Take-home Exercise 1: Demographic structures and distribution of Singapore in 2024

Belinda Lim

2025-05-01

# **1 Introduction**

## **1.1 Overview**

According to [World Health Organisation](https://www.who.int/data/gho/indicator-metadata-registry/imr-details/1119), an economically dependent population is defined as the sum of the population under 15 years of age plus the population 65 years of age and over, for a given country, territory, or geographic area, at a specific point in time.

Economically productive population is defined as the population between 15 and 64 years of age, for the same country, territory, or geographic area, at the same specific point in time.

With this definition, we aim to conduct an exploratory data analysis on the demographic structures and distribution of Singapore in 2024, focusing on the three groups of population, to understand their characteristics for better policy making.

# **2 Data and Data Pre-processing**

## **2.1 Data Source**

The data for Singapore Residents by Planning Area / Subzone, Single Year of Age and Sex, June 2024 are downloaded from [Department of Statistics, Singapore (DOS)](https://www.singstat.gov.sg/find-data/search-by-theme/population/geographic-distribution/latest-data). The dataset comprise 6 variables and 60,424 records.

## **2.2 Loading data and libraries**

|  |
| --- |
| Note |
| It is assumed that you already have [**pacman**](https://cran.r-project.org/web/packages/pacman/index.html) package installed. If not, please go ahead install pacman first. The code chunk below uses p\_load() of pacman package to check if packages listed below are installed in the computer. If they are, then they will be launched into R. |

pacman::p\_load(tidyverse, patchwork, scales, sf, plotly)

We import Demographic\_data.csv into R environment by using [*read\_csv()*](https://readr.tidyverse.org/reference/read_delim.html) function of [**readr**](https://readr.tidyverse.org/) package. **readr** is one of the tidyverse package.

pop\_data <- read\_csv("data/Demographic\_data.csv", show\_col\_types = FALSE)

The attributes found in the pop\_data are:

* Planning Areas (PA): There are 55 PAs in Singapore designated by URA, such as Ang Mo Kio, Bedok, etc.
* Subzones (SZ): PAs are further sub-divided into 332 SZs, such as Ang Mo Kio Town Centre, Cheng San etc. under Ang Mo Kio PA.
* Age: 0 to 89 and 90\_and\_Over
* Sex: Male and Female
* Pop: Number of Population
* Time: 2024

Under Age, there is also a category “90\_and\_Over”. For data visualisation, the category is converted to the numeric number “90” instead. The following codes will convert and ensure that all figures under “Age” and “Pop” are numeric prior to further data analysis.

pop\_data <- pop\_data %>%  
 mutate(  
 Age = if\_else(Age == "90\_and\_Over", "90", Age),   
 Age = as.numeric(Age),   
 Pop = as.numeric(Pop)   
 )

# **3 Exploratory Data Analysis and Insights**

## **3.1 Singapore Population Age Pyramid and Life Expectancy**

### 3.1.1 Data preparation and Visualisation

First we will bin the ages into 5-year blocks for greater visibility by age groups in the age pyramid.

pop\_data <- pop\_data %>%  
mutate(AgeGroup = cut(Age, breaks = c(seq(0, 90, by = 5), Inf),  
 labels = c(paste(seq(0,85,5), seq(4,89,5), sep="-"), "90+"),  
 right = FALSE))

The population data is grouped by age group and sex and the data for male population is converted such that it is on the left.

age\_pyramid <- pop\_data %>%  
 group\_by(AgeGroup, Sex) %>%  
 summarise(Population = sum(Pop), .groups = 'drop') %>%  
 mutate(  
 Population = ifelse(Sex == "Males", -Population, Population),   
 Sex = factor(Sex, levels = c("Males", "Females"))  
 )

The code for the Age Pyramid plot is:

pyramid\_plot <- ggplot(age\_pyramid, aes(x = Population, y = AgeGroup, fill = Sex)) +  
 geom\_bar(stat="identity") +  
 scale\_x\_continuous(  
 labels = function(x) scales::comma(abs(x)),  
 breaks = scales::pretty\_breaks()) +  
 labs(title="Age Pyramid (Population Number)", x= NULL, y="Age Group") +  
 scale\_fill\_manual(values = c("Males" = "lightblue", "Females" = "pink")) +  
 theme\_minimal() +  
 theme(legend.position="bottom") +  
 theme(plot.title = element\_text(hjust = 0.5,size = 10, face = "bold"))

The proportion of males and females are also compared for each age group. The code to compute the porportion is as follows:

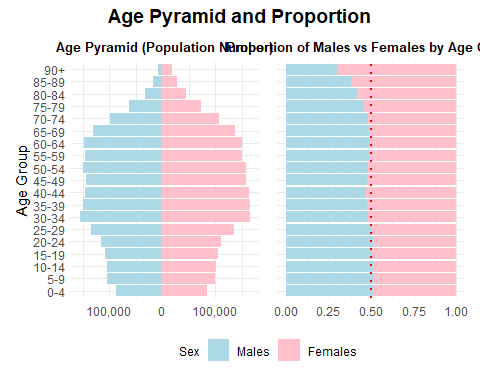
proportion\_data <- pop\_data %>%  
 group\_by(AgeGroup, Sex) %>%  
 summarise(Population = sum(Pop, na.rm = TRUE), .groups = "drop") %>%  
 group\_by(AgeGroup) %>%  
 mutate(  
 TotalInGroup = sum(Population),  
 Proportion = Population / TotalInGroup  
 )

The code for the Proportion plot is:

proportion\_plot <- ggplot(proportion\_data, aes(x = Proportion, y = AgeGroup, fill = Sex)) +  
 geom\_bar(stat = "identity") +  
 geom\_vline(xintercept = 0.5, color = "red", linetype = "dotted", linewidth = 1) +   
 labs(  
 title = "Proportion of Males vs Females by Age Group",  
 x = NULL,  
 y = NULL  
 ) +  
 scale\_fill\_manual(  
 values = c("Males" = "lightblue", "Females" = "pink"),  
 guide = "none"   
) +  
 theme\_minimal() +  
 theme(  
 axis.text.x = element\_text(hjust = 0.5),  
 plot.title = element\_text(hjust = 0.5,size = 10, face = "bold"),  
 legend.position = "none"  
 )

The 2 plots are combined using the code below.

### The plot



proportion\_plot <- proportion\_plot +  
 theme(axis.text.y = element\_blank(), axis.ticks.y = element\_blank())  
  
combined\_plot <- wrap\_plots(  
 list(pyramid\_plot, proportion\_plot),  
 ncol = 2,  
 widths = c(1, 1.5),  
 guides = "collect",  
 align = "hv",  
 axis = "y"  
) +  
 plot\_annotation(  
 title = "Age Pyramid and Proportion",  
 theme = theme(  
 plot.title = element\_text(hjust = 0.5, size = 14, face = "bold")  
 )  
 ) &  
 theme(  
 legend.position = "bottom",  
 legend.box = "horizontal",  
 legend.justification = "center",  
 legend.box.just = "center",  
 legend.text = element\_text(size = 9),  
 legend.title = element\_text(size = 9)  
 )  
  
combined\_plot

### 3.1.2 Insights

Singapore’s population age pyramid can be described as a [constrictive pyramid](https://populationeducation.org/what-are-different-types-population-pyramids/), which is typical for countries with higher socio-economic development. There are low birth and mortality rates, contributing to an ageing population. The largest population groups belong to ages 30-64. As population ages, the proportion of economically productive population will decline.

The population’s proportion shows that there is gender balance in the younger population aged 30 and below. Notably, there is increasingly higher proportion of females than males from age 70 onwards. This suggests that females tend to have longer life expectancy than males.

With ageing population, it is important to focus on preventive healthcare measures for a healthier population. This enables better quality of life and delayed retirement age for older Singaporeans to contribute economically. As there are higher proportion of elderly females, more resources such as hospital beds or geriatric care can be catered for the females.

## **3.2 Population demographics by Planning Areas**

There are 55 Planning Areas (PA) designated by [URA](https://data.gov.sg/collections/1817/view) where each planning area is served by a town centre and several neighbourhood commercial/shopping centres. Each area is supported by smaller subzones (SZ), which are usually centred on a focal point such as neighbourhood centre or activity node.

Using data visualisation, we can identify the top 10 PAs with the highest percentage of population aged 65 and above.

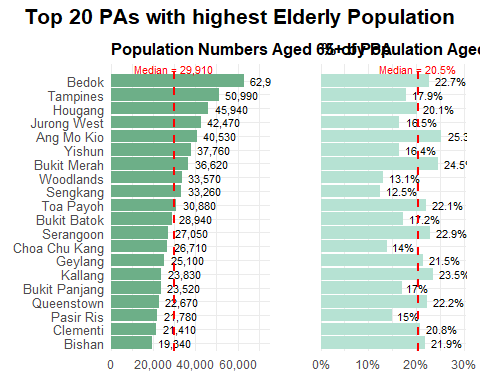
### 3.2.1 Data preparation and Visualisation

The following codes are used to:

* Compute Total population per PA
* Compute Total Population aged 65 and above by PA
* Merge both data and compute percentage of Population aged 65 and above by PA

pop\_total <- pop\_data %>%  
 group\_by(PA) %>%  
 summarise(TotalPop = sum(Pop, na.rm = TRUE))  
  
pop\_65plus <- pop\_data %>%  
 filter(Age >= 65) %>%  
 group\_by(PA) %>%  
 summarise(Pop65Plus = sum(Pop, na.rm = TRUE))  
  
pop\_summary <- left\_join(pop\_total, pop\_65plus, by = "PA") %>%  
 mutate(  
 Pop65Plus = replace\_na(Pop65Plus, 0),  
 PctOver65 = Pop65Plus / TotalPop \* 100,  
 PctOver65 = ifelse(is.nan(PctOver65), NA, round(PctOver65, 1))  
 )

### The plot



### The code

plot\_data <- pop\_summary %>%  
 arrange(desc(Pop65Plus)) %>%  
 slice\_head(n = 20) %>%  
 mutate(  
 PA = factor(PA, levels = rev(PA)),   
 LabelPop = format(Pop65Plus, big.mark = ","),  
 LabelPct = paste0(round(PctOver65, 1), "%")  
 )  
  
median\_pop <- median(plot\_data$Pop65Plus, na.rm = TRUE)  
median\_pct <- median(plot\_data$PctOver65, na.rm = TRUE)  
  
pop\_plot <- ggplot(plot\_data, aes(x = Pop65Plus, y = PA)) +  
 geom\_col(fill = "#6DAF88")+  
 geom\_text(aes(label = LabelPop), hjust = -0.2, size = 2.8) +  
 geom\_vline(xintercept = median\_pop, color = "red", linetype = "dashed", linewidth = 0.8) +  
annotate("text", x = median\_pop, y = 21.2, label = paste("Median =", scales::comma(median\_pop)), color = "red", size = 3, vjust = 1) +  
 labs(title = "Population Numbers Aged 65+ by PA", x = NULL, y = NULL) +  
 theme\_minimal() +  
 theme(  
 plot.title = element\_text(size = 12, face = "bold"),  
 axis.text.y = element\_text(size = 10),  
 plot.margin = margin(5, 30, 5, 5)   
 ) +  
 scale\_x\_continuous(labels = scales::comma, expand = expansion(mult = c(0, 0.2)))  
  
pct\_plot <- ggplot(plot\_data, aes(x = PctOver65, y = PA)) +  
 geom\_col(fill = "#b6e2d3") +  
 geom\_text(aes(label = LabelPct), hjust = -0.2, size = 2.8) +  
 geom\_vline(xintercept = median\_pct, color = "red", linetype = "dashed", linewidth = 0.8) +  
 annotate("text", x = median\_pct, y = 21.2, label = paste0("Median = ", round(median\_pct, 1), "%"), color = "red", size = 3, vjust = 1)+  
 labs(title = "% of Population Aged 65+ by PA", x = NULL, y = NULL) +  
 theme\_minimal() +  
 theme(  
 plot.title = element\_text(size = 12, face = "bold"),  
 axis.text.y = element\_blank(),  
 axis.ticks.y = element\_blank(),  
 plot.margin = margin(5, 5, 5, 5) # optional  
 ) +  
 scale\_x\_continuous(labels = scales::label\_percent(scale = 1), expand = expansion(mult = c(0, 0.2)))  
  
final\_chart <- pop\_plot + pct\_plot +  
 plot\_layout(ncol = 2, widths = c(1.1, 1)) +  
 plot\_annotation(  
 title = "Top 20 PAs with highest Elderly Population",  
 theme = theme(  
 plot.title = element\_text(hjust = 0.5, face = "bold", size = 16),  
 plot.subtitle = element\_text(hjust = 0.5, size = 11)  
 )  
 )  
  
final\_chart

### 3.2.2 Insights

There are 10 PAs with elderly population aged 65 and above exceeding Singapore’s median. As the population size vary across PAs, it is important to compare the numbers and percentage of elderly population in each PA to understand the ageing profile. From both charts, we conclude that there are 4 PAs, namely Bedok, Ang Mo Kio, Bukit Merah and Toa Payoh that have both higher number and percentage of elderly population higher than the median. These are critical ageing towns that requires necessary infrastructure, healthcare and eldercare services to support mobility and active lifestyle.

Towns like Tampines and Hougang also have high numbers of elderly population but with percentage of elderly population slightly below the median. These towns have more demographic mix, with younger populations residing. Given the high absolute number of elderly population, it is important to plan ahead and scale up infrastructure, healthcare and eldercare services too.

## **3.3 Dependency Ratio**

Dependency ratio measures the average number of economically dependent population per 100 economically productive population, for a given country, territory, or geographic area, at a specific point in time.

### 3.3.1 Data preparation and Visualisation

Given the definitions mentioned under the overview, we create and sum up the population numbers in the following groups

* Those Age\_DR < 15 are “Dependent\_0\_14”
* Those Age\_DR >= 65 are “Dependent\_65\_plus”
* Those Age\_DR >= 15 & Age\_DR <= 64 are “Productive”

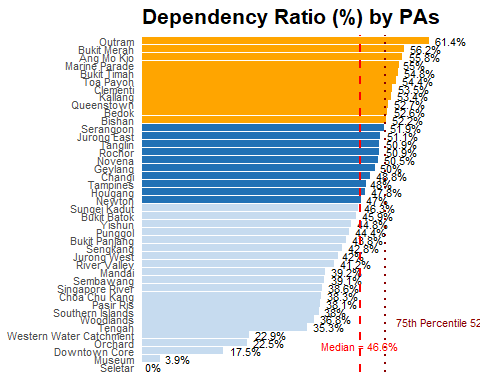
The codes for the groupings and data preparation are:

pop\_data <- read\_csv("data/Demographic\_data.csv") %>%  
 mutate(  
 Age\_DR = if\_else(Age == "90\_and\_Over", "90", as.character(Age)),  
 Age\_DR = as.numeric(Age\_DR),  
 Pop\_DR = as.numeric(Pop),  
 AgeGroup\_DR = case\_when(  
 Age\_DR < 15 ~ "Dependent\_0\_14",  
 Age\_DR >= 65 ~ "Dependent\_65\_plus",  
 Age\_DR >= 15 & Age\_DR <= 64 ~ "Productive",  
 TRUE ~ NA\_character\_  
 )  
 )  
  
pop\_summary\_DR <- pop\_data %>%  
 filter(!is.na(AgeGroup\_DR)) %>%  
 group\_by(PA, AgeGroup\_DR) %>%  
 summarise(Pop\_DR = sum(Pop\_DR, na.rm = TRUE), .groups = "drop")  
  
# Prepare wide table for Dependency Ratio calculation  
pop\_wide\_DR <- pop\_summary\_DR %>%  
 pivot\_wider(  
 names\_from = AgeGroup\_DR,  
 values\_from = Pop\_DR,  
 values\_fill = 0  
 ) %>%  
 mutate(  
 Dependent\_DR = Dependent\_0\_14 + Dependent\_65\_plus,  
 Productive\_DR = Productive,  
 DependencyRatio = round((Dependent\_DR / Productive\_DR) \* 100, 1)  
 ) %>%  
 filter(Productive\_DR > 0)

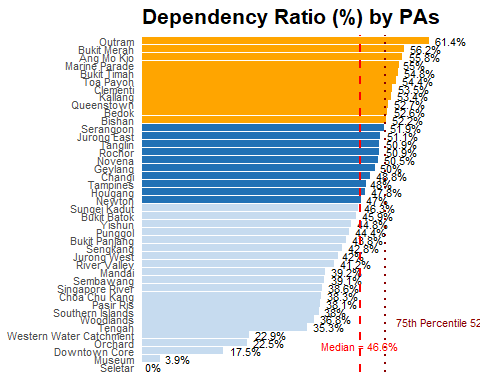
The Dependency ratio for the PA is computed with the following code:

# Calculate median and 75th percentile  
median\_DR <- median(pop\_wide\_DR$DependencyRatio)  
q75\_DR <- quantile(pop\_wide\_DR$DependencyRatio, 0.75)  
  
# Create 3-category fill group  
pop\_wide\_DR <- pop\_wide\_DR %>%  
 mutate(  
 FillGroup = case\_when(  
 DependencyRatio > q75\_DR ~ "Above 75th",  
 DependencyRatio > median\_DR ~ "Between Median–75th",  
 TRUE ~ "Below Median"  
 )  
 )

### The plot



### The code



### 3.3.2 Insights

The dependency ratio reflects the number of economically-dependent population (i.e. children under 15 and elderly above 65) supported by economically-productive population. This is usually measured at the national level, but the breakdown by PAs reveals variations in the demographic distribution. There are 11 PAs with dependency ratio exceeding 75th percentile (52.1%). As these towns have more children and elderly, it is recommended to priortise more mixed use infrastructure, co-located elder/ childcare facilities and community services catered for both groups. From housing perspective, Housing & Development Board can consider building more multi-generational flats for families to stay in the same neighbourhood.

There are also 10 PAs with dependency ratio between 50th to 75th percentile (46.6% to 52.1%). These are subsequent PAs that will benefit from better mixed-use urban planning. As infrastructure and facilities take years to build, the Government can intervene early to futureproof these neighbourhood and support the ageing population.

# **4 Conclusion and Future Study**

The exploratory data analysis summarises Singapore’s demographic structure in 2024, focusing on the demograhic characteristics on age composition, life expectancy, elderly population distribution, and dependency ratio.

Future studies can consider having time-series analysis to analyse population changes and trends over time. It is also useful to supplement the dependency ratio analysis with household incomes of the population in the PAs to under the economical pressure that the economically-productive population are facing.

# **5 References**

* Department of Statistics Singapore (2024). *Singapore Residents by Planning Area / Subzone, Single Year of Age and Sex, June 2024*. Retrieved from: <https://www.singstat.gov.sg>
* Population Education. *What are the Different Types of Population Pyramids?* Retrieved from: <https://populationeducation.org/what-are-different-types-population-pyramids/>
* Urban Redevelopment Authority (URA). *Master Plan 2014 – Planning Area Boundary*. Retrieved from: <https://www.ura.gov.sg>
* World Health Organization (2023). *Indicator Metadata Registry: Economically Dependent Population*. Retrieved from: <https://www.who.int/data/gho/indicator-metadata-registry/imr-details/1119>

# **6 Extra visualisation beyond the scope of Take-Home Exercise**

The Singapore’s map shapefile is downloaded from [URA’s Master Plan 2014 Planning Area Boundary](https://data.gov.sg/collections/1701/view).

sg\_map <- st\_read("data/Map/MP14\_PLNG\_AREA\_WEB\_PL.shp")

Reading layer `MP14\_PLNG\_AREA\_WEB\_PL' from data source   
 `C:\Users\Belinda\OneDrive - Singapore Management University\Desktop\Visual Analytics\belindalim\ISSS608-VAA\Take-home\_Ex\Take-home\_Ex01\data\Map\MP14\_PLNG\_AREA\_WEB\_PL.shp'   
 using driver `ESRI Shapefile'  
Simple feature collection with 55 features and 12 fields  
Geometry type: MULTIPOLYGON  
Dimension: XY  
Bounding box: xmin: 2667.538 ymin: 15748.72 xmax: 56396.44 ymax: 50256.33  
Projected CRS: SVY21

The percentage of Population aged 65 and above by PA is computed and joined to the map.

sg\_map <- sg\_map %>%  
 mutate(PA = toupper(trimws(PLN\_AREA\_N)))  
  
pop\_summary <- pop\_summary %>%  
 mutate(PA = toupper(trimws(PA)))  
  
sg\_map\_joined <- sg\_map %>%  
 left\_join(pop\_summary, by = "PA")

The choropleth map with percentage of elderly population aged 65 and above is shown below:

### The plot

p <- ggplot(sg\_map\_joined) +  
 geom\_sf(aes(  
 fill = PctOver65,  
 text = paste0("Planning Area: ", PA, "\nAge 65+ Population: ", PctOver65, "%")  
 ), color = "white", linewidth = 0.3) +  
 scale\_fill\_gradient(  
 low = "#ffe0b2",   
 high = "#b30000",   
 na.value = "grey90") +  
 labs(  
 title = "% of Population Aged 65+ by PA",  
 fill = "% Aged 65+"  
 ) +  
 theme\_void() + theme(panel.border = element\_blank())   
  
ggplotly(p, tooltip = "text")

### The code

p <- ggplot(sg\_map\_joined) +  
 geom\_sf(aes(  
 fill = PctOver65,  
 text = paste0("Planning Area: ", PA, "\nAge 65+ Population: ", PctOver65, "%")  
 ), color = "white", linewidth = 0.3) +  
 scale\_fill\_gradient(  
 low = "#ffe0b2",   
 high = "#b30000",   
 na.value = "grey90") +  
 labs(  
 title = "Residents Aged 65+ by Planning Area (2024)",  
 fill = "% Aged 65+"  
 ) +  
 theme\_void() + theme(panel.border = element\_blank())   
  
ggplotly(p, tooltip = "text")