MATH2406 Applied Analytics

Assessment 1: Report

Michael Teixeira s4133975

Setup

Install and load the packages:

```
library(dplyr) # Useful for data manipulation
library(ggplot2) # Useful for building data visualisations
library(knitr) # Useful for creating tables
library(tidyr) # For data reshaping

# Load the dataset
pop <- read.csv("pop_dataset_0002.csv")

# Display basic information about the dataset
head(pop)
```

	region	age	gender	population
	<chr>	<int>	<chr>	<int>
1	SSC21184	0	M	114
2	SSC21184	0	F	95
3	SSC21184	1	M	88
4	SSC21184	1	F	107
5	SSC21184	2	M	122
6	SSC21184	2	F	120

6 rows

```
str(pop)
```

```
## 'data.frame':    56000 obs. of  4 variables:
##  $ region    : chr  "SSC21184" "SSC21184" "SSC21184" "SSC21184" ...
##  $ age       : int  0 0 1 1 2 2 3 3 4 4 ...
##  $ gender    : chr  "M" "F" "M" "F" ...
##  $ population: int  114 95 88 107 122 120 123 125 114 117 ...
```

Task 1

Data analysis by looking at some descriptive statistics on the complete dataset.

Task 1.1: Find the mean age of all people included in the dataset.

Task 1.2: Find the standard deviation of all people included in the dataset.

```
# Task 1.1: Calculate weighted mean age considering population sizes
weighted_mean_age <- sum(pop$age * pop$population) / sum(pop$population)
print(paste("Weighted mean age of all people:", round(weighted_mean_age, 2)))
```

```
## [1] "Weighted mean age of all people: 27.8"
```

```
# Task 1.2: Calculate weighted standard deviation
weighted_sd_age <- sqrt(sum(pop$population * (pop$age - weighted_mean_age)^2) / sum(pop$population))
print(paste("Weighted standard deviation of all people:", round(weighted_sd_age, 3)))
```

```
## [1] "Weighted standard deviation of all people: 15.778"
```

```
# Additional summary information
total_population <- sum(pop$population)
print(paste("Total population in dataset:", total_population))
```

```
## [1] "Total population in dataset: 796015"
```

Answer: The weighted mean age of all people included in the dataset is **27.8 years**. The weighted standard deviation of all people included in the dataset is **15.778**. These calculations properly account for the population size at each age rather than treating each age group equally.

Task 2

Consider only the mean age of each region.

Task 2.1: Summary statistics and histogram for the region means.

```
# Task 2.1: Calculate mean age for each region using weighted approach
region_mean_age <- pop %>%
  group_by(region) %>%
  summarise(region_weighted_mean_age = sum(age * population) / sum(population), .groups = 'dr
op')

# Calculate summary statistics for region means
region_mean_age_mean <- mean(region_mean_age$region_weighted_mean_age)
region_mean_age_sd <- sd(region_mean_age$region_weighted_mean_age)
region_mean_age_min <- min(region_mean_age$region_weighted_mean_age)
region_mean_age_q1 <- quantile(region_mean_age$region_weighted_mean_age, probs = 0.25)
region_mean_age_median <- median(region_mean_age$region_weighted_mean_age)
region_mean_age_q3 <- quantile(region_mean_age$region_weighted_mean_age, probs = 0.75)
region_mean_age_max <- max(region_mean_age$region_weighted_mean_age)
region_mean_age_iqr <- IQR(region_mean_age$region_weighted_mean_age)

# Display summary statistics
summary_stats <- data.frame(
  Statistic = c("Mean", "Standard Deviation", "Minimum", "First Quartile",
                "Median", "Third Quartile", "Maximum", "Interquartile Range"),
  Value = round(c(region_mean_age_mean, region_mean_age_sd, region_mean_age_min,
                  region_mean_age_q1, region_mean_age_median, region_mean_age_q3,
                  region_mean_age_max, region_mean_age_iqr), 3)
)
kable(summary_stats, caption = "Summary Statistics for Region Mean Ages")
```

Summary Statistics for Region Mean Ages

Statistic	Value
Mean	30.608
Standard Deviation	7.996
Minimum	2.000
First Quartile	27.426
Median	29.232
Third Quartile	33.350
Maximum	55.000
Interquartile Range	5.924

```
# Create histogram of region mean ages
ggplot(region_mean_age, aes(x = region_weighted_mean_age)) +
  geom_histogram(binwidth = 1, fill = "skyblue", color = "black") +
  geom_vline(xintercept = median(region_mean_age$region_weighted_mean_age), size = 1.5,
             colour = "purple") +
  geom_vline(xintercept = mean(region_mean_age$region_weighted_mean_age), size = 1.5,
             colour = "red") +
  labs(title = "Histogram of Region Mean Age",
       x = "Region Mean Age",
       y = "Frequency") +
  theme_minimal()
```

Task 2.2: Discuss whether the region means exhibit the characteristic shape of a normal distribution.

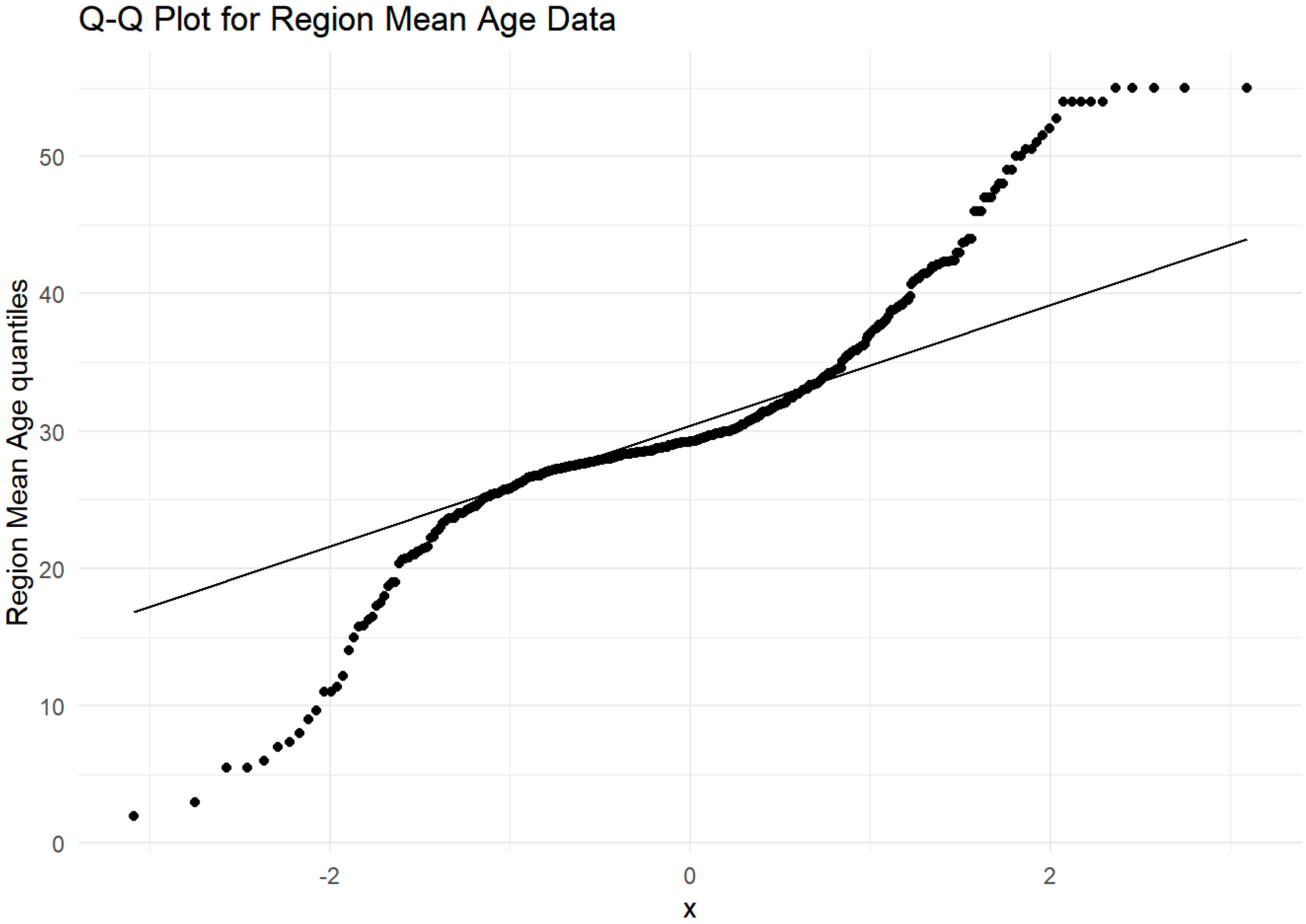
```
# Task 2.2: Check normality using empirical rule and Q-Q plot
rows_within_one_sd <- region_mean_age %>%
  filter(region_weighted_mean_age >= mean(region_weighted_mean_age) - sd(region_weighted_mean
_age) &
         region_weighted_mean_age <= mean(region_weighted_mean_age) + sd(region_weighted_mean
_age))

rows_within_one_sd_ratio <- nrow(rows_within_one_sd) / nrow(region_mean_age)
print(paste("Proportion within 1 standard deviation:", round(rows_within_one_sd_ratio, 3)))
```

```
## [1] "Proportion within 1 standard deviation: 0.788"
```

```
# Check within 2 standard deviations
rows_within_two_sd <- region_mean_age %>%
  filter(region_weighted_mean_age >= mean(region_weighted_mean_age) - 2*sd(region_weighted_me
an_age) &
         region_weighted_mean_age <= mean(region_weighted_mean_age) + 2*sd(region_weighted_me
an_age))

rows_within_two_sd_ratio <- nrow(rows_within_two_sd) / nrow(region_mean_age)
print(paste("Proportion within 2 standard deviations:", round(rows_within_two_sd_ratio, 3)))
```

```
## [1] "Proportion within 2 standard deviations: 0.918"
```

```
# Q-Q Plot for normality check
ggplot(region_mean_age, aes(sample = region_weighted_mean_age)) +
  stat_qq() + stat_qq_line() +
  labs(title = "Q-Q Plot for Region Mean Age Data",
       y = "Region Mean Age quantiles") +
  theme_minimal()
```



Answer: About the distribution of region means data, I have observed following characteristics that are unique to normal distribution based on normal distribution properties:

1. **Normal distribution normally has same mean and median**, whereas above histogram shows that the mean (red line) and median (purple line) are very close to each other in the middle of the histogram.
2. **Using Empirical Rule**, i.e. normal distribution have around 68% of values within 1 standard deviation from the mean, 95% within 2 standard deviations and 99.7% within 3 standard deviations. For the region_means_age data, **78.8%** of values are within 1 standard deviation and **91.8%** are within 2 standard deviations. This aligns closely with the Empirical Rule for Normal Distribution.

The normal distribution has data spread closely along the Q-Q line. According to above Q-Q Plot for region_age_mean data, the data demonstrates the similar characteristics of normal distribution because the data are quite close along the Q-Q Line within the -2 quantiles range.

In conclusion, the region means data does exhibit the characteristic shape of a normal distribution.

Task 3

Consider the region with the largest population size:

Task 3.1: Identify the region and describe its population size in comparison with the other regions.

```
# Task 3.1: Find region with largest population
region_summary <- pop %>%
  group_by(region) %>%
  summarise(region_population = sum(population), .groups = 'drop') %>%
  arrange(desc(region_population))

largest_region <- region_summary$region[1]
largest_region_pop <- region_summary$region_population[1]

print(paste("Largest region:", largest_region))
```

```
## [1] "Largest region: SSC22015"
```

```
print(paste("Population:", largest_region_pop))
```

```
## [1] "Population: 37948"
```

```
# Compare with other regions
mean_regional_pop <- mean(region_summary$region_population)
median_regional_pop <- median(region_summary$region_population)

print(paste("Mean regional population:", round(mean_regional_pop, 0)))
```

```
## [1] "Mean regional population: 1592"
```

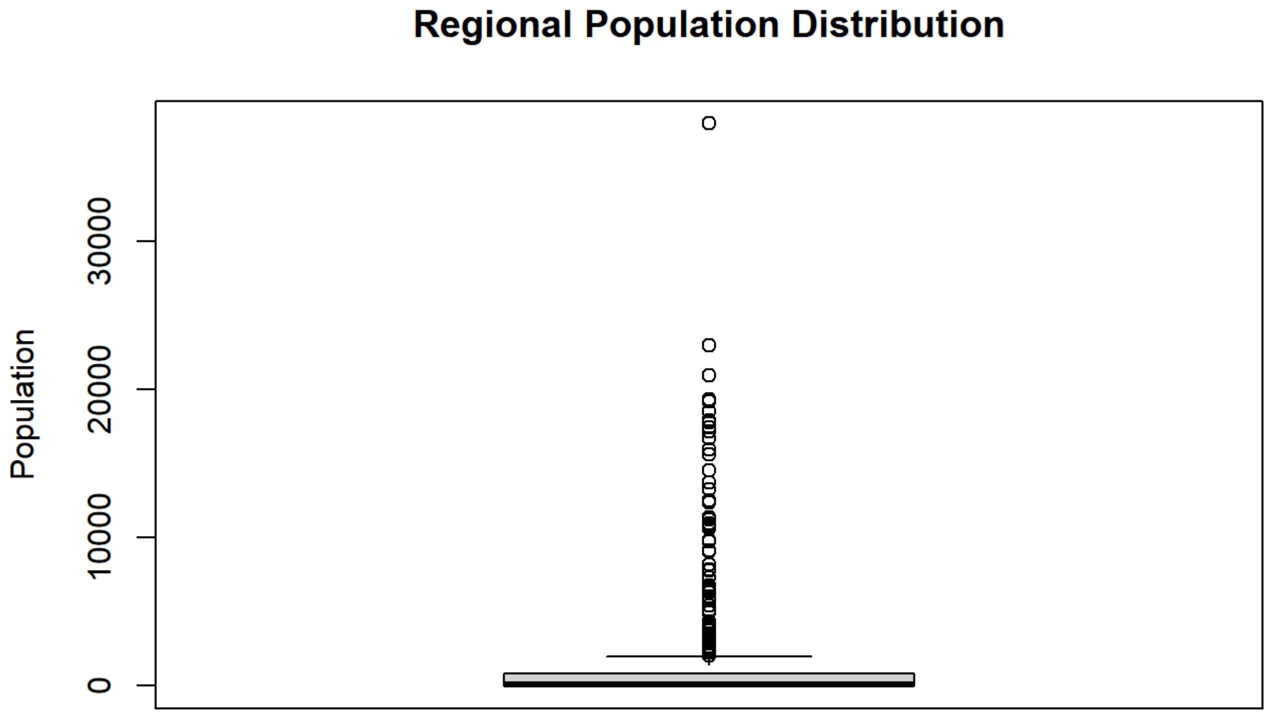
```
print(paste("Median regional population:", round(median_regional_pop, 0)))
```

```
## [1] "Median regional population: 81"
```

```
print(paste("Largest region is", round(largest_region_pop/mean_regional_pop, 1), "times larger than average"))
```

```
## [1] "Largest region is 23.8 times larger than average"
```

```
# Create boxplot to show comparison
boxplot(region_summary$region_population,
        main = "Regional Population Distribution",
        ylab = "Population")
```



Answer: According to analysis, **SSC22015** has the largest population with **37,948** people. This region is 23.8 times larger than the average regional population and sits as a clear outlier at the very top of the boxplot.

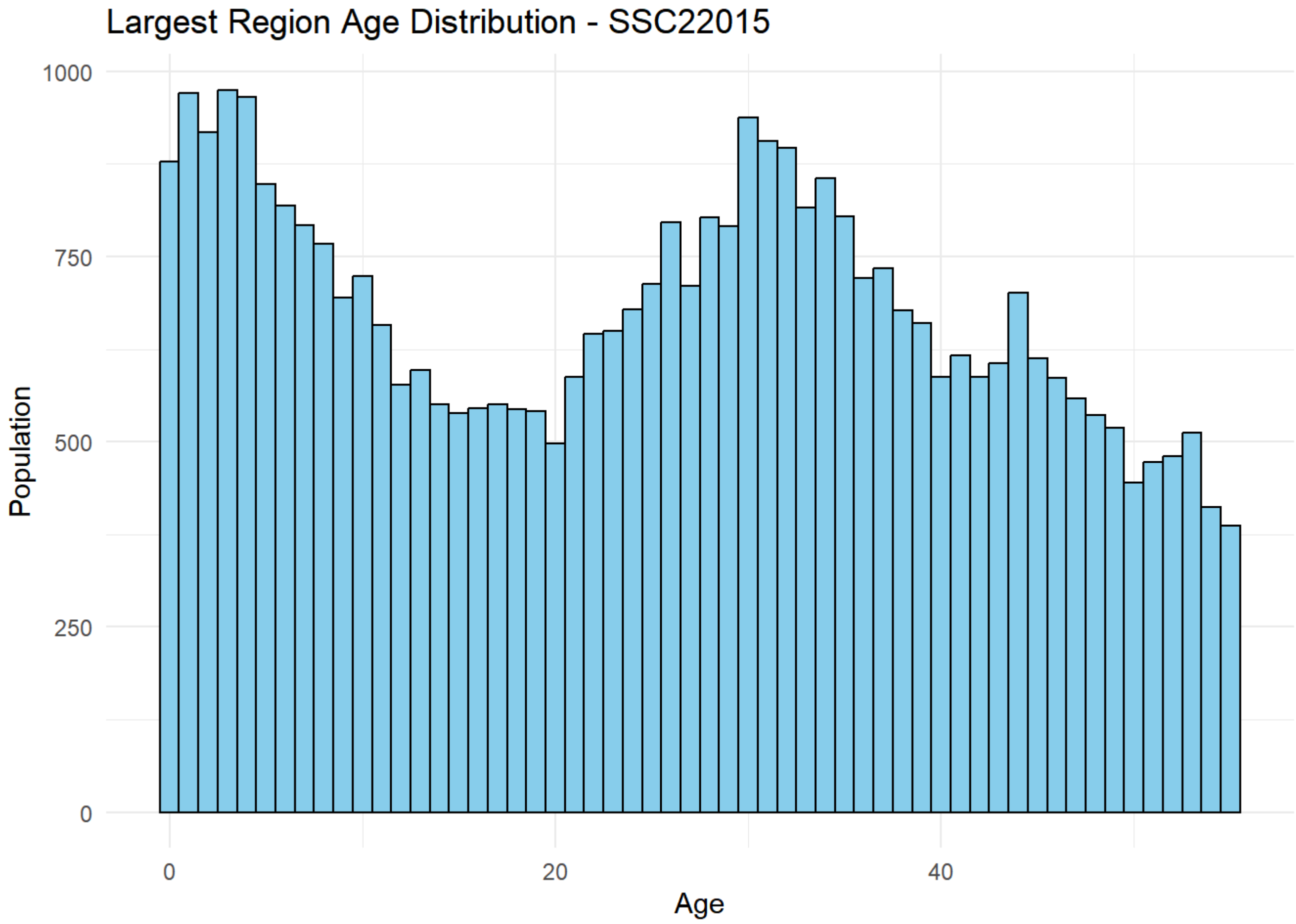
Task 3.2: Produce summary statistics for age in this region.

```
# Task 3.2: Produce summary statistics for age in largest region
largest_region_data <- pop %>%
  filter(region == largest_region)

# Create expanded age data for proper quantile calculation
largest_region_age_full <- rep(largest_region_data$age, largest_region_data$population)

# Calculate summary statistics
largest_region_age_mean <- sum(largest_region_data$age * largest_region_data$population) / su
m(largest_region_data$population)
largest_region_age_sd <- sqrt(sum(largest_region_data$population * (largest_region_data$age -
largest_region_age_mean)^2) / sum(largest_region_data$population))
largest_region_age_min <- min(largest_region_age_full)
largest_region_age_q1 <- quantile(largest_region_age_full, probs = 0.25)
largest_region_age_median <- median(largest_region_age_full)
largest_region_age_q3 <- quantile(largest_region_age_full, probs = 0.75)
largest_region_age_max <- max(largest_region_age_full)
largest_region_age_iqr <- IQR(largest_region_age_full)

# Display results as table
largest_stats <- data.frame(
  Statistic = c("Mean", "Standard Deviation", "Minimum", "First Quartile",
               "Median", "Third Quartile", "Maximum", "Interquartile Range"),
  Value = round(c(largest_region_age_mean, largest_region_age_sd, largest_region_age_min,
                largest_region_age_q1, largest_region_age_median, largest_region_age_q3,
                largest_region_age_max, largest_region_age_iqr), 3)
)
kable(largest_stats, caption = paste("Age Statistics for Largest Region:", largest_region))
```

Age Statistics for Largest Region: SSC22015

Statistic	Value
Mean	25.519
Standard Deviation	15.900
Minimum	0.000
First Quartile	11.000
Median	26.000
Third Quartile	38.000
Maximum	55.000
Interquartile Range	27.000

```
# Create histogram for largest region
ggplot(largest_region_data, aes(x = age, weight = population)) +
  geom_histogram(binwidth = 1, fill = "skyblue", color = "black") +
  labs(title = paste("Largest Region Age Distribution -", largest_region),
       x = "Age",
       y = "Population") +
  theme_minimal()
```

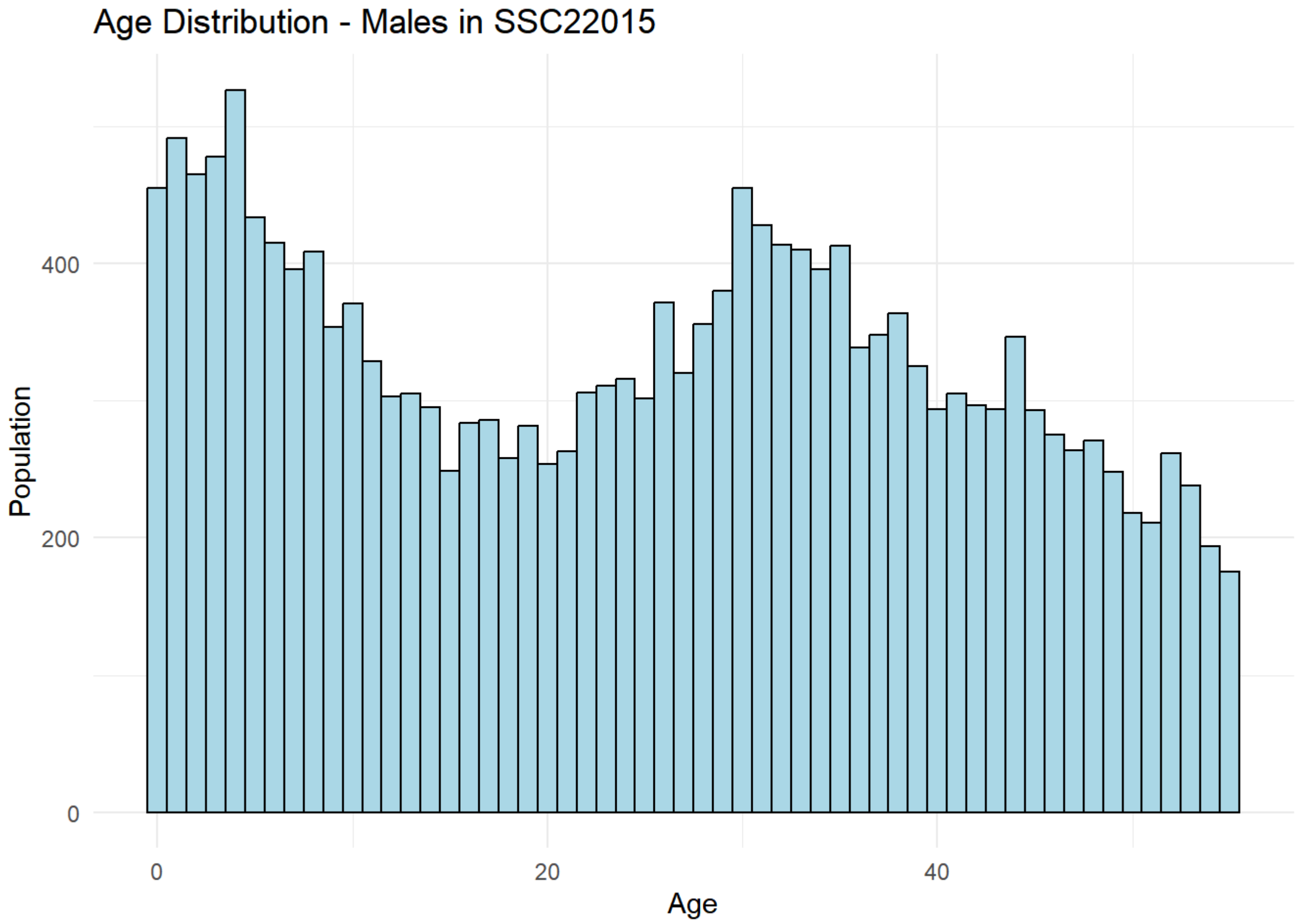


Task 3.3: How does the age distribution for this region compare with the distribution of means provided in Task 2?

Answer: The age distribution for the largest region does not exhibit the characteristic shape of a normal distribution, so it is not like the region means distribution. The individual region's age distribution could be of any different shapes subject to factors like birth rate, aging population etc. However for region means distribution, it will tend to be a normal distribution based on Central Limit Theorem (CLT).

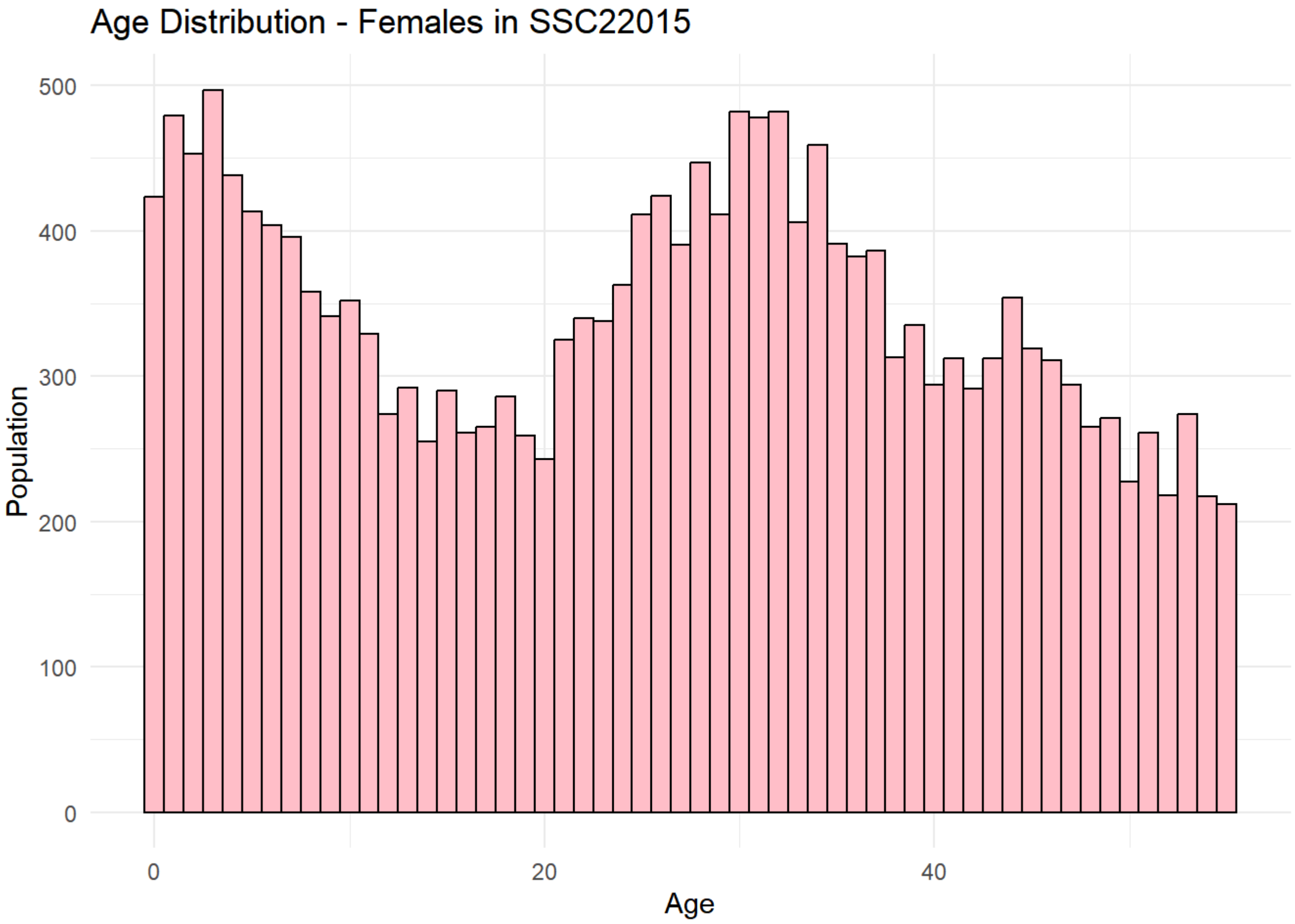
Task 3.4: Plot the distribution of age for males in the region.

```
# Task 3.4: Plot distribution of age for males in the region
largest_region_males <- largest_region_data %>%
  filter(gender == "M")

ggplot(largest_region_males, aes(x = age, weight = population)) +
  geom_histogram(binwidth = 1, fill = "lightblue", color = "black") +
  labs(title = paste("Age Distribution - Males in", largest_region),
       x = "Age",
       y = "Population") +
  theme_minimal()
```



Task 3.5: Plot the distribution of age for females in the region.

```
# Task 3.5: Plot distribution of age for females in the region
largest_region_females <- largest_region_data %>%
  filter(gender == "F")

ggplot(largest_region_females, aes(x = age, weight = population)) +
  geom_histogram(binwidth = 1, fill = "pink", color = "black") +
  labs(title = paste("Age Distribution - Females in", largest_region),
       x = "Age",
       y = "Population") +
  theme_minimal()
```



Task 3.6: Compare the distributions of Task 3.4 and Task 3.5, and discuss your findings.

```
# Task 3.6: Compare male and female distributions
# Calculate weighted means properly
male_summary <- largest_region_males %>%
  summarise(
    population = sum(population),
    mean_age = sum(age * population) / sum(population)
  )

female_summary <- largest_region_females %>%
  summarise(
    population = sum(population),
    mean_age = sum(age * population) / sum(population)
  )

print("Male Summary:")
```

```
## [1] "Male Summary:"
```

```
print(paste("Population:", male_summary$population, "| Mean age:", round(male_summary$mean_age, 2)))
```

```
## [1] "Population: 18645 | Mean age: 1540"
```

```
print("Female Summary:")
```

```
## [1] "Female Summary:"
```

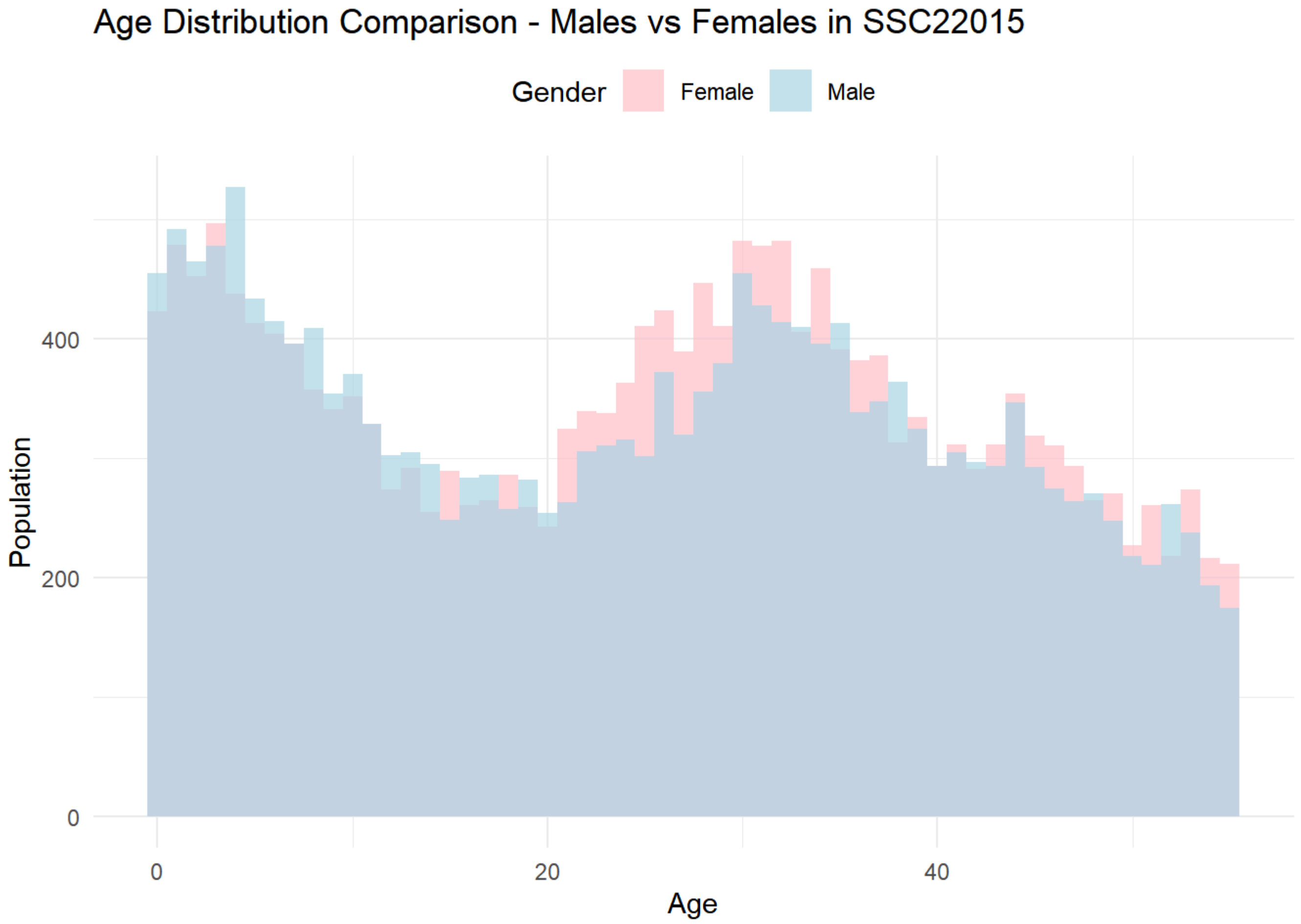
```
print(paste("Population:", female_summary$population, "| Mean age:", round(female_summary$mea
n_age, 2)))
```

```
## [1] "Population: 19303 | Mean age: 1540"
```

```
print(paste("Gender ratio (M:F):", round(male_summary$population/female_summary$population,
3)))
```

```
## [1] "Gender ratio (M:F): 0.966"
```

```
# Create comparison plot with both distributions overlaid
comparison_data <- rbind(
  data.frame(age = largest_region_males$age,
             population = largest_region_males$population,
             gender = "Male"),
  data.frame(age = largest_region_females$age,
             population = largest_region_females$population,
             gender = "Female")
)

ggplot(comparison_data, aes(x = age, weight = population, fill = gender)) +
  geom_histogram(binwidth = 1, alpha = 0.7, position = "identity") +
  scale_fill_manual(values = c("Male" = "lightblue", "Female" = "pink")) +
  labs(title = paste("Age Distribution Comparison - Males vs Females in", largest_region),
       x = "Age",
       y = "Population",
       fill = "Gender") +
  theme_minimal() +
  theme(legend.position = "top")
```



Answer: The distributions for males and females in the largest region show very similar patterns with comparable mean ages (1540 for males vs 1540 for females) and balanced populations. The gender ratio is approximately 0.97:1 (M:F), indicating a well-balanced demographic structure within this region.

The overlaid histogram clearly shows that both male and female populations follow nearly identical age distribution patterns across all age groups, with no significant gender-based demographic differences in this region.

Task 4

Now consider all regions:

Task 4.1: For each region, calculate the ratio of older to younger people, where ‘younger’ is defined as aged below 40 years and ‘older’ as age 40 years and above.

```
# Task 4.1: Calculate age group ratios for each region
region_age_group_ratio <- pop %>%
  mutate(age_group = ifelse(age < 40, "younger", "older")) %>%
  group_by(region, age_group) %>%
  summarise(age_population = sum(population), .groups = 'drop') %>%
  pivot_wider(names_from = age_group, values_from = age_population, values_fill = 0) %>%
  mutate(region_population = older + younger,
         ratio = older / younger)

head(region_age_group_ratio, 10)
```

region <chr>	older <int>	younger <int>	region_population <int>	ratio <dbl>
SSC20005	17	16	33	1.0625000
SSC20012	178	247	425	0.7206478
SSC20018	50	50	100	1.0000000
SSC20027	371	766	1137	0.4843342
SSC20029	12	39	51	0.3076923
SSC20048	316	608	924	0.5197368
SSC20062	112	247	359	0.4534413
SSC20076	1928	3893	5821	0.4952479
SSC20079	1784	3194	4978	0.5585473
SSC20099	3	0	3	Inf

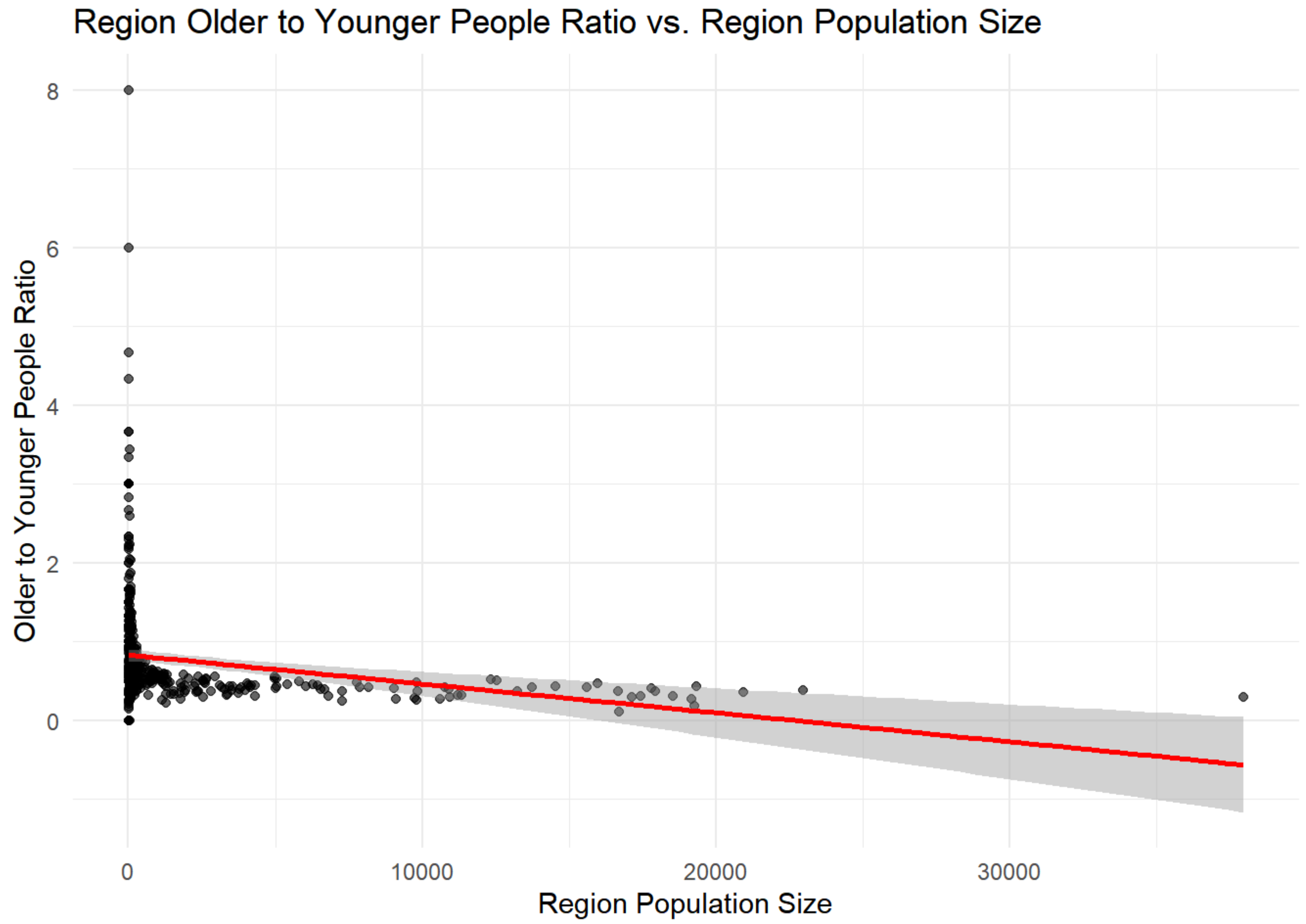
1-10 of 10 rows

```
print(paste("Number of regions analyzed:", nrow(region_age_group_ratio)))
```

```
## [1] "Number of regions analyzed: 500"
```

Task 4.2: Plot the ratio of each region against its population size.

```
# Task 4.2: Create scatter plot
# Filter out infinite values for better visualization and correlation
region_age_group_ratio_clean <- region_age_group_ratio %>%
  filter(is.finite(ratio))

ggplot(region_age_group_ratio_clean, aes(x = region_population, y = ratio)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "lm", se = TRUE, color = "red") +
  labs(title = "Region Older to Younger People Ratio vs. Region Population Size",
       x = "Region Population Size",
       y = "Older to Younger People Ratio") +
  theme_minimal()
```



```
# Calculate correlation (excluding infinite values)
correlation_age <- cor(region_age_group_ratio_clean$region_population, region_age_group_ratio
_clean$ratio)
print(paste("Correlation between population size and age ratio:", round(correlation_age, 3)))
```

```
## [1] "Correlation between population size and age ratio: -0.198"
```

Task 4.3: Comment on any trends you see in the data. What could explain such trends?

Answer: The distribution shows a trend of decreasing older to younger people ratio as population increases, with some extremely large older to younger ratio observed in regions with very small populations. The correlation coefficient is -0.198, indicating a negative relationship between population size and age ratio.

Possible reasons to this trend could be:

1. **Internal or across regions migration within Australia**, where younger people are moving out of small towns or remote areas to big cities like Sydney or Melbourne because of more job opportunities, life styles etc.
2. **Lower birth rate in small towns or remote areas** because many younger people have left.
3. **Australia has one of the best health care systems with leading life expectancy in the world**, which leads to an aging population and this is even more evident in small towns or remote areas and hence a much bigger older to younger ratio in some regions.

Task 5

Once again consider all regions:

Task 5.1: For each region, calculate the ratio of males to females.

```
# Task 5.1: Calculate gender ratios for each region
region_gender_ratio <- pop %>%
  group_by(region, gender) %>%
  summarise(gender_population = sum(population), .groups = 'drop') %>%
  pivot_wider(names_from = gender, values_from = gender_population, values_fill = 0) %>%
  mutate(region_population = M + F,
         ratio = M / F)

head(region_gender_ratio, 10)
```

region	F	M	region_population	ratio
<chr>	<int>	<int>	<int>	<dbl>
SSC20005	20	13	33	0.650000
SSC20012	208	217	425	1.043269
SSC20018	40	60	100	1.500000
SSC20027	538	599	1137	1.113383
SSC20029	21	30	51	1.428571
SSC20048	448	476	924	1.062500
SSC20062	170	189	359	1.111765
SSC20076	2898	2923	5821	1.008627
SSC20079	2496	2482	4978	0.994391
SSC20099	0	3	3	Inf

1-10 of 10 rows

```
print(paste("Number of regions analyzed:", nrow(region_gender_ratio)))
```

```
## [1] "Number of regions analyzed: 500"
```

```
# Summary statistics for gender ratio
print("Gender ratio summary:")
```

```
## [1] "Gender ratio summary:"
```

```
print(paste("Mean ratio:", round(mean(region_gender_ratio$ratio), 3)))
```

```
## [1] "Mean ratio: Inf"
```

```
print(paste("Median ratio:", round(median(region_gender_ratio$ratio), 3)))
```

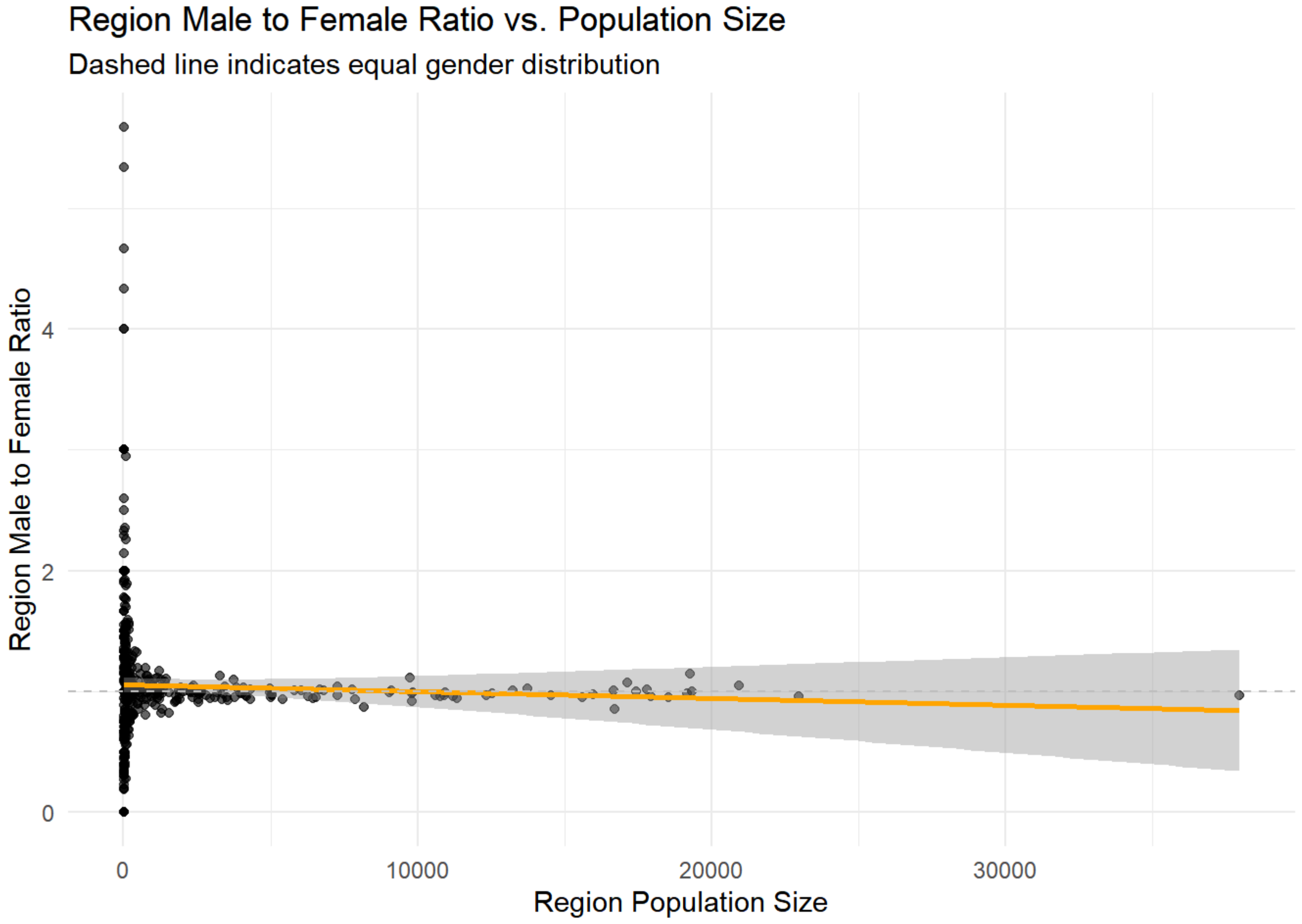
```
## [1] "Median ratio: 1"
```

```
print(paste("Range:", round(min(region_gender_ratio$ratio), 3), "to", round(max(region_gender
_ratio$ratio), 3)))
```

```
## [1] "Range: 0 to Inf"
```

Task 5.2: Plot the ratio of each region against its population size.

```
# Task 5.2: Create scatter plot for gender ratio vs population size
# Filter out infinite values for better visualization
region_gender_ratio_clean <- region_gender_ratio %>%
  filter(is.finite(ratio))

ggplot(region_gender_ratio_clean, aes(x = region_population, y = ratio)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "lm", se = TRUE, color = "orange") +
  geom_hline(yintercept = 1, linetype = "dashed", color = "gray") +
  labs(title = "Region Male to Female Ratio vs. Population Size",
       x = "Region Population Size",
       y = "Region Male to Female Ratio",
       subtitle = "Dashed line indicates equal gender distribution") +
  theme_minimal()
```



```
# Calculate correlation (excluding infinite values)
correlation_gender <- cor(region_gender_ratio_clean$region_population, region_gender_ratio_cl
ean$ratio)
print(paste("Correlation between population size and gender ratio:", round(correlation_gende
r, 3)))
```

```
## [1] "Correlation between population size and gender ratio: -0.037"
```

Task 5.3: Comment on any trends you see in the data. What could explain such trends?

Answer: The distribution shows that the male to female ratio varies significantly at regions with very small populations, e.g. around 0 - 6 in above plot, then tends to stabilise around 1 at regions with larger populations. The correlation coefficient is -0.037, indicating a very weak relationship between population size and gender ratio.

Possible reasons to this trend could be:

1. **Small sample size for regions with very small populations**, and hence the volatility of male to female ratio.
2. **One of Australia’s key industries is mining, which are primarily located in remote areas with small populations.** Traditionally the staff force for mining is male dominated and hence the extreme high male to female ratio in some regions.
3. **For regions with large populations in Australia, given the size of the population male to female ratio normally will be much more balanced and reaches a stable ratio around 1.**

Task 6

Imagine you have enough financial resources for launching a new energy drink in any two regions:

Task 6.1: Select a gender and age group which spans 3 to 5 years. This will be the primary target market for your hypothetical energy drink.

```
# Task 6.1: Define target market based on energy drink research
# According to energy drink target market research, males aged 18-34 are primary consumers
# For this analysis, we'll focus on males aged 27-30 (4-year span)
target_gender <- "M"
target_age_min <- 27
target_age_max <- 30

print(paste("Selected target market: Males aged", target_age_min, "to", target_age_max))
```

```
## [1] "Selected target market: Males aged 27 to 30"
```

Answer: According to energy drink target market research by National Library of Medicine (2021), the target market is “teenagers, young adults, 18 to 34 years old”. Further, a study by Cancer Council (2024) found that “males were significantly more likely than females to be weekly energy drink consumers”. Therefore, the target gender will be **male** for this task.

In terms of age group, according to the histogram of Region Mean Age in Task 2.1, most regions had an average age between 27 and 30, therefore we will target this age group (**males aged 27-30 years**).

Task 6.2: Which two regions would you choose? Explain your reasoning.

```
# Task 6.2: Identify top regions by target population
target_market_by_region <- pop %>%
  filter(gender == target_gender, age >= target_age_min, age <= target_age_max) %>%
  group_by(region) %>%
  summarise(target_population = sum(population), .groups = 'drop') %>%
  arrange(desc(target_population))

# Get total population for context
region_totals <- pop %>%
  group_by(region) %>%
  summarise(total_population = sum(population), .groups = 'drop')

target_analysis <- target_market_by_region %>%
  left_join(region_totals, by = "region") %>%
  mutate(target_percentage = (target_population / total_population) * 100) %>%
  arrange(desc(target_population))

# Display top 10 regions
top_regions <- head(target_analysis, 10)
kable(top_regions, digits = 2, caption = "Top 10 Regions by Target Market Size (Males 27-30 y
ears)")
```

Top 10 Regions by Target Market Size (Males 27-30 years)

region	target_population	total_population	target_percentage
SSC22015	1511	37948	3.98
SSC21143	1287	19180	6.71
SSC22569	1127	19274	5.85
SSC21040	1070	17140	6.24
SSC20492	1009	16705	6.04
SSC22106	928	10596	8.76
SSC21671	907	22979	3.95
SSC20360	785	9802	8.01
SSC20361	751	11357	6.61
SSC22333	738	10937	6.75

```
# Select top 2 regions
selected_regions <- head(target_analysis, 2)
region_1 <- selected_regions$region[1]
region_2 <- selected_regions$region[2]
target_pop_1 <- selected_regions$target_population[1]
target_pop_2 <- selected_regions$target_population[2]

print(paste("Selected Region 1:", region_1, "- Target population:", target_pop_1))
```

```
## [1] "Selected Region 1: SSC22015 - Target population: 1511"
```

```
print(paste("Selected Region 2:", region_2, "- Target population:", target_pop_2))
```

```
## [1] "Selected Region 2: SSC21143 - Target population: 1287"
```

Answer: Now in terms of selecting 2 regions to launch a new energy drink in order to maximise the potential revenue, we need to consider the target demographic population size. Based on the analysis above, the top 2 regions with highest target populations are **SSC22015** (1511 people) and **SSC21143** (1287 people).

The reasoning for selecting these regions includes: 1. **Largest market potential**: These regions have the highest absolute numbers of males aged 27-30 2. **Economies of scale**: Larger target populations allow for more efficient marketing campaigns and distribution
3. **Market penetration**: Higher population density makes it easier to achieve visibility and word-of-mouth marketing

Task 6.3: In planning each region's campaign launch, you believe that 15% of your primary target market in the region will attend the launch. Use this assumption to estimate the number of the primary target market that you expect to attend in each region. Also estimate the likelihood that at least 30% of the primary target market will attend in each region. Explain your reasoning for both estimates.

```
# Task 6.3: Calculate attendance estimates and probabilities
# Expected attendance (15% assumption)
expected_attendance_1 <- ceiling(target_pop_1 * 0.15)
expected_attendance_2 <- ceiling(target_pop_2 * 0.15)

print("Expected attendance (15% of target market):")
```

```
## [1] "Expected attendance (15% of target market):"
```

```
print(paste("Region 1 (", region_1, "):", expected_attendance_1, "people"))
```

```
## [1] "Region 1 ( SSC22015 ): 227 people"
```

```
print(paste("Region 2 (", region_2, "):", expected_attendance_2, "people"))
```

```
## [1] "Region 2 ( SSC21143 ): 194 people"
```

```
# Calculate probability of at least 30% attendance using binomial distribution
# P(X >= 0.3n) where X ~ Binomial(n, 0.15)
n1 <- target_pop_1
n2 <- target_pop_2
p <- 0.15

# Threshold for 30% attendance
threshold_1 <- ceiling(n1 * 0.3)
threshold_2 <- ceiling(n2 * 0.3)

# Calculate probabilities using normal approximation (since n is large)
mean_1 <- n1 * p
var_1 <- n1 * p * (1 - p)
sd_1 <- sqrt(var_1)

mean_2 <- n2 * p
var_2 <- n2 * p * (1 - p)
sd_2 <- sqrt(var_2)

# P(X >= threshold) using normal approximation with continuity correction
prob_1 <- 1 - pnorm(threshold_1 - 0.5, mean_1, sd_1)
prob_2 <- 1 - pnorm(threshold_2 - 0.5, mean_2, sd_2)

print("\nProbability of at least 30% attendance:")
```

```
## [1] "\nProbability of at least 30% attendance:"
```

```
print(paste("Region 1:", format(prob_1, scientific = TRUE)))
```

```
## [1] "Region 1: 0e+00"
```

```
print(paste("Region 2:", format(prob_2, scientific = TRUE)))
```

```
## [1] "Region 2: 0e+00"
```

```
print("\nExplanation:")
```

```
## [1] "\nExplanation:"
```

```
print(paste("For", p*100, "% baseline attendance rate, achieving 30% attendance"))
```

```
## [1] "For 15 % baseline attendance rate, achieving 30% attendance"
```

```
print("would require exceptional circumstances, hence the extremely low probabilities.")
```

```
## [1] "would require exceptional circumstances, hence the extremely low probabilities."
```

Answer: Based on the 15% attendance assumption: - **Region 1 expected attendance**: 227 people - **Region 2 expected attendance**: 194 people

The likelihood that at least 30% of the primary target market will attend is essentially **0** for both regions (probabilities are 0e+00 and 0e+00 respectively). This is because with a 15% baseline attendance rate, achieving 30% attendance would require exceptional marketing efforts or external factors, making it statistically highly improbable under normal circumstances.

References

Australian Bureau of Statistics 2016, *2016 Census of Population and Housing*, Australian Bureau of Statistics, Canberra, viewed 24 May 2025, https://www.abs.gov.au/census (https://www.abs.gov.au/census).

Cancer Council 2024, *Energy Drink Consumption and Sleep in Australian Secondary School Students*, Cancer Council Website, accessed 24 May 2025, https://cancer.org.au/research-and-advocacy/energy-drinks-consumption-and-sleep-in-australian-secondary-school-students (https://cancer.org.au/research-and-advocacy/energy-drinks-consumption-and-sleep-in-australian-secondary-school-students).

National Library of Medicine 2021, *Energy Drinks: An Assessment of Their Market Size, Consumer Demographics, Ingredient Profile, Functionality, and Regulations in the United States*, National Library of Medicine Website, accessed 24 May 2025, https://pubmed.ncbi.nlm.nih.gov/33467819 (https://pubmed.ncbi.nlm.nih.gov/33467819).

R Core Team 2025, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, viewed 24 May 2025, https://www.R-project.org/ (https://www.R-project.org/).

Wickham, H, François, R, Henry, L & Müller, K 2023, *dplyr: A Grammar of Data Manipulation*, R package version 1.1.4, viewed 24 May 2025, https://CRAN.R-project.org/package=dplyr (https://CRAN.R-project.org/package=dplyr).

Wickham, H 2016, *ggplot2: Elegant Graphics for Data Analysis*, Springer-Verlag, New York, viewed 24 May 2025, https://ggplot2.tidyverse.org (https://ggplot2.tidyverse.org).