COMP4010 - Homework 2

2024-03-31

Homework 2

```
library(tidyverse)
```

```
## Warning: package 'ggplot2' was built under R version 4.3.3
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.0     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(ggridges)
```

```
## Warning: package 'ggridges' was built under R version 4.3.3
```

```
library(showtext)
```

```
## Warning: package 'showtext' was built under R version 4.3.3
```

```
## Loading required package: sysfonts
```

```
## Warning: package 'sysfonts' was built under R version 4.3.3
```

```
## Loading required package: showtextdb
```

```
## Warning: package 'showtextdb' was built under R version 4.3.3
```

```
library(scales)
```

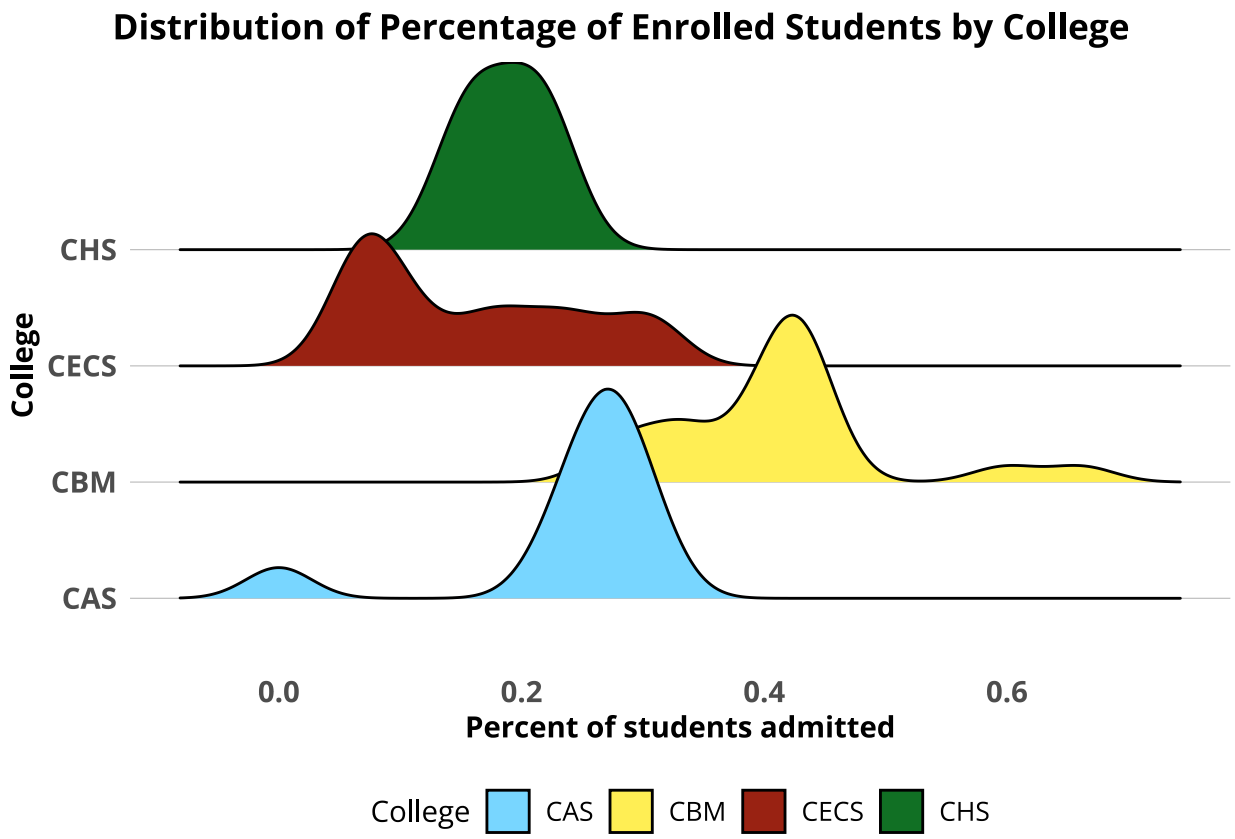
```
##
## Attaching package: 'scales'
##
## The following object is masked from 'package:purrr':
##
##     discard
##
## The following object is masked from 'package:readr':
##
##     col_factor
```

```r
font_add_google("Open Sans")
showtext_auto()
main_font = "Open Sans"

college_palette <- c("#78d6ff", "#ffee54", "#992212", "#117024")

theme_vinuni <- function(base_size = 11, base_family = main_font,
                         base_line_size = base_size / 22,
                         base_rect_size = base_size / 22){
  theme_minimal(
    base_family = base_family,
    base_size = base_size,
    base_line_size = base_line_size,
    base_rect_size = base_rect_size
  ) +
  theme(
    plot.title.position = "plot",
    plot.title = element_text(hjust = 0.5, face="bold"),
    plot.subtitle = element_text(hjust = 0.5, face="bold"),
    legend.position = "bottom",
    panel.grid.major.x = element_blank(),
    panel.grid.minor.x = element_blank(),
    panel.grid.major.y = element_line(color = "#bbbbbb", linewidth = 0.2),
    panel.grid.minor.y = element_blank(),
    axis.title = element_text(face = "bold"),
    axis.text = element_text(size = rel(1.0), face = "bold"),
    legend.text = element_text(size = rel(0.9)),
    scale_fill_manual(values = college_palette)
  )
}
```

Task 1.

```r
df <- read.csv("college_data_normalized.csv")

ggplot(df, aes(x = pct, y = college, fill = college)) + geom_density_ridges() +
  labs(x = "Percent of students admitted", y = "College",
  title = "Distribution of Percentage of Enrolled Students by College") +
  scale_fill_manual(values = college_palette, name = "College") +
  theme_vinuni()
```

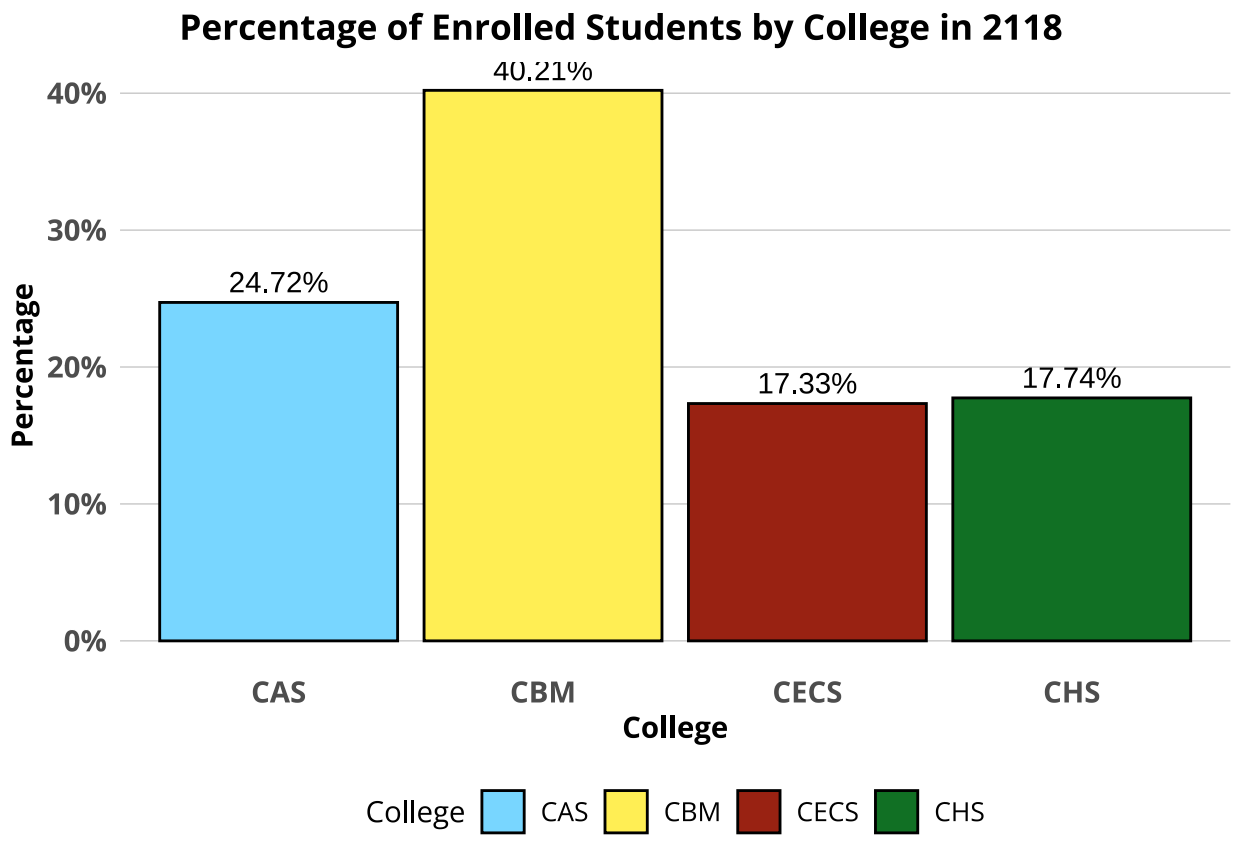
```
## Picking joint bandwidth of 0.0271
```



CECS often admits the smallest percentage of candidates, while CBM oftens admits the largest.

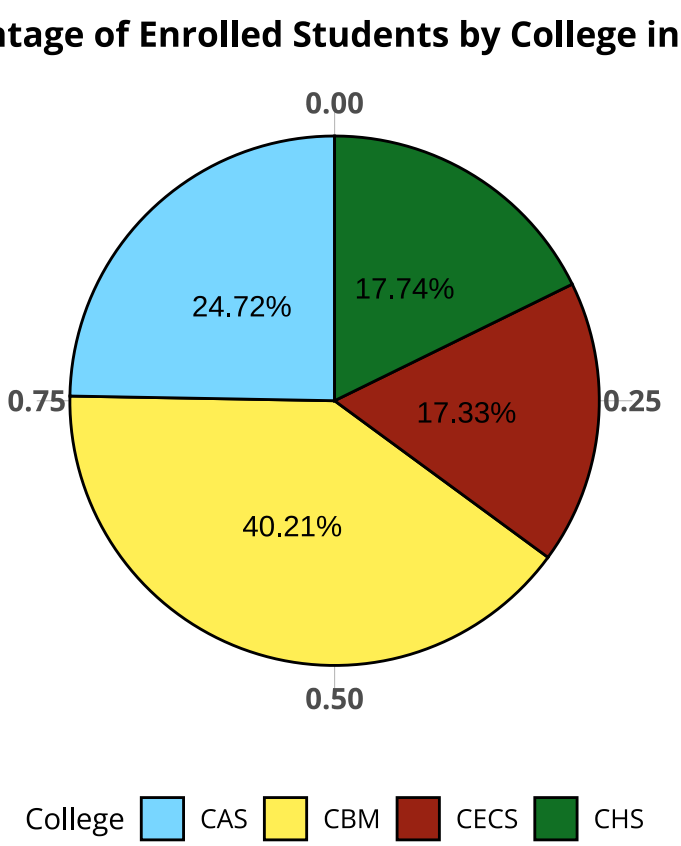
Task 2.

```r
df_2118 <- df[df$year == 2118,]

ggplot(df_2118, aes(x = college, y = pct, fill = college)) +
  geom_bar(stat = "identity", color = "black") +
  geom_text(aes(label = scales::percent(pct)), vjust = -0.5) +
  scale_fill_manual(values = college_palette, name = "College") +
  theme_vinuni() +
  scale_y_continuous(labels = label_percent()) +
  labs(x = "College", y = "Percentage",
  title = "Percentage of Enrolled Students by College in 2118")
```



Task 3.

```
df_2118 <- df[df$year == 2118,]

ggplot(df_2118, aes(x = "", y = pct, fill = college)) +
  geom_bar(stat = "identity", color = "black") +
  coord_polar(theta = "y") +
  geom_text(aes(label = scales::percent(pct)), position = position_stack(vjust = 0.5)) +
  scale_fill_manual(values = college_palette, name = "College") +
  theme_vinuni() +
  theme(axis.title = element_blank()) +
  labs(title = "Percentage of Enrolled Students by College in 2118")
```

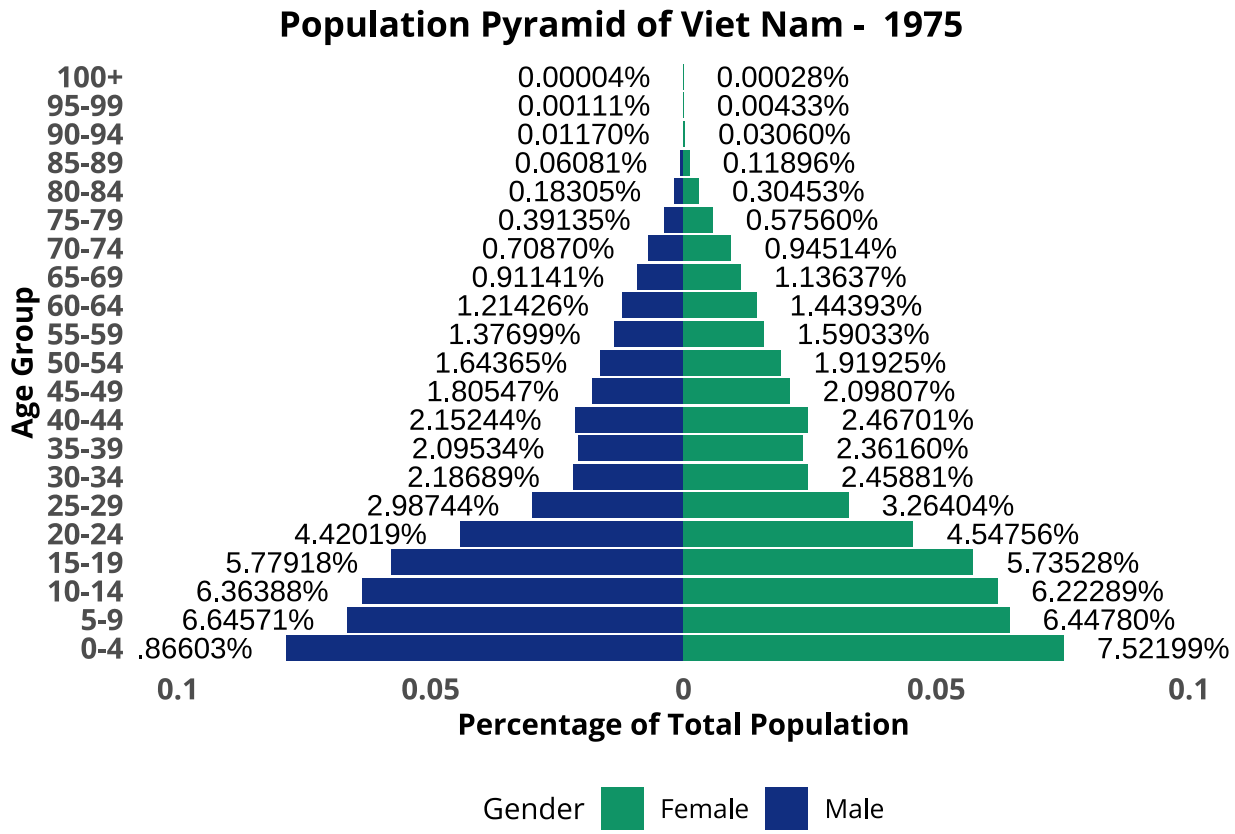


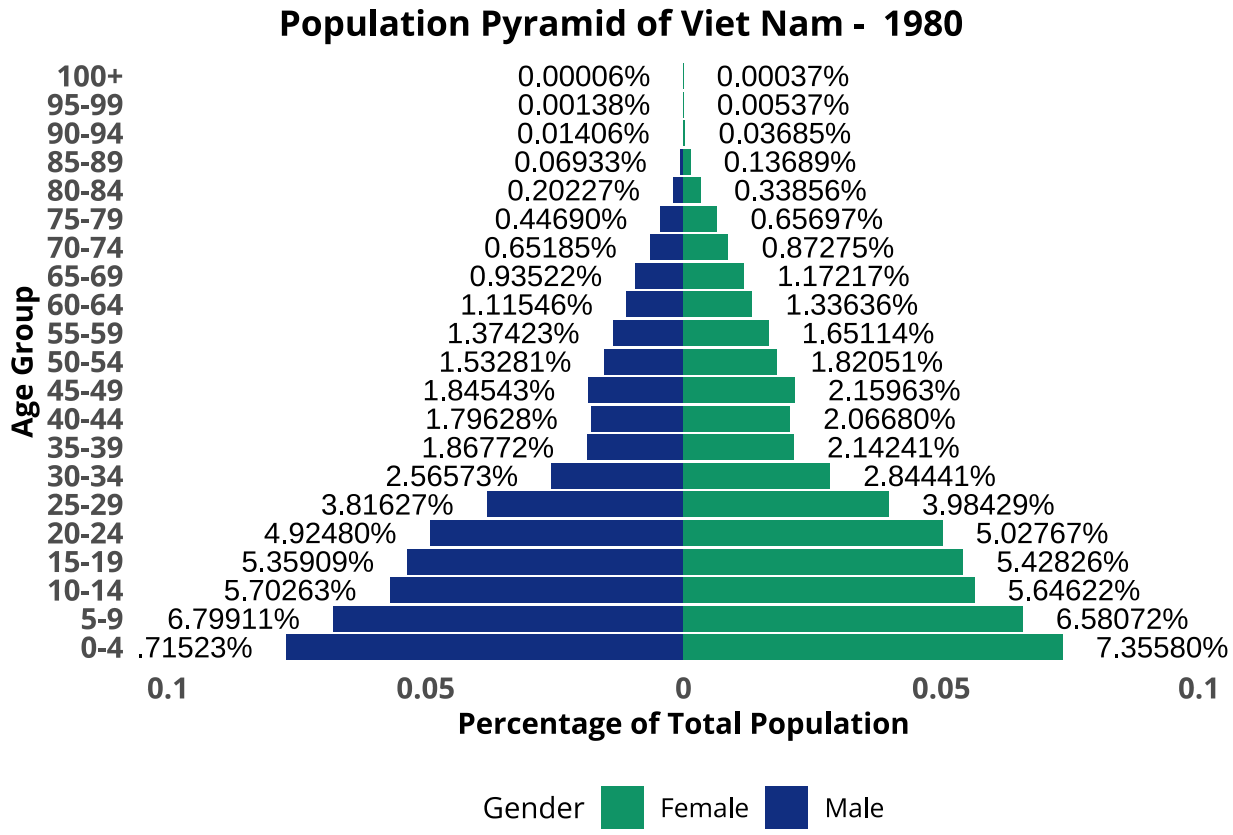
Both charts convey the information effectively, but the bar chart makes it easier to compare magnitudes, making it preferable.

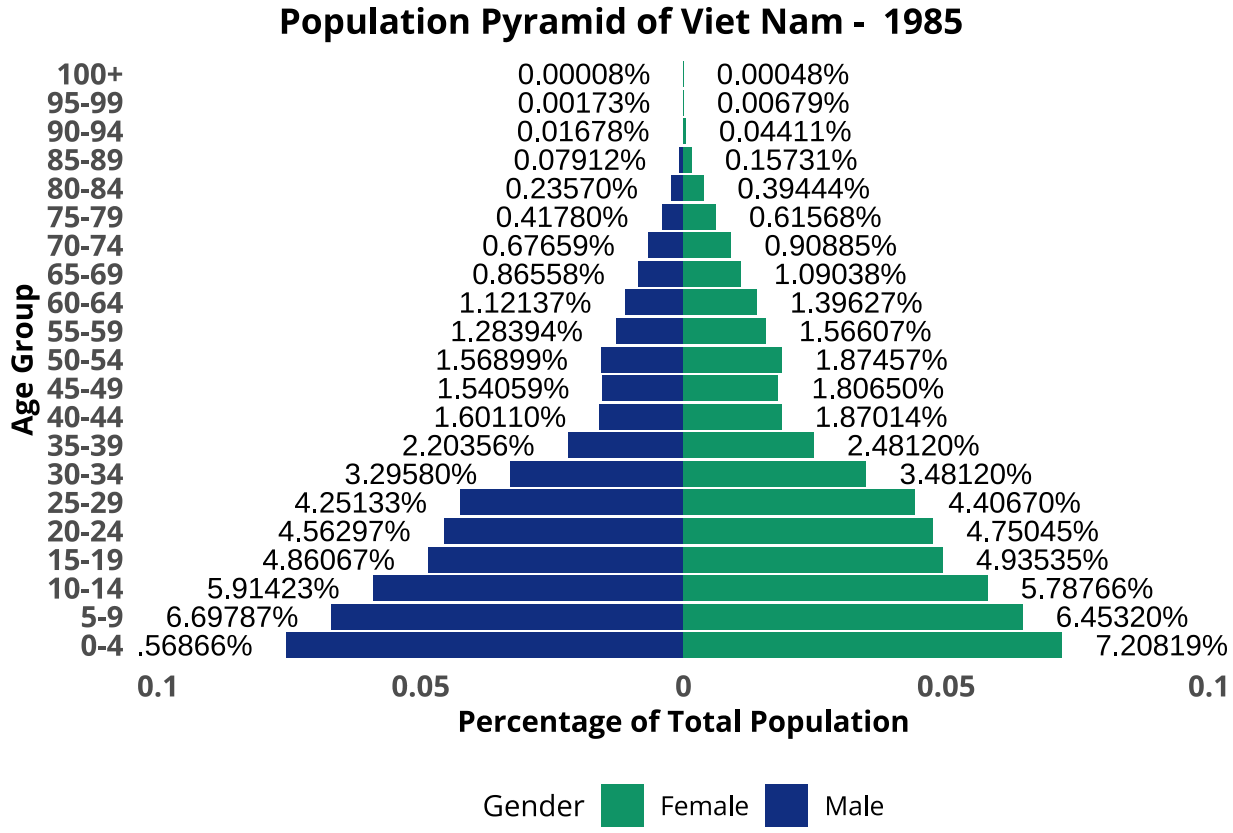
Task 4.

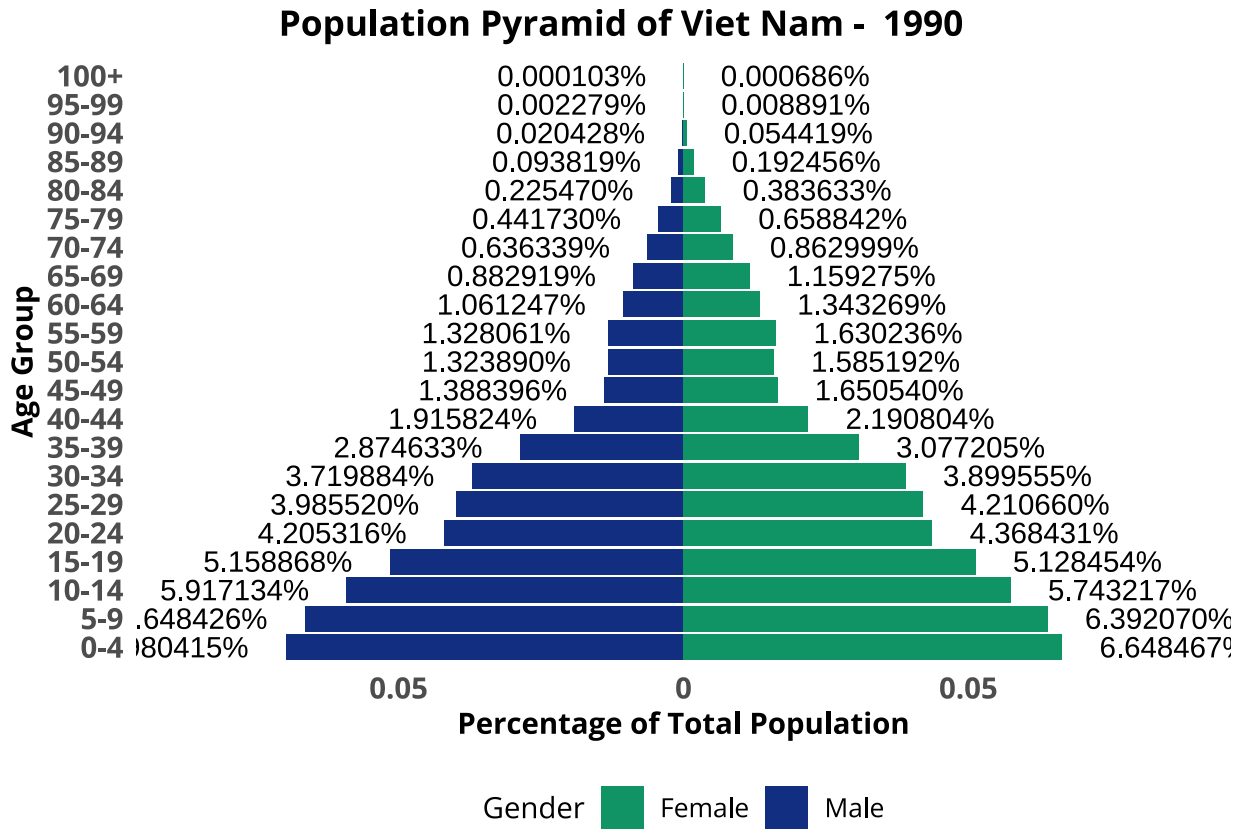
```
file_list <- list.files(pattern = "Viet Nam-")

for (file in file_list){
  file_df = read.csv(file)
  year = substring(file, 10, 13)
  file_df$Age <- factor(file_df$Age, levels = c("0-4", "5-9", "10-14", "15-19", "20-24", "25-29", "30-34
  file_df <- file_df%>%
  pivot_longer(
    cols = c("M","F"),
    names_to = "Sex",
    values_to = "Population"
  )
  file_df$Population = file_df$Population / sum(file_df$Population)


  plot <- ggplot(file_df, aes(x = Age, y = ifelse(Sex == "F", Population, Population * -1))) +
    geom_bar(aes(fill = Sex), stat = "identity") +
    geom_text(aes(hjust = ifelse(Sex == "M", 1.25, -0.25), label = scales::percent(abs(Population)))) +
    scale_y_continuous(
    labels = abs,
```

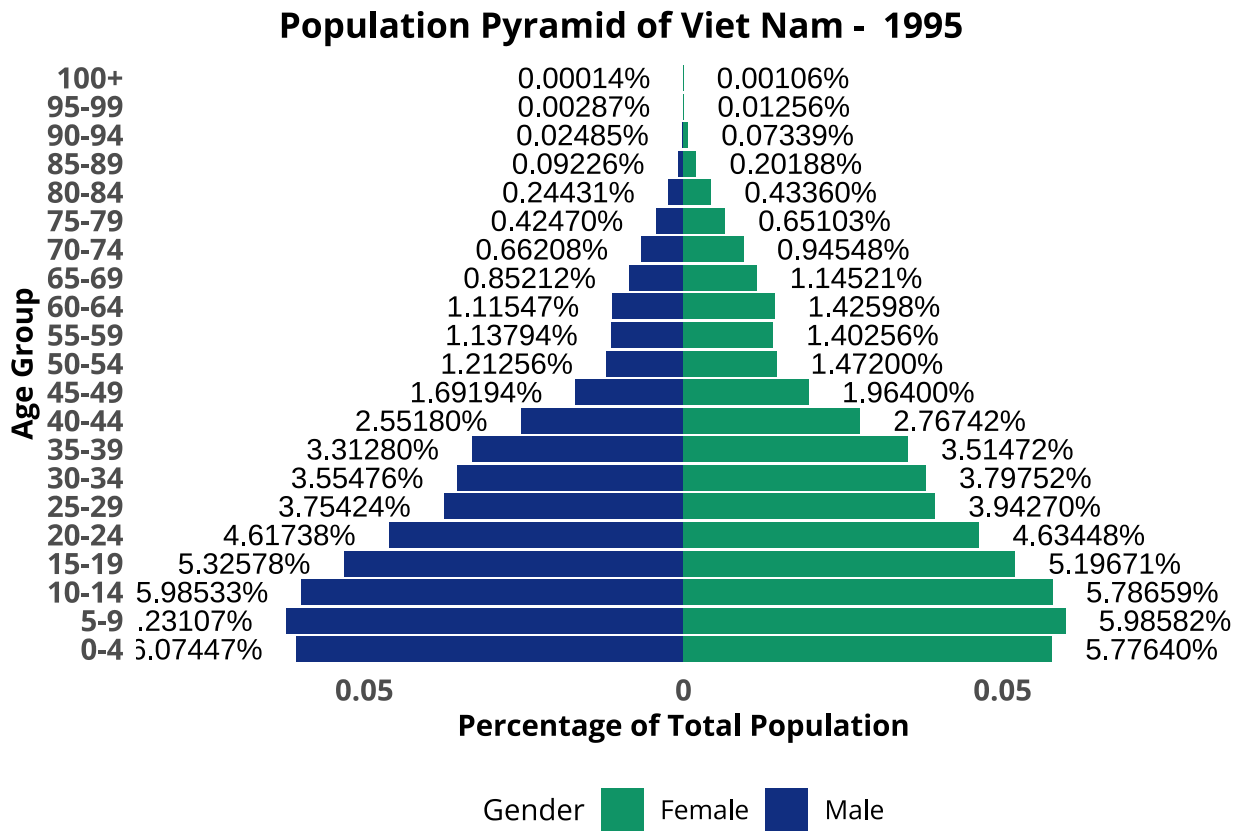
```
    limits = max(file_df$Population) * c(-1.25, 1.25)
    ) +
    scale_fill_manual(labels = c("Female", "Male"), values = c("#109466", "#112e80"), name = "Gender") +
    theme_vinuni() +
    theme(panel.grid.major.y = element_blank()) +
    labs(x = "Age Group", y = "Percentage of Total Population",
    title = paste("Population Pyramid of Viet Nam - ", year)) +
    coord_flip()

  print(plot)
}
```

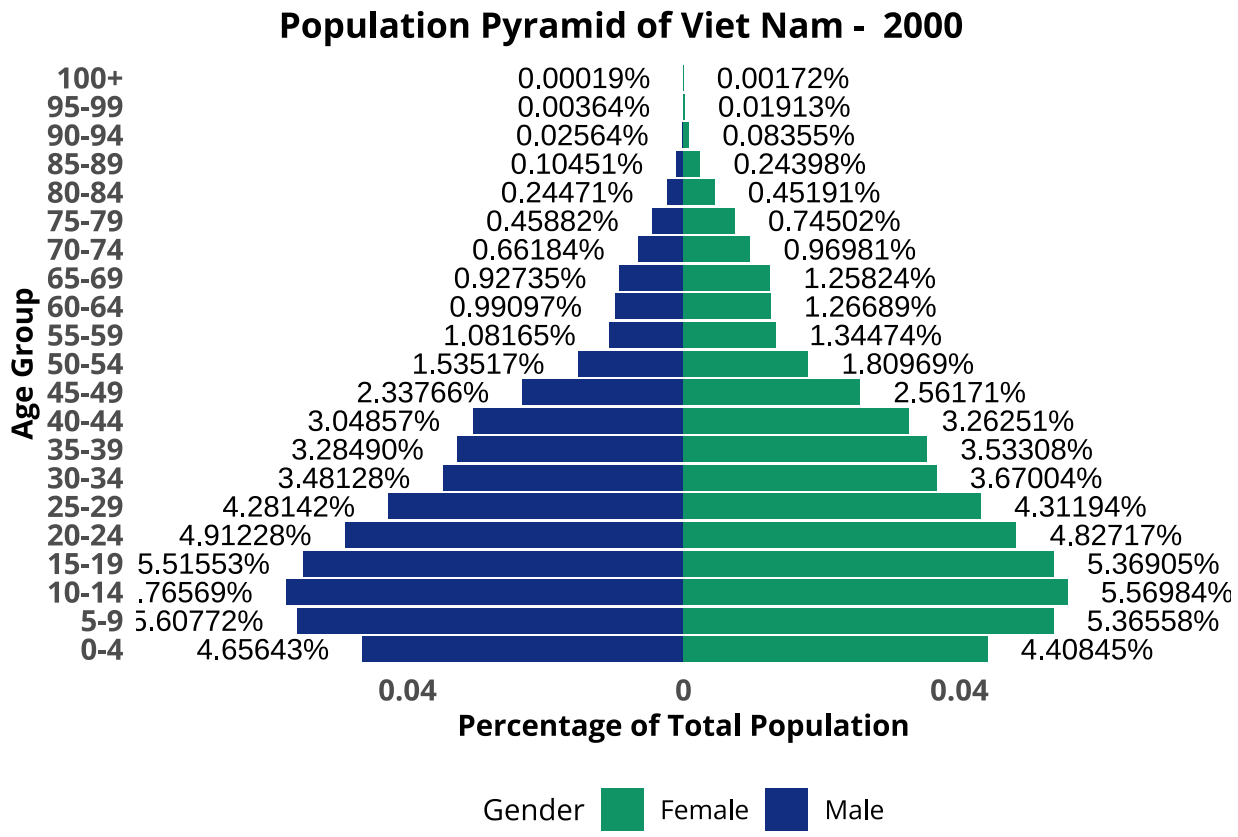


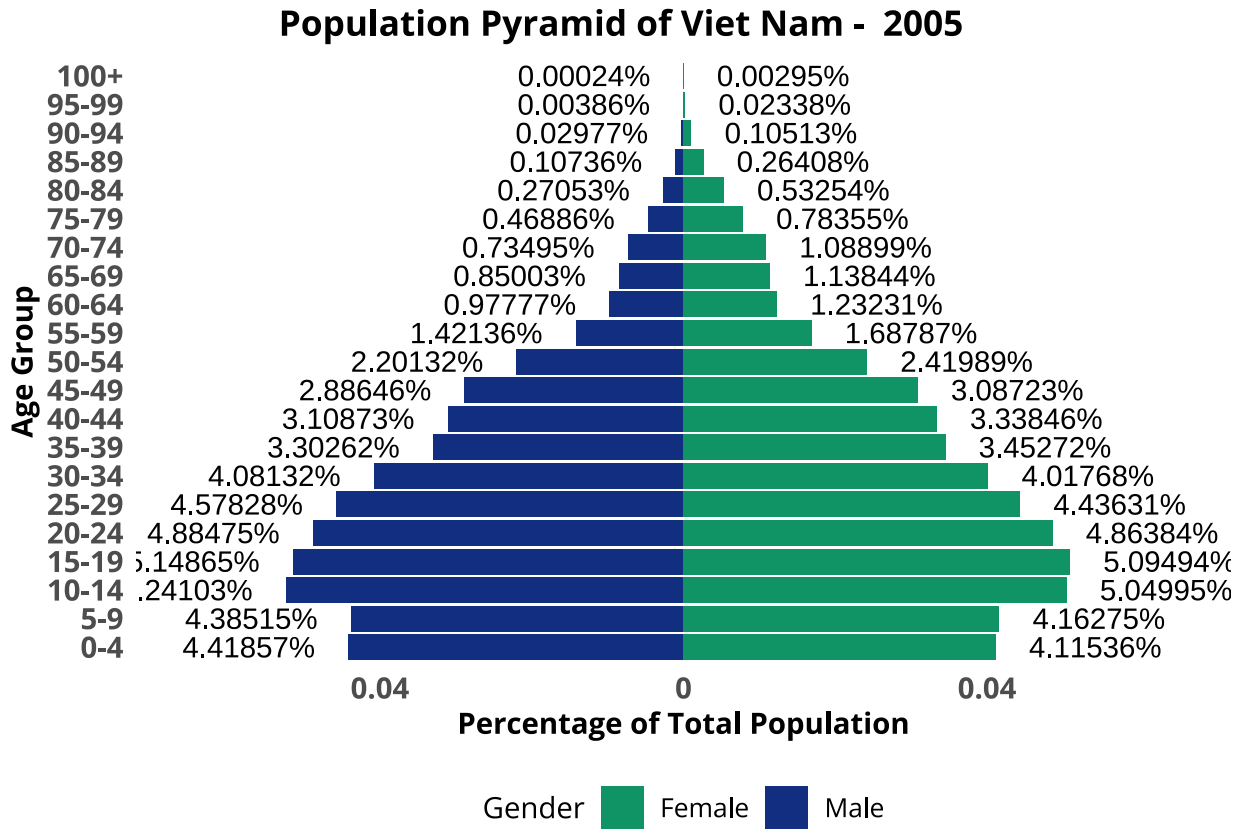

Population Pyramid of Viet Nam - 1980
Age Group
100+
95-99
90-94
85-89
80-84
75-79
70-74
65-69
60-64
55-59
50-54
45-49
40-44
35-39
30-34
25-29
20-24
15-19
10-14
5-9
0-4
0.00006%
0.00037%
0.00138%
0.00537%
0.01406%
0.03685%
0.06933%
0.13689%
0.20227%
0.33856%
0.44690%
0.65697%
0.65185%
0.87275%
0.93522%
1.17217%
1.11546%
1.33636%
1.37423%
1.65114%
1.53281%
1.82051%
1.84543%
2.15963%
1.79628%
2.06680%
1.86772%
2.14241%
2.56573%
2.84441%
3.81627%
3.98429%
4.92480%
5.02767%
5.35909%
5.42826%
5.70263%
5.64622%
6.79911%
6.58072%
.71523%
7.35580%
0.1
0.05
0
0.05
0.1
Percentage of Total Population
Gender
Female
Male

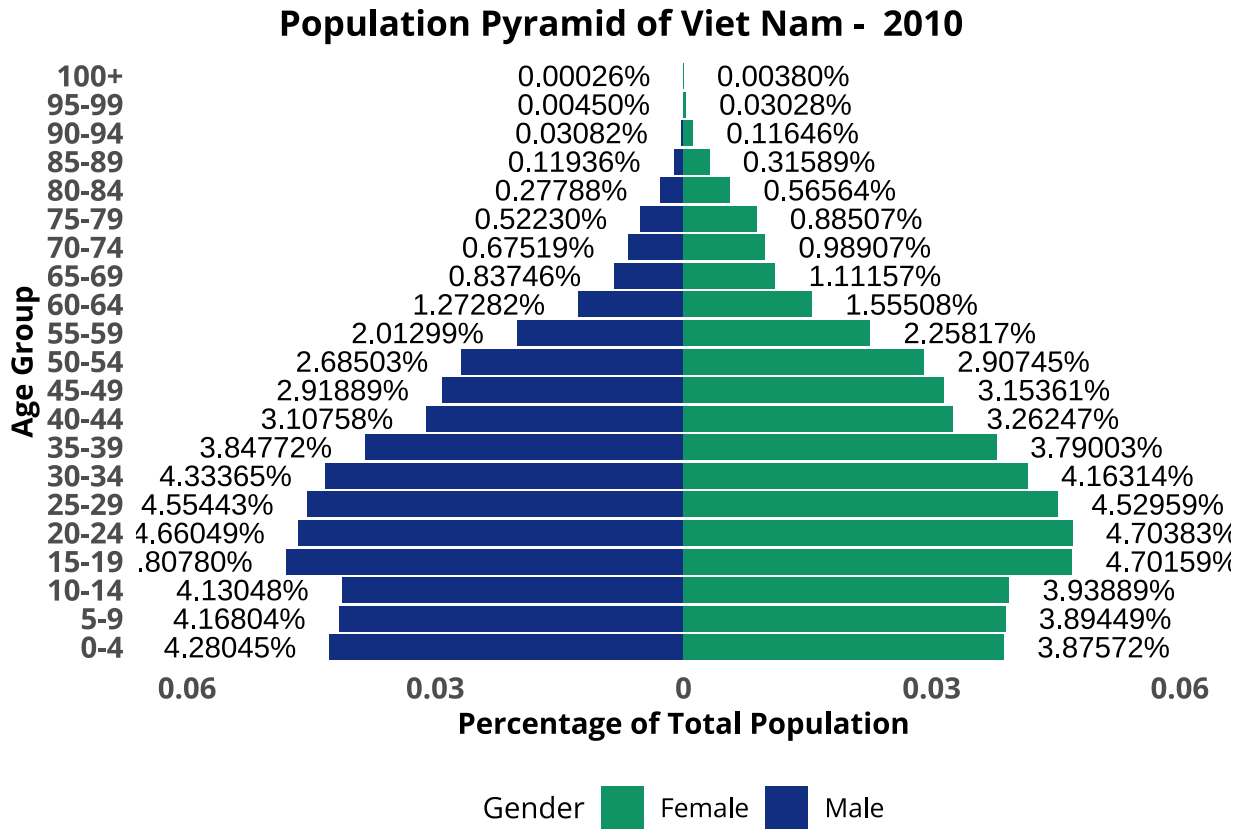

Population Pyramid of Viet Nam - 1985
Age Group
100+
95-99
90-94
85-89
80-84
75-79
70-74
65-69
60-64
55-59
50-54
45-49
40-44
35-39
30-34
25-29
20-24
15-19
10-14
5-9
0-4
0.00008%
0.00048%
0.00173%
0.00679%
0.01678%
0.04411%
0.07912%
0.15731%
0.23570%
0.39444%
0.41780%
0.61568%
0.67659%
0.90885%
0.86558%
1.09038%
1.12137%
1.39627%
1.28394%
1.56607%
1.56899%
1.87457%
1.54059%
1.80650%
1.60110%
1.87014%
2.20356%
2.48120%
3.29580%
3.48120%
4.25133%
4.40670%
4.56297%
4.75045%
4.86067%
4.93535%
5.91423%
5.78766%
6.69787%
6.45320%
.56866%
7.20819%
0.1
0.05
0
0.05
0.1
Percentage of Total Population
Gender
Female
Male

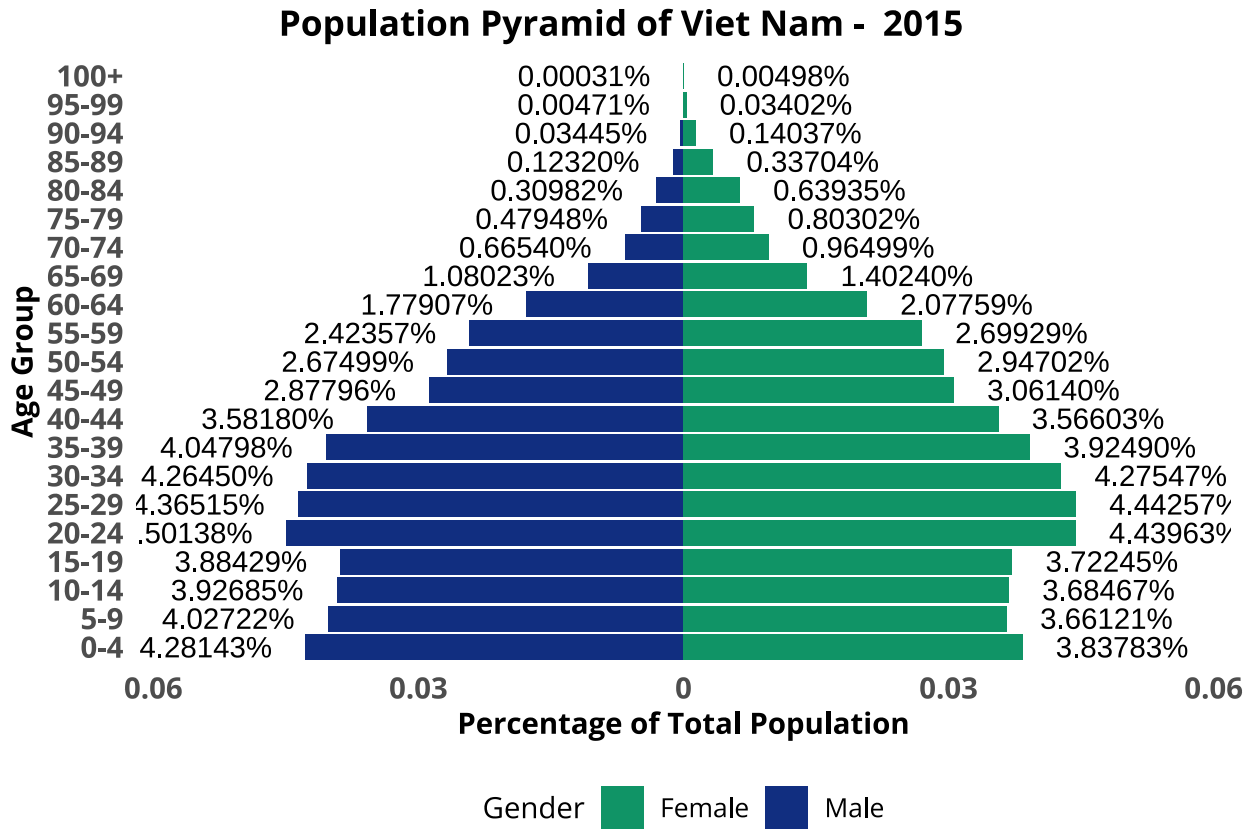

Population Pyramid of Viet Nam - 1990
Age Group
100+
95-99
90-94
85-89
80-84
75-79
70-74
65-69
60-64
55-59
50-54
45-49
40-44
35-39
30-34
25-29
20-24
15-19
10-14
5-9
0-4
0.000103%
0.002279%
0.020428%
0.093819%
0.225470%
0.441730%
0.636339%
0.882919%
1.061247%
1.328061%
1.323890%
1.388396%
1.915824%
2.874633%
3.719884%
3.985520%
4.205316%
5.158868%
5.917134%
.648426%
80415%
0.000686%
0.008891%
0.054419%
0.192456%
0.383633%
0.658842%
0.862999%
1.159275%
1.343269%
1.630236%
1.585192%
1.650540%
2.190804%
3.077205%
3.899555%
4.210660%
4.368431%
5.128454%
5.743217%
6.392070%
6.648467%
0.05
0
0.05
Percentage of Total Population
Gender
Female
Male

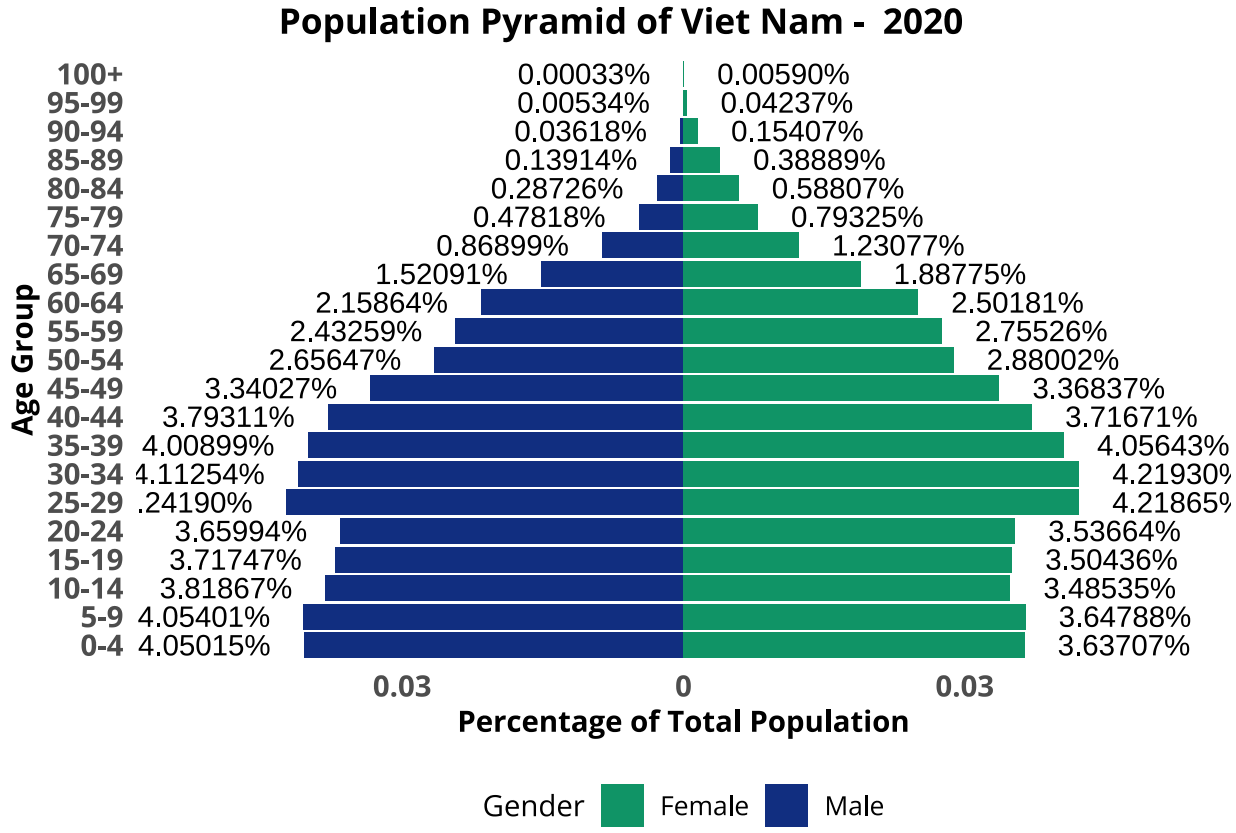

Population Pyramid of Viet Nam - 1995
Age Group
100+
95-99
90-94
85-89
80-84
75-79
70-74
65-69
60-64
55-59
50-54
45-49
40-44
35-39
30-34
25-29
20-24
15-19
10-14
5-9
0-4
0.00014%
0.00287%
0.02485%
0.09226%
0.24431%
0.42470%
0.66208%
0.85212%
1.11547%
1.13794%
1.21256%
1.69194%
2.55180%
3.31280%
3.55476%
3.75424%
4.61738%
5.32578%
5.98533%
.23107%
.07447%
0.00106%
0.01256%
0.07339%
0.20188%
0.43360%
0.65103%
0.94548%
1.14521%
1.42598%
1.40256%
1.47200%
1.96400%
2.76742%
3.51472%
3.79752%
3.94270%
4.63448%
5.19671%
5.78659%
5.98582%
5.77640%
0.05
0
0.05
Percentage of Total Population
Gender
Female
Male

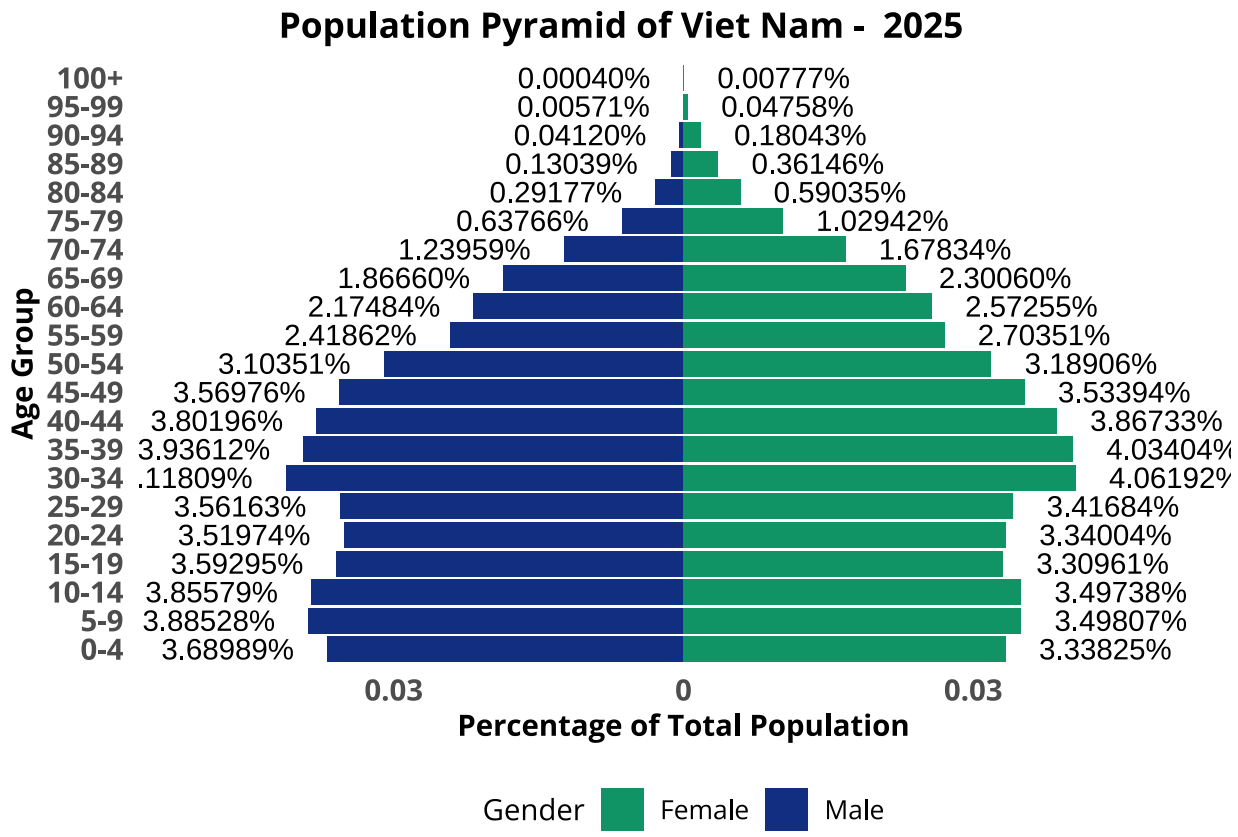

Population Pyramid of Viet Nam - 2000
Age Group
100+
95-99
90-94
85-89
80-84
75-79
70-74
65-69
60-64
55-59
50-54
45-49
40-44
35-39
30-34
25-29
20-24
15-19
10-14
5-9
0-4
0.00019%
0.00172%
0.00364%
0.01913%
0.02564%
0.08355%
0.10451%
0.24398%
0.24471%
0.45191%
0.45882%
0.74502%
0.66184%
0.96981%
0.92735%
1.25824%
0.99097%
1.26689%
1.08165%
1.34474%
1.53517%
1.80969%
2.33766%
2.56171%
3.04857%
3.26251%
3.28490%
3.53308%
3.48128%
3.67004%
4.28142%
4.31194%
4.91228%
4.82717%
5.51553%
5.36905%
.76569%
5.56984%
5.60772%
5.36558%
4.65643%
4.40845%
0.04
0
0.04
Percentage of Total Population
Gender
Female
Male


Population Pyramid of Viet Nam - 2005
Age Group
100+
95-99
90-94
85-89
80-84
75-79
70-74
65-69
60-64
55-59
50-54
45-49
40-44
35-39
30-34
25-29
20-24
15-19
10-14
5-9
0-4
0.00024%
0.00295%
0.00386%
0.02338%
0.02977%
0.10513%
0.10736%
0.26408%
0.27053%
0.53254%
0.46886%
0.78355%
0.73495%
1.08899%
0.85003%
1.13844%
0.97777%
1.23231%
1.42136%
1.68787%
2.20132%
2.41989%
2.88646%
3.08723%
3.10873%
3.33846%
3.30262%
3.45272%
4.08132%
4.01768%
4.57828%
4.43631%
4.88475%
4.86384%
5.14865%
5.09494%
5.24103%
5.04995%
4.38515%
4.16275%
4.41857%
4.11536%
0.04
0
0.04
Percentage of Total Population
Gender
Female
Male


Population Pyramid of Viet Nam - 2010
Age Group
100+
95-99
90-94
85-89
80-84
75-79
70-74
65-69
60-64
55-59
50-54
45-49
40-44
35-39
30-34
25-29
20-24
15-19
10-14
5-9
0-4
0.00026%
0.00380%
0.00450%
0.03028%
0.03082%
0.11646%
0.11936%
0.31589%
0.27788%
0.56564%
0.52230%
0.88507%
0.67519%
0.98907%
0.83746%
1.11157%
1.27282%
1.55508%
2.01299%
2.25817%
2.68503%
2.90745%
2.91889%
3.15361%
3.10758%
3.26247%
3.84772%
3.79003%
4.33365%
4.16314%
4.55443%
4.52959%
4.66049%
4.70383%
.80780%
4.70159%
4.13048%
3.93889%
4.16804%
3.89449%
4.28045%
3.87572%
0.06
0.03
0
0.03
0.06
Percentage of Total Population
Gender
Female
Male


Population Pyramid of Viet Nam - 2015
Age Group
100+ 0.00031% 0.00498%
95-99 0.00471% 0.03402%
90-94 0.03445% 0.14037%
85-89 0.12320% 0.33704%
80-84 0.30982% 0.63935%
75-79 0.47948% 0.80302%
70-74 0.66540% 0.96499%
65-69 1.08023% 1.40240%
60-64 1.77907% 2.07759%
55-59 2.42357% 2.69929%
50-54 2.67499% 2.94702%
45-49 2.87796% 3.06140%
40-44 3.58180% 3.56603%
35-39 4.04798% 3.92490%
30-34 4.26450% 4.27547%
25-29 4.36515% 4.44257%
20-24 .50138% 4.43963%
15-19 3.88429% 3.72245%
10-14 3.92685% 3.68467%
5-9 4.02722% 3.66121%
0-4 4.28143% 3.83783%
0.06
0.03
0
0.03
0.06
Percentage of Total Population
Gender
Female
Male

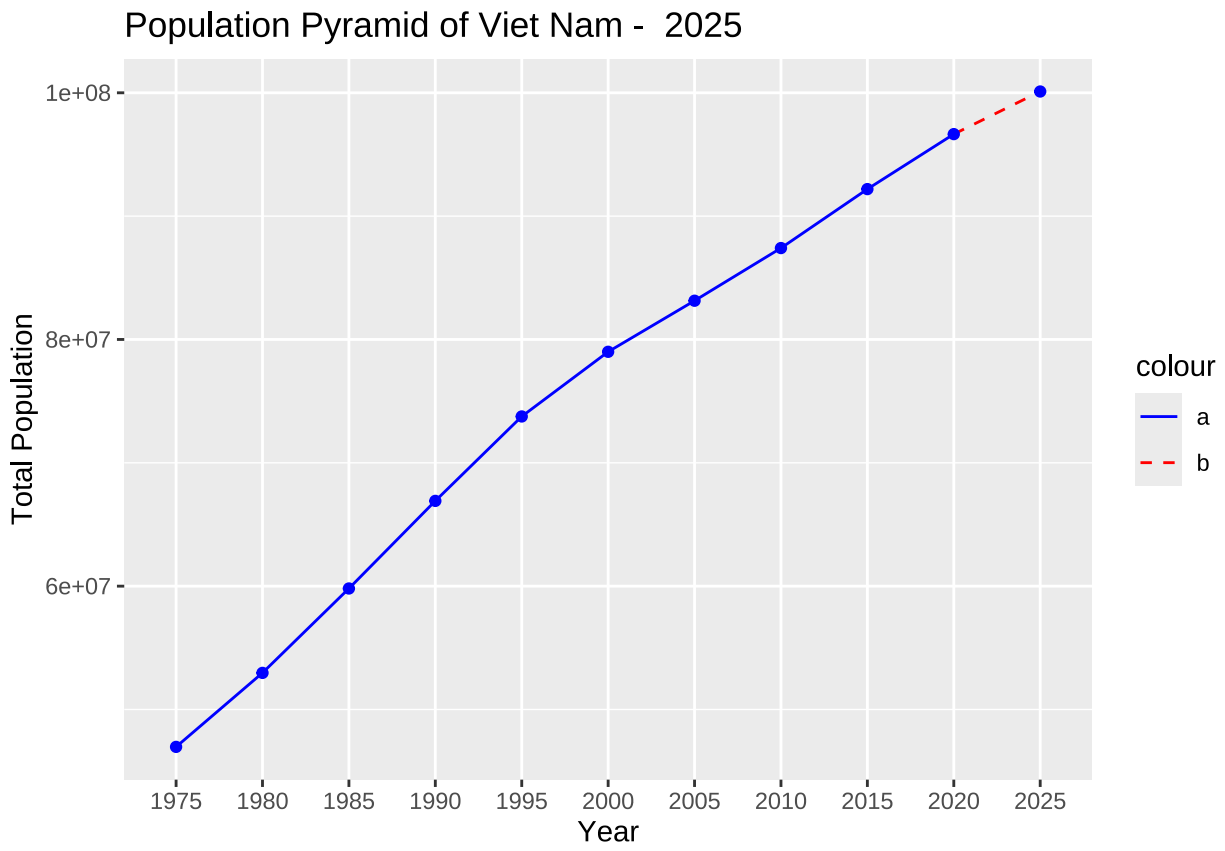

Population Pyramid of Viet Nam - 2020
Age Group
100+
95-99
90-94
85-89
80-84
75-79
70-74
65-69
60-64
55-59
50-54
45-49
40-44
35-39
30-34
25-29
20-24
15-19
10-14
5-9
0-4
0.00033%
0.00534%
0.03618%
0.13914%
0.28726%
0.47818%
0.86899%
1.52091%
2.15864%
2.43259%
2.65647%
3.34027%
3.79311%
4.00899%
4.11254%
.24190%
3.65994%
3.71747%
3.81867%
4.05401%
4.05015%
0.00590%
0.04237%
0.15407%
0.38889%
0.58807%
0.79325%
1.23077%
1.88775%
2.50181%
2.75526%
2.88002%
3.36837%
3.71671%
4.05643%
4.21930%
4.21865%
3.53664%
3.50436%
3.48535%
3.64788%
3.63707%
0.03
0
0.03
Percentage of Total Population
Gender
Female
Male



The population pyramid provides how the population is distributed by age, which allows insights into things like percentage of dependents (senior citizens and children) and gender ratio. Comparing the pyramid over time tells us the mortality and fertility rates.looking at the 11 charts above tells us that Vietnam has an aging population, which is related to lower fertility and mortality rates. There are multiple possible causes, including higher quality of life and people's tendency to have fewer children later for reasons such as chasing their careers.

Task 5.

```
file_list <- list.files(pattern = "Viet Nam-")
total_df = data.frame(Year = c(), Population = c(), stringsAsFactors = FALSE)
for (file in file_list){
  file_df = read.csv(file)
  year = substring(file, 10, 13)
  total_df <- rbind(total_df, data.frame(Year = year, Population = sum(file_df$M) + sum(file_df$F)))
}
ggplot(total_df, aes(x = Year, y = Population, group = 1)) +
  geom_line(aes(color = "a"), data = subset(total_df, Year <= 2020), linetype = 1) +
  geom_line(aes(color = "b"), data = subset(total_df, Year >= 2020), linetype = 2) +
  geom_point(color = "blue") +
  labs(x = "Year", y = "Total Population",
  title = paste("Population Pyramid of Viet Nam - ", year)) +
  scale_color_manual(values = c("a" = "blue", "b" = "red"))
```



Task 6.

Despite the amount of information that can be learned from the gravity map, this visualization suffers from a poor choice of color scale. First of all, the lack of a legend makes it unclear whether red or blue shades are associated with higher gravity. In heatmaps, warm colors are used to represent higher magnitude, but this is not always the case for all visualizations. Another glaring issue with the rainbow color scale is accessibility. In particular, colorblindness can severely interfere with one's ability to interpret the visualization. Even without colorblindness, the usage of a wide range of colors makes it difficult for viewers to distinguish between different gravity levels and compare regions. Compared to the rainbow scale, another color scheme such as grayscale may make it easier to perceive the dfferent gravity levels in the map.

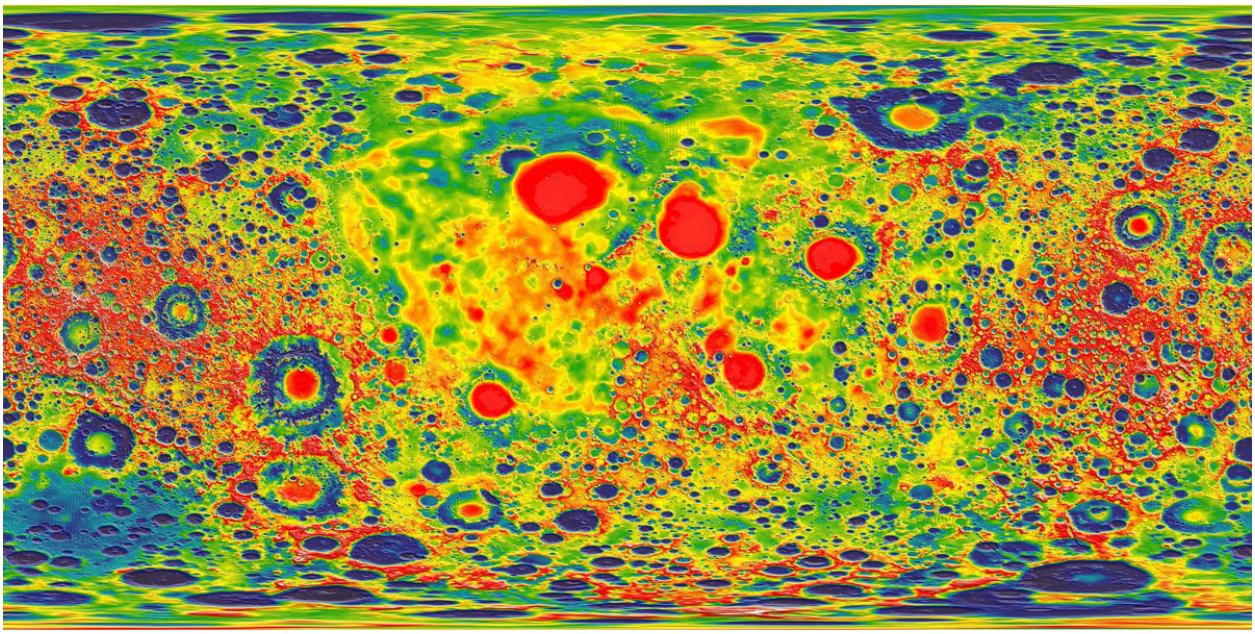


Figure 1: The moon's gravitational pull, mapped with a rainbow scale by NASA. This map allows scientists to learn about the moon's structure. It reveals geographical features such as craters or volcanic landforms, and provides insights regarding potential subsurface features.