Take-home Exercise 1

Stefanie Felicia

2025-05-03

# 1. Background

Singapore is a country home to more than 5 million population. Though it may be a small country, but understanding its demographic structure and distribution is crucial for assessing its development and planning needs.

Understanding demographic structure and distribution is vital as helps us to see how a country is doing. The distribution and structure of a country can reveal how developed a country is, through birth rate, death rates, economy and literacy.

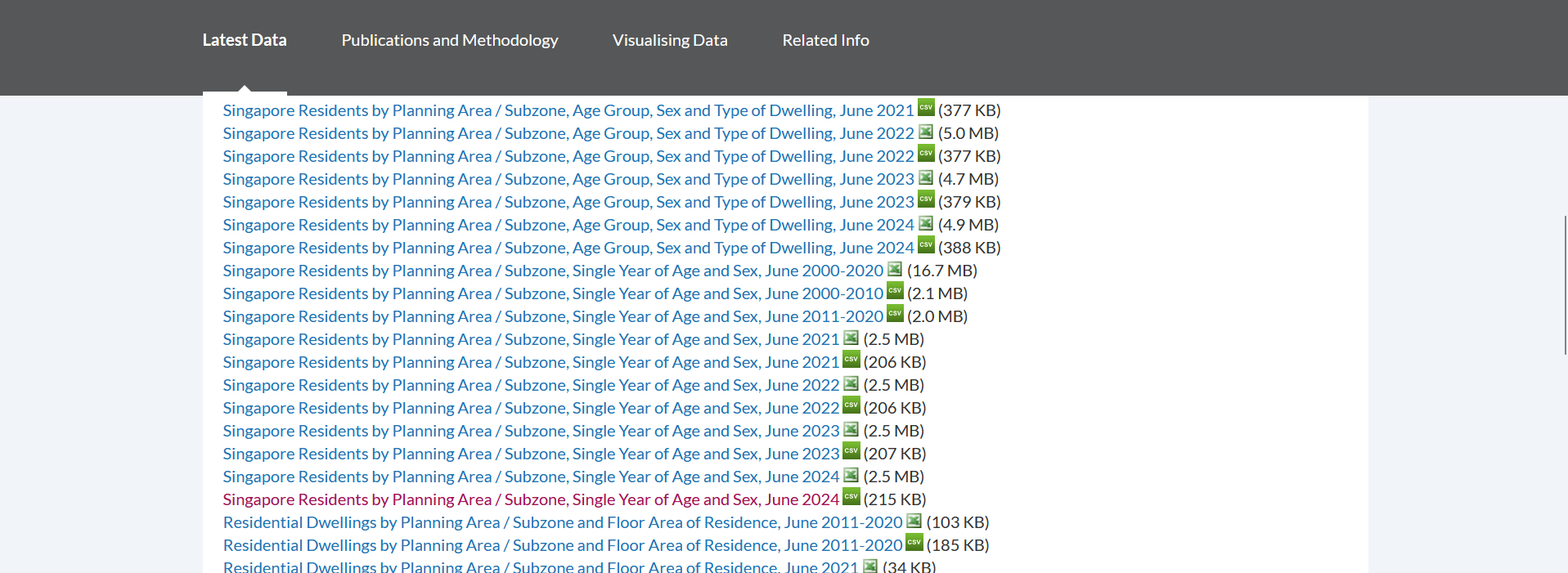
Additionally, understanding demographics also helps governments, and businesses in order to design policies, allocate resources and infrastructure development and provide public services that are effectively targeted.

## 1.2 Objective

In this take home exercise, we desire to create demographic structures and distribution of Singapore in 2024.

## 1.3 The Data

The dataset was extracted from Singstat under the name Singapore Residents by Planning Area / Subzone, Single Year of Age and Sex, June 2024.



It comprised of 6 variables and 60K records. The dataset recorded the demographics and count of residents who lived in each sub zone and planning area in June 2024.

## 1.4 Load Required Libraries

pacman::p\_load(dplyr, tidyverse, scales, forcats, knitr)

the library used for this exercise are:

| Package Name | Description |
| --- | --- |
| dplyr | Grammar of data manipulation |
| tidyverse | Provide key data transformation functions |
| scales | Provides internal scaling infrastructure used by ggplot2 to a general framework |
| forcats | Tools for working with factors (categorical variable) |
| knitr | Provides a tool for dynamic report generation |

## 1.5 Import Data

population\_data <- read\_csv("respopagesex2024.csv")

Rows: 60424 Columns: 6  
── Column specification ────────────────────────────────────────────────────────  
Delimiter: ","  
chr (4): PA, SZ, Age, Sex  
dbl (2): Pop, Time  
  
ℹ Use `spec()` to retrieve the full column specification for this data.  
ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

As can be seen above, the data set comprise of 6 variables: 4 variables are character data type (namely, PA, SZ, age and sex), and 2 variables are double data type (number of population, time). Here is the a snippet of the data set:

head(population\_data)

# A tibble: 6 × 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

# 2. Data pre-processing

First, we will look if there are duplicates found in the data set.

nrow(population\_data[duplicated(population\_data), ])

[1] 0

As we found no duplicate records in the data set, next we will check for missing values in the data set.

sapply(population\_data, function(x) sum(is.na(x)))

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

As there’s no missing values in each column, we will move on to column transformation.

## 2.1 Column Transformation

## Age

Currently, the Age variable is in string data type and we want to change it to numeric data type.

population\_data$Age[population\_data$Age == "90\_and\_Over"] <- "90"  
typeof(population\_data$Age)

[1] "character"

population\_data$Age <- as.numeric(population\_data$Age)

## Sex

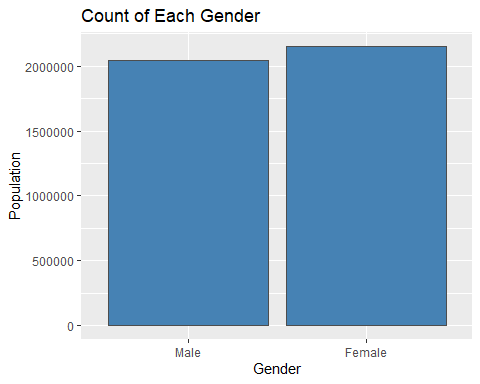
We also want to implement dummy encoding to the Sex variable from “Males” to 0 and “Females” to 1:

population\_data$Sex[population\_data$Sex == "Females"] <- "1"  
population\_data$Sex[population\_data$Sex == "Males"] <- "0"  
population\_data$Sex <- as.numeric(population\_data$Sex)

## 2.2 Data Exploration

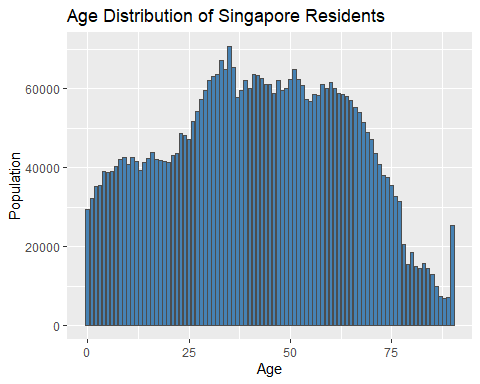
### 2.2.1. Frequency of Genders

gender\_distribution <- population\_data %>%  
 group\_by(Sex) %>%  
 summarise(Pop = sum(Pop), .groups = "drop")  
  
ggplot(gender\_distribution, aes(x = factor(Sex),y=Pop)) +  
 geom\_bar(stat = "identity", color = "grey30", fill = "steelblue") +  
 scale\_x\_discrete(labels = c("0" = "Male", "1" = "Female")) +  
 labs(x = "Gender", y = "Population", title = "Count of Each Gender")



### 2.2.2 Distribution of Age

age\_distribution <- population\_data %>%  
 group\_by(Age) %>%  
 summarise(Pop = sum(Pop), .groups = "drop")  
  
  
  
ggplot(age\_distribution, aes(x = Age, y = Pop)) +  
 geom\_col(fill = "steelblue", color = "gray30") +  
 labs(title = "Age Distribution of Singapore Residents",  
 x = "Age",  
 y = "Population")



### 2.2.3. Number of Subzone each Planning Area Has

pa\_count <- population\_data %>%  
 distinct(PA, SZ) %>%   
 count(PA, name = "number\_of\_subzones") %>%  
 arrange(desc(number\_of\_subzones))  
  
kable(pa\_count, caption = "Number of Subzones per Planning Area")

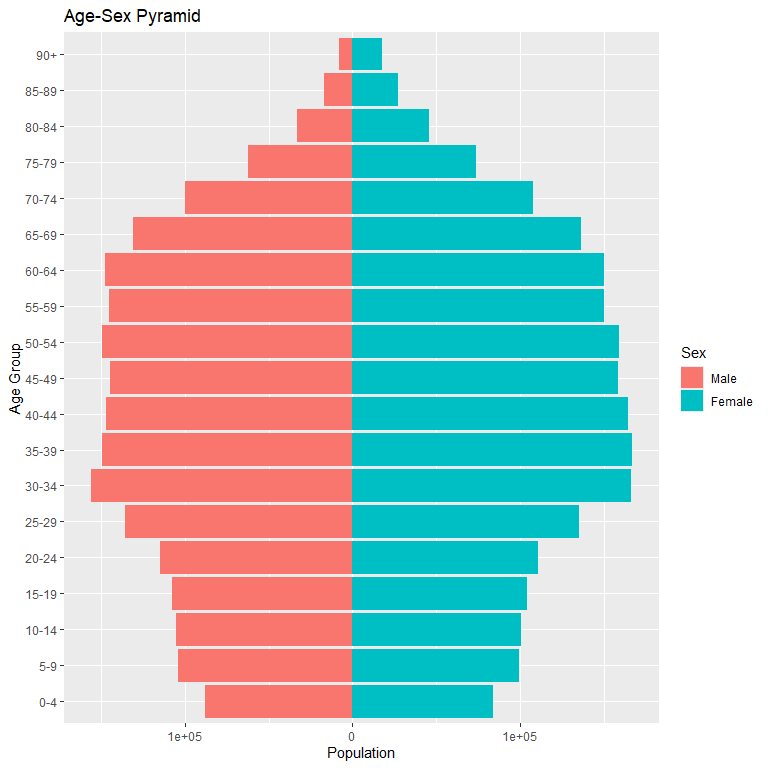
Number of Subzones per Planning Area

| PA | number\_of\_subzones |
| --- | --- |
| Bukit Merah | 17 |
| Queenstown | 15 |
| Downtown Core | 13 |
| Ang Mo Kio | 12 |
| Toa Payoh | 12 |
| Jurong East | 11 |
| Hougang | 10 |
| Rochor | 10 |
| Bukit Batok | 9 |
| Clementi | 9 |
| Jurong West | 9 |
| Kallang | 9 |
| Sembawang | 9 |
| Woodlands | 9 |
| Yishun | 9 |
| Bedok | 8 |
| Bukit Timah | 8 |
| Pasir Ris | 8 |
| Bukit Panjang | 7 |
| Punggol | 7 |
| Sengkang | 7 |
| Serangoon | 7 |
| Choa Chu Kang | 6 |
| Newton | 6 |
| Tengah | 6 |
| Tuas | 6 |
| Geylang | 5 |
| Marine Parade | 5 |
| Novena | 5 |
| Paya Lebar | 5 |
| Pioneer | 5 |
| River Valley | 5 |
| Sungei Kadut | 5 |
| Tampines | 5 |
| Boon Lay | 4 |
| Outram | 4 |
| Seletar | 4 |
| Simpang | 4 |
| Tanglin | 4 |
| Bishan | 3 |
| Changi | 3 |
| Mandai | 3 |
| Museum | 3 |
| Orchard | 3 |
| Singapore River | 3 |
| Western Islands | 3 |
| Western Water Catchment | 3 |
| Southern Islands | 2 |
| Central Water Catchment | 1 |
| Changi Bay | 1 |
| Lim Chu Kang | 1 |
| Marina East | 1 |
| Marina South | 1 |
| North-Eastern Islands | 1 |
| Straits View | 1 |

# 3. Visualization

## 3.1 Age and Gender Distribution of Singapore Residents

## Graph



## Code

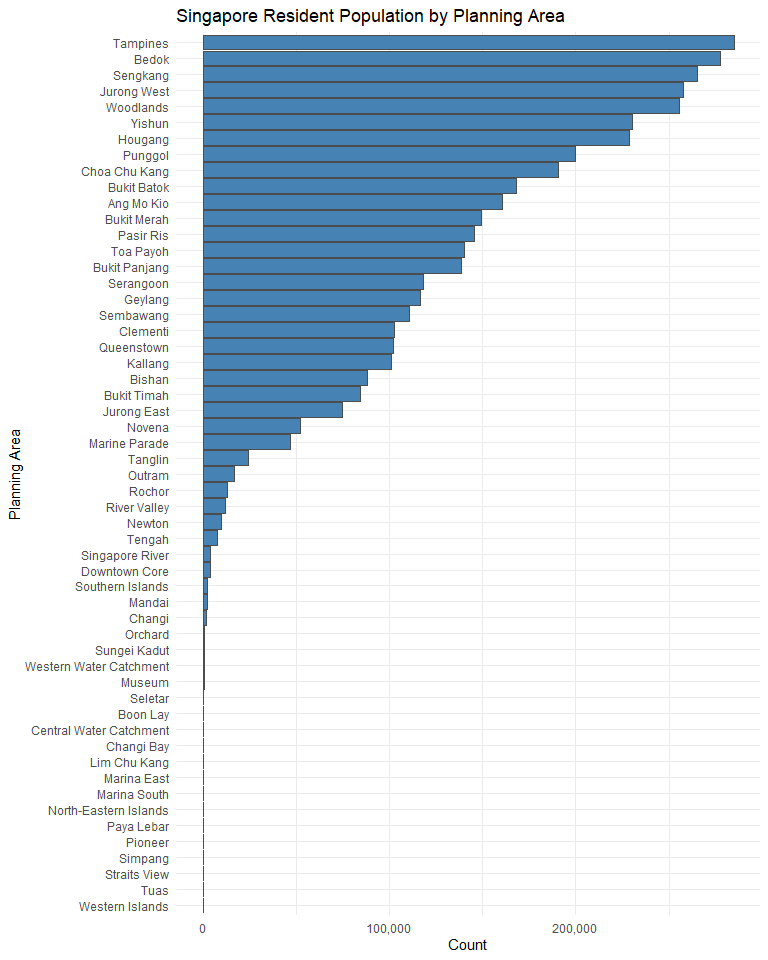
population\_data <- population\_data %>%  
 mutate(age\_group = cut(Age,   
 breaks = c(seq(0, 90, 5), Inf),   
 right = FALSE,   
 include.lowest = TRUE,  
 labels = c(paste(seq(0, 85, 5), seq(4, 89, 5), sep = "-"), "90+")),  
 Sex = factor(Sex, levels = c(0, 1), labels = c("Male", "Female")),  
 PopAdj = ifelse(Sex == "Male", -Pop, Pop)  
 )  
agg\_data <- population\_data %>%  
 group\_by(age\_group, Sex) %>%  
 summarise(Pop = sum(PopAdj), .groups = 'drop')  
ggplot(agg\_data, aes(x = age\_group, y = Pop, fill = Sex)) +  
 geom\_bar(stat = "identity") +  
 coord\_flip() +  
 labs(title = "Age-Sex Pyramid", x = "Age Group", y = "Population") +  
 scale\_y\_continuous(labels = abs)

|  |
| --- |
| Insight from the visualization |
| * Singapore has a higher economically active population (aged 15 – 65) than elderly (aged 65+) and young dependents (aged 0 – 14) in 2024. Thus, we can conclude that Singapore’s population has a stationary pyramid * Singapore also has roughly similar proportion of male to female population as well as similar age distribution between male and female. * The population also has a high number of elderly, which can indicate an advanced healthcare system and low mortality rate. |

## 3. 2 Population and Demographics of Residents Living in Each Planning Area

## Population of Residents

pop\_pa <- population\_data %>%  
 group\_by(PA) %>%  
 summarise(total\_pop = sum(Pop))  
pop\_pa$PA <- fct\_rev(factor(pop\_pa$PA, levels = pop\_pa$PA[order(pop\_pa$total\_pop, decreasing = TRUE)]))  
ggplot(data = pop\_pa, aes(x = PA, y = total\_pop)) +  
 geom\_col(color = "gray30", fill = "steelblue") +  
 labs(  
 title = "Singapore Resident Population by Planning Area",  
 x = "Planning Area",  
 y = "Count"  
 ) +  
 scale\_y\_continuous(labels = label\_comma())+ theme\_minimal() +  
 coord\_flip()



## Demographics of Residents Living in Each Planning Area

population\_binned <- population\_data %>%  
 mutate(age\_group = cut(  
 Age,  
 breaks = c(seq(0, 90, 5), Inf),   
 labels = c(paste(seq(0, 85, 5), seq(4, 89, 5), sep = "-"), "90+"),  
 right = FALSE,  
 include.lowest = TRUE  
 )) %>%  
 group\_by(PA, age\_group) %>%  
 summarise(total\_pop = sum(Pop), .groups = "drop")  
population\_binned\_wide <- population\_binned %>%  
 pivot\_wider(names\_from = age\_group, values\_from = total\_pop, values\_fill = 0)  
population\_binned\_wide %>%  
 kable(  
 caption = "Population by Planning Area and Age Group")

Population by Planning Area and Age Group

| PA | 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+ |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Ang Mo Kio | 4470 | 5950 | 6500 | 7420 | 7900 | 9310 | 10340 | 9730 | 10750 | 11280 | 12110 | 11720 | 12440 | 12010 | 11160 | 7970 | 5100 | 2760 | 1530 |
| Bedok | 9010 | 11490 | 12230 | 13370 | 14540 | 16510 | 18450 | 18310 | 19020 | 19840 | 20680 | 19890 | 21330 | 20560 | 17330 | 12220 | 6790 | 3870 | 2220 |
| Bishan | 2990 | 3800 | 4120 | 4450 | 4610 | 5640 | 6630 | 5910 | 5660 | 5920 | 6460 | 5970 | 6710 | 6600 | 5220 | 3450 | 2110 | 1210 | 750 |
| Boon Lay | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Bukit Batok | 8070 | 8250 | 7680 | 8340 | 9340 | 10690 | 13710 | 13670 | 11430 | 11390 | 12160 | 12270 | 12450 | 11400 | 8220 | 4700 | 2450 | 1380 | 790 |
| Bukit Merah | 5110 | 5730 | 6290 | 6410 | 6360 | 8020 | 9940 | 10590 | 10970 | 11390 | 11220 | 10320 | 10420 | 10390 | 9480 | 7350 | 4860 | 2830 | 1710 |
| Bukit Panjang | 5450 | 6350 | 6950 | 7510 | 8230 | 9730 | 10680 | 10190 | 9720 | 9260 | 10130 | 10120 | 10890 | 9390 | 6450 | 3850 | 2070 | 1160 | 600 |
| Bukit Timah | 3430 | 5720 | 5310 | 5100 | 4460 | 4340 | 4720 | 5730 | 6390 | 6560 | 6680 | 5430 | 5000 | 4570 | 3980 | 2990 | 1880 | 1240 | 720 |
| Central Water Catchment | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Changi | 130 | 150 | 120 | 100 | 80 | 90 | 130 | 180 | 150 | 160 | 140 | 100 | 100 | 80 | 40 | 30 | 10 | 20 | 20 |
| Changi Bay | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Choa Chu Kang | 7680 | 9030 | 9370 | 10450 | 12720 | 16260 | 15460 | 13160 | 12580 | 12270 | 14100 | 15970 | 14880 | 11210 | 7410 | 4070 | 2150 | 1210 | 660 |
| Clementi | 4110 | 5150 | 5210 | 5150 | 4910 | 5290 | 7110 | 7810 | 8220 | 8010 | 7670 | 6580 | 6330 | 6020 | 6060 | 4570 | 2600 | 1410 | 750 |
| Downtown Core | 90 | 90 | 70 | 90 | 110 | 270 | 560 | 640 | 540 | 460 | 390 | 300 | 190 | 160 | 110 | 50 | 0 | 10 | 40 |
| Geylang | 4670 | 4680 | 4470 | 4800 | 5080 | 7020 | 10770 | 9570 | 8350 | 7980 | 8020 | 7780 | 8530 | 8320 | 6650 | 4580 | 2820 | 1670 | 1060 |
| Hougang | 8080 | 10020 | 9960 | 10600 | 12010 | 15270 | 17470 | 17090 | 15980 | 15120 | 16300 | 16540 | 18400 | 17210 | 12640 | 7870 | 4470 | 2450 | 1300 |
| Jurong East | 1980 | 2540 | 2990 | 3540 | 4320 | 4890 | 4960 | 4740 | 4860 | 4990 | 5580 | 5530 | 5950 | 6130 | 5290 | 3290 | 1670 | 870 | 470 |
| Jurong West | 8850 | 11590 | 13400 | 15080 | 16270 | 17860 | 18230 | 17410 | 19430 | 20000 | 19960 | 19250 | 18110 | 16380 | 12550 | 7270 | 3670 | 1710 | 890 |
| Kallang | 3350 | 3990 | 4040 | 4080 | 4470 | 5500 | 7230 | 7400 | 7420 | 7590 | 7720 | 7080 | 7490 | 7220 | 6290 | 4580 | 2920 | 1730 | 1090 |
| Lim Chu Kang | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Mandai | 70 | 110 | 90 | 130 | 150 | 150 | 140 | 150 | 130 | 160 | 180 | 160 | 180 | 130 | 90 | 70 | 30 | 0 | 10 |
| Marina East | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Marina South | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Marine Parade | 1680 | 2440 | 2500 | 2320 | 2270 | 2310 | 2600 | 3110 | 3570 | 3810 | 3910 | 3250 | 3140 | 2770 | 2410 | 2020 | 1340 | 920 | 580 |
| Museum | 10 | 10 | 0 | 0 | 0 | 40 | 50 | 100 | 110 | 90 | 100 | 20 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Newton | 480 | 630 | 620 | 550 | 380 | 440 | 590 | 830 | 890 | 910 | 870 | 690 | 590 | 480 | 360 | 300 | 140 | 90 | 70 |
| North-Eastern Islands | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Novena | 1950 | 2950 | 2880 | 2540 | 2280 | 2830 | 3460 | 3950 | 4230 | 4360 | 4280 | 3520 | 3190 | 2900 | 2480 | 1840 | 1210 | 800 | 500 |
| Orchard | 0 | 10 | 40 | 50 | 20 | 10 | 70 | 60 | 120 | 120 | 130 | 60 | 70 | 40 | 70 | 0 | 0 | 0 | 0 |
| Outram | 410 | 440 | 700 | 650 | 510 | 870 | 1040 | 970 | 1180 | 1370 | 1240 | 1190 | 1350 | 1330 | 1260 | 940 | 590 | 380 | 320 |
| Pasir Ris | 4900 | 6370 | 7100 | 7910 | 9830 | 12360 | 10980 | 8900 | 9220 | 9780 | 10670 | 13090 | 12590 | 8980 | 5720 | 3390 | 1940 | 1080 | 670 |
| Paya Lebar | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Pioneer | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Punggol | 12650 | 15780 | 14210 | 11140 | 8950 | 9500 | 16160 | 20880 | 21610 | 17490 | 14080 | 10060 | 8540 | 6970 | 5370 | 3540 | 1700 | 840 | 410 |
| Queenstown | 3930 | 4130 | 4470 | 4440 | 4630 | 5870 | 7720 | 7520 | 7480 | 7680 | 7820 | 7020 | 6660 | 6010 | 5490 | 4620 | 3220 | 2140 | 1190 |
| River Valley | 470 | 700 | 660 | 660 | 560 | 530 | 750 | 910 | 1210 | 1220 | 1150 | 860 | 630 | 580 | 440 | 360 | 160 | 30 | 90 |
| Rochor | 400 | 390 | 420 | 510 | 520 | 770 | 910 | 1070 | 1020 | 1020 | 1040 | 880 | 900 | 910 | 860 | 600 | 400 | 200 | 220 |
| Seletar | 0 | 0 | 0 | 10 | 10 | 0 | 0 | 0 | 0 | 30 | 70 | 20 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Sembawang | 6220 | 6350 | 6160 | 6590 | 6780 | 7020 | 9560 | 9360 | 8710 | 8900 | 8810 | 7330 | 6500 | 5050 | 3330 | 1870 | 1180 | 600 | 350 |
| Sengkang | 13110 | 16530 | 16640 | 15290 | 14050 | 15640 | 20370 | 23410 | 24470 | 21810 | 19680 | 16110 | 15180 | 13250 | 9300 | 5560 | 2940 | 1470 | 740 |
| Serangoon | 3660 | 4540 | 5130 | 5750 | 6380 | 7920 | 8900 | 8050 | 7510 | 7680 | 8320 | 8280 | 9080 | 9370 | 7410 | 4870 | 2740 | 1690 | 970 |
| Simpang | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Singapore River | 240 | 340 | 250 | 210 | 160 | 150 | 260 | 430 | 470 | 450 | 420 | 260 | 220 | 140 | 100 | 80 | 20 | 0 | 0 |
| Southern Islands | 140 | 180 | 170 | 160 | 110 | 90 | 120 | 180 | 230 | 270 | 290 | 220 | 200 | 110 | 80 | 20 | 0 | 10 | 0 |
| Straits View | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Sungei Kadut | 10 | 0 | 50 | 0 | 30 | 60 | 40 | 40 | 30 | 40 | 80 | 30 | 60 | 50 | 40 | 20 | 0 | 10 | 10 |
| Tampines | 14650 | 14190 | 12560 | 12840 | 14340 | 20130 | 26990 | 23000 | 19710 | 17200 | 17440 | 19170 | 21740 | 20330 | 14550 | 8230 | 4250 | 2340 | 1290 |
| Tanglin | 1130 | 1600 | 1580 | 1280 | 1080 | 1160 | 1430 | 1880 | 2140 | 2150 | 2080 | 1650 | 1410 | 1170 | 1040 | 870 | 430 | 280 | 180 |
| Tengah | 850 | 300 | 290 | 230 | 200 | 890 | 1700 | 850 | 650 | 400 | 360 | 300 | 290 | 290 | 200 | 90 | 20 | 20 | 10 |
| Toa Payoh | 6390 | 6330 | 5740 | 5840 | 6020 | 7760 | 12270 | 11240 | 9620 | 9560 | 9830 | 9030 | 9520 | 9250 | 7730 | 5800 | 4130 | 2520 | 1450 |
| Tuas | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Western Islands | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Western Water Catchment | 10 | 40 | 50 | 20 | 120 | 40 | 30 | 60 | 80 | 50 | 10 | 20 | 50 | 10 | 0 | 0 | 0 | 0 | 0 |
| Woodlands | 10500 | 11850 | 12810 | 14940 | 18490 | 22000 | 20010 | 17490 | 16910 | 17480 | 19930 | 21040 | 18490 | 14130 | 9130 | 5240 | 2770 | 1550 | 750 |
| Yishun | 10070 | 11980 | 11450 | 11280 | 12150 | 15010 | 18710 | 18920 | 17980 | 15660 | 15840 | 15940 | 17570 | 14800 | 10790 | 6240 | 3350 | 1700 | 880 |

|  |
| --- |
| Demographics of Residents per Planning Area |
| * Most residents live in Tampines, Bedok, Sengkang, Jurong West, or Woodlands. With each planning area having more than 200,000 residents living in the planning area. This could be * There are 13 out of 55 planning areas no residents live in. * A lot of younger population (aged 0 – 14) lIve in these planning area: Tampines, Sengkang. Punggol, Woodlands, Yishun or Jurong West * The following planning areas have higher number of elderly who lives there: Bedok, Bukit Merah, Ang Mo Kio, Toa Payoh, Hougang |

# 4. References

1. How to Count Duplicates in R (With Examples)