Take-Home Exercise 1 Part 2

David Chiam & Liu Chih Yuan

2025-05-10

|  |
| --- |
| Commenting on David Chiam’s EDA 3.2 |
|  |

# **1. Overview**

## **Setting the scene**

A local online media company that publishes daily content on digital platforms is planning to release an article on **demographic structures and distribution of Singapore in 2024**.

## **The task**

Assuming the role of the graphical editor of the media company, the task was to prepare **at most three data visualisation** for the exercise.

Hence in this exercise, I curated three key visualisations as part of dats exploration:

* **Generational Distribution by Region:** A paired bar chart highlights both absolute and relative generational makeup across Singapore’s five planning regions, revealing the Central and East as home to proportionally older populations.
* **Age Demographics by Planning Area:** A boxplot with overlaid generation bands showcases median age and interquartile ranges, emphasising regions with mature population profiles like Central and West Singapore.
* **Top Planning Areas by Age Band:** A set of ranked bar charts and accompanying pie charts depict the top 15 planning areas by population size in three age segments—youth (0–24), adults (25–54), and seniors (55+). These are complemented with gender breakdowns to visualise demographic imbalances.

# **2. Getting started, Data Loading and Preparation**

This section documents the process of preparing the dataset for analysis and visualization. It covers the loading of R packages, initial inspection of the dataset, and the necessary wrangling and transformation steps applied.

### **2.1 Load required packages 📦**

Before importing the data, we load several R packages from the **tidyverse** ecosystem and supporting libraries for data wrangling and visualization. These packages streamline data processing and produce publication-ready plots using ggplot2. This setup ensure that we have both the flexibility and power to craft expressive, publication-quality demographic visualizations for the report.

📌 **Summary of Package Roles**

| Library | Description |
| --- | --- |
| [**tidyverse**](https://www.tidyverse.org/) | A collection of core packages designed for data science, used extensively for data preparation and wrangling |
| [**knitr**](https://yihui.org/knitr/) | Generates dynamic R markdown or Quarto reports |
| [**patchwork**](https://patchwork.data-imaginist.com/) | Combine multiple ggplots into a single layout |
| [**ggthemes**](https://yutannihilation.github.io/allYourFigureAreBelongToUs/ggthemes/) | Provides extra themes (e.g. economist, wsj) |
| [**scales**](https://scales.r-lib.org/) | For data/axes label formatting and annotations for ggplot2 |
| [**ggtext**](https://wilkelab.org/ggtext/) | Supports improved and rich text rendering for ggplot2 |
| [**ggalt**](https://yonicd.github.io/ggalt/) | A compendium of ‘geoms’, ‘coords’, ‘stats’, scales and fonts for ggplots2 |
| [**ggridges**](https://cran.r-project.org/web/packages/ggridges/vignettes/introduction.html) | For ridgeline plots that create the impression of a mountain range. They can be useful for visualising changes in distributions over time or space. |
| [**ggpubr**](https://github.com/kassambara/ggpubr) | For publication-ready plots. |
| [**ggextra**](https://github.com/daattali/ggExtra) | For adding marginal plots to ggplot2 |
| [**cowplot**](https://github.com/wilkelab/cowplot) | Provides precise figure alignment and plot annotation |
| [**ggnewscale**](https://eliocamp.github.io/ggnewscale/) | Enables multiple fill/colour scales in a single plot. For defining scales in ggplot2. |
| ggrepel |  |

We load the following R packages using the pacman::p\_load() :

pacman::p\_load(  
 tidyverse, # Core data manipulation and ggplot2 plotting  
 knitr, # Dynamic report generation and rendering  
 patchwork, # Combines multiple ggplot2 plots  
 ggthemes, # Additional themes and visual styles  
 scales, # Formatting axes, scales, and labels  
 ggtext, # Rich text formatting in ggplot titles/labels  
 ggalt, # Dumbbell, lollipop, and other advanced geoms  
 ggridges, # Ridgeline plots to visualize distributions  
 ggpubr, # Publication-ready ggplot2 enhancements  
 ggExtra, # Marginal plots for added insight  
 cowplot, # Publication-quality figure alignment  
 ggnewscale, # Allows multiple color/fill scales in one ggplot  
 ggrepel # Allows  
)

### **2.2** Loading the Data and Taking a Glimpse of the Dataset **📂**

The dataset, *Singapore Residents by Planning Area / Subzone, Single Year of Age and Sex, June 2024*, published by the [Department of Statistics, Singapore (DOS)](https://www.singstat.gov.sg/) was loaded using the read\_csv() function from the readr package in the tidyverse collection.

# Load the dataset  
df <- read\_csv("data/respopagesex2024.csv")

The dataset contains **60,424 rows** and **6 columns**. Each row represents the count of residents for a specific combination of planning area, subzone, age, sex, and year.

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,…

🧾 **Variable Descriptions**

| Column | Type | Description |
| --- | --- | --- |
| PA | Categorical | Planning Area (e.g., Ang Mo Kio, Bedok) |
| SZ | Categorical | Subzone within each Planning Area |
| Age | Categorical (to be converted) | Age in single years (e.g., 0–100+) |
| Sex | Categorical | Gender: Male / Female |
| Pop | Continuous (Integer) | Number of residents (rounded to nearest 10) |
| Time | Identifier / Constant | Time period (always 2024 in this dataset) |

### **2.3** **Wrangling, Cleaning & Transformation 🧹**

To ensure the data is ready for exploratory visualization and demographic analysis, a series of wrangling and transformation steps are applied. These steps are logically grouped and reusable across multiple visualization tasks.

#### ✅ Convert **Age** to Integer

The Age column is stored as a character string (e.g., "100+"), which can prevent proper sorting, numeric comparisons, and calculations. I converted it to an integer, coercing "100+" to 100, and remove any missing or malformed entries.

This step ensures that age values are usable in:

* Population pyramids
* Median age calculations
* Age band groupings

df <- df %>%  
 mutate(Age = as.integer(Age)) %>% # Convert to numeric (100+ becomes NA)  
 drop\_na(Age) # Remove missing values caused by conversion

#### ✅ Create AgeBand for Grouped Analysis

Working with 101 separate age values (0–100) may clutter visualizations and dilute insight. Therefore, I bin the ages into meaningful age bands for easier demographic comparison and comparative visualization. These age bands are especially helpful for:

* Grouped bar charts
* Age structure comparisons across planning areas

df <- df %>%  
 mutate(AgeBand = cut(Age,  
 breaks = c(0, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55,  
 60, 65, 70, 75, 80, 85, 90, 95, 100, Inf),  
 labels = c("0-4", "5-9", "10-14", "15-19", "20-24", "25-29",  
 "30-34", "35-39", "40-44", "45-49", "50-54",  
 "55-59", "60-64", "65-69", "70-74", "75-79",  
 "80-84", "85-89", "90-94", "95-99", "100+"),  
 right = FALSE))

#### ✅ Aggregate for Exploratory Plots & Key Insights

The data was group by Planning Area, Sex, and Age Band to prepare for visual comparisons across demographic segments. Using the code below, we can compute the **aggregated population totals** across:

* **Age × Sex**: for national pyramids
* **Planning Area**: for ranking and spatial comparisons

# For population pyramid (Age-Sex)  
pyramid\_data <- df %>%  
 group\_by(Age, Sex) %>%  
 summarise(Pop = sum(Pop), .groups = "drop")  
  
# For top 5 planning areas  
top5\_pa <- df %>%  
 group\_by(PA) %>%  
 summarise(Pop = sum(Pop), .groups = "drop") %>%  
 slice\_max(order\_by = Pop, n = 5)

#### ✅ Add Region Classification Based on Planning Area

To enhance geographical analysis, we classify each Planning Area (PA) into one of Singapore’s five official regions (1) Central, (2) East, (3) North, (4) North-East, and (5) West. This grouping follows Singapore’s **Urban Redevelopment Authority (URA) Master Plan** and allows for cleaner regional comparisons, faceted plots, and high-level demographic summaries.

df <- df %>%  
 mutate(  
 PA = str\_trim(str\_to\_upper(PA)),  
 Region = case\_when(  
 PA %in% c("BISHAN", "BUKIT MERAH", "BUKIT TIMAH", "CENTRAL AREA", "DOWNTOWN CORE",  
 "GEYLANG", "KALLANG", "MARINA EAST", "MARINA SOUTH", "MUSEUM", "NEWTON",  
 "NOVENA", "ORCHARD", "OUTRAM", "PAYA LEBAR", "QUEENSTOWN", "RIVER VALLEY",  
 "ROCHOR", "SINGAPORE RIVER", "SOUTHERN ISLANDS", "STRAITS VIEW", "TANGLIN",  
 "TOA PAYOH", "MARINE PARADE") ~ "Central",  
 PA %in% c("BEDOK", "CHANGI", "CHANGI BAY", "PASIR RIS", "TAMPINES") ~ "East",  
 PA %in% c("CENTRAL WATER CATCHMENT", "MANDAI", "SEMBABWANG", "SIMPANG",  
 "SUNGEI KADUT", "WOODLANDS", "YISHUN", "SEMBAWANG") ~ "North",  
 PA %in% c("ANG MO KIO", "HOUGANG", "NORTH-EASTERN ISLANDS", "PUNGGOL",  
 "SELETAR", "SENGKANG", "SERANGOON") ~ "North-East",  
 PA %in% c("BOON LAY", "BUKIT BATOK", "BUKIT PANJANG", "CHOA CHU KANG",  
 "CLEMENTI", "JURONG EAST", "JURONG WEST", "LIM CHU KANG",  
 "PIONEER", "TENGAH", "TUAS", "WESTERN ISLANDS",  
 "WESTERN WATER CATCHMENT") ~ "West",  
 TRUE ~ NA\_character\_  
 )  
 )

#### ✅ Compute the Weighted Median Age

As the simple median of the Age column would be misleading, as it treats each age as equally populated. Instead, we calculate a **weighted median** that reflects **actual population counts** per age to give a statistically more accurate median age for the **entire Singapore resident population**..

df\_median <- df %>%  
 arrange(Age) %>%  
 mutate(  
 cum\_pop = cumsum(Pop),  
 total\_pop = sum(Pop)  
 ) %>%  
 filter(cum\_pop >= total\_pop / 2) %>%  
 slice(1)  
  
median\_age <- df\_median$Age

#### ✅ Generation Band Mapping

Create Generation bands based on categories.

df <- df %>%  
 mutate(  
 GenerationBand = case\_when(  
 Age <= 24 ~ "Alpha & Gen Z (0–24)",  
 Age <= 34 ~ "Young Adults (25–34)",  
 Age <= 54 ~ "Working Adults (35–54)",  
 Age <= 64 ~ "Silver Generation (55–64)",  
 Age <= 74 ~ "Merdeka Generation (65–74)",  
 TRUE ~ "Pioneer Generation (75+)"  
 ),  
 GenerationBand = factor(GenerationBand, levels = c(  
 "Alpha & Gen Z (0–24)", "Young Adults (25–34)", "Working Adults (35–54)",  
 "Silver Generation (55–64)", "Merdeka Generation (65–74)", "Pioneer Generation (75+)"  
 ))  
 )

#### ✅ Create Aggregated Summary Table

df\_summary <- df %>%  
 group\_by(Region, PA, AgeBand, Sex) %>%  
 summarise(Total = sum(Pop), .groups = "drop")

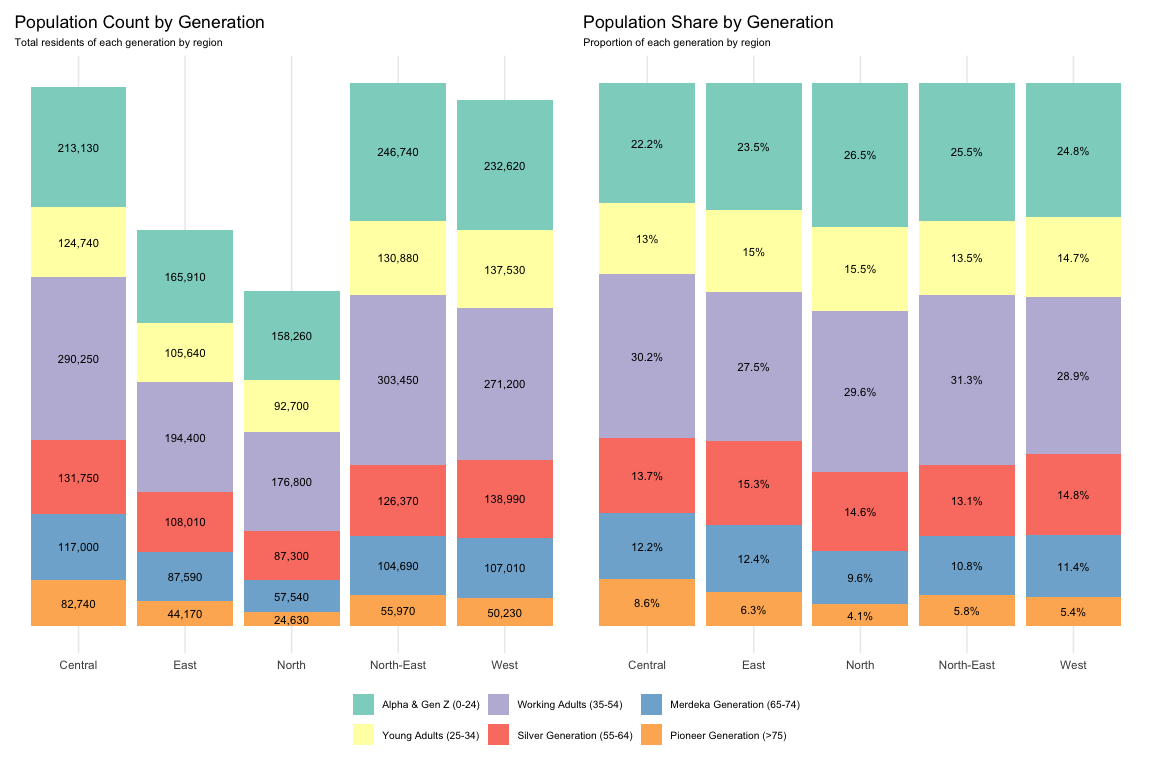
# 3. Exploratory Data Analysis (EDA) 📊

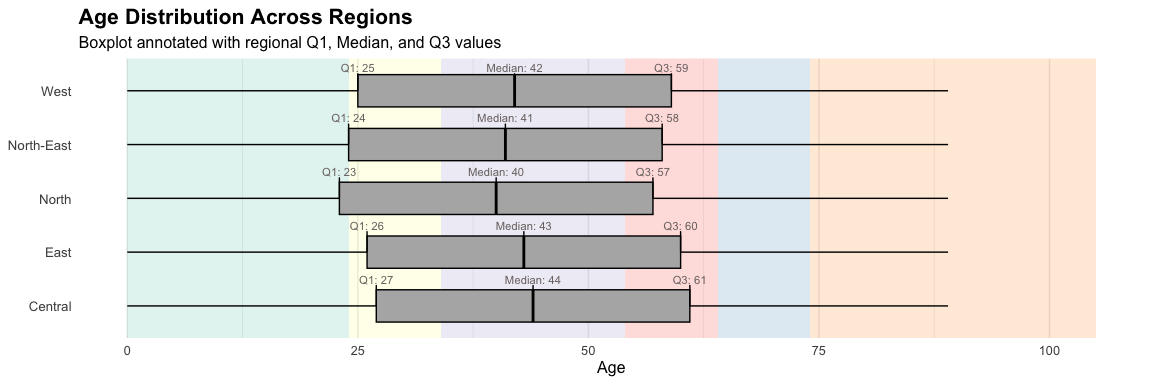
### 🔍 **3.2 How are age groups distributed across the five regions?**

**Goal**: To explore the demographic composition across Singapore’s five regions using meaningful generational groupings, this visualisation compares the proportion of residents by age cohort. It addresses how youth, workforce, and senior populations differ geographically.

**Method**: Using the 2024 DOS dataset, residents are grouped by region and classified into six age bands based on Singapore’s social-policy generations: *Alpha & Gen Z (0–24yo), Young Adults (25–34yo), Working Adults (35–54yo), Silver Generation (55–64yo), Merdeka Generation (65–74yo), and Pioneer Generation (>75yo)*. Percentages are calculated within each region and plotted using a colour-safe, perceptually consistent palette to enhance visual clarity and accessibility.

## The Combined Plot



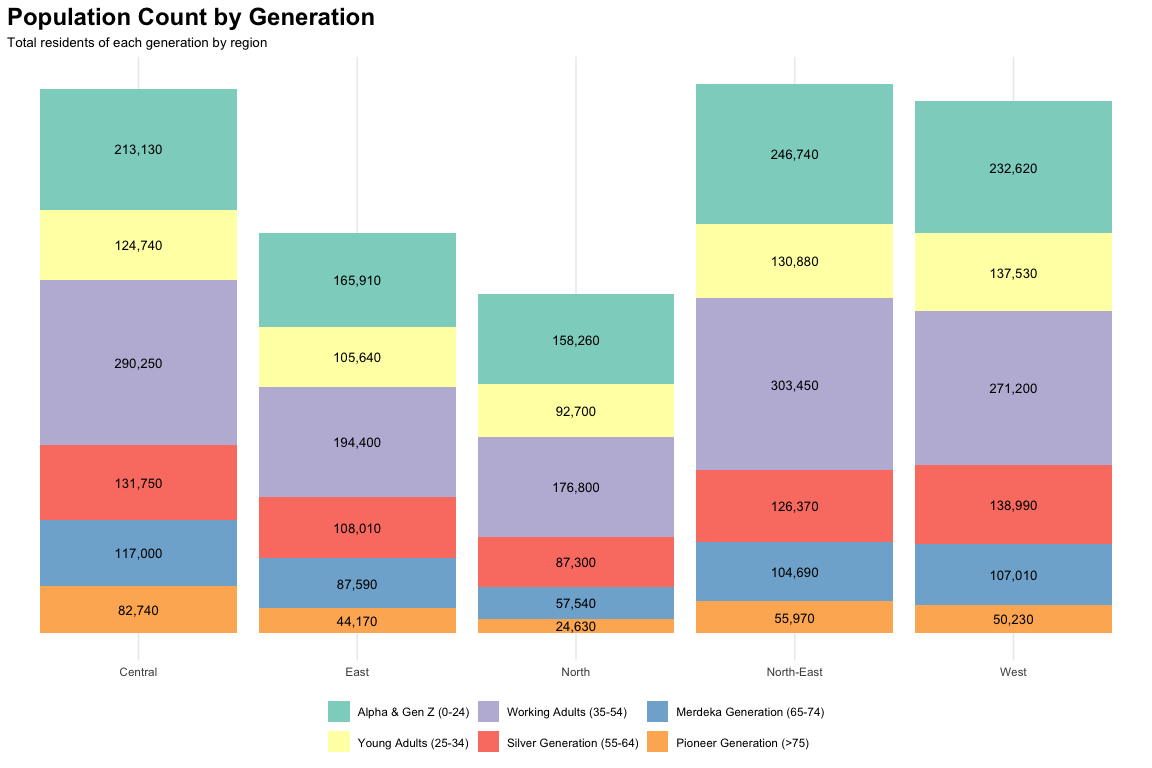


^*For the visuals on the Age Distribution Across Regions, the plot’s background has been coloured to show the Generation Bands across age spectrum of the X-axis.*

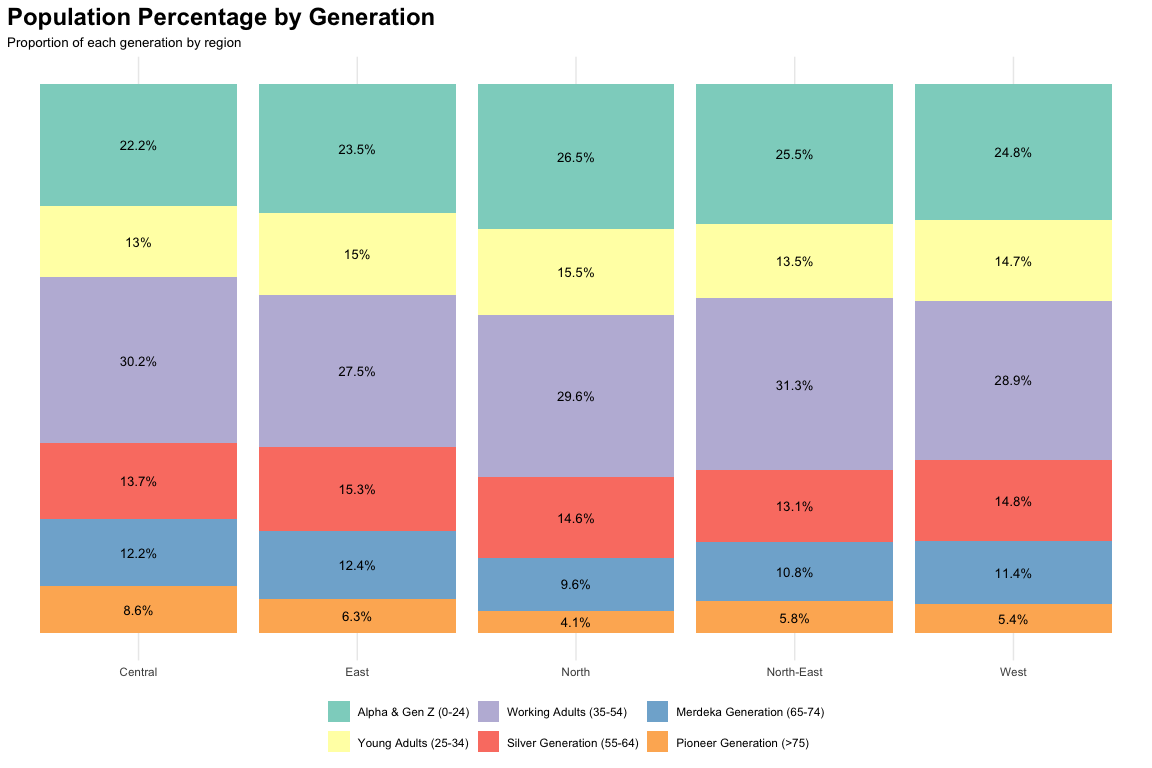
>>> *Toggle to the other tabs to view the clearer version of each plot.*

This visualisation examines the generational distribution of Singapore’s resident population across the five main planning regions, using both absolute population counts and proportional shares. The left panel highlights the total number of residents in each generation, while the right panel normalises these figures to show the internal demographic mix within each region.

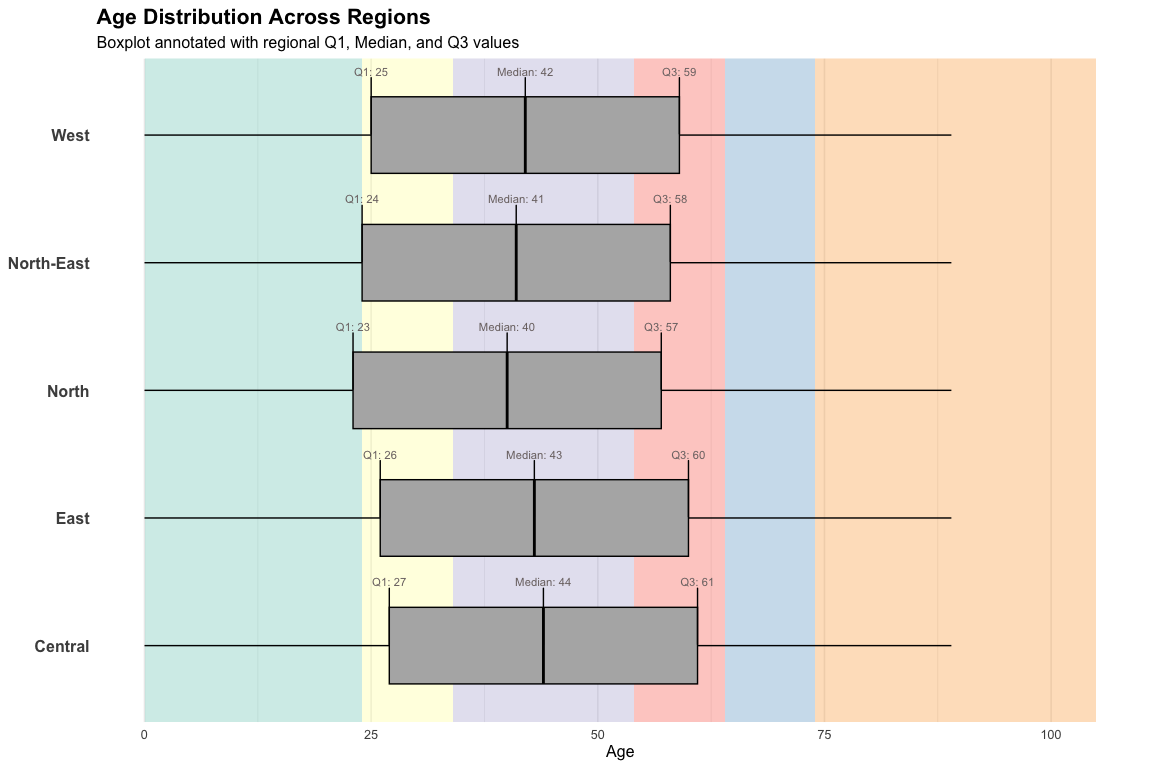
## Population Count by Generation



## Population Percentage by Generation



## Bar Plot of Age Distribution across Regions



## The Code Chunk

# Data prep with generation bands  
df\_generations <- df %>%  
 mutate(  
 AgeGroup = case\_when(  
 Age <= 24 ~ "Alpha & Gen Z (0-24)",  
 Age <= 34 ~ "Young Adults (25-34)",  
 Age <= 54 ~ "Working Adults (35-54)",  
 Age <= 64 ~ "Silver Generation (55-64)",  
 Age <= 74 ~ "Merdeka Generation (65-74)",  
 TRUE ~ "Pioneer Generation (>75)"  
 ),  
 AgeGroup = factor(AgeGroup, levels = c(  
 "Alpha & Gen Z (0-24)",  
 "Young Adults (25-34)",  
 "Working Adults (35-54)",  
 "Silver Generation (55-64)",  
 "Merdeka Generation (65-74)",  
 "Pioneer Generation (>75)"  
 ))  
 ) %>%  
 group\_by(Region, AgeGroup) %>%  
 summarise(Pop = sum(Pop), .groups = "drop") %>%  
 group\_by(Region) %>%  
 mutate(Perc = Pop / sum(Pop) \* 100)  
  
# --- Plot A: Count-based ---  
plot\_counts <- ggplot(df\_generations, aes(x = Region, y = Pop, fill = AgeGroup)) +  
 geom\_bar(stat = "identity") +  
 geom\_text(aes(label = comma(Pop)), position = position\_stack(vjust = 0.5),  
 size = 2.2, color = "black") +  
 scale\_fill\_brewer(palette = "Set3") +  
 scale\_y\_continuous(labels = NULL, breaks = NULL) + # Remove y-axis labels  
 labs(  
 title = "Population Count by Generation",  
 subtitle = "Total residents of each generation by region",  
 x = NULL, y = NULL,  
 fill = NULL  
 ) +  
 theme\_minimal() +  
 theme(  
 legend.position = "none",  
 plot.subtitle = element\_text(size = 8)  
 )  
  
  
# --- Plot B: Percentage-based ---  
plot\_perc <- ggplot(df\_generations, aes(x = Region, y = Perc, fill = AgeGroup)) +  
 geom\_bar(stat = "identity") +  
 geom\_text(aes(label = paste0(round(Perc, 1), "%")),  
 position = position\_stack(vjust = 0.5), size =2.2, color = "black") +  
 scale\_fill\_brewer(palette = "Set3") +  
 scale\_y\_continuous(labels = NULL, breaks = NULL) + # Remove y-axis labels  
 labs(  
 title = "Population Share by Generation",  
 subtitle = "Proportion of each generation by region",  
 x = NULL, y = NULL,  
 fill = NULL  
 ) +  
 theme\_minimal() +  
 theme(  
 legend.position = "bottom",  
 plot.subtitle = element\_text(size = 8),  
 legend.text = element\_text(size = rel(0.7)) # 30% smaller  
 )  
  
# Combine plots  
plot\_counts + plot\_perc + plot\_layout(guides = "collect") &   
 theme(legend.position = "bottom", legend.text = element\_text(size = rel(0.7)))

Insights gleamed from these plots in section 3.2

These visualisations compares Singapore’s resident population by generation across the five main planning regions, using both absolute ounts (left panel) and proportional shares (right panel). **Working adults (ages 35–54)** form the largest demographic in all regions, especially in the **East and North**. However, the **Central region** stands out for its **higher concentration of older residents**, particularly from the **Merdeka and Pioneer generations**, both in absolute numbers and as a proportion of its population.

Conversely, the **North-East and West** regions show a higher share of **younger residents under 35**, reflecting newer housing developments and growing family clusters. While younger generations are more evenly spread in these regions, older cohorts are disproportionately concentrated in the **Central, East, and West**.

Supporting this, the **boxplots and age summaries** reveal **higher median ages** and **tighter interquartile ranges** in these ageing regions—highlighting a more mature demographic profile that warrants targeted **urban planning, healthcare, and ageing-in-place strategies**.

# 4. Comments by Liu Chih Yuan

## Good points

1. Clear Breakdown by Generation
   * The stacked bar plots clearly distinguished generations using multiple colors and labels, making it easy to digest
2. Good Use of Facets (Counts vs Percentage)
   * Shows both raw counts and percentages side by side for good comparison
3. Boxplot with Annotated Summary Stats
   * Labels on age distribution boxplot aids for easier interpretation

## Improvements

1. Font Size Too Small for Lables
   * *Issue:* text is hard to read
   * **Fix:** increase font size to improve readability
2. Color Palette Not Very Readable
   * *Issue:* hard to read and not colorblind friendly
   * **Fix:** use color palette such as viridis or OkabeIto
3. Overcrowded Text in Stacked Bars
   * *Issue:* readability of text inside stacked bars are not good
   * **Fix:** display labels for only larger segments

## Revised Version of EDA 3.2

