MATH2406 Applied Analytics

Assessment 1: Report

Thaddeus Lee s3933533

Setup

```
library(dplyr) # Useful for data manipulation
library(ggplot2) # Useful for building data visualisations
library(knitr) # Useful for creating nice tables
library(magrittr)
library(ggplot2)
library(tinytex)
```

Question 1

1.1 Mean Age of People in Data Set. In calculating the **Mean**, we first calculate the total population in our data frame labelled **population_df**, by sub-setting the population variable and using the **sum** function. Using the **group_by** and **summarise** functions on the age variable, we then extract the age frequency for each age in the population of our data set. We assign this to a new data frame labelled **age_stats**.

we also create a new column by dividing the **Age Frequency** by **Total Population** which shows the proportion of each age among the total population.

Mean is then calculated by multiplying the sum of age frequencies and age proportion to arrive at a weighted population age mean of 27.8003.

1.2 Standard Deviation of People in Dataset. Using the information that we have created in our **age_stats** data frame, we can calculate the standard deviation age of population. We do this by first using the computationally efficient expression for calculating the variance which gives us 248.9465.

The **sqrt** function on the variance then gives the standard deviation of 15.7780.

```
################################################################################
#  Question 1
################################################################################

## Read CSV File
population_df <- read.csv("pop_dataset_0002.csv", header = TRUE)


# Dataframe Summary


population_df$region <- as.factor(population_df$region)
population_df$gender <- as.factor(population_df$gender)
```

```
#### Check ran for uniqueness

# unique(population_df$region)
# unique(population_df$gender)

#summary(population_df)
head(population_df)
```

```
##     region age gender population
## 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
```

```
################################################################################
#          1.1 - Mean age of all people in data set                            #
################################################################################

TotalPopulation <- sum(population_df$population)

age_stats <- population_df %>% group_by(age) %>%
  summarise(age_freq = sum(population, na.rm = TRUE),
            age_prop = age_freq/TotalPopulation) %>% ungroup()


# Mean
mu <- sum(age_stats$age * age_stats$age_prop)

################################################################################
#                   1.2 Standard Deviation                                     #
################################################################################

# Calculate the Variance
# We use the more computationally efficient expression for the variance:
sigma_sq <- sum(age_stats$age * age_stats$age * age_stats$age_prop) - mu^2
round(sigma_sq,3)
```

```
## [1] 248.946
```

```
# Standard Deviation
sd <- sqrt(sigma_sq)
round(sd,3)
```

```
## [1] 15.778
```

Question 2

Summary Statistics for Each Region In calculating Summary statistics of age for each region, we use the **aggregate** (*Data Science Made Simple, 2022*) function to generate a new data frame labelled **pop_df2**

which displays age and frequency of those ages per region. **Mutate** function is then used to create a column **REGION_POPULATION** which gives us a population total for each region.

We then use the **mutate** function to again create a **PROPORTION** column from age frequency divided by total population.

With Proportion calculated, we can then create a new column with the **mutate** function, and multiply the **AGE** column with the **PROPORTION** column. This will provide us with a colum with the mean age of each region. This enables use to calculate the Summary statistics based on weighted age means.

- Min = 2
- First Quartile = 27.42578
- Median = 29.23158
- Third Quartile = 33.35013
- Maximum = 55
- Mean = 27.80027
- Standard Deviation = 1.560479

Histogram - Distribution of Age Means Please refer to code chunk below.

Distribution of Region Means Looking at the Histogram plotted in the question above, it would appear that the distribution of Region Age Means is normally distributed given that the plot looks symmetrical..

In checking if the distribution of Region Age Means shares the characteristics of a normal distribution, we conducted 2 tests.

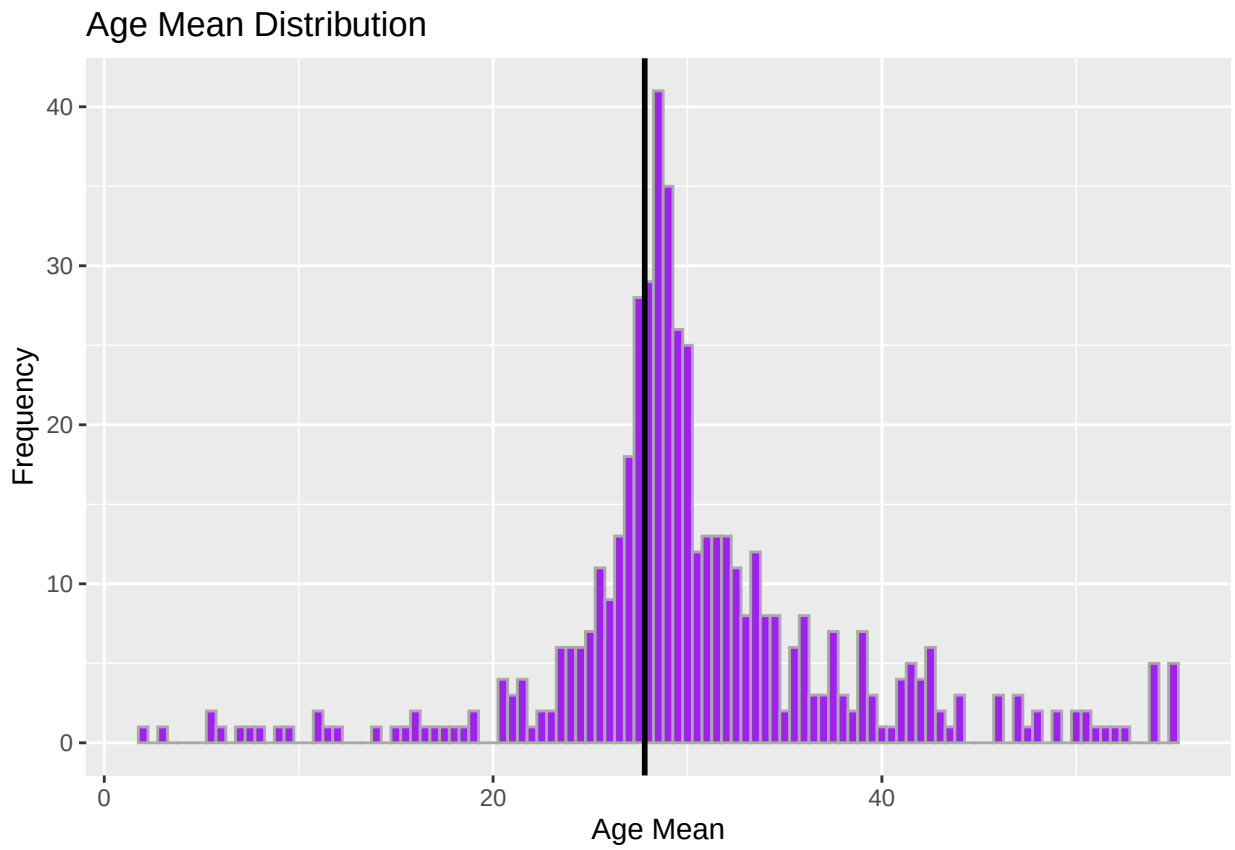
The first of which is generating a **QQ plot** from the distribution. Plotting the distribution on the QQ plot, we can observe that the points do not lie consistently on the straight line and bend out at each tail end and slightly in the middle. This tells us that the Distribution of Age Means, is not a normal distribution.

Our second test for normal distribution, is the **Shapiro-Wilk Normality test**. Using the test in the code below, we can observe a p value of 2.2e-16. Given that this p value is insignificantly small, smaller than 0.05 (*LAERD.com 2022*), we can see that the distribution of region age means is not normally distributed as the histogram would initially suggest.

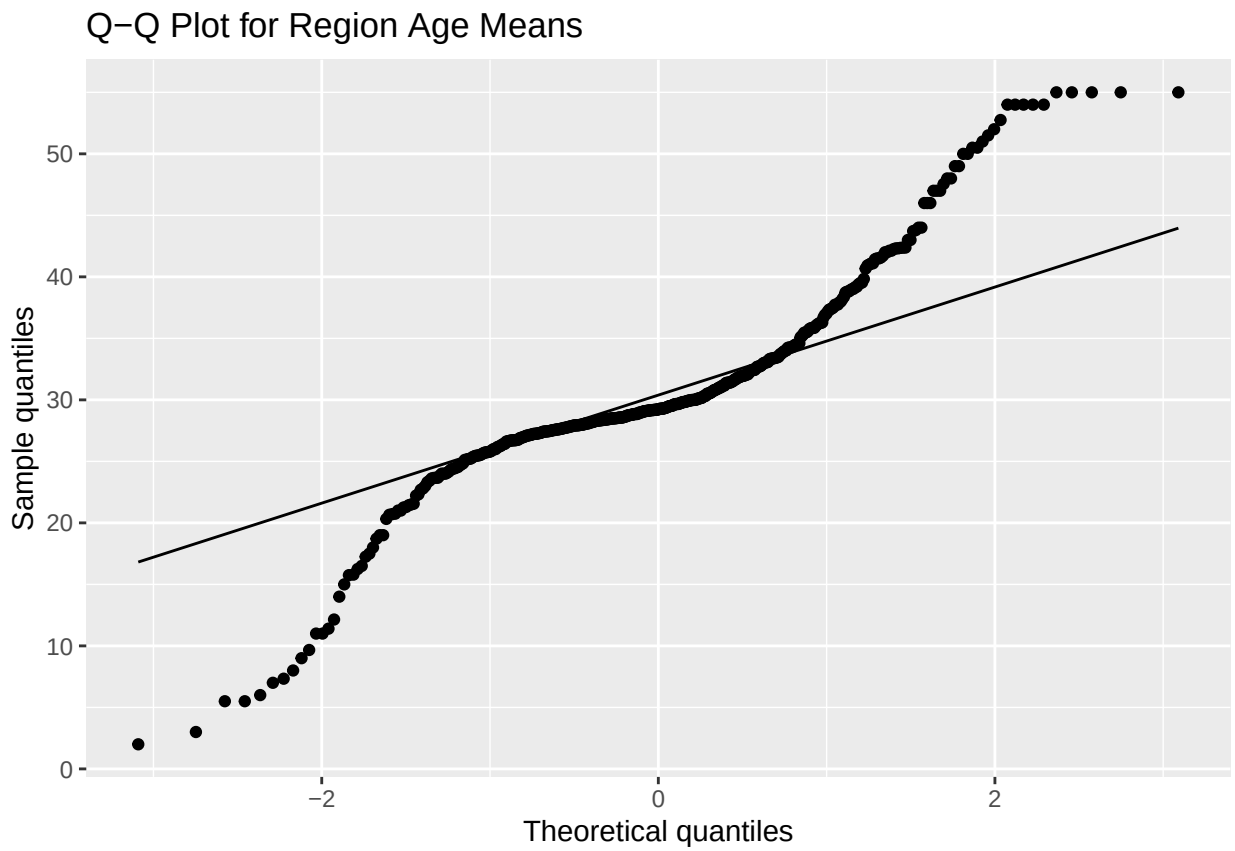
```
##############################################################################
#  Question 2
##############################################################################


# Creates dataframe that displays age frequency of male and female in each region.
pop_df2 <- aggregate(population_df$population,
                     by = list(population_df$region,
                               population_df$age), FUN=sum, na.rm = TRUE)


# Rename column names for pop_df2
pop_df2 <- rename(pop_df2, "REGION" = "Group.1")
pop_df2 <- rename(pop_df2, "AGE" = "Group.2")
pop_df2 <- rename(pop_df2, "AGE_FREQUENCY" = "x")

# arrange pop_df2 by REGION
pop_df2 <- arrange(pop_df2, REGION)
```

```r
# Mutate region population
pop_df2 %<>% group_by(REGION) %>% mutate(REGION_POPULATION = sum(AGE_FREQUENCY)) %>%
  ungroup()



# calculate proportion for region
pop_df2 %<>% group_by(REGION) %>% mutate(PROPORTION = AGE_FREQUENCY/REGION_POPULATION) %>%
  ungroup()


# caculate mean for each region
pop_df2 %<>% group_by(REGION) %>% mutate(REGION_MEAN = sum(AGE*PROPORTION)) %>%
  ungroup()

AGE_MEAN_BY_REGION <- select(pop_df2,
                             c("REGION", "REGION_MEAN", "REGION_POPULATION"))
AGE_MEAN_BY_REGION <- distinct(AGE_MEAN_BY_REGION)

AGE_MEAN_BY_REGION %<>% mutate(REGION_PROPORTION = REGION_POPULATION/TotalPopulation)


##############################################################################
#  **(WEIGHTED)**    2.1-2.8 - Summary Statistics - Region Age Means          #
##############################################################################
# Based of means of each region

WEIGHTED_SUMMARY_AGE_MEAN_BY_REGION <- AGE_MEAN_BY_REGION %>%
  summarise(Min = min(REGION_MEAN, na.rm = TRUE),
                  Q1 = quantile(REGION_MEAN, probs = 0.25, na.rm = TRUE),
                  Median = median(REGION_MEAN, na.rm = TRUE),
                  Q3 = quantile(REGION_MEAN, probs = 0.75, na.rm = TRUE),
                  Max = max(REGION_MEAN, na.rm = TRUE),
                  Mean = sum(REGION_MEAN * REGION_PROPORTION),
                  SD = sqrt(sum(REGION_MEAN * REGION_MEAN * REGION_PROPORTION) - Mean^2),
                  n = n(),
                  Missing = sum(is.na(REGION_MEAN)))




##############################################################################
#               2.9 - Histogram - Distribution of Age Region Means           #
##############################################################################
# x - axis displays the age means
# y - axis shows the frequency number of times it occurs


plot_AGE_MEAN_BY_REGION <- ggplot(AGE_MEAN_BY_REGION, aes(x = REGION_MEAN)) +
  geom_histogram(binwidth = 0.5, colour = "dark grey", fill = "purple") +
  geom_vline(aes(xintercept= sum(REGION_MEAN * REGION_PROPORTION)),
             color="black", size = 1) +
```

```
  labs(x = "Age Mean", y = "Frequency", title = "Age Mean Distribution")

plot_AGE_MEAN_BY_REGION
```




```
#https://github.com/rstudio/cheatsheets/blob/main/data-visualization-2.1.pdf

################################################################################
#                2.10 - Checking for Normal Distribution                       #
################################################################################


plot2_AGE_MEAN_BY_REGION <- ggplot(data.frame(AGE_MEAN_BY_REGION),
                                   aes(sample = REGION_MEAN)) + stat_qq() + stat_qq_line() +
  labs(x  = 'Theoretical quantiles', y = 'Sample quantiles',
       title = 'Q-Q Plot for Region Age Means')

plot2_AGE_MEAN_BY_REGION
```




```
#   Shapiro-Wilk normality test
z.scores <- AGE_MEAN_BY_REGION$REGION_MEAN
shapiro.test(z.scores %>% sample(500))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  z.scores %>% sample(500)
## W = 0.90793, p-value < 2.2e-16
```

Question 3

3.1 Identify the Region with the Largest Population Size In finding the region with the largest population size, we use a similar approach to question 2, but instead create a the data frame which we have labelled **REGION_POPULATION_TOTAL** to show population by region as well as the proportion of the region population to the total population.

From the code below, We then use the **max** function to return the highest population number for a region, we then determine the region with the highest population is **SSC22015** with a population of 37948.

In comparison to the other regions, looking at the population summary statistics provided from **POPULATION SUMMARY**, as below:

- Min = 3
- First Quartile = 21

- Median = 81
- Third Quartile = 797.5
- Max = 37948
- Mean = 11917.35
- Standard Deviation = 8747.124

When compared to the other regions in our data set, **SSC22015** is over three times more than the region population mean. The median population among other regions is 81, makeing **SSC22015** almost 470 times larger than the median region.

3.2 Provide Summary Statistics for the Region plot_AGE_MEAN_BY_REGION (Age) From the code below, we have calculated the summary statistics for region **SSC22015** as follows:

- Min = 0
- First Quartile = 13.75
- Median = 27.5
- Third Quartile = 41.25
- Max = 55
- Mean = 25.51882
- Standard Deviation = 15.8999

3.3 How does the age distribution for the region compare with the distribution of means provided in Q2? Firstly, it is notable to mention that the age of our population data is truncated to end at 55.

Looking at the distribution of ages for region **SSC22015**, we can observe that there is a high proportion of people in the region aged between 25 and 35, and 0 to 5.

When compared to the Distribution of Age Means we calculated and graphed earlier in question 2, we can similarly observe that a high frequency of age means is centered between the ages of 25 and 35. However, in the Distribution of Means, we can also observe that the age means decrease and flatten out at the younger and older ages.

That region **SSC22015** has a high amount of its population aged between 0 and 5, would suggest it has a higher frequency of people aged 0-5 compared to other regions. This is further supported when we compare the summary statistics of the Distribution of means to that of region **SSC22015**, where the age is 27.8 compared to **SSC22015** which is 25.5188.8, and the median of the mean age distribution is 27.8, compared to **SSC22015** which is 27.5.

3.4 Plot the distribution of age for males in the region. Please refer to code below.

3.5 Plot the distribution of age for females in the region. Please refer to code below.

3.6 Compare the Distributions of Q3.4 and Q3.5 and Discuss your Findings From our analysis of region **SSC22015**, the region has 18645 males, and 19303 females.

In comparing the age distribution between Men and Women of the region, we can observe that the pattern in the distribution of ages from the charts below, are very similar to each other, however the number of women in their 30s are slightly higher than the number of men in their 30s.

```
################################################################################
#  Question 3
################################################################################


population_df2 <- population_df
### Region with largest population size

# Finding region with largest population size
# Below code provides each region's population and proportion.
REGION_POPULATION_TOTAL <- population_df2 %<>%
  group_by(region) %>%
  summarise(POPULATION = sum(population),
            POPULATION_PROP = POPULATION/TotalPopulation) %>% ungroup()

max(REGION_POPULATION_TOTAL$POPULATION)
```

```
## [1] 37948
```

```
#returns 37948

LARGEST_REGION_POPULATION <- REGION_POPULATION_TOTAL %>% filter(POPULATION == 37948)
LARGEST_REGION_POPULATION
```

```
## # A tibble: 1 x 3
##   region   POPULATION POPULATION_PROP
##   <fct>         <int>           <dbl>
## 1 SSC22015      37948          0.0477
```

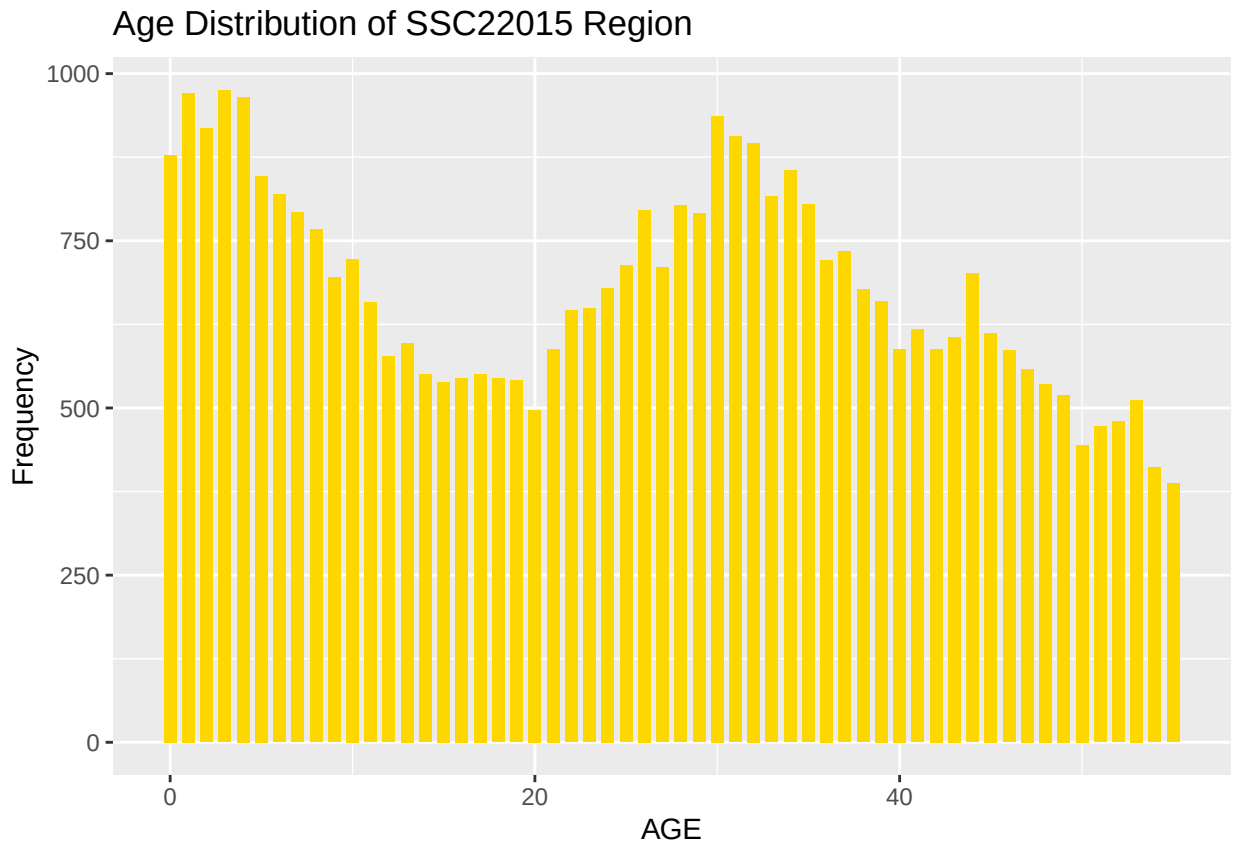
```
# Region with largest population size is SSC22015

SSC22015_Region <- population_df %>% filter(region == "SSC22015")
summary(SSC22015_Region)
```

```
##       region         age        gender   population
##  SSC22015:112   Min.   : 0.00   F:56   Min.   :175.0
##  SSC20005:  0   1st Qu.:13.75   M:56   1st Qu.:280.2
##  SSC20012:  0   Median :27.50          Median :325.0
##  SSC20018:  0   Mean   :27.50          Mean   :338.8
##  SSC20027:  0   3rd Qu.:41.25          3rd Qu.:404.5
##  SSC20029:  0   Max.   :55.00          Max.   :527.0
##  (Other) :  0
```

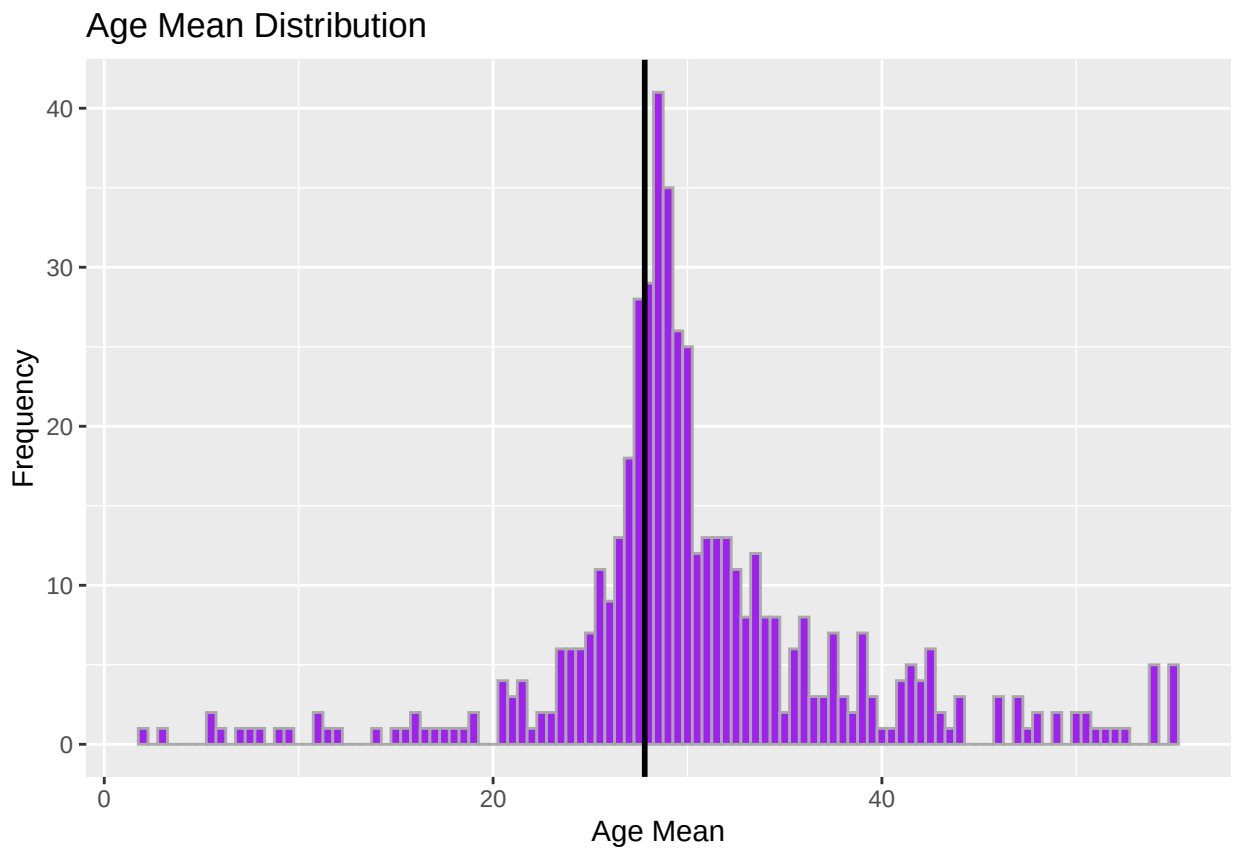
```
#Sum of total population
SSC22015_TOTAL <- sum(SSC22015_Region$population)
SSC22015_TOTAL
```

```
## [1] 37948
```

```r
##############################################################################
#                          3.1 POPULATION                                    #
##############################################################################


POPULATION_SUMMARY <- REGION_POPULATION_TOTAL %>%
  summarise(Min = min(POPULATION, na.rm = TRUE),
                  Q1 = quantile(POPULATION, probs = 0.25, na.rm = TRUE),
                  Median = median(POPULATION, na.rm = TRUE),
                  Q3 = quantile(POPULATION, probs = 0.75, na.rm = TRUE),
                  Max = max(POPULATION, na.rm = TRUE),
                  Mean = sum(POPULATION * POPULATION_PROP),
                  SD = sqrt(sum(POPULATION * POPULATION * POPULATION_PROP) - Mean^2),
                  n = n(),
                  Missing = sum(is.na(POPULATION)))

largest_region_compared_to_median <- 37948/81
##############################################################################
#                    3.2 summary stats for Region - AGE                      #
##############################################################################

#Age

# Create data frame that shows age, age frequency and age proportion for SSC22015 region
SSC22015_AGES <- SSC22015_Region %>%
  group_by(age) %>%
  summarise(age_frequency = sum(population, na.rm = TRUE),
            age_prop = age_frequency/SSC22015_TOTAL) %>% ungroup()

# Summary Statistics for Region SSC22015
AGE_SUMMARY_SSC22015 <- SSC22015_AGES %>%
  summarise(Min = min(age, na.rm = TRUE),
                  Q1 = quantile(age, probs = 0.25, na.rm = TRUE),
                  Median = median(age, na.rm = TRUE),
                  Q3 = quantile(age, probs = 0.75, na.rm = TRUE),
                  Max = max(age, na.rm = TRUE),
                  Mean = sum(age * age_prop),
                  SD = sqrt(sum(age * age * age_prop) - Mean^2),
                  n = n(),
                  Missing = sum(is.na(age)))


##############################################################################
#                  3.3 Plot of Age Distribution for Region                   #
##############################################################################
# Bar Graph
plot_SSC22015_AGES <- ggplot(SSC22015_AGES, aes(x = age, y = age_frequency)) +
  geom_bar(stat = "identity", width = .7, fill = "gold", size = 1) +
  labs(x = "AGE", y = "Frequency", title = "Age Distribution of SSC22015 Region")


plot_SSC22015_AGES
```



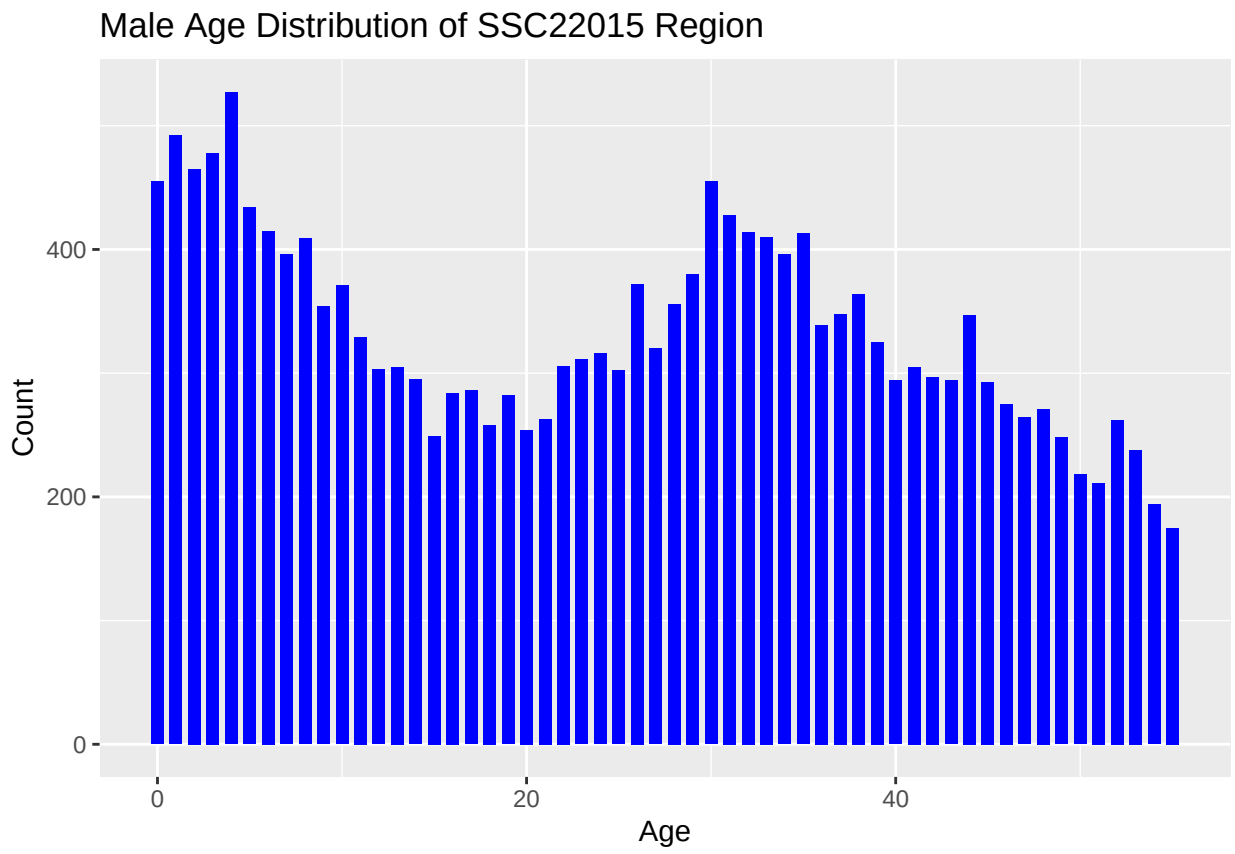

```
# Plot from Question 2.9
plot_AGE_MEAN_BY_REGION
```




```
##############################################################################
#             3.4 Plot the distribution of age for males in the region.      #
##############################################################################

# Distribution of Ages - Male
SSC22015_MALE_AGES <- population_df %>% filter(region == "SSC22015" & gender == "M")
sum(SSC22015_MALE_AGES$population)
```

```
## [1] 18645
```

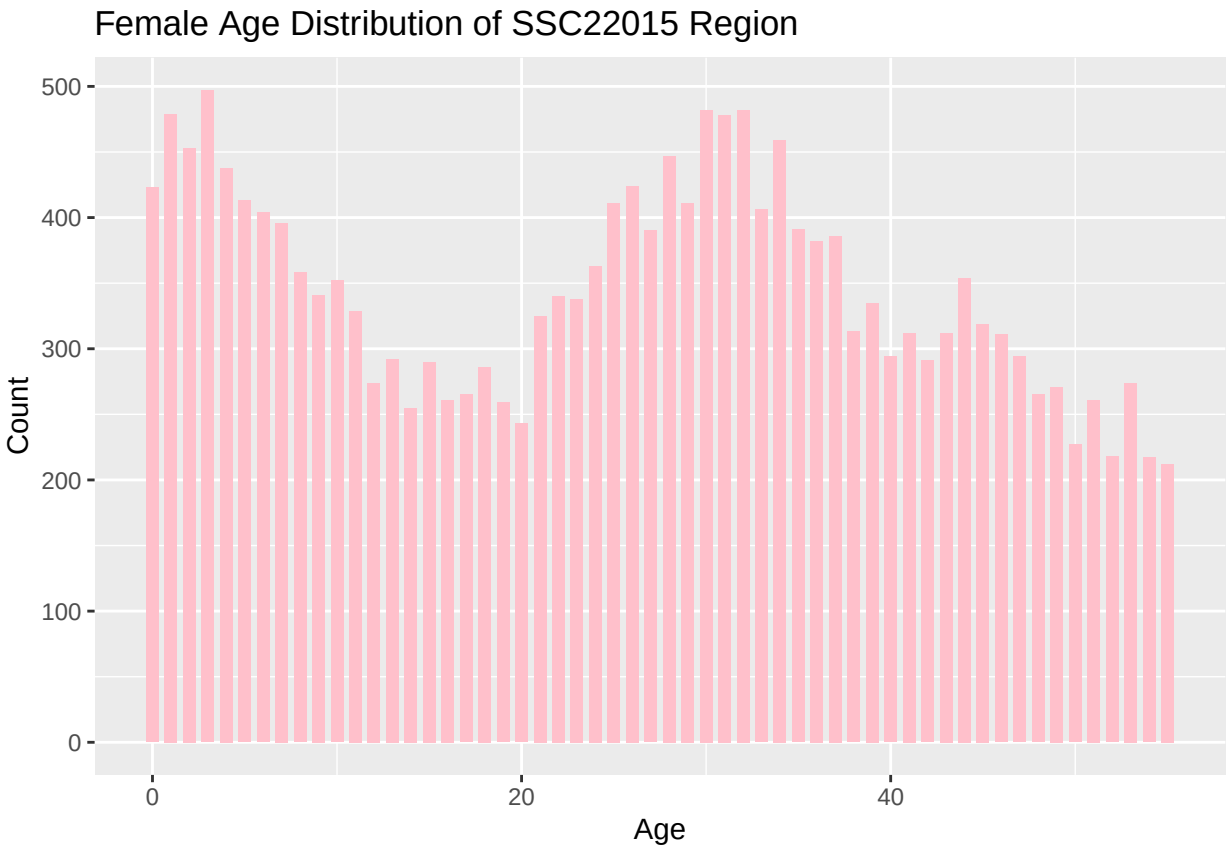
```
# create plots
plot_SSC22015_MALE_AGES <- ggplot(SSC22015_MALE_AGES, aes(x = age, y = population)) +
  geom_bar(stat = "identity", width = .7, fill = "blue", size = 1) +
  labs(x = "Age", y = "Count", title = "Male Age Distribution of SSC22015 Region")


plot_SSC22015_MALE_AGES
```




```
##############################################################################
#             3.5 Plot the distribution of age for females in the region     #
##############################################################################

# Distribution of Ages - Female
SSC22015_FEMALE_AGES <- population_df %>% filter(region == "SSC22015" & gender == "F")
sum(SSC22015_FEMALE_AGES$population)
```

```
## [1] 19303
```

```
# create plots
plot_SSC22015_FEMALE_AGES <- ggplot(SSC22015_FEMALE_AGES, aes(x = age, y = population)) +
  geom_bar(stat = "identity", width = .7, fill = "pink", size = 1) +
  labs(x = "Age", y = "Count", title = "Female Age Distribution of SSC22015 Region")


plot_SSC22015_FEMALE_AGES
```



Question 4

4.1 Plot the Ratio of Older to Younger People In the code below, we manipulate the population data frame to create two data frames. **OLDER_DF** a data frame with 3 variables that shows total population per region and a count of those who are aged 40+. The second data frame **YOUNGER_DF** is similar, however the 3rd variable is a count of those aged less than 40 in each region.

We then merge the two data frames into **AGE_DF** and we create a column for our ratio labelled **OLD_YOUNG_RATIO** which is the result of **COUNT_OLDER / COUNT YOUNGER**.

In providing an analysis of trends in the Old to Young ratio against the population, a scatter plot is used.

4.2 Comment on any trends you see in the data. What could explain such trends? The ratio calculated in **AGE_DF** above is a continuous, but the value for region population is discrete.

To analyse the trends in the ratio compared to the population we produce a scatter plot.

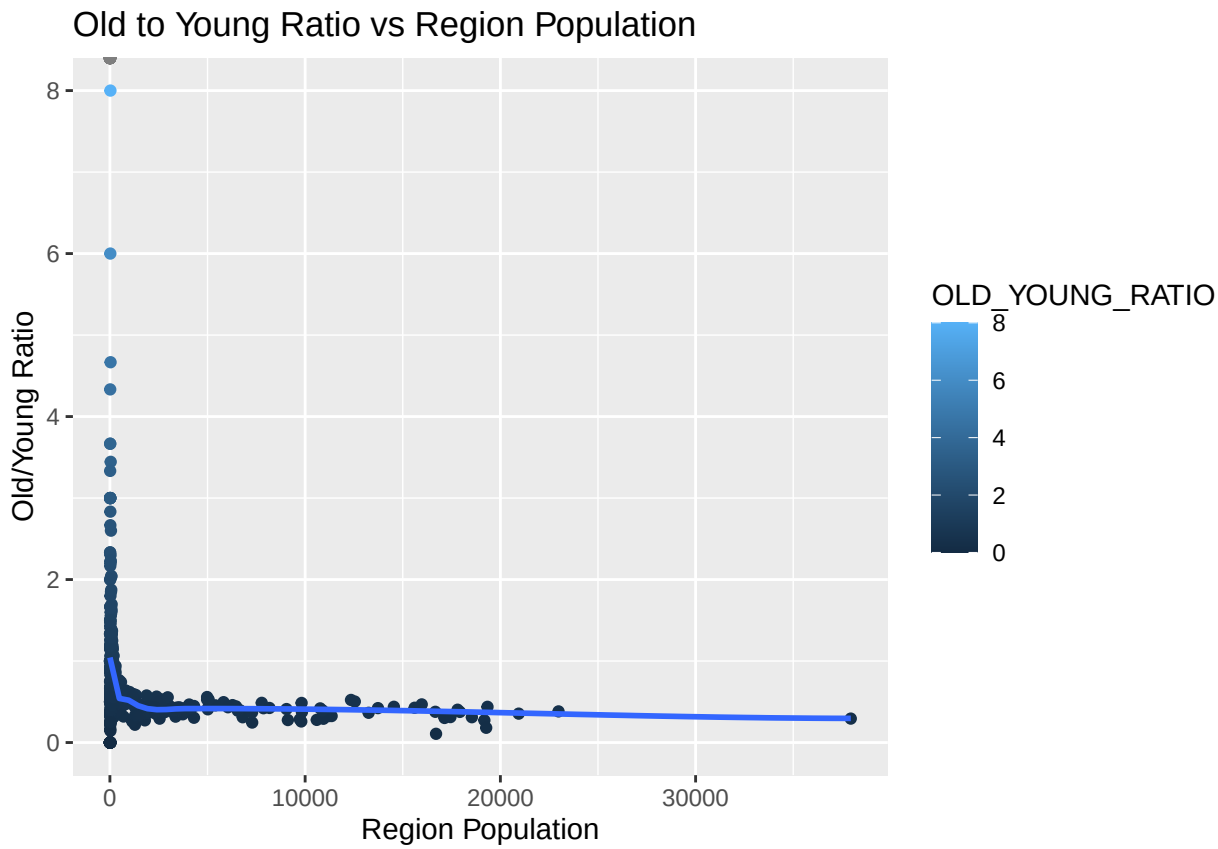
The scatter plot below shows a concentration of plots where region populations are lower. In the areas with lower populations we can observe higher number of plots where the "older to younger" ratio is higher. Some of the lower populated regions have the ratio as high as 6 and above.

As population size increases, we see that the "older to younger" ratio trends downwards. A hypothesis for the downward trend in the ratio as population size increases, could be higher rates of birth in higher populated regions compared to lower populated regions, and even migration from areas with lower populations to higher ones, which would affect the ratio. From the data, lower populations have aging populations.

Given the results from question 5 below, we also observe that regions with smaller population size tend to have a higher ratio of men to women which would also explain why these areas would have higher ratio of older to younger people.

```
##############################################################################
#              4.1 Plot the ratio of older to younger people                 #
##############################################################################

#create new data frame

pop_df3 <- aggregate(population_df$population,
                     by = list(population_df$region, population_df$age),
                     FUN=sum, na.rm = TRUE)

# Rename column names for pop_df2
pop_df3 <- rename(pop_df3, "REGION" = "Group.1")
pop_df3 <- rename(pop_df3, "AGE" = "Group.2")
pop_df3 <- rename(pop_df3, "AGE_FREQUENCY" = "x")

# Create column with age frequency
pop_df3 %<>% group_by(REGION) %>% mutate(REGION_POPULATION = sum(AGE_FREQUENCY)) %>%
  ungroup()


# Create data frame filtering people 40+ years
OLDER_df <- pop_df3 %>%  filter(AGE > 39)

# Create data frame filtering people below 40 years
YOUNGER_df <- pop_df3 %>%  filter(AGE <= 39)



##  mutate new column for counts
# OLDER
OLDER_df %<>% group_by(REGION) %>% mutate(COUNT_OLDER = sum(AGE_FREQUENCY)) %>%
  ungroup()

# YOUNGER
YOUNGER_df %<>% group_by(REGION) %>% mutate(COUNT_YOUNGER = sum(AGE_FREQUENCY)) %>%
  ungroup()


# Select / drop columns in new data frames
# OLDER
OLDER_df <- subset(OLDER_df, select = -c(AGE, AGE_FREQUENCY))
YOUNGER_df <- subset(YOUNGER_df, select = -c(AGE, REGION_POPULATION, AGE_FREQUENCY))

# Remove duplicate rows
OLDER_df <- OLDER_df %>%  distinct()
YOUNGER_df <- YOUNGER_df %>%  distinct()


# Merge OLDER and YOUNGER dataframes
AGE_DF <- merge(x = OLDER_df, y = YOUNGER_df, by = "REGION", all.x = TRUE)
```

```
AGE_DF %<>% mutate(OLD_YOUNG_RATIO = COUNT_OLDER/COUNT_YOUNGER) %>% ungroup()
##############################################################################
#  Plot - Old/ Young Ratio vs Region Polulation
##############################################################################
plot_AGE_DF <- ggplot(data = AGE_DF, aes(x = REGION_POPULATION, y = OLD_YOUNG_RATIO)) +
  geom_point(aes(col = OLD_YOUNG_RATIO)) + geom_smooth(method = "loess", se = F) +
  labs(x = "Region Population", y = "Old/Young Ratio", title = "Old to Young Ratio vs Region Population

plot_AGE_DF
```



Question 5

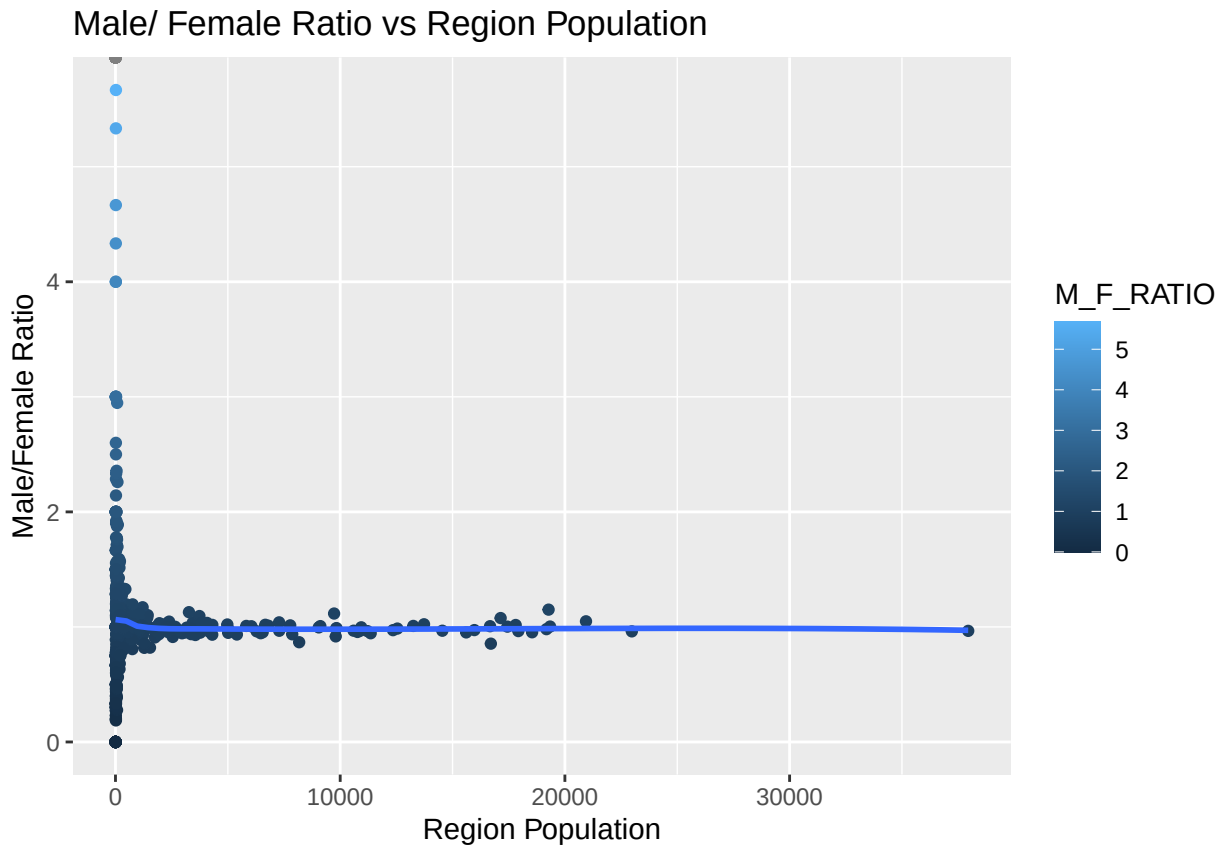
5.1 Plot the ratio against the population by region. Similar to our creation for ratios in question 4, we create a *MALE_DF **data frame that contains a count of males in each region and** FEMALE_DF **which contains a count for females in each region. We then merge both data frames together creating** GENDER_DF**, remove unnecessary columns and duplicates and create a new column,** M-F_RATIO **which is calculated by** COUNT_MALE / COUNT_FEMALE**. This can be seen in the code below.

5.2 Comment on any trends you see in the data. What could explain such trends? In analysing trends in **Male/ Female** ratio against region population, we create a scatter plot which can be seen in the code below.

From the scatter plot generated below, we can observe where population range up to 2000, that the ratio for the number of males compared to females can vary greatly. In regions with low populations the ratio of men compared to women can be 4 and above. As population sizes increase from 2000 on wards, we can see that the ratio is balanced around 1, which we can see by the straight blue line. Regions with an imbalance between the number of men compared to women and vice versa likely have lower birth rates which could be hypothetical reason for the trend. From question 4, we observed that regions with smaller populations have a higher ratio of older to younger people, which would also explain the imbalance men to women.

```
#############################################################################
#            5.1 Plot the ratio against the population by region.           #
#############################################################################
#create new data frame

pop_df4 <- aggregate(population_df$population,
                     by = list(population_df$region,
                               population_df$gender), FUN=sum, na.rm = TRUE)

# Rename column names for pop_df2
pop_df4 <- rename(pop_df4, "REGION" = "Group.1")
pop_df4 <- rename(pop_df4, "GENDER" = "Group.2")
pop_df4 <- rename(pop_df4, "GENDER_FREQUENCY" = "x")

# Create column with age frequency
pop_df4 %<>% group_by(REGION) %>% mutate(REGION_POPULATION = sum(GENDER_FREQUENCY)) %>%
  ungroup()


# Create data frame filtering people 40+ years
MALE_DF <- pop_df4 %>%  filter(GENDER == "M")

# Create data frame filtering people below 40 years
FEMALE_DF <- pop_df4 %>%  filter(GENDER == "F")



##  mutate new column for counts
# MALE
MALE_DF %<>% group_by(REGION) %>% mutate(COUNT_MALE = sum(GENDER_FREQUENCY)) %>%
  ungroup()

# FEMALE
FEMALE_DF %<>% group_by(REGION) %>% mutate(COUNT_FEMALE = sum(GENDER_FREQUENCY)) %>%
  ungroup()


# Select / drop columns in new data frames:
MALE_DF <- subset(MALE_DF, select = -c(GENDER, GENDER_FREQUENCY))
FEMALE_DF <- subset(FEMALE_DF, select = -c(GENDER, REGION_POPULATION, GENDER_FREQUENCY))

#Remove duplicates and only display distinct rows.
MALE_DF <- MALE_DF %>%  distinct()
FEMALE_DF <- FEMALE_DF %>%  distinct()
```

```
# Merge MALE and FEMALE data frames by region:
GENDER_DF <- merge(x = MALE_DF, y = FEMALE_DF, by = "REGION", all.x = TRUE)

GENDER_DF %<>% mutate(M_F_RATIO = COUNT_MALE/COUNT_FEMALE) %>% ungroup()


##############################################################################
#  Plot - Male/ Female Ratio vs Region Population
##############################################################################
plot_GENDER_DF <- ggplot(data = GENDER_DF, aes(x = REGION_POPULATION, y = M_F_RATIO)) +
  geom_point(aes(col = M_F_RATIO)) + geom_smooth(method = "loess", se = F) +
  labs(x = "Region Population", y = "Male/Female Ratio", title = "Male/ Female Ratio vs Region Populati

plot_GENDER_DF
```




```
##############################################################################
#              Find Infinite/ Missing values in GENDER_DF                    #
##############################################################################
infinite_count2 <- sum(is.infinite(GENDER_DF$M_F_RATIO))
```

Question 6

6.1 Select a gender and age group which spans 3 to 5 years In selecting the demographics for marketing a new energy drink, we looked at the main consumers for energy drinks.

According to the article "20 Must Know Energy Drink Statistics and Facts for 2022" on Med Alert Help (*Cikaric, D 2021*), men between the ages of 18 to 24 make up the majority of consumers of energy drinks. This age group of men account for 34% of sales. Consumers of energy drinks, enjoy the flavours that are made available, as well as the benefits of increased energy levels and physical and mental alertness. The energy drinks market is expected to grow at a compounded annual rate of 7.3% by 2026. Therefore, for our purposes, we will make our primary target Men between the ages 18 to 23.

6.2 Region Selection In selecting the two regions to launch our energy drink, we will find the regions with the greatest amount of people in our target demographic (Men between 18 - 23).

To do this, we use our original **populaiton_df** data frame and we filter the data frame for men between the ages of 18 and 23 and assign this to a new data frame labelled **edrink**.

We then create a new column named **primary_demographic** that contains the number of people in our demographic for each region.

Using the **arrange** function, We can observe that our two largest regions containing our targer market are:

- **SSC20492** - with 2311 people
- **SSC22015** - with 1363 people

6.3 Energy Drink Launch Attendance and Poisson Distribution As we are modelling the occurrence of discrete events over a specific period or space, we will use the Poisson Distribution. In our scenario, we will apply it to the number of people attending the energy drink launch.

It is expected that 15% of the target demographic from regions **SSC20492** and **SSC22015** will attend. From the code below we calculate the expected number of people, and therefore, assumed lamda values of each region:

- **Region SSC20492:** 347
- **Region SSC22015:** 204

Though we expect 15% of our target demographic will attend, we want to know the probability that 30% of our target demographic from each region will attend. For each region, 30% of our target demographic and thus x number of attendees is:

- **Region SSC20492:** 693
- **Region SSC22015:** 409

Using the **dpois** function in the code below, the likelihood that 30% of the target demographic from each region will attend is:

- **Region SSC20492:** 1.844746579840165e-60, or 0% rounded
- **Region SSC22015:** 5.87806273854965e-37, or 0% rounded

```
##############################################################################
#                          6.2 Region Selection                              #
##############################################################################

# Create new data frame.
# The below data frame contains data for our target demographic.
# Males between ages 18 and 22:
```

```r
edrink <- population_df %>% filter(gender == "M" & age > 17 & age <= 22)


# Create count of
edrink %<>% group_by(region) %>% mutate(primary_demographic = sum(population)) %>% ungroup()

edrink <- subset(edrink, select = -c(age, gender, population))
edrink <- distinct(edrink)
edrink %>% arrange(desc(primary_demographic))
```

```
## # A tibble: 500 x 2
##    region   primary_demographic
##    <fct>                  <int>
##  1 SSC20492                2311
##  2 SSC22015                1363
##  3 SSC21143                1256
##  4 SSC21125                1132
##  5 SSC21671                1019
##  6 SSC20865                 891
##  7 SSC20911                 823
##  8 SSC21178                 767
##  9 SSC21040                 757
## 10 SSC21743                 742
## # ... with 490 more rows
```

```r
# https://www.datanovia.com/en/lessons/reorder-data-frame-rows-in-r/
# We can see that our two largest regions are:
# SSC20492 - 2311 people
# SSC22015 - 1363 people


# 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 numbers
#of the primary target market that you expect to attend in each region. Also estimate the
#likelihood that 30% of the primary target market will attend in each region. Explain your
#reasoning for both estimates.

# SSC20492 - Calculate 15% expected attendance.
SSC20492_attend15 <- 2311 * .15
SSC20492_attend15 <- round(SSC20492_attend15,0)

# SSC22015 - Calculate 15% expected attendance.
SSC22015_attend15 <- 1363 * .15
SSC22015_attend15 <- round(SSC22015_attend15,0)


# Estimate that 30% of those in our target demographic well attend launch event.
SSC20492_attend30 <- 2311 * .30
SSC20492_attend30 <- round(SSC20492_attend30,0)


SSC22015_attend30 <- 1363 * .30
SSC22015_attend30 <- round(SSC22015_attend30,0)
```

```r
# DPOIS SSC20492 where P(X=693)
lambda_SSC20492 <- SSC20492_attend15 # Typical number of video streaming requests per second
x1 <-SSC20492_attend30

ans1 <- dpois(x1, lambda_SSC20492)
round(ans1, 3)
```

```
## [1] 0
```

```r
# DPOIS SSC22015 where P(X=409)
lambda_SSC22015 <- SSC22015_attend15 # Typical number of video streaming requests per second
x2 <-SSC22015_attend30

ans2 <- dpois(x2, lambda_SSC22015)
round(ans1, 3)
```

```
## [1] 0
```

```r
citation("tinytex")
```

```
##
## To cite the 'tinytex' package in publications use:
##
##   Yihui Xie (2022). tinytex: Helper Functions to Install and Maintain
##   TeX Live, and Compile LaTeX Documents. R package version 0.38.
##
##   Yihui Xie (2019) TinyTeX: A lightweight, cross-platform, and
##   easy-to-maintain LaTeX distribution based on TeX Live. TUGboat 40
##   (1): 30--32. https://tug.org/TUGboat/Contents/contents40-1.html
##
## To see these entries in BibTeX format, use 'print(<citation>,
## bibtex=TRUE)', 'toBibtex(.)', or set
## 'options(citation.bibtex.max=999)'.
```

References:

datasciencemadesimple.com 2022, *Aggregate() Function in R*, Data Science Made Simple, viewed 20 May 2022, https://www.datasciencemadesimple.com/aggregate-function-in-r/

LAERD Statistics, 2022, *Testing for Normality using SPSS Statistics*, LAERD Statistics, viewed 21 May 2022, https://statistics.laerd.com/spss-tutorials/testing-for-normality-using-spss-statistics.php

Cikaric, D, 2021, *20 Must-Know Energy Drink Statistics and Facts for 2022* Med Alert Help, viewed 2021 May 2002, https://medalerthelp.org/blog/energy-drink-statistics/

Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2022). dplyr: A Grammar of Data Manipulation. R package version 1.0.8. https://CRAN.R-project.org/package=dplyr

H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

Yihui Xie (2021). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.37.

Yihui Xie (2015) Dynamic Documents with R and knitr. 2nd edition. Chapman and Hall/CRC. ISBN 978-1498716963

Yihui Xie (2014) knitr: A Comprehensive Tool for Reproducible Research in R. In Victoria Stodden, Friedrich Leisch and Roger D. Peng, editors, Implementing Reproducible Computational Research. Chapman and Hall/CRC. ISBN 978-1466561595

Stefan Milton Bache and Hadley Wickham (2022). magrittr: A Forward-Pipe Operator for R. R package version 2.0.2. https://CRAN.R-project.org/package=magrittr

H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

Yihui Xie (2022). tinytex: Helper Functions to Install and Maintain TeX Live, and Compile LaTeX Documents. R package version 0.38.

Yihui Xie (2019) TinyTeX: A lightweight, cross-platform, and easy-to-maintain LaTeX distribution based on TeX Live. TUGboat 40 (1): 30–32. https://tug.org/TUGboat/Contents/contents40-1.html