Take-Home Exercise 1

Liu Chih Yuan

2025-04-29

# Exploratory Analysis of Age and Gender Distribution Across Singapore Planning Areas (2024)

## 1. Overview

The 2024 dataset from Singapore’ Department of Statistics provides resident population data by planning area, subzone, age, and sex.

### 1.1 Objective

To perform structured exploratory data analysis to uncover insights on demographic distribution across regions.

## 2. Getting Started

### 2.1 Load Packages

pacman::p\_load(tidyverse, ggrepel, ggthemes, patchwork, ggridges, scales)

The R packages used in this EDA are as follows:

1. tidyverse core R package for data science (contains essential packages such as ggplot2)
2. ggrepel for ggplot2 to repel overlapping text labels
3. ggthemes extra ggplot themes
4. patchwork combine ggplot
5. ggridges for ridgeline plots
6. scales customer number formatting

### 2.2 Import Data

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

common\_theme <- theme\_minimal(base\_size = 16) +  
 theme(  
 axis.text = element\_text(size = 18),  
 axis.title = element\_text(size = 20),  
 plot.title = element\_text(size = 18, face = "bold"),  
 legend.text = element\_text(size = 17),  
 legend.title = element\_text(size = 16)  
 )

#### 2.2.1 Check Missing Values

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 <int> 10, 10, 10, 10, 10, 10, 10, 10, 30, 10, 20, 10, 20, 30, 30, 10, 3…  
$ Time <int> 2024, 2024, 2024, 2024, 2024, 2024, 2024, 2024, 2024, 2024, 2024,…

colSums(is.na(df))

PA SZ Age Sex Pop Time   
 0 0 0 0 0 0

### 2.3 Data Processing

#### 2.3.1 Checking Column Types

str(df)

'data.frame': 60424 obs. of 6 variables:  
 $ PA : chr "Ang Mo Kio" "Ang Mo Kio" "Ang Mo Kio" "Ang Mo Kio" ...  
 $ SZ : chr "Ang Mo Kio Town Centre" "Ang Mo Kio Town Centre" "Ang Mo Kio Town Centre" "Ang Mo Kio Town Centre" ...  
 $ Age : chr "0" "0" "1" "1" ...  
 $ Sex : chr "Males" "Females" "Males" "Females" ...  
 $ Pop : int 10 10 10 10 10 10 10 10 30 10 ...  
 $ Time: int 2024 2024 2024 2024 2024 2024 2024 2024 2024 2024 ...

We see column Age is “chr” (character) instead of numeric, let’s check why by finding the unique value of column Age

sort(unique(df$Age))

[1] "0" "1" "10" "11" "12"   
 [6] "13" "14" "15" "16" "17"   
[11] "18" "19" "2" "20" "21"   
[16] "22" "23" "24" "25" "26"   
[21] "27" "28" "29" "3" "30"   
[26] "31" "32" "33" "34" "35"   
[31] "36" "37" "38" "39" "4"   
[36] "40" "41" "42" "43" "44"   
[41] "45" "46" "47" "48" "49"   
[46] "5" "50" "51" "52" "53"   
[51] "54" "55" "56" "57" "58"   
[56] "59" "6" "60" "61" "62"   
[61] "63" "64" "65" "66" "67"   
[66] "68" "69" "7" "70" "71"   
[71] "72" "73" "74" "75" "76"   
[76] "77" "78" "79" "8" "80"   
[81] "81" "82" "83" "84" "85"   
[86] "86" "87" "88" "89" "9"   
[91] "90\_and\_Over"

Here most likely it’s the 90\_and\_over causing it to be a str not int

#### 2.3.2 Creating a Numeric Age Column

df <- df %>%  
 mutate(  
 AgeNum = suppressWarnings(  
 ifelse(Age == "90\_and\_Over", 90, as.numeric(Age))  
 )  
 )

Since we observed people over 90 years old are categorized 90\_and\_above instead of actual numbers, for the ease of plotting we hereby create a new column AgeNum

#### 2.3.3 Creating Age Grouping

df <- df %>%  
 mutate(  
 AgeGroup = case\_when(  
 AgeNum <= 12 ~ "Child",  
 AgeNum <= 24 ~ "Youth",  
 AgeNum <= 64 ~ "Adult",  
 TRUE ~ "Senior"  
 )  
 )

We create a new column AgeGroup for future EDA purposes

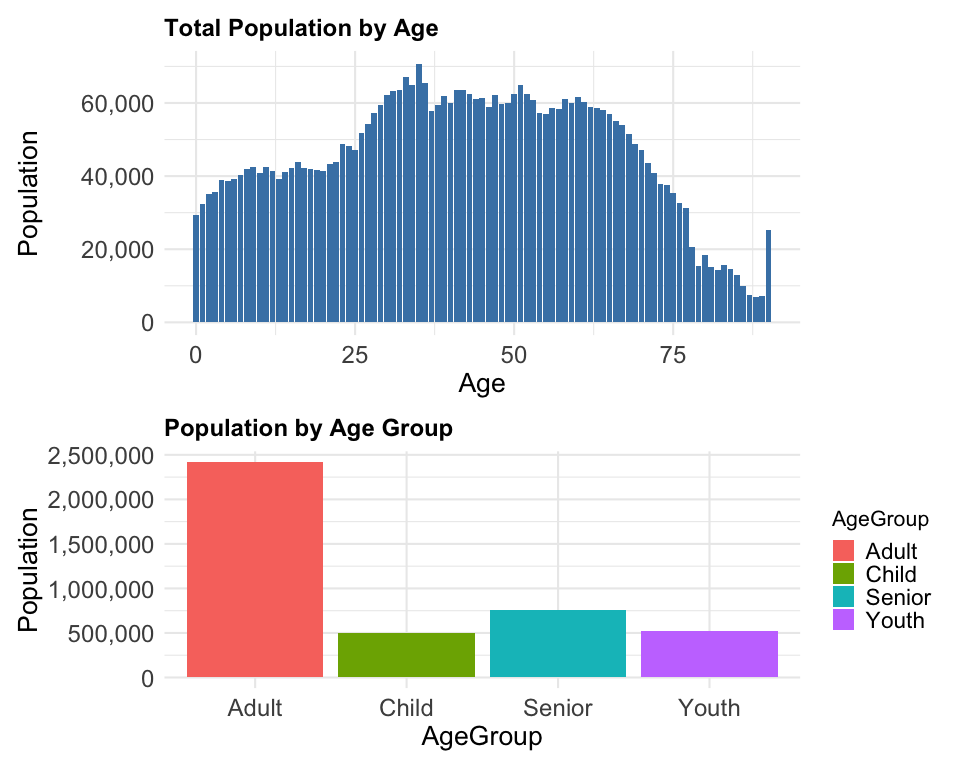
str(df)

'data.frame': 60424 obs. of 8 variables:  
 $ PA : chr "Ang Mo Kio" "Ang Mo Kio" "Ang Mo Kio" "Ang Mo Kio" ...  
 $ SZ : chr "Ang Mo Kio Town Centre" "Ang Mo Kio Town Centre" "Ang Mo Kio Town Centre" "Ang Mo Kio Town Centre" ...  
 $ Age : chr "0" "0" "1" "1" ...  
 $ Sex : chr "Males" "Females" "Males" "Females" ...  
 $ Pop : int 10 10 10 10 10 10 10 10 30 10 ...  
 $ Time : int 2024 2024 2024 2024 2024 2024 2024 2024 2024 2024 ...  
 $ AgeNum : num 0 0 1 1 2 2 3 3 4 4 ...  
 $ AgeGroup: chr "Child" "Child" "Child" "Child" ...

## 3. Exploratory Data Analysis

### EDA 1: Population by Age / Age Group

p1 <- ggplot(df, aes(x = AgeNum, y = Pop)) +  
 stat\_summary(fun = sum, geom = "bar", fill = "steelblue") +  
 labs(title = "Total Population by Age", x = "Age", y = "Population") +  
 scale\_y\_continuous(labels = label\_comma()) +  
 common\_theme  
  
p2 <- df %>%  
 group\_by(AgeGroup) %>%  
 summarise(Pop = sum(Pop)) %>%  
 ggplot(aes(x = AgeGroup, y = Pop, fill = AgeGroup)) +  
 geom\_bar(stat = "identity") +  
 labs(title = "Population by Age Group", y = "Population") +  
 scale\_y\_continuous(labels = label\_comma()) +  
 common\_theme  
  
(p1 / p2) + plot\_layout(heights = c(1.2, 1))

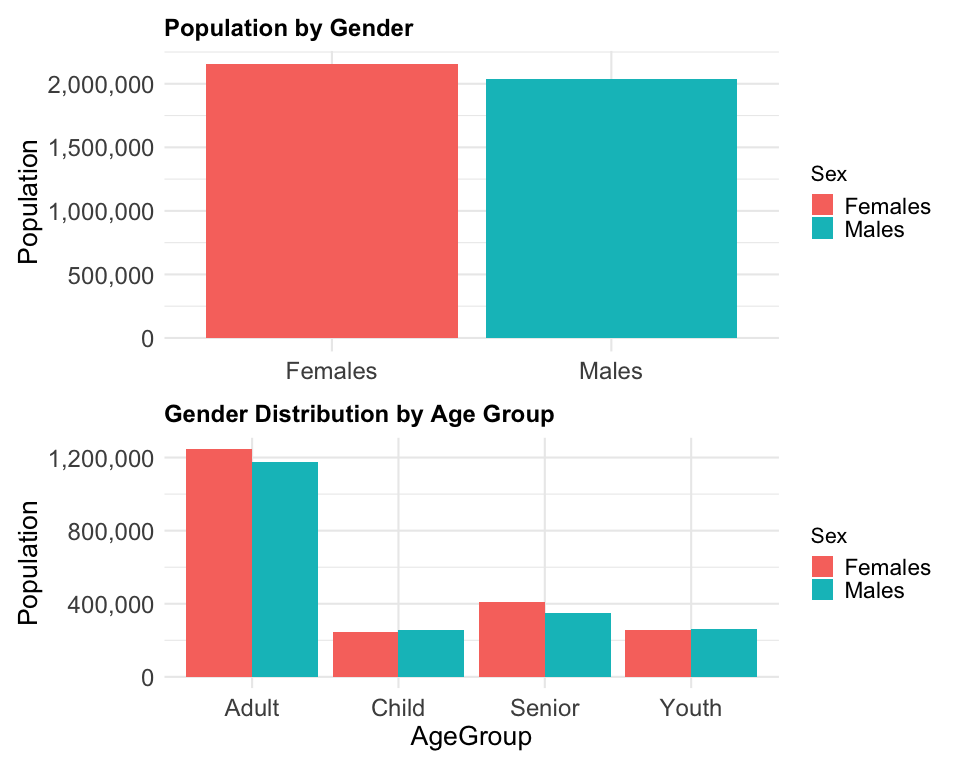


**Insights:**

1. Most residents fall between **ages 25 to 54**
2. **Youth population is shrinking**, suggesting long-term labor sustainability issues
3. **Senior population (65+) rising**, indicating growing need for eldercare and aging population

### EDA 2: Gender Analysis

p3 <- ggplot(df, aes(x = Sex, y = Pop, fill = Sex)) +  
 stat\_summary(fun = sum, geom = "bar") +  
 labs(title = "Population by Gender", x = NULL, y = "Population") +  
 scale\_y\_continuous(labels = label\_comma()) +  
 common\_theme  
  
p4 <- df %>%  
 group\_by(Sex, AgeGroup) %>%  
 summarise(Pop = sum(Pop)) %>%  
 ggplot(aes(x = AgeGroup, y = Pop, fill = Sex)) +  
 geom\_bar(stat = "identity", position = "dodge") +  
 labs(title = "Gender Distribution by Age Group", y = "Population") +  
 scale\_y\_continuous(labels = label\_comma()) +  
 common\_theme  
  
(p3 / p4) + plot\_layout(heights = c(1.2, 1))

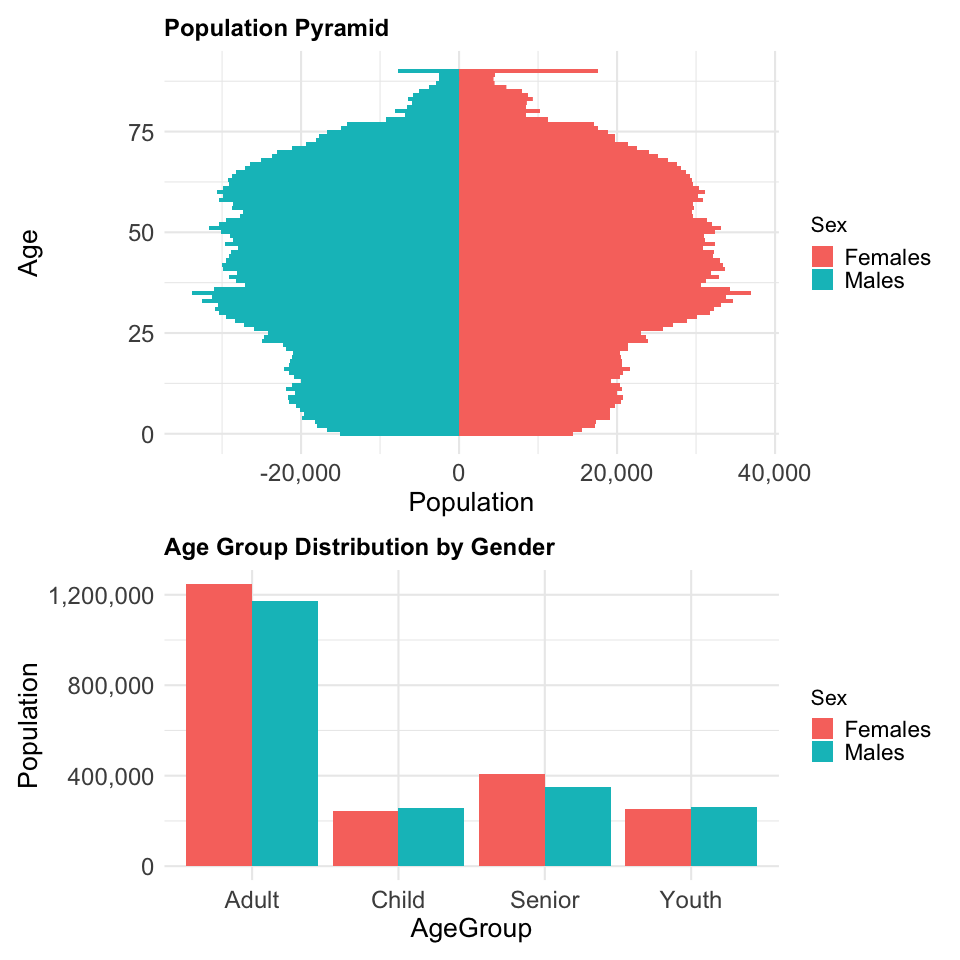


**Insights:**

1. Gender balance is nearly equal overall
2. **Female dominates** in the senior age group, likely due to higher life expectancy

### EDA 3: Population Structure by Age / Gender

df\_pyramid <- df %>%  
 filter(AgeNum <= 90) %>%  
 mutate(Pop = ifelse(Sex == "Males", -Pop, Pop))  
  
p6 <- ggplot(df\_pyramid, aes(x = AgeNum, y = Pop, fill = Sex)) +  
 geom\_col(width = 1) +  
 coord\_flip() +  
 labs(title = "Population Pyramid", x = "Age", y = "Population") +  
 scale\_y\_continuous(labels = label\_comma()) +  
 common\_theme  
  
p7 <- df %>%  
 group\_by(Sex, AgeGroup) %>%  
 summarise(Pop = sum(Pop)) %>%  
 ggplot(aes(x = AgeGroup, y = Pop, fill = Sex)) +  
 geom\_bar(stat = "identity", position = "dodge") +  
 labs(title = "Age Group Distribution by Gender", y = "Population") +  
 scale\_y\_continuous(labels = label\_comma()) +  
 common\_theme  
  
(p6 / p7) + plot\_layout(heights = c(1.3, 1))



**Insights:**

1. Pyramid shows **narrowing base wider top**, typical for aging societies
2. **Adults dominate** across both genders, seniors are the second largest group

## 4. Conculsion

1. Singapore faces a demographic shift towards aging, requiring proactive planning
2. Uneven population spread across subzones and planning ares calls for smart urban development
3. This EDA provides clear insights and serves as baseline for policy design, urban planning, and future modelling