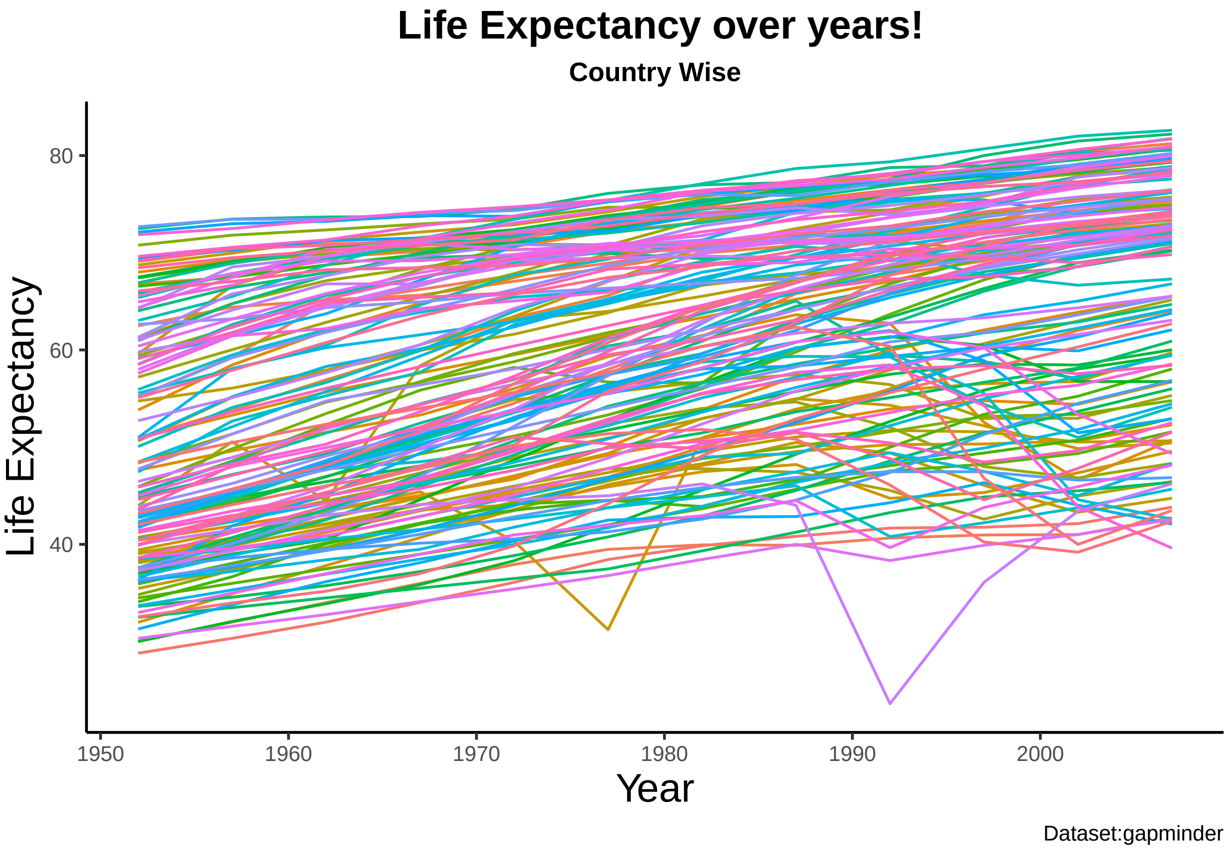
Assignment 2: Data Visualization

Rohit Jain | 194161020

Question 1:

Load the data gapminder and analyze different columns of data.Plot life expectancy over time for each country. Do you think the plot is meaningful? Justify your answer in markdown

```r
p <- ggplot(data = gapminder,
            mapping = aes(x = year,
                          y = lifeExp, col=country))
p + geom_line() +
  guides(col=FALSE) +
  theme_classic() +
  labs(x="Year", y="Life Expectancy",
       title=" Life Expectancy over years!",
       subtitle="Country Wise",
       caption="Dataset:gapminder")+
  theme
```



The plot is not very meaningful since we cannot differentiate the particular pattern for any country. The plot looks rough and messy and hence not very meaningful.

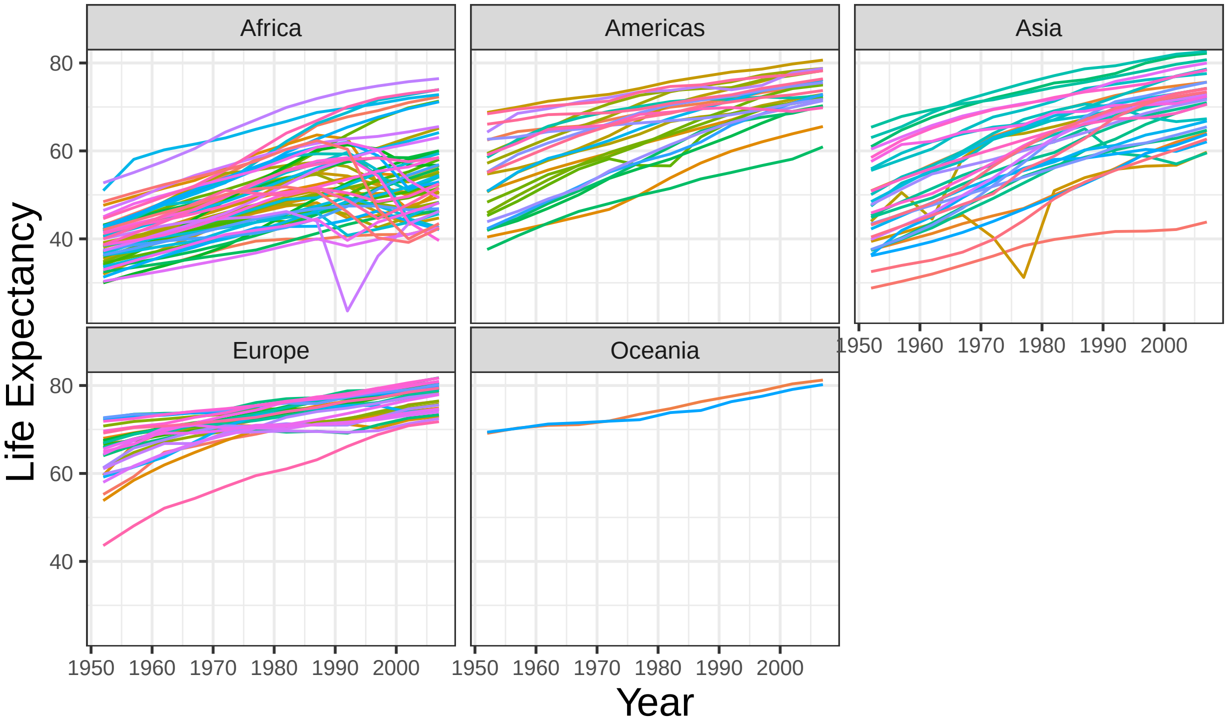
Question 2

I. Can you improve the plot from the first question by faceting the data? Which is the most appropriate variable (column) to facet the data? Plot and justify.

```
p <- ggplot(data = gapminder,
            mapping = aes(x = year,
                          y = lifeExp, col=country))
p + geom_line() +
  facet_wrap(~continent)+
  guides(col=FALSE)+
  theme_bw()+
  theme+
  labs(x="Year", y="Life Expectancy",
       caption="Dataset:gapminder",
       subtitle="Continent Wise",
       title=" Life Expectancy over years for each country!")
```

Life Expectancy over years for each country!

Continent Wise



Dataset:gapminder

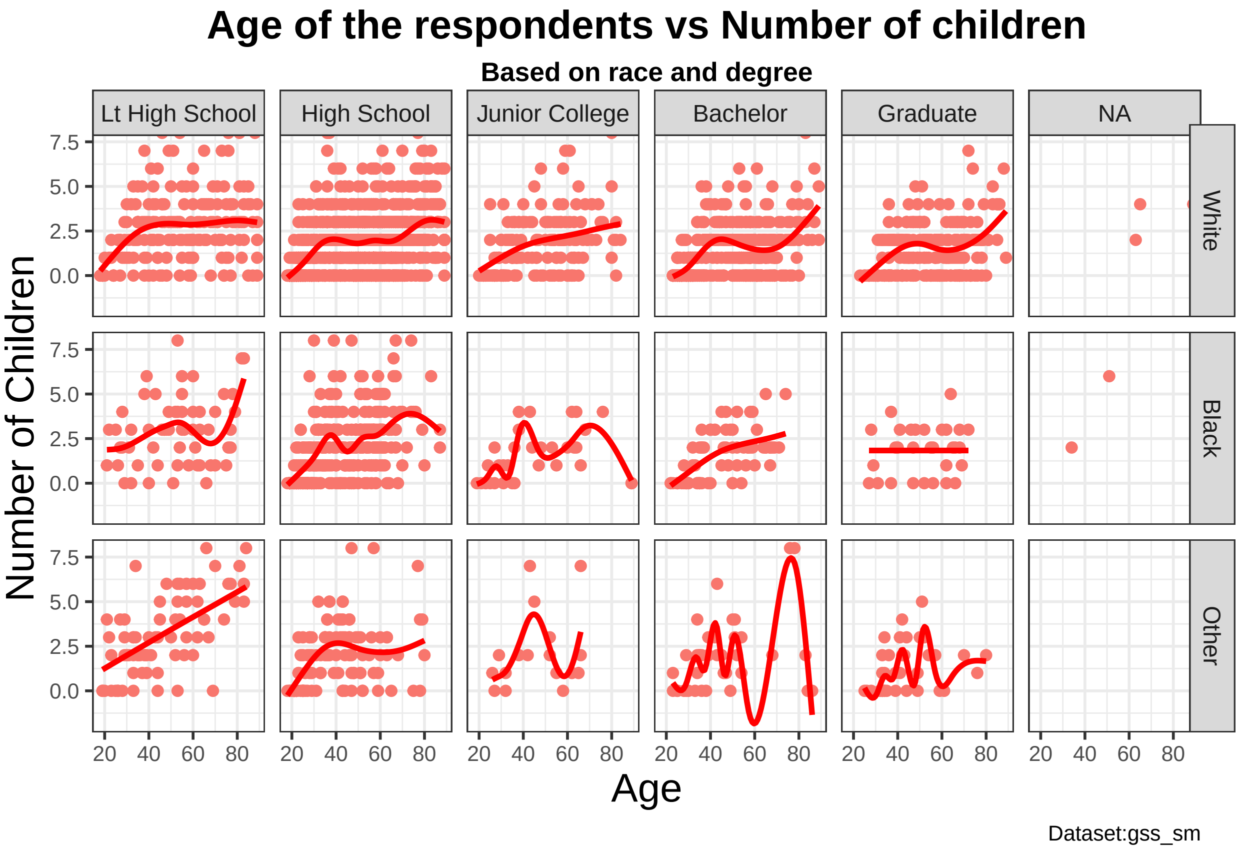
"Continent" turns out to be the best column for faceting the data into groups because it is categorical in nature while other remaining columns have continuous values. Categorical variable makes it easy for grouping the data together.

II. We can facet the data based on more than one variable, Use gss sm data to plot a smoothed scatter plot showing the relationship between the age of the respondent and the number of children they have. Facet this relation based on race and degree. Also, in markdown, describe your observation from the plot in brief.

```
p<-ggplot(data=gss_sm, aes(x=age, y=childs))

p+ geom_point(aes(col=""))+
  geom_smooth(se=FALSE, col="red")+
  facet_grid(race~degree)+
   guides(col=FALSE) +
  theme_bw()+
  theme+
  labs(x="Age", y="Number of Children",
       caption="Dataset:gss_sm",
       subtitle="Based on race and degree",
       title="Age of the respondents vs Number of children")
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

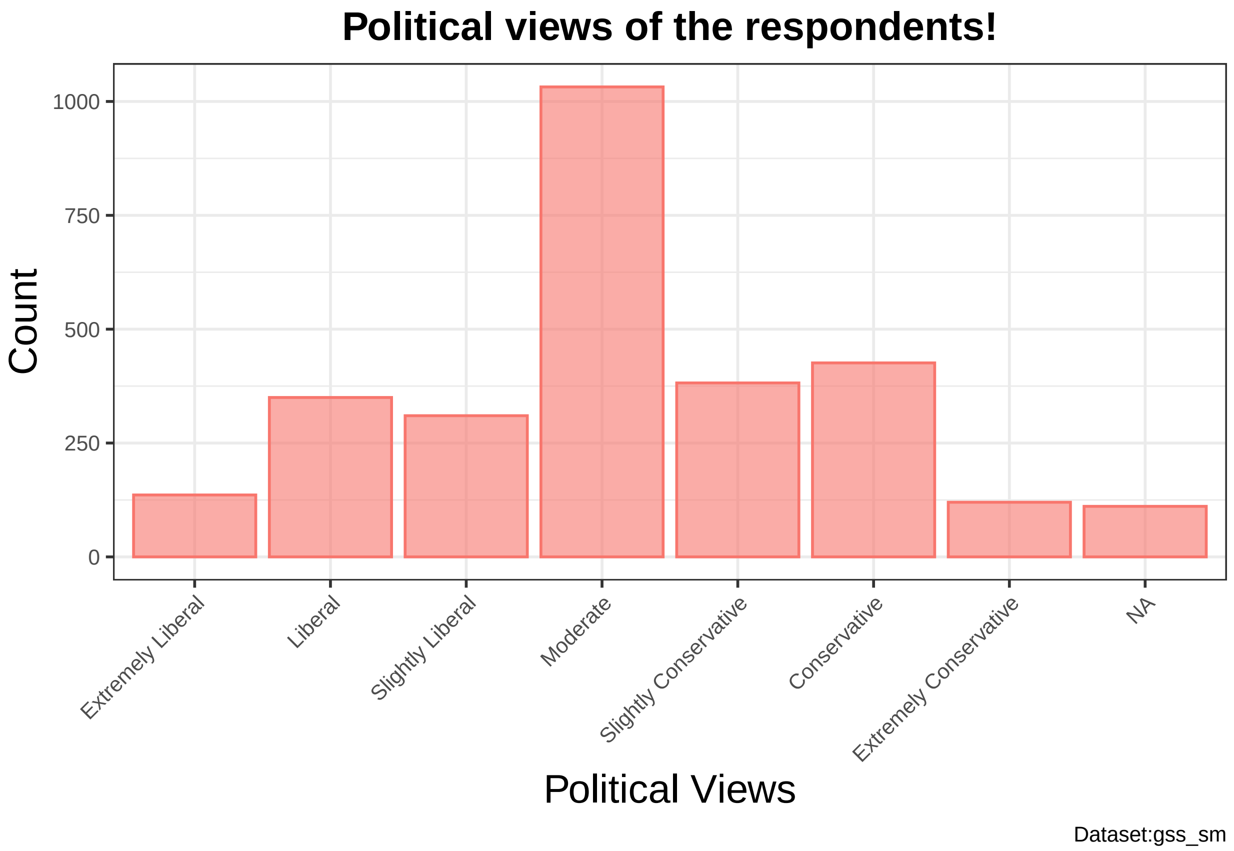


In this graph we can see the relationship between the age of the respondent and the number of children they have, grouped based on race and degree. The first plot is basically about the respondents who belong to white race and have left High School.

Question 3

gss sm data contains the political view (polviews) variable. Plot a bar graph, with the bar in the chart represented by different political views.

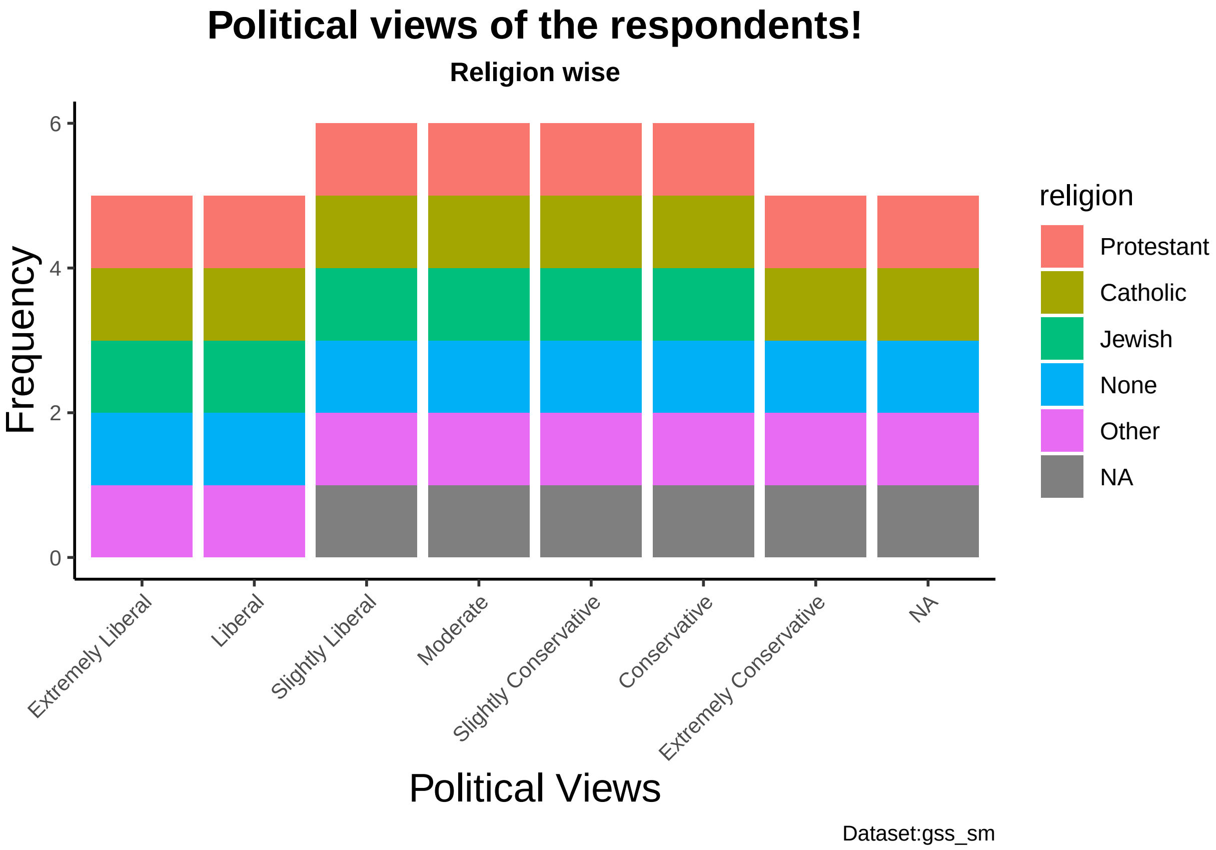
```
p<-ggplot(data=gss_sm, aes(x=polviews))
p+geom_bar(aes(col="",fill="" ),alpha=0.6)+
  guides(col=FALSE,fill=FALSE)+
  theme_bw()+
  theme+
  labs(x="Political Views", y="Count",
       title="Political views of the respondents!",
       caption="Dataset:gss_sm")+
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



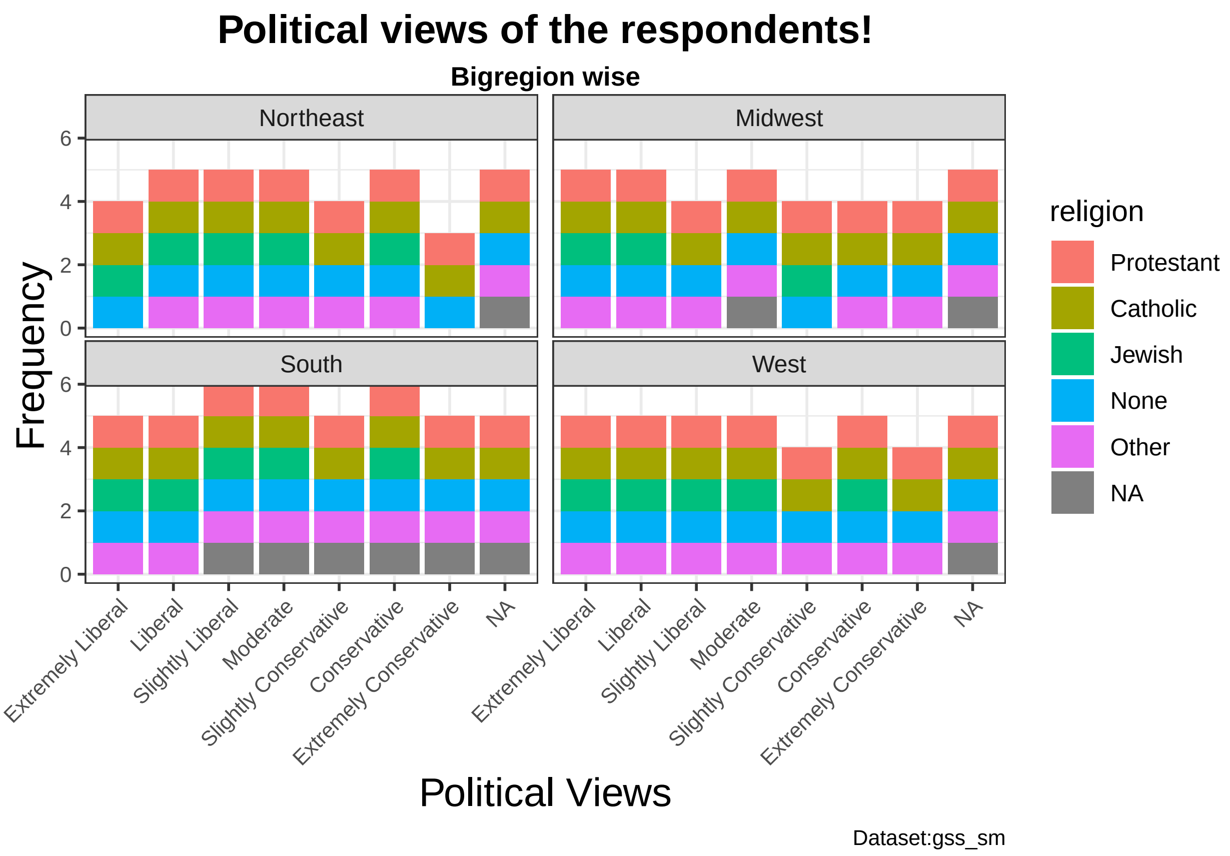
Question 4

Again using gss sm data, visualize the frequency plot, with the bars representing different political views and each bar in the graph, is further categorized by a different religion. Also, visualize frequency plot faceted by variable bigregion

```
p <- ggplot(data = gss_sm,
            mapping = aes(x = polviews,
                          fill = religion))
p + geom_bar(aes(y=..prop..))+
  theme_classic()+
  theme+
  labs(x="Political Views", y="Frequency",
       title="Political views of the respondents!",
       subtitle="Religion wise",
       caption="Dataset:gss_sm")+
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



```
p <- ggplot(data = gss_sm,
            mapping = aes(x = polviews, fill = religion))
p + geom_bar(aes(y=..prop..))+
  facet_wrap(~bigregion)+
  theme_bw()+
  theme+
  labs(x="Political Views", y="Frequency",
       title="Political views of the respondents!", subtitle="Bigregion wise",
       caption="Dataset:gss_sm")+
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

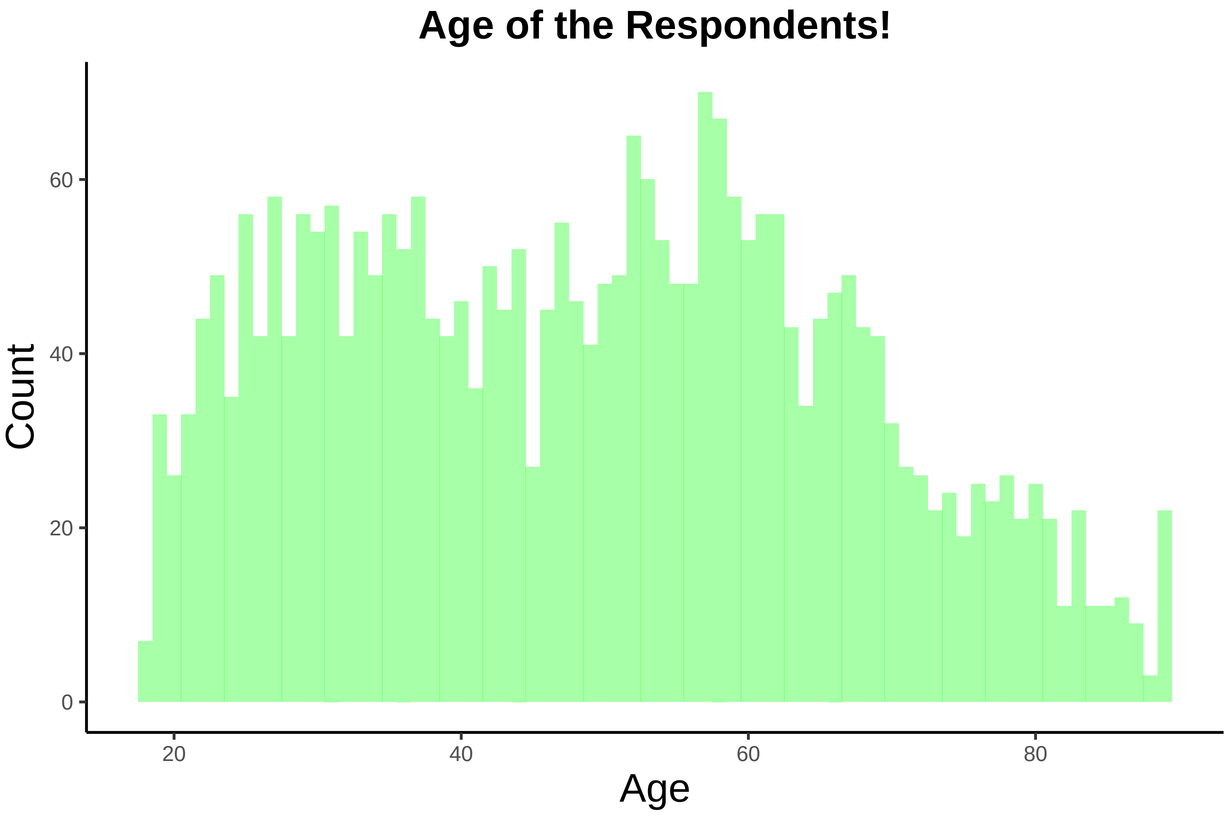


Question 5

Plot following from the gss sm data.

I. Histogram. Showcasing ages into bins.

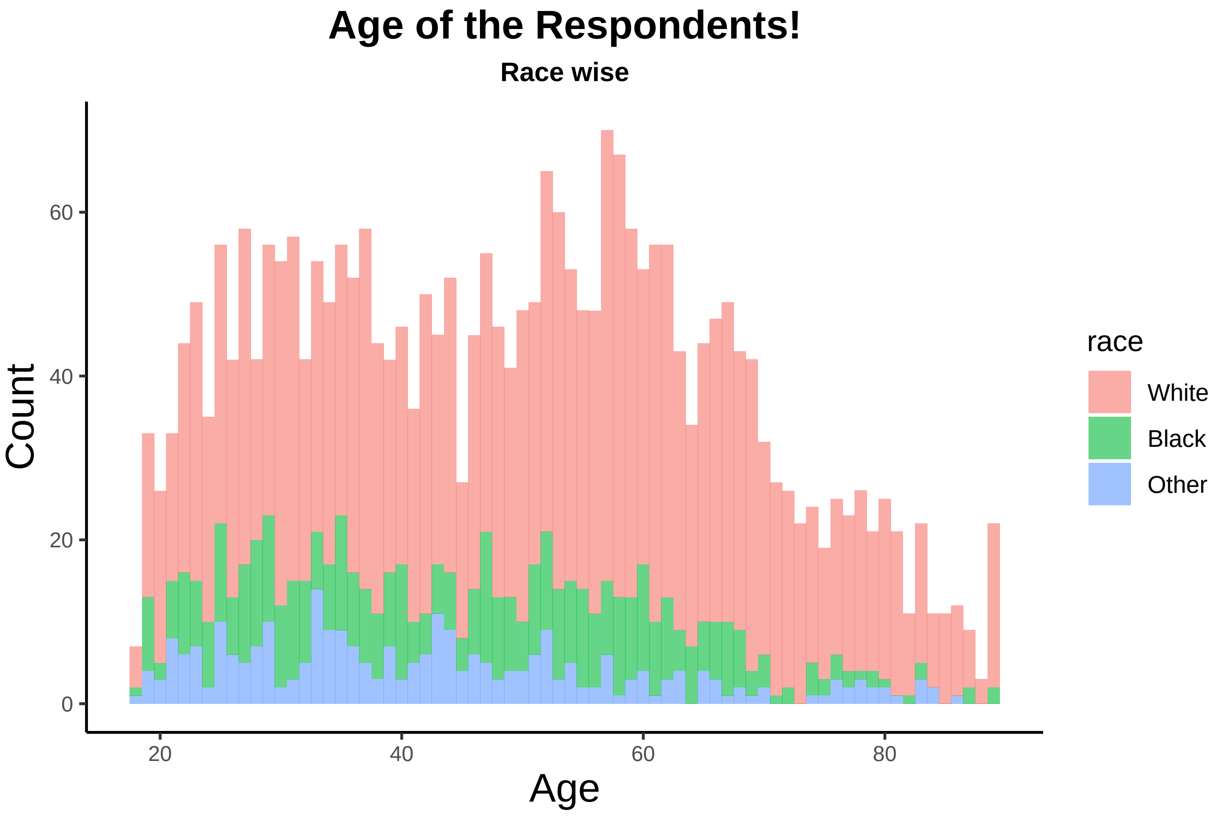
```
ggplot(data = gss_sm) +
  geom_histogram(mapping = aes(x = age),
               binwidth = 1, fill="green", alpha=0.35)+
  labs(x="Age", y="Count",
       title="Age of the Respondents!",
       caption="Dataset:gss_sm")+theme_classic()+theme
```



Dataset:gss_sm

** II. Modify the previous histogram, dividing the observation based on variable race. Each race should be represented in a different color.**

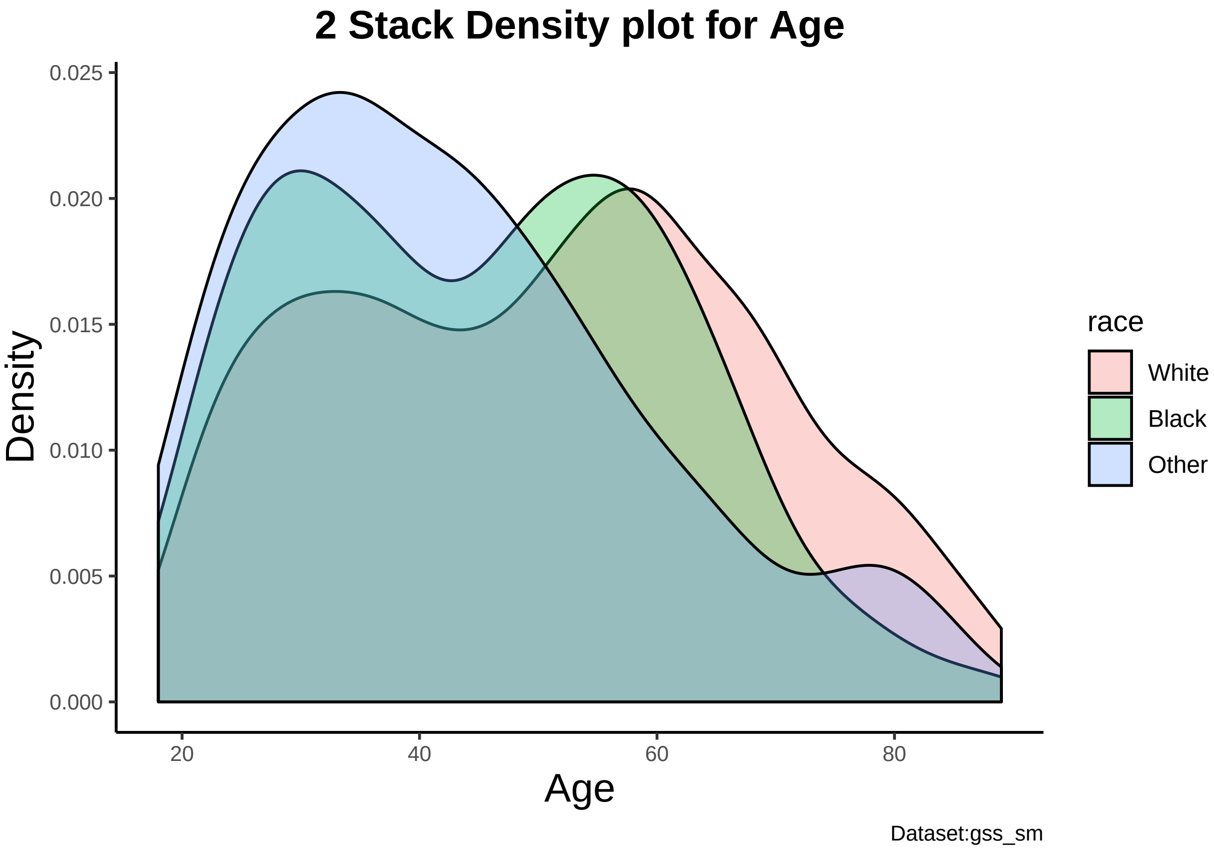
```
ggplot(data = gss_sm) +
  geom_histogram(mapping = aes(x = age, fill=race),
                 binwidth = 1, alpha=0.6)+
  labs(x="Age", y="Count", title="Age of the Respondents!",
       subtitle="Race wise", caption="Dataset:gss_sm")+theme_classic()+theme
```



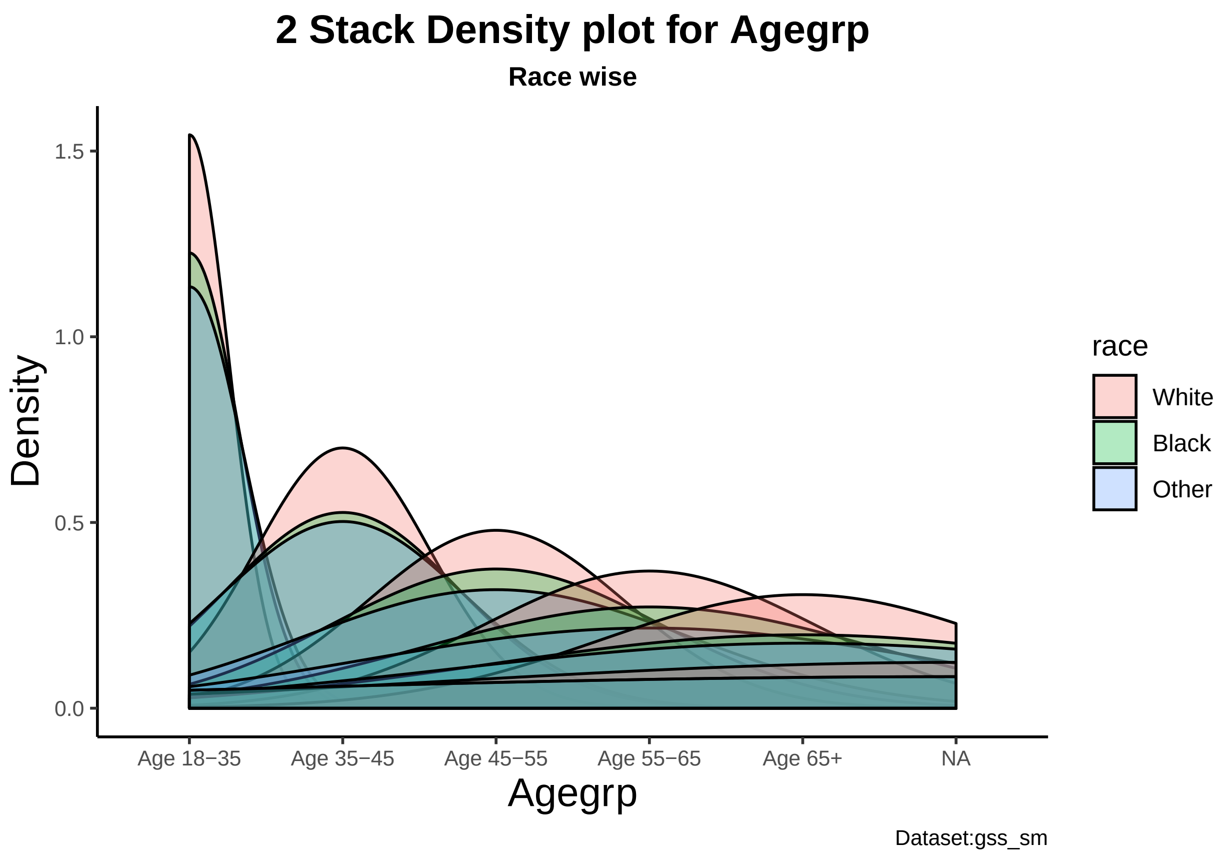
Dataset:gss_sm

III. 2 Stacked Density plots, dividing the observation based on the variable race: – In x-axis, choose variable age – In x-axis, choose variable agegrp. Discuss which variable among the two age and agegrp, is more appropriate for plots like a histogram and density plots and why?

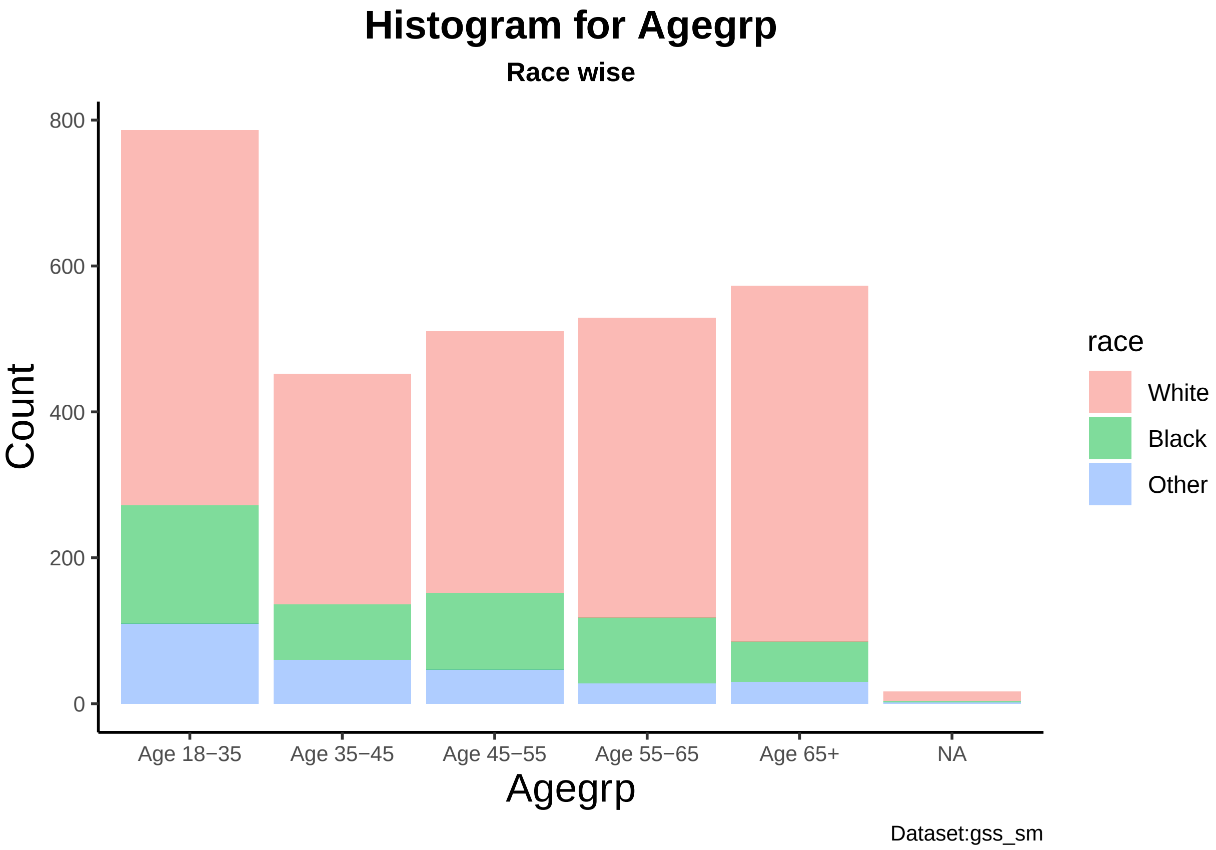
```
p <- ggplot(data = gss_sm,
            mapping = aes(x = age, fill=race))
p + geom_density(alpha = 0.3)+
  theme_classic()+
  theme+
  labs(x="Age", y="Density",
       title="2 Stack Density plot for Age",
       caption="Dataset:gss_sm")
```



```
p <- ggplot(data = gss_sm,
            mapping = aes(x = agegrp, fill=race))
p + geom_density(alpha = 0.3)+
  theme_classic()+
  theme+
  labs(x="Agegrp", y="Density",
       title="2 Stack Density plot for Agegrp",
       subtitle = "Race wise",
       caption="Dataset:gss_sm")
```



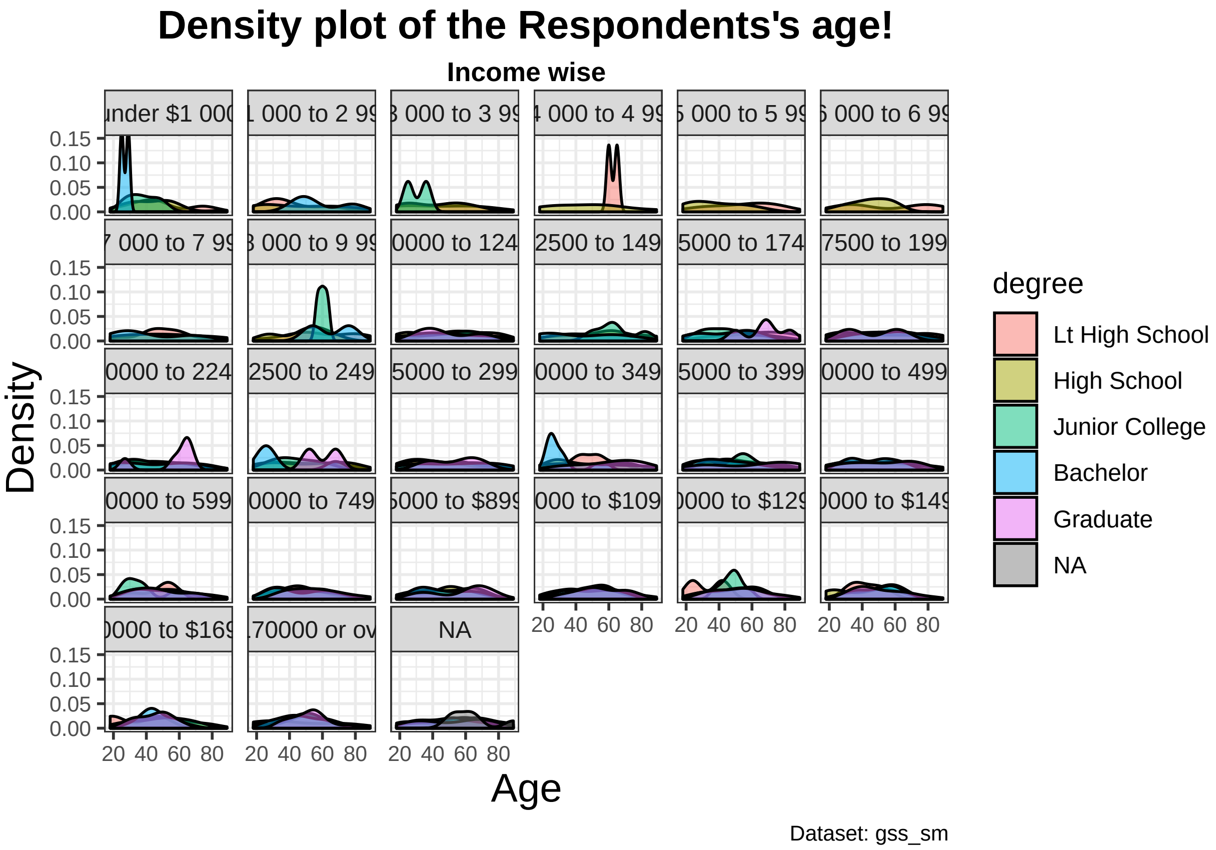
```
p <- ggplot(data = gss_sm,
            mapping = aes(x = agegrp, fill=race))
p + geom_histogram(alpha = 0.5,stat="count")+
  theme_classic()+
  theme+
  labs(x="Agegrp", y="Count",
       title="Histogram for Agegrp",
       subtitle = "Race wise",
       caption="Dataset:gss_sm")
```



Agegrp is a discrete variable. It serves as a suitable variable for the histogram, while when plotted as a density plot , it doesnot give discriminative observation.

IV. Density plot – x-axis: age. – Division based on variable degree. – Faceting based on variable income16.

```
ggplot(data = gss_sm,mapping = aes(x = age )) +
  geom_density(aes(fill=degree),alpha=0.5)+
  facet_wrap(~income16)+
  theme_bw()+
  theme+
  labs(x="Age", y="Density",
       title="Density plot of the Respondents's age!",
       caption="Dataset: gss_sm",
       subtitle="Income wise")
```


Density plot of the Respondents's age!
Income wise
ınder $1 000
1 000 to 2 99
3 000 to 3 99
4 000 to 4 99
5 000 to 5 99
6 000 to 6 99
7 000 to 7 99
8 000 to 9 99
0000 to 124
2500 to 149
5000 to 174
7500 to 199
0000 to 224
2500 to 249
5000 to 299
0000 to 349
5000 to 399
0000 to 499
0000 to 599
0000 to 749
5000 to $899
000 to $109
0000 to $129
0000 to $149
0000 to $169
70000 or ov
NA
Density
0.15
0.10
0.05
0.00
Age
20 40 60 80
degree
Lt High School
High School
Junior College
Bachelor
Graduate
NA
Dataset: gss_sm