Age\_Group) %>%  
 summarise(Pop = sum(Pop, na.rm = TRUE), .groups = "drop") %>%  
 group\_by(PA) %>%  
 mutate(  
 Total = sum(Pop),  
 Ratio = Pop / Total  
 ) %>%  
 filter(Total > 0) %>%  
 select(-Total) %>%  
 ungroup() %>%  
 group\_by(PA) %>%  
 mutate(Middle\_Ratio = Ratio[Age\_Group == "Middle"]) %>%  
 ungroup() %>%  
 arrange(desc(Middle\_Ratio)) %>%  
 mutate(PA = factor(PA, levels = unique(PA))) %>%  
 select(-Middle\_Ratio)  
  
# p5  
v1\_p5 <- v1 %>%  
 group\_by(Age\_Group) %>%  
 summarise(Pop = sum(Pop), .groups = "drop") %>%  
 mutate(  
 Prop = Pop / sum(Pop),  
 Percent = paste0(Age\_Group, ": ",   
 round(Prop \* 100, 1),   
 "%"),  
 xmin = cumsum(lag(Prop, default = 0)),  
 xmax = cumsum(Prop),  
 label\_x = (xmin + xmax) / 2  
 )

#### 2.5.3.2 data prepare for Visualization 3.2

#### Description

v2\_p6 calculates the total population of males and females across different age groups, along with their respective proportions of the total population. It assigns mirrored percentage values by gender (positive for males, negative for females) to prepare for the construction of a population pyramid. This approach facilitates the comparison of gender distribution across age groups.

v2\_p7 calculates the gender-specific population count for each Planning Area (PA) and computes the proportion of each gender within the area’s total population. For visualization purposes, Percent\_plot is mirrored (positive for males, negative for females), and the PA variable is reordered based on total population to support meaningful comparative plotting.

v2\_p8 computes the elderly population ratio (aged ≥ 65) for both males and females in each Planning Area. It mirrors the values for plotting (positive for males, negative for females), filters out regions with zero elderly population, and selects the top and bottom 10 PA based on elderly population share. These selected areas are used to compare gender differences in aging patterns.

#### The Code

# p6  
v2\_p6 <- v2 %>%  
 group\_by(Sex, Age\_Group) %>%  
 summarise(Pop = sum(Pop), .groups = "drop") %>%  
 mutate(  
 Percent = Pop / sum(Pop) \* 100,  
 Percent\_plot = ifelse(Sex == "Males",   
 Percent,   
 -Percent)  
 )  
  
# p7  
v2\_p7 <- v2 %>%  
 group\_by(PA, Sex) %>%  
 summarise(Pop = sum(Pop), .groups = "drop")  
  
total\_pa <- v2\_p7 %>%  
 group\_by(PA) %>%  
 summarise(Total = sum(Pop), .groups = "drop")  
  
v2\_p7 <- left\_join(v2\_p7,   
 total\_pa,   
 by = "PA") %>%  
 mutate(  
 Percent = Pop / Total \* 100,  
 Percent\_plot = ifelse(Sex == "Males",   
 Percent,   
 -Percent)  
 ) %>%  
 group\_by(PA) %>%  
 filter(sum(Pop) > 0) %>%  
 ungroup() %>%  
 mutate(PA = fct\_reorder(PA, Total))  
  
# p8  
v2\_p8 <- v2 %>%  
 group\_by(PA, Sex) %>%  
 summarise(  
 Total = sum(Pop, na.rm = TRUE),  
 Elderly = sum(Pop[Age >= 65], na.rm = TRUE),  
 .groups = "drop"  
 ) %>%  
 mutate(  
 Elderly\_Rate = ifelse(Total > 0,   
 Elderly / Total \* 100, 0),  
 Elderly\_Rate\_plot = ifelse(Sex == "Males",   
 Elderly\_Rate,   
 -Elderly\_Rate)  
 )  
  
valid\_PA <- v2\_p8 %>%  
 group\_by(PA) %>%  
 summarise(Total\_Elderly = sum(Elderly)) %>%  
 filter(Total\_Elderly > 0) %>%  
 pull(PA)  
  
v2\_p8 <- v2\_p8 %>% filter(PA %in% valid\_PA)  
  
pa\_avg <- v2\_p8 %>%  
 group\_by(PA) %>%  
 summarise(Avg\_Elderly\_Rate = mean(Elderly\_Rate), .groups = "drop")  
  
selected\_PA <- bind\_rows(  
 slice\_max(pa\_avg,   
 Avg\_Elderly\_Rate,   
 n = 10),  
 slice\_min(pa\_avg,   
 Avg\_Elderly\_Rate,   
 n = 10)  
) %>%  
 arrange(desc(Avg\_Elderly\_Rate)) %>%  
 pull(PA)  
  
v2\_p8 <- v2\_p8 %>%  
 filter(PA %in% selected\_PA) %>%  
 mutate(PA = factor(PA, levels = rev(selected\_PA)))

#### 2.5.3.3 data prepare for Visualization 3.3

#### Description

v3\_p9This dataset aggregates the total population by region and calculates each region’s percentage share of the overall population. A new Label column is also created to display the region name, total population , and its percentage share.

v3\_p10 expands the original region-level aggregated population data into individual-level rows, where each row represents a single person.

v3\_p11 draws approximately 0.07% of the data by region and sex as a sample, in order to support visually cleaner and more efficient distribution visualizations.

v3\_points extracts a further small sample (0.02%) from the full individual-level dataset, stratified by Region and Age. A numeric mapping of Region is added (Region\_numeric) to enable vertical jittering in visualizations by numerical y-position.

#### The Code

v3\_p9 <- v3\_region %>%  
 group\_by(Region) %>%  
 summarise(Total\_Pop = sum(Pop, na.rm = TRUE)) %>%  
 mutate(  
 Percent = Total\_Pop / sum(Total\_Pop) \* 100,  
 Label = paste0(Region, "\n", comma(Total\_Pop), "\n", round(Percent, 1), "%")  
 )  
  
v3\_p10 <- v3\_region %>%  
 uncount(weights = Pop)  
  
v3\_p11 <- v3\_p10 %>%  
 group\_by(Region, Sex) %>%  
 sample\_frac(0.0007)  
  
v3\_ridges <- v3\_p10   
v3\_points <- v3\_p10 %>%  
 group\_by(Region, Age) %>%  
 sample\_frac(0.0002)  
region\_levels <- levels(fct\_rev(factor(v3\_ridges$Region)))  
v3\_points <- v3\_points %>%  
 mutate(Region\_numeric = match(Region, region\_levels))

### 2.5.4 Preview pre-processed data

Use head() to display the first 200 rows of the data. By viewing the processed data, we can observe that new variables have been added to the corresponding data sets.

#### data

head(data, 200)

# A tibble: 200 × 6  
 PA SZ Age Sex Pop Time  
 <chr> <chr> <chr> <chr> <dbl> <dbl>  
 1 Ang Mo Kio Ang Mo Kio Town Centre 0 Males 10 2024  
 2 Ang Mo Kio Ang Mo Kio Town Centre 0 Females 10 2024  
 3 Ang Mo Kio Ang Mo Kio Town Centre 1 Males 10 2024  
 4 Ang Mo Kio Ang Mo Kio Town Centre 1 Females 10 2024  
 5 Ang Mo Kio Ang Mo Kio Town Centre 2 Males 10 2024  
 6 Ang Mo Kio Ang Mo Kio Town Centre 2 Females 10 2024  
 7 Ang Mo Kio Ang Mo Kio Town Centre 3 Males 10 2024  
 8 Ang Mo Kio Ang Mo Kio Town Centre 3 Females 10 2024  
 9 Ang Mo Kio Ang Mo Kio Town Centre 4 Males 30 2024  
10 Ang Mo Kio Ang Mo Kio Town Centre 4 Females 10 2024  
# ℹ 190 more rows

#### v1

head(v1,200)

# A tibble: 200 × 7  
 PA SZ Age Sex Pop Time Age\_Group  
 <chr> <chr> <dbl> <chr> <dbl> <dbl> <chr>   
 1 Ang Mo Kio Ang Mo Kio Town Centre 0 Males 10 2024 Youth   
 2 Ang Mo Kio Ang Mo Kio Town Centre 0 Females 10 2024 Youth   
 3 Ang Mo Kio Ang Mo Kio Town Centre 1 Males 10 2024 Youth   
 4 Ang Mo Kio Ang Mo Kio Town Centre 1 Females 10 2024 Youth   
 5 Ang Mo Kio Ang Mo Kio Town Centre 2 Males 10 2024 Youth   
 6 Ang Mo Kio Ang Mo Kio Town Centre 2 Females 10 2024 Youth   
 7 Ang Mo Kio Ang Mo Kio Town Centre 3 Males 10 2024 Youth   
 8 Ang Mo Kio Ang Mo Kio Town Centre 3 Females 10 2024 Youth   
 9 Ang Mo Kio Ang Mo Kio Town Centre 4 Males 30 2024 Youth   
10 Ang Mo Kio Ang Mo Kio Town Centre 4 Females 10 2024 Youth   
# ℹ 190 more rows

#### v2

head(v2,200)

# A tibble: 200 × 7  
 PA SZ Age Sex Pop Time Age\_Group  
 <chr> <chr> <dbl> <chr> <dbl> <dbl> <fct>   
 1 Ang Mo Kio Ang Mo Kio Town Centre 0 Males 10 2024 00\_to\_04   
 2 Ang Mo Kio Ang Mo Kio Town Centre 0 Females 10 2024 00\_to\_04   
 3 Ang Mo Kio Ang Mo Kio Town Centre 1 Males 10 2024 00\_to\_04   
 4 Ang Mo Kio Ang Mo Kio Town Centre 1 Females 10 2024 00\_to\_04   
 5 Ang Mo Kio Ang Mo Kio Town Centre 2 Males 10 2024 00\_to\_04   
 6 Ang Mo Kio Ang Mo Kio Town Centre 2 Females 10 2024 00\_to\_04   
 7 Ang Mo Kio Ang Mo Kio Town Centre 3 Males 10 2024 00\_to\_04   
 8 Ang Mo Kio Ang Mo Kio Town Centre 3 Females 10 2024 00\_to\_04   
 9 Ang Mo Kio Ang Mo Kio Town Centre 4 Males 30 2024 00\_to\_04   
10 Ang Mo Kio Ang Mo Kio Town Centre 4 Females 10 2024 00\_to\_04   
# ℹ 190 more rows

#### v3\_region

head(v3\_region,200)

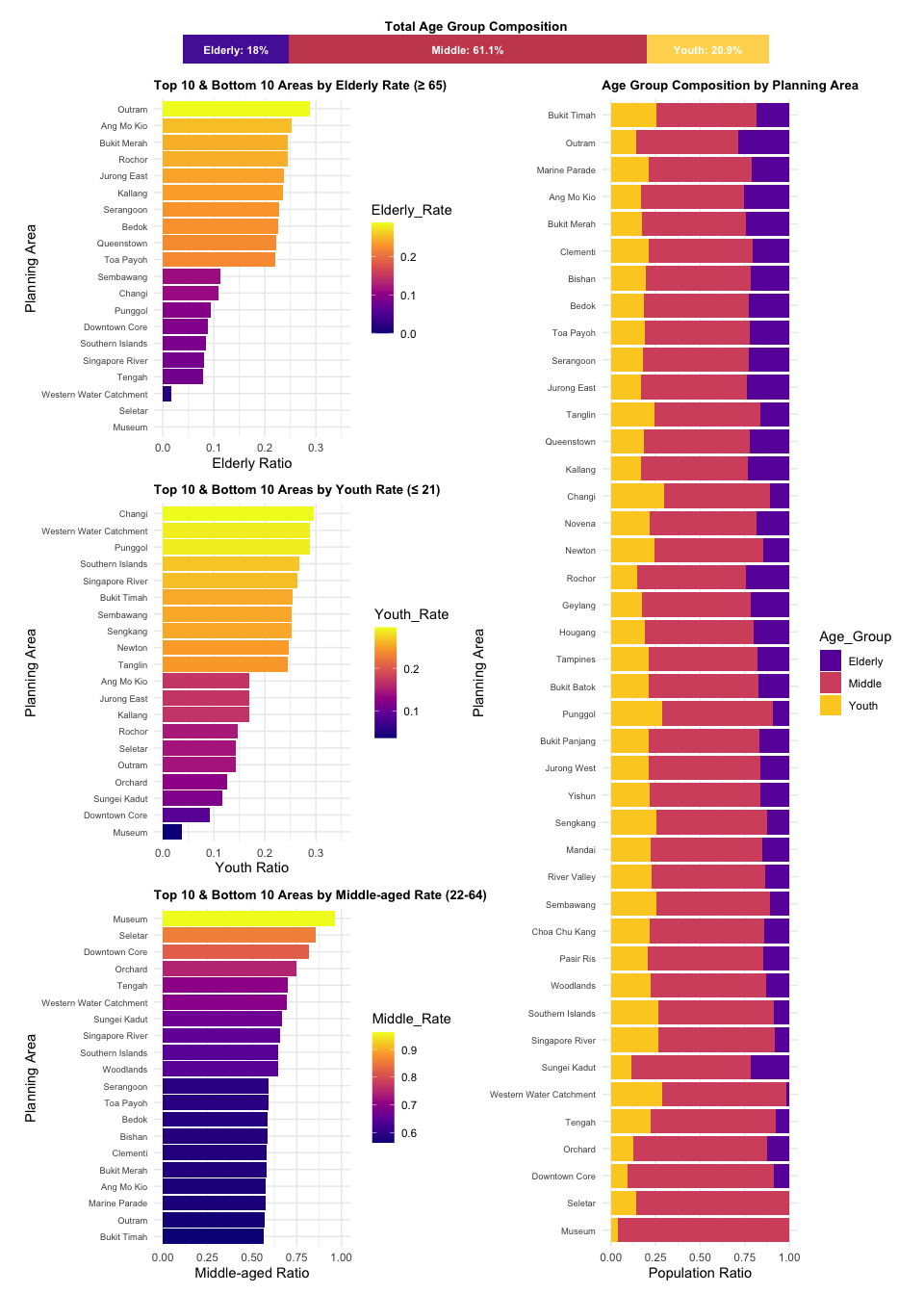
# A tibble: 200 × 7  
 PA SZ Age Sex Pop Time Region   
 <chr> <chr> <dbl> <chr> <dbl> <dbl> <chr>   
 1 Ang Mo Kio Ang Mo Kio Town Centre 0 Males 10 2024 North-East  
 2 Ang Mo Kio Ang Mo Kio Town Centre 0 Females 10 2024 North-East  
 3 Ang Mo Kio Ang Mo Kio Town Centre 1 Males 10 2024 North-East  
 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  
 7 Ang Mo Kio Ang Mo Kio Town Centre 3 Males 10 2024 North-East  
 8 Ang Mo Kio Ang Mo Kio Town Centre 3 Females 10 2024 North-East  
 9 Ang Mo Kio Ang Mo Kio Town Centre 4 Males 30 2024 North-East  
10 Ang Mo Kio Ang Mo Kio Town Centre 4 Females 10 2024 North-East  
# ℹ 190 more rows

# 3 Data Visualization

## 3.1 Population Age Structure Across Singapore Planning Areas

This data visualisation presents the age structure of Singapore’s resident population across different planning areas, divided into three categories: Youth (aged 21 and below), Middle-aged (22 to 64), and Elderly (65 and above).

### The Plot



### The Code

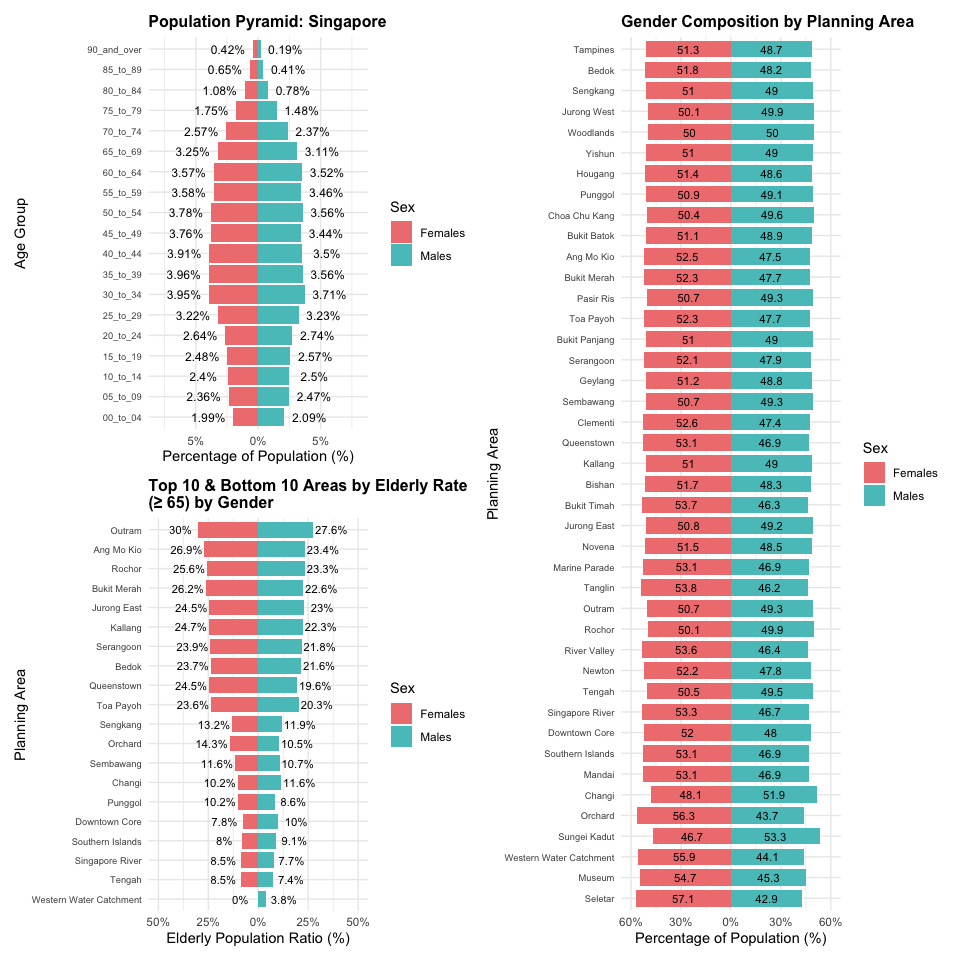
# common theme  
common\_theme <- theme(  
 plot.title = element\_text(size = 10,   
 face = "bold",   
 hjust = 0),  
 axis.text.y = element\_text(size = 7),  
 legend.position = "right"  
)  
  
# elderly plot  
p1 <- ggplot(top\_bottom\_elderly, aes(x = Elderly\_Rate,   
 y = PA,   
 fill = Elderly\_Rate)) +  
 geom\_col() +  
 xlim(0, 0.35) +  
 scale\_fill\_viridis\_c(option = "C") +  
 labs(  
 title = "Top 10 & Bottom 10 Areas by Elderly Rate (≥ 65)",  
 x = "Elderly Ratio",  
 y = "Planning Area"  
 ) +  
 theme\_minimal() +  
 common\_theme  
  
# youth plot  
p2 <- ggplot(top\_bottom\_youth, aes(x = Youth\_Rate,   
 y = PA,   
 fill = Youth\_Rate)) +  
 geom\_col() +  
 xlim(0, 0.35) +  
 scale\_fill\_viridis\_c(option = "C") +  
 labs(  
 title = "Top 10 & Bottom 10 Areas by Youth Rate (≤ 21)",  
 x = "Youth Ratio",  
 y = "Planning Area"  
 ) +  
 theme\_minimal() +  
 common\_theme  
  
# middle plot  
p3 <- ggplot(top\_bottom\_middle, aes(x = Middle\_Rate, y = PA, fill = Middle\_Rate)) +  
 geom\_col() +  
 xlim(0, 1) +  
 scale\_fill\_viridis\_c(option = "C") +  
 labs(  
 title = "Top 10 & Bottom 10 Areas by Middle-aged Rate (22-64)",  
 x = "Middle-aged Ratio",  
 y = "Planning Area"  
 ) +  
 theme\_minimal() +  
 common\_theme  
  
# Plot for each area  
p4 <- ggplot(v1\_p4, aes(x = PA,   
 y = Ratio,   
 fill = Age\_Group)) +  
 geom\_bar(stat = "identity") +  
 coord\_flip() +  
 labs(  
 title = "Age Group Composition by Planning Area",  
 x = "Planning Area",  
 y = "Population Ratio"  
 ) +  
 scale\_fill\_viridis\_d(option = "C",   
 begin = 0.2,   
 end = 0.9) +  
 theme\_minimal() +  
 common\_theme  
  
# Total Plot   
p5 <- ggplot(v1\_p5) +  
 geom\_rect(aes(xmin = xmin, xmax = xmax, ymin = 0, ymax = 1, fill = Age\_Group)) +  
 geom\_text(aes(x = label\_x,   
 y = 0.5,   
 label = Percent),   
 color = "white", size = 3,   
 fontface = "bold") +  
 scale\_fill\_manual(values = c("Youth" = "#FDD75E",   
 "Middle" = "#C94C5C",   
 "Elderly" = "#5626A6")) +  
 coord\_cartesian(ylim = c(0, 1)) +  
 theme\_void() +  
 ggtitle("Total Age Group Composition") +  
 theme(  
 aspect.ratio = 0.05,   
 plot.title = element\_text(hjust = 0.5,   
 size = 10,   
 face = "bold"),  
 plot.margin = margin(0.2, 0.5, 0.2, 0.5,   
 unit = "cm"),  
 legend.position = "none"  
 )  
  
  
  
p5 / ((p1 / p2 / p3) | p4)

|  |
| --- |
| Insight |
| * Nationally, the middle-aged group forms the majority (61.1%), followed by Youth (20.9%) and Elderly (18%). * The top ten planning areas with the highest proportion of elderly residents include Outram, Ang Mo Kio, and Bukit Merah, suggesting these may be aging communities with long-established residential zones. * Youth populations are more concentrated in areas such as Changi, Punggol, and Southern Islands, possibly due to lower living costs for younger individuals or proximity to educational institutions. In contrast, central areas like Museum and Downtown Core have the lowest youth representation. * Middle-aged individuals dominate areas such as Museum, Seletar, and Downtown Core, where their proportion reaches up to 90%. This may indicate these are employment hubs where middle-aged individuals with stronger spending power tend to reside. * Overall, the visualisation reveals clear spatial disparities in age composition across planning areas, highlighting urban ageing, youth clustering in peripheral areas, and middle-aged dominance in central or transitional zones. |

## 3.2 Gender-Based Demographics of Singapore Residents

This visualisation presents the age and gender structure of Singapore’s population across different planning areas, with each chart highlighting population distribution by gender.

### The Plot



### The Code

# common theme  
common\_theme\_v2 <- theme(  
 plot.title = element\_text(size = 12,   
 face = "bold",   
 hjust = 0),  
 axis.text.y = element\_text(size = 7),  
 legend.position = "right"  
)  
  
# p6   
p6 <- ggplot(v2\_p6, aes(x = Percent\_plot,   
 y = Age\_Group,   
 fill = Sex)) +  
 geom\_col(width = 0.9) +  
 geom\_text(  
 data = subset(v2\_p6,   
 Sex == "Females"),  
 aes(label = paste0(round(abs(Percent), 2), "%")),  
 nudge\_x = -2,   
 size = 3.2,  
 color = "black"  
 ) +  
 geom\_text(  
 data = subset(v2\_p6,   
 Sex == "Males"),  
 aes(label = paste0(round(abs(Percent), 2), "%")),  
 nudge\_x = 2,   
 size = 3.2,  
 color = "black"  
 ) +  
   
 scale\_x\_continuous(labels = function(x) paste0(abs(x), "%")) +  
 scale\_fill\_manual(values = c("Males" = "#58c3c3",   
 "Females" = "#f08080")) +  
 labs(  
 title = "Population Pyramid: Singapore",  
 x = "Percentage of Population (%)",  
 y = "Age Group",  
 fill = "Sex"  
 ) +  
 coord\_cartesian(xlim = c(-8, 8)) +   
 theme\_minimal() +  
 common\_theme\_v2  
  
#p7  
p7 <- ggplot(v2\_p7,   
 aes(x = Percent\_plot,   
 y = PA,   
 fill = Sex)) +  
 geom\_col(width = 0.8) +  
 geom\_text(aes(label = round(abs(Percent), 1)),  
 position = position\_stack(vjust = 0.5),  
 size = 3, color = "black") +  
 scale\_x\_continuous(limits = c(-60, 60),  
 labels = function(x) paste0(abs(x), "%")) +  
 scale\_fill\_manual(values = c("Males" = "#58c3c3",   
 "Females" = "#f08080")) +  
 labs(  
 title = "Gender Composition by Planning Area",  
 x = "Percentage of Population (%)",  
 y = "Planning Area",  
 fill = "Sex"  
 ) +  
 theme\_minimal() +  
 common\_theme\_v2  
  
# p8  
p8 <- ggplot(v2\_p8, aes(x = Elderly\_Rate\_plot,   
 y = PA,   
 fill = Sex)) +  
 geom\_col(width = 0.8) +  
 geom\_text(  
 data = subset(v2\_p8,   
 Sex == "Females"),  
 aes(label = paste0(round(abs(Elderly\_Rate), 1), "%")),  
 nudge\_x = -9,  
 size = 3,  
 color = "black"  
 ) +  
   
 geom\_text(  
 data = subset(v2\_p8,   
 Sex == "Males"),  
 aes(label = paste0(round(abs(Elderly\_Rate), 1), "%")),  
 nudge\_x = 9,  
 size = 3,  
 color = "black"  
 ) +  
 scale\_x\_continuous(labels = function(x) paste0(abs(x), "%")) +  
 scale\_fill\_manual(values = c("Males" = "#58c3c3",   
 "Females" = "#f08080")) +  
 labs(  
 title = "Top 10 & Bottom 10 Areas by Elderly Rate\n(≥ 65) by Gender",  
 x = "Elderly Population Ratio (%)",  
 y = "Planning Area",  
 fill = "Sex"  
 ) +  
 coord\_cartesian(xlim = c(-50, 50)) +   
 theme\_minimal() +  
 common\_theme\_v2  
  
  
(p6 / p8) | p7

|  |
| --- |
| Insight |
| * The population pyramid in the top-left shows the proportion of males and females across various age groups. The gender ratio is relatively balanced below age 64, but females significantly outnumber males in the 65 and above age groups, suggesting a larger elderly female population likely due to longer life expectancy. * The chart on the right displays the gender composition by planning area. In most areas, females slightly outnumber males, with the highest female proportions found in Orchard (56.3%), Western Water Catchment (55.9%), and Seletar (57.1%). In contrast, areas like Sungei Kadut and Changi have higher male proportions, possibly due to the dominance of male-oriented industries in those regions. * The bottom-left chart presents the top 10 and bottom 10 areas by elderly population ratio (≥65), disaggregated by gender. Outram, Ang Mo Kio, and Rochor have the highest elderly ratios, and females consistently make up a larger share of the elderly population, further highlighting gendered ageing trends. * Overall, the visualisation reveals a gender imbalance among the elderly, with females forming the majority, and illustrates regional differences in gender and age distribution—providing important insights for planning in healthcare, housing, and social services. |

## 3.3 Population Structure Across Singapore Regions

This visualization presents the population size, gender, and age distribution across the five main regions of Singapore: Central, East, North, North-East, and West.

### The Plot



### The Code

# theme  
common\_theme\_v3 <- theme\_minimal() +  
 theme(  
 plot.title = element\_text(face = "bold",   
 size = 14,   
 hjust = 0.5),  
 axis.title = element\_text(size = 12),  
 axis.text = element\_text(size = 10),  
 legend.position = "top"  
 )  
  
  
# p9  
p9 <- ggplot(v3\_p9, aes(x = 2,   
 y = Total\_Pop,   
 fill = Region)) +  
 geom\_col(width = 1,   
 color = "white") +  
 coord\_polar(theta = "y") +  
 xlim(c(1, 2.5)) +   
 geom\_text(  
 aes(label = Label),  
 position = position\_stack(vjust = 0.5),  
 size = 3  
 ) +  
 labs(title = "Population Distribution by Region") +  
 theme\_void() +  
 theme(  
 plot.title = element\_text(hjust = 0.5, face = "bold"),  
 legend.position = "none"  
 ) +  
 scale\_fill\_brewer(palette = "Set3")  
  
   
   
  
# p10  
p10 <- ggplot(v3\_p10,   
 aes(x = Age,   
 y = Region,   
 fill = Sex)) +  
 geom\_density\_ridges(  
 alpha = 0.5,  
 scale = 1.2,  
 color = "white",  
 position = "identity"  
 ) +  
 scale\_fill\_manual(values = c("Males" = "#58c3c3",   
 "Females" = "#f08080")) +  
 scale\_x\_continuous(breaks = seq(0, 90, by = 10), limits = c(0, 90)) +   
 labs(  
 title = "Age Distribution by Region",  
 x = NULL,  
 y = NULL,  
 fill = NULL  
 ) +  
 theme\_minimal() +  
 common\_theme\_v3   
  
# p11  
p11 <- ggplot() +  
 geom\_boxplot(  
 data = v3\_p10,  
 aes(x = Region,   
 y = Age,   
 color = Sex),  
 position = position\_dodge(width = 0.75),  
 ) +  
 geom\_point(  
 data = v3\_p11,  
 aes(x = Region,   
 y = Age,   
 color = Sex),  
 position = position\_jitterdodge(jitter.width = 0.3,   
 dodge.width = 0.75),  
 size = 0.5,  
 alpha = 0.4  
 ) +  
 scale\_color\_manual(values = c("Males" = "#58c3c3",   
 "Females" = "#f08080")) +  
 scale\_y\_continuous(breaks = seq(0, 90, by = 10), limits = c(0, 90))+  
 labs(  
 title = "Age Distribution by Region and Gender",  
 x = NULL,  
 y = NULL,  
 color = NULL  
 ) +  
 theme\_minimal() +  
 common\_theme\_v3  
  
  
p12 <- ggplot(v3\_ridges, aes(x = Age, y = fct\_rev(Region),   
 fill = Region)) +  
 geom\_density\_ridges(scale = 0.9,   
 alpha = 0.6,   
 color = "black") +  
  
 geom\_jitter(data = v3\_points,  
 aes(x = Age,   
 y = Region\_numeric + 0.25,   
 color = Region),  
 height = 0.2,   
 width = 0,   
 size = 0.05,   
 alpha = 0.6,  
 inherit.aes = FALSE) +  
   
 scale\_fill\_brewer(palette = "Set3") +  
 scale\_color\_brewer(palette = "Dark2") +  
 scale\_x\_continuous(breaks = seq(0, 90, by = 10), limits = c(0, 90)) +   
 labs(  
 title = "Age Distribution Across Regions in Singapore",  
 x = NULL,  
 y = NULL  
 ) +  
 theme\_minimal(base\_size = 12) +  
 theme(  
 legend.position = "none",  
 panel.grid.major.y = element\_blank(),  
 panel.grid.minor = element\_blank(),  
 axis.title.x = element\_text(hjust = 1),  
 plot.title = element\_text(face = "bold", size = 14),  
 )  
  
  
a <- (p9 + p10) / (p11+p12)  
a

|  |
| --- |
| Insight |
| * The pie chart in the top-left shows that Singapore’s population is relatively evenly distributed across regions, with the North-East and Central areas having the highest populations, while the North has the smallest. * The top-right visualisation displays the age distribution by gender across different regions. The age density curve for the Central region skews noticeably to the right, peaking above age 50, indicating a higher proportion of elderly residents. In the West and North-East, the 25 to 45 age group is more prominent. Across all regions, females aged 75 and above outnumber males, consistent with higher female life expectancy. * The bottom-left chart illustrates the age distribution by gender within each region. The Central region stands out as the most aged. The East and North show narrower boxplots, indicating more concentrated age distributions, possibly reflecting more stable and mature communities. Overall, similar interquartile ranges suggest a generally balanced age structure across regions. * The bottom-right chart further highlights intra-regional differences, with higher elderly population densities in the Central and Eastern regions. These patterns suggest that although all regions span a wide age range, some may be aging more rapidly, which could influence regional planning and healthcare needs. |

# 4 Conclusion

Based on a comprehensive visual analysis of the population structure across Singapore’s planning areas, the study reveals significant spatial disparities in demographic composition. In terms of age, the middle-aged population (22 to 64 years) forms the majority at 61.1%, though this proportion varies considerably across districts. Elderly populations are predominantly concentrated in long-established central areas such as Outram and Ang Mo Kio, while younger populations are more commonly found in peripheral or transitional areas like Punggol and Changi. Middle-aged residents are mainly located in economically active zones such as the Downtown Core and Museum area, reflecting functional differences among regions.

Gender structure analysis shows that although the sex ratio among younger age groups is relatively balanced, females significantly outnumber males in the elderly population, reflecting longer female life expectancy. Regionally, the Central area is characterized by both a high degree of ageing and a female-majority population. In contrast, the West and North regions exhibit more balanced gender distributions, which may be associated with the presence of male-dominated, labor-intensive industries.

In conclusion, the analysis highlights two key demographic trends—population ageing and gender imbalance—and underscores the need for regionally differentiated strategies in urban planning, healthcare resource allocation, and the provision of social services.

# 5 Reference

* Kam, Tin Seong. (2023). Visualising Distribution. *R for Visual Analytics.* <https://r4va.netlify.app/chap09>
* Hadley Wickham (2023) [**ggplot2: Elegant Graphics for Data Analysis**](https://ggplot2-book.org/). Online 3rd edition.
* Winston Chang (2013) [**R Graphics Cookbook 2nd edition**](https://r-graphics.org/). Online version.