Assignment2

```
library(gapminder)
library(tidyverse)
```

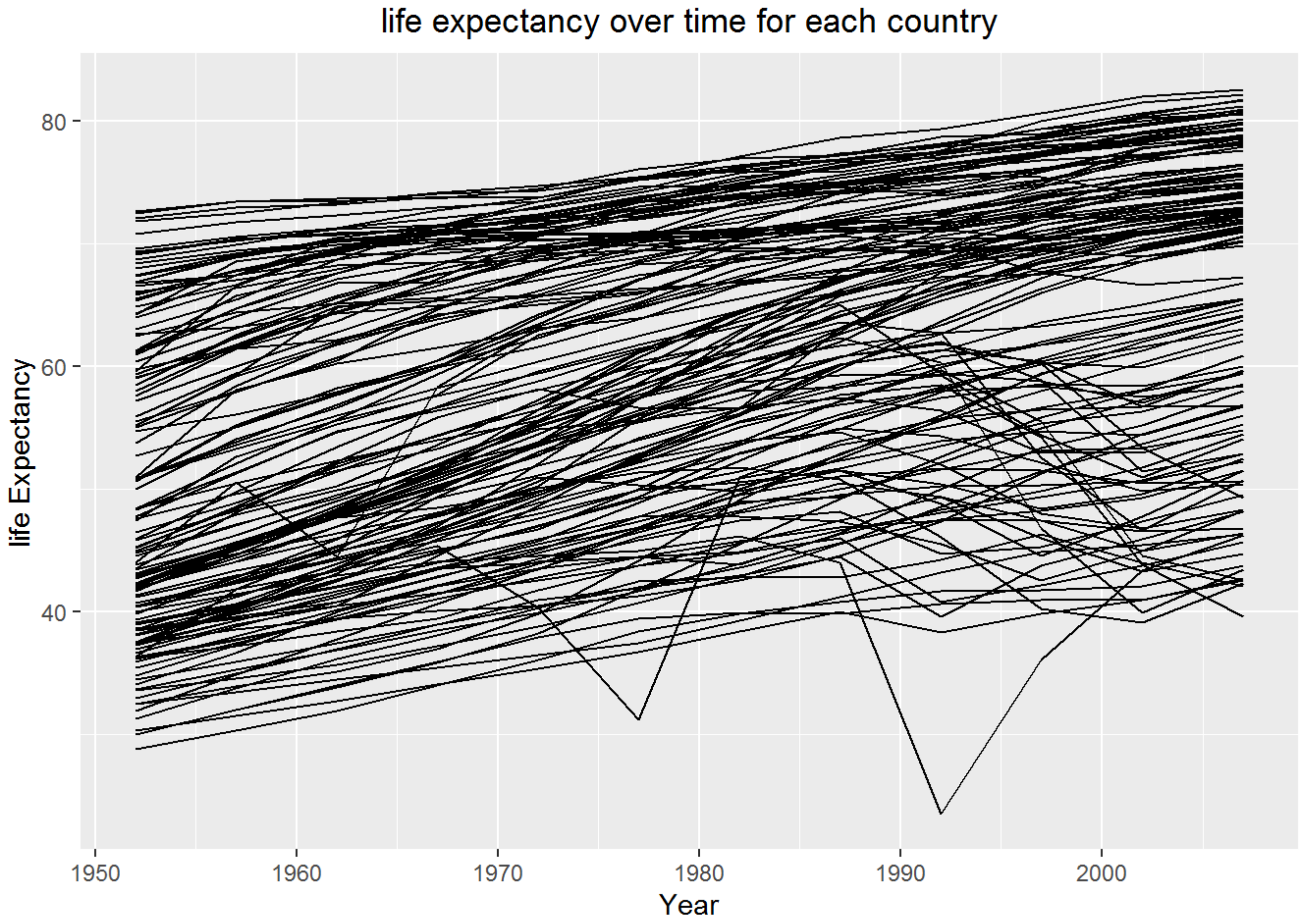
```
## -- Attaching packages ------------------------------------------------------ tidyverse 1.3.0
--
```

```
## <U+2713> ggplot2 3.2.1     <U+2713> purrr   0.3.3
## <U+2713> tibble  2.1.3     <U+2713> dplyr   0.8.3
## <U+2713> tidyr   1.0.2     <U+2713> stringr 1.4.0
## <U+2713> readr   1.3.1     <U+2713> forcats 0.4.0
```

```
## -- Conflicts --------------------------------------------------------- tidyverse_conflicts()
--
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

##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.

```
library(gapminder)
library(ggplot2)

p <- ggplot(data = gapminder,mapping = aes(x = year,y =lifeExp))
p + geom_line(aes(group =country))+labs(x="Year",y="life Expectancy",title="life expectancy over
time for each country")+theme(plot.title = element_text(hjust = 0.50))
```



Ans: No, it is difficult to infer anything. Therefore, there is a need to provide some extra information to graph beforehand, so that it can visualize better. Even though this graph is not much informative, one can see, each line represents the countries life expectancy over the years.

##Question 2 As discussed in section1, faceting the data using extra variable means making small multiple plots for the same data.

1. 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.

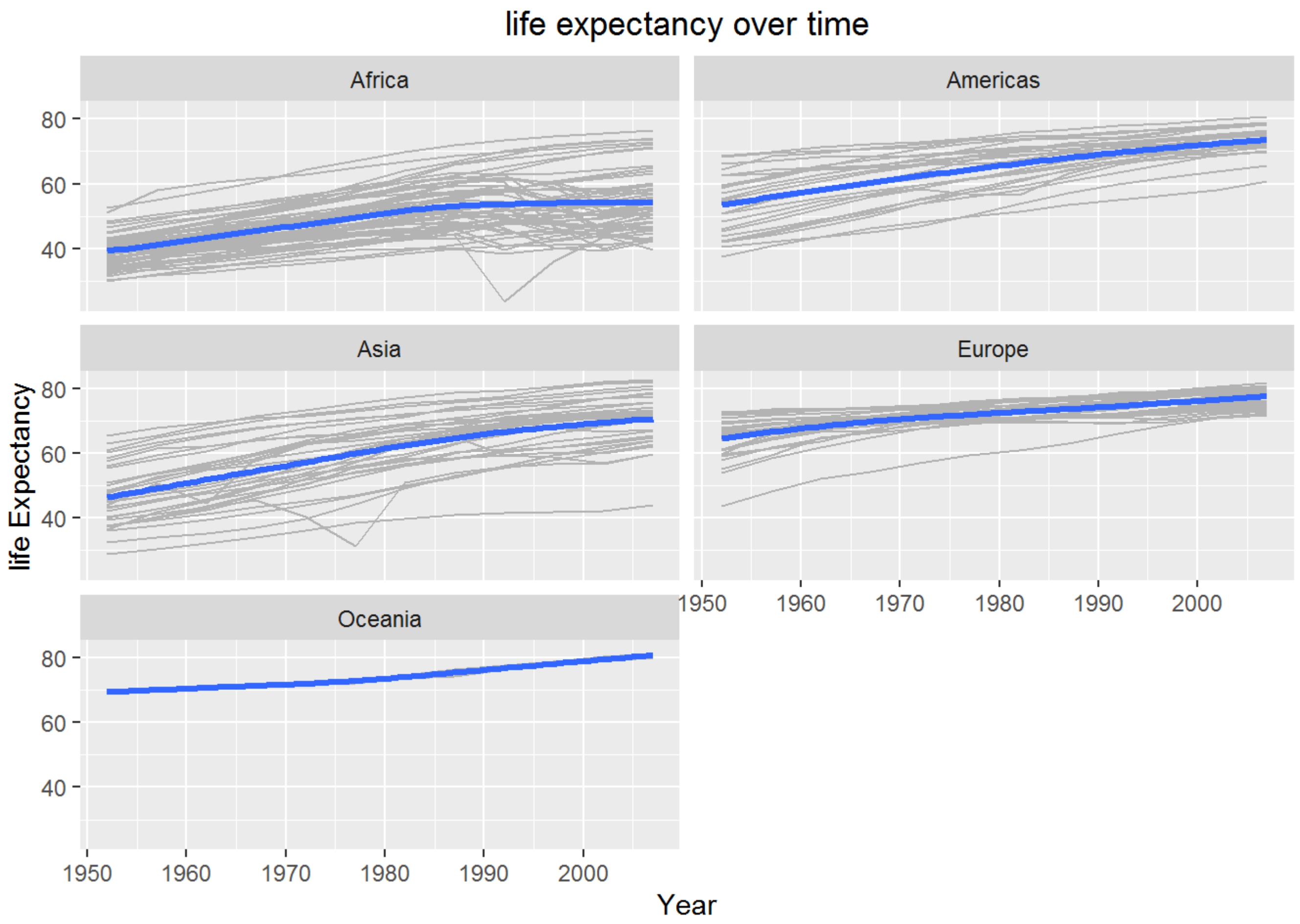
Ans: Continent is the most appropriate variable to facet the data.

Continent is a categorical variable.

If we divide the plot based on the continent then we can easily analyze life expectancy of each country

facet_wrap() split our plot.

```
p <- ggplot(data = gapminder,mapping = aes(x = year,y =lifeExp))

p + geom_line(aes(group = country),color="gray70") +
  facet_wrap(~ continent,nrow=3,ncol=2)+geom_smooth(size=1.1,method="loess",se=FALSE)+
  labs(x="Year",y="life Expectancy",title="life expectancy over time")+
  theme(plot.title = element_text(hjust = 0.50))
```



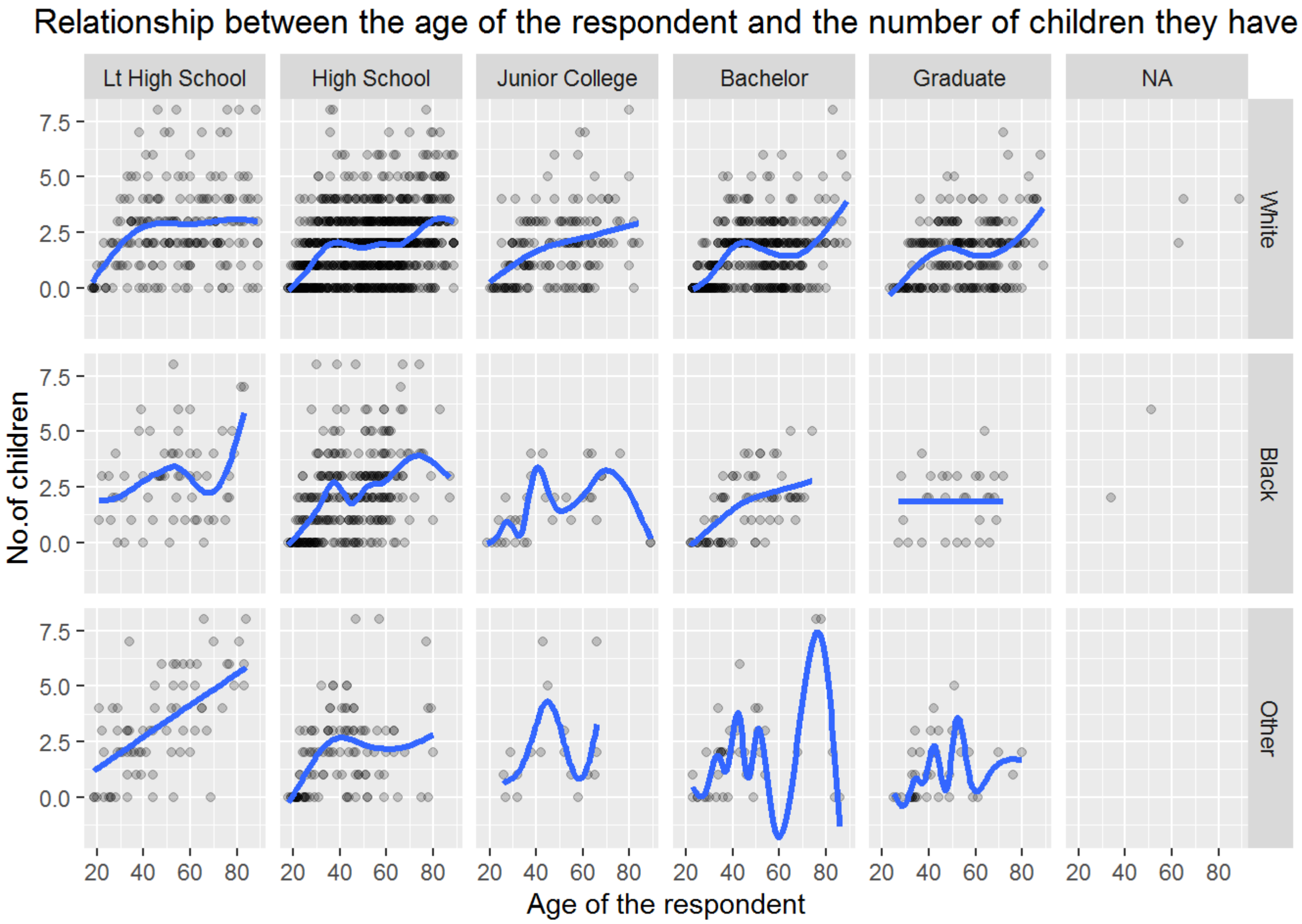
2.We can facet the data based on more than one variable, Use gss sm data to plot a smoothed scatter plot showing the relationship between the age of the respondent and the number of children they have. Facet this relation based on race and degree. Also,in markdown, describe your observation from the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

```
library(socviz)

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

p + geom_point(alpha=0.2) +
    geom_smooth(size=1.1,se=FALSE) +
    facet_grid(race~degree)+labs(x="Age of the respondent",y="No.of children ",title="Relationsh
ip between the age of the respondent and the number of children they have")+
  theme(plot.title = element_text(hjust = 0.50))
```

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

```
## Warning: Removed 18 rows containing non-finite values (stat_smooth).
```

```
## Warning: Computation failed in `stat_smooth()`:
## x has insufficient unique values to support 10 knots: reduce k.

## Warning: Computation failed in `stat_smooth()`:
## x has insufficient unique values to support 10 knots: reduce k.
## Warning: Removed 18 rows containing missing values (geom_point).
```



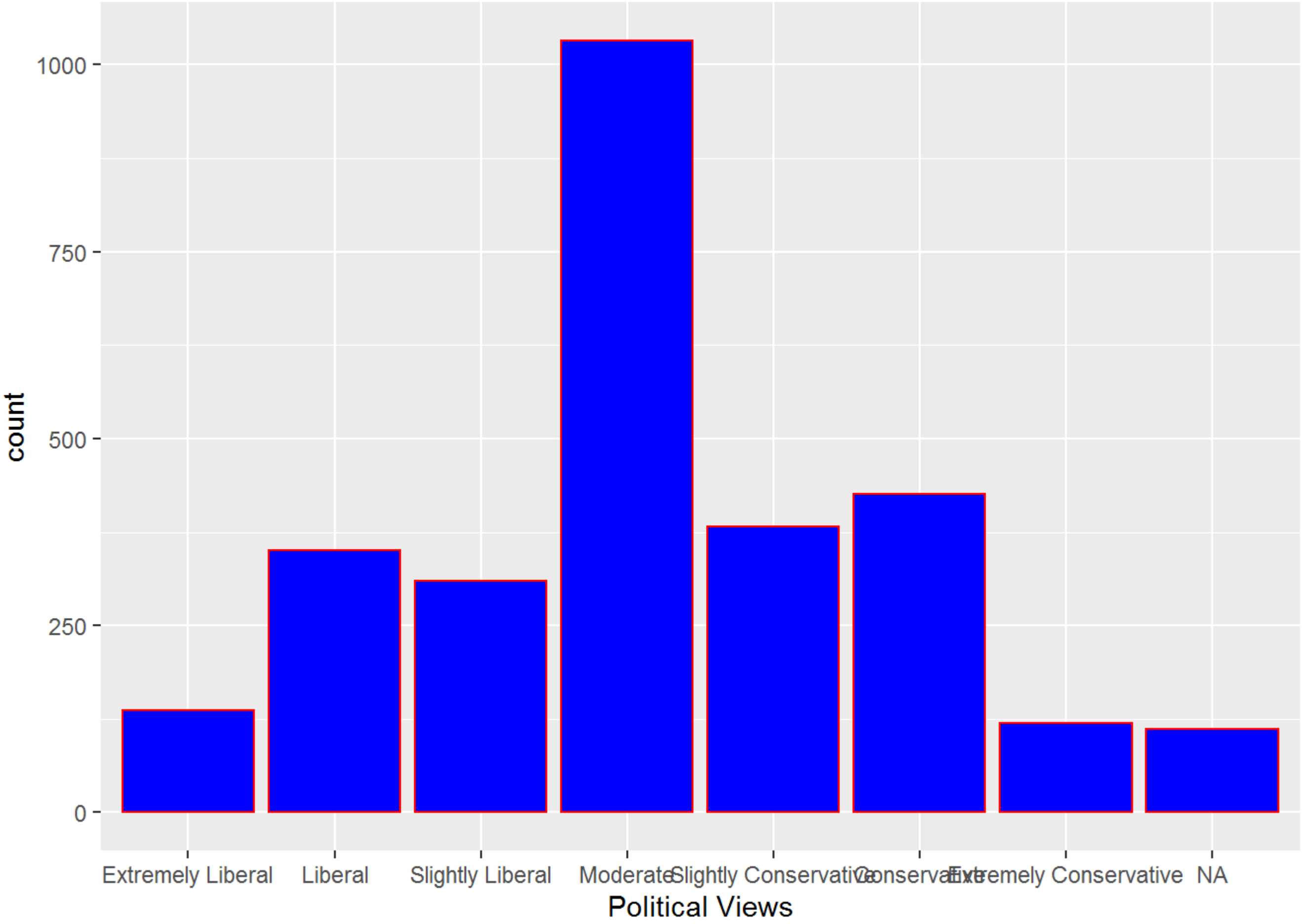
Faceting on two categorical variables. Each panel plots the relationship between age and number of children, with the facets breaking out the data by race (in the rows) and degree (in the columns).

Observation:

1.Number of white respondent are higher then black respondent and other respondent. 2.There are slightly higher no. of respondent who have passed High School. 3.There is very High no. of white people having bachelor or graduate degree as compared to black or other category. 4.

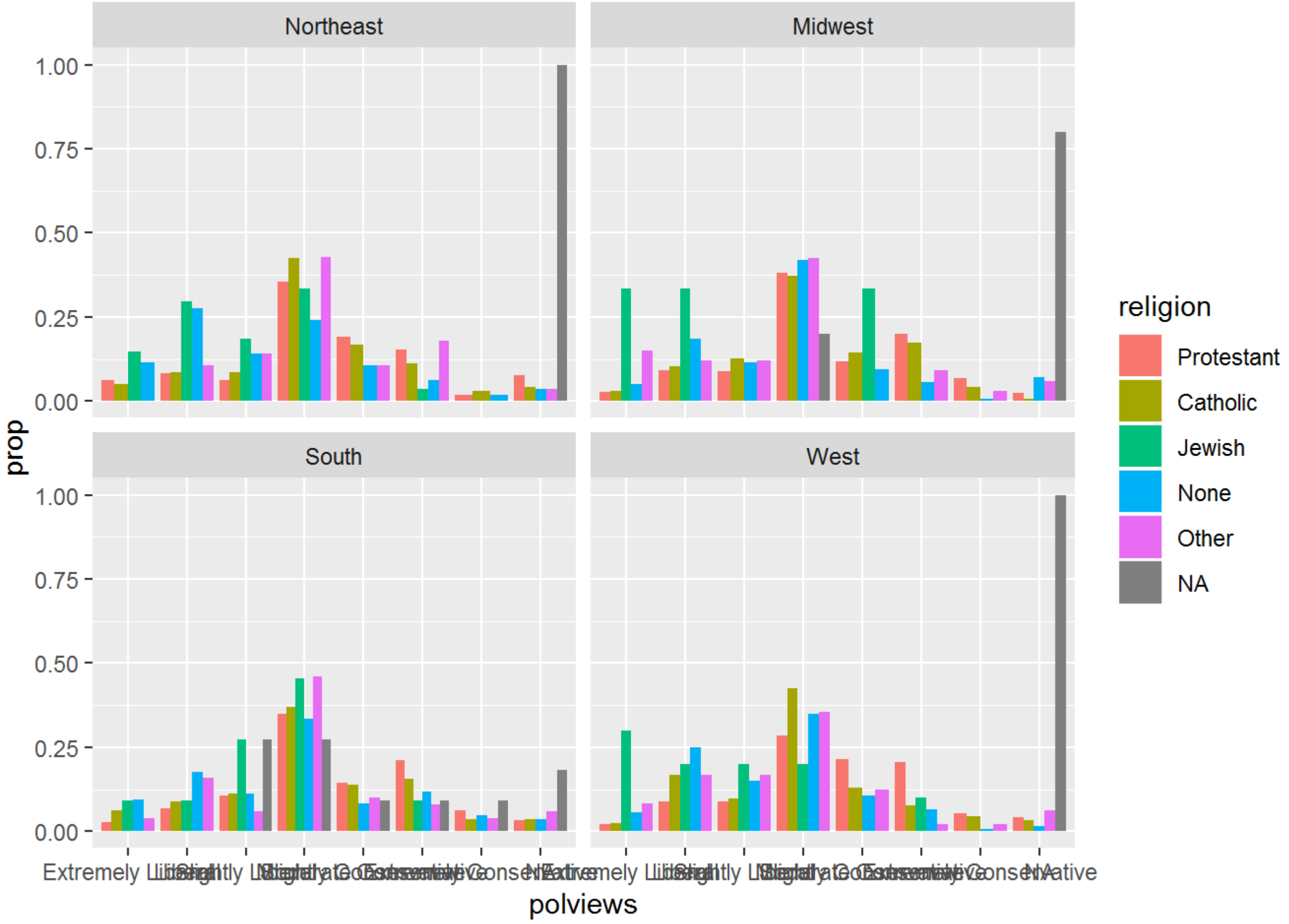
##Question 3 gss_sm data contains the political view (polviews) variable. Plot a bar graph, with the bar in the chart represented by different political views.

```
library(socviz)
ggplot(data=gss_sm)+geom_bar(aes(polviews),fill="blue",col="red")+labs(x="Political Views")
```



##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(position="dodge",mapping = aes(y = ..prop..,group=religion))+
  facet_wrap(~bigregion)
```

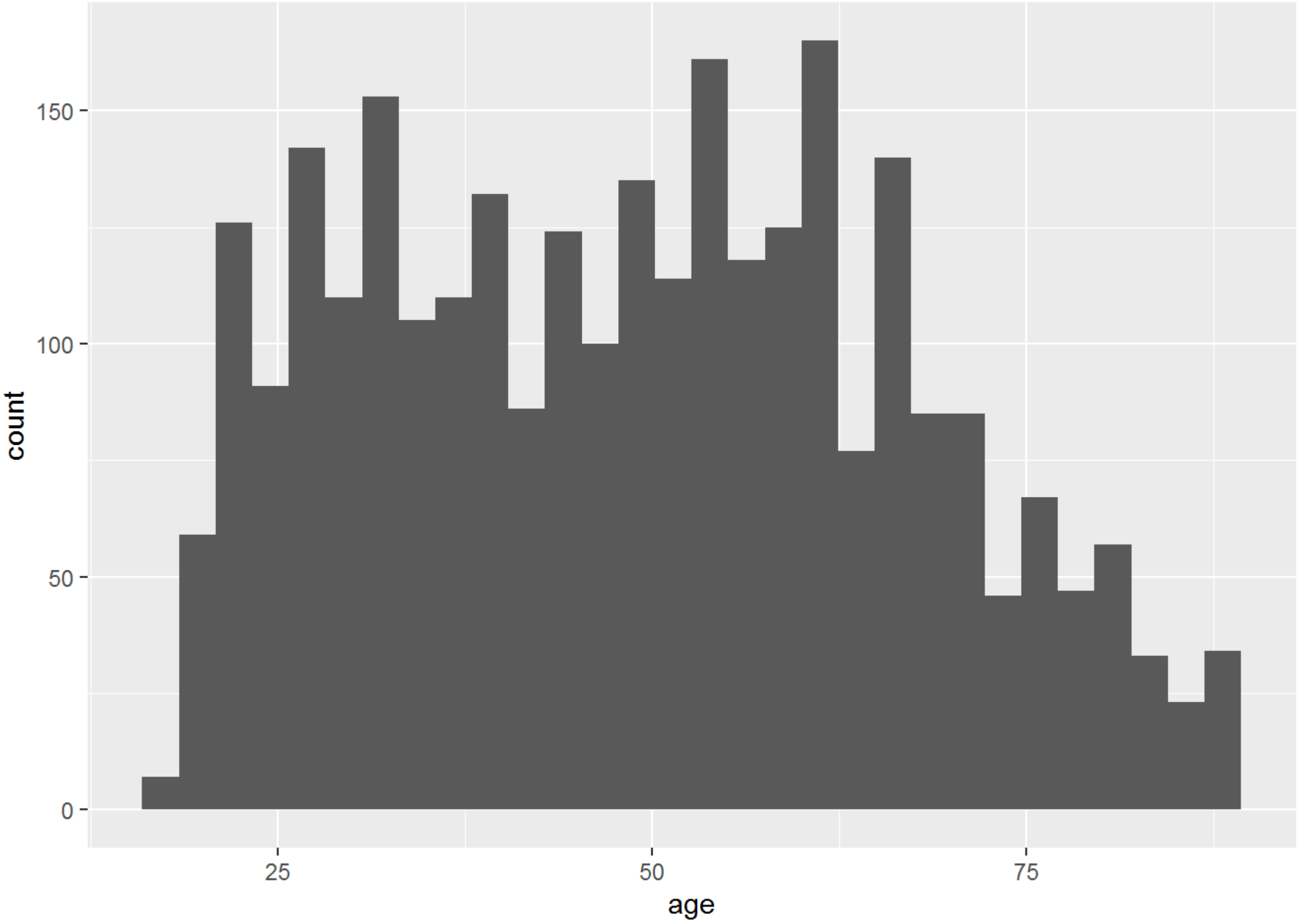


##Question 5 Plot following from the gss_sm data.

1.Histogram. Showcasing ages into bins.

```
ggplot(data=gss_sm)+geom_histogram(aes(age),bins=30)
```

```
## Warning: Removed 10 rows containing non-finite values (stat_bin).
```

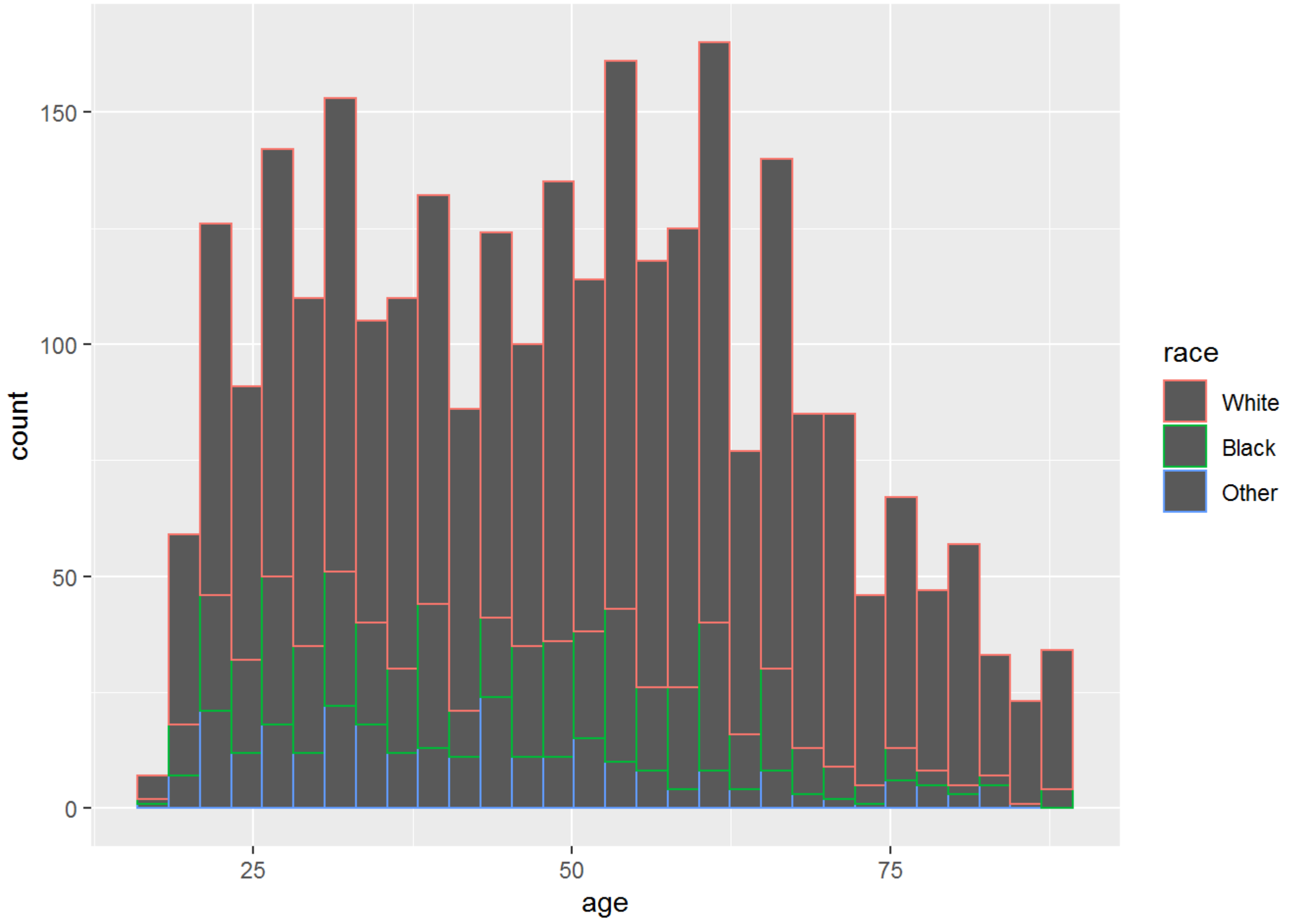


2.Modify the previous histogram, dividing the observation based on variable race. Each race should be represented in a different color.

```
ggplot(data=gss_sm)+geom_histogram(aes(age,col=race))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 10 rows containing non-finite values (stat_bin).
```

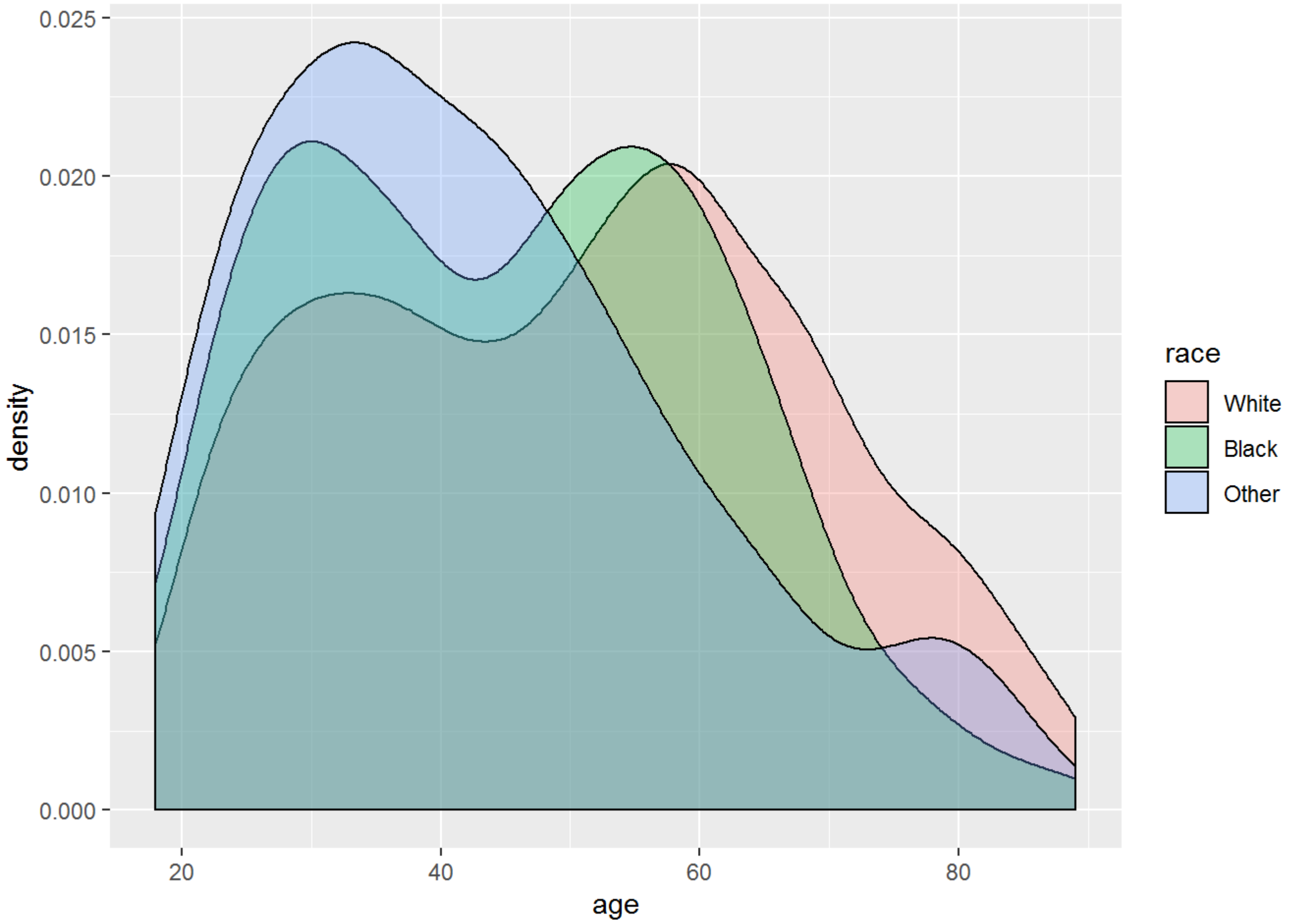


3. Stacked Density plots, dividing the observation based on the variable race: In x-axis, choose variable age In x-axis, choose variable agegrp.

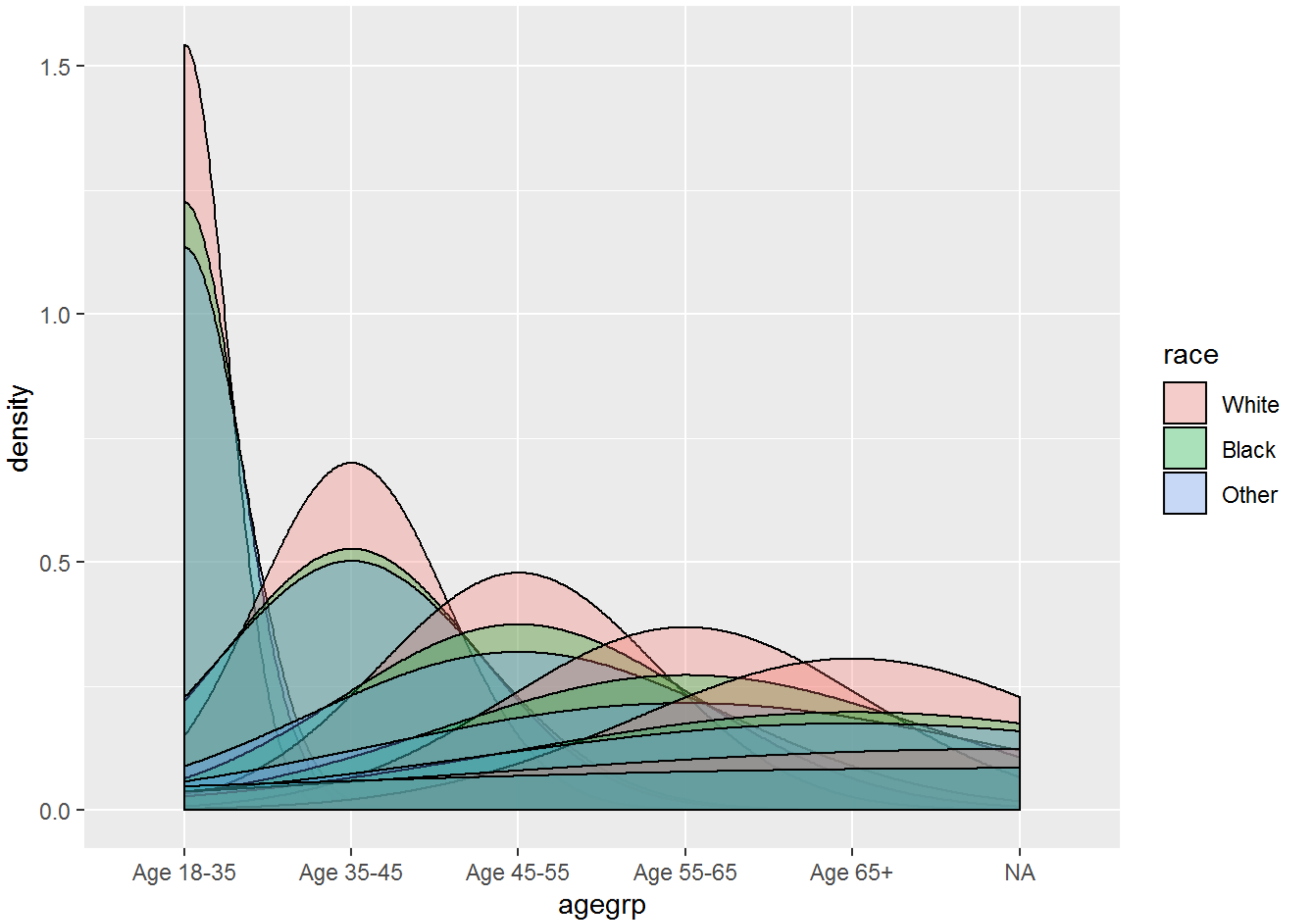
Discuss which variable among the two age and agegrp, is more appropriate for plots like a histogram and density plots and why?

```
#
# p <- ggplot(data=gss_sm,
#              mapping = aes(x =age,fill=race))
# p + geom_histogram(alpha = 0.3)

p <- ggplot(data=gss_sm,
            mapping = aes(x =age,fill=race))
p + geom_density(alpha = 0.3)
```

```
## Warning: Removed 10 rows containing non-finite values (stat_density).
```



```
#
# q <- ggplot(data=gss_sm,
#             mapping = aes(x =agegrp))
# q + geom_histogram(alpha = 0.3)

q <- ggplot(data=gss_sm,
            mapping = aes(x =agegrp,fill=race))
q + geom_density(alpha = 0.3)
```

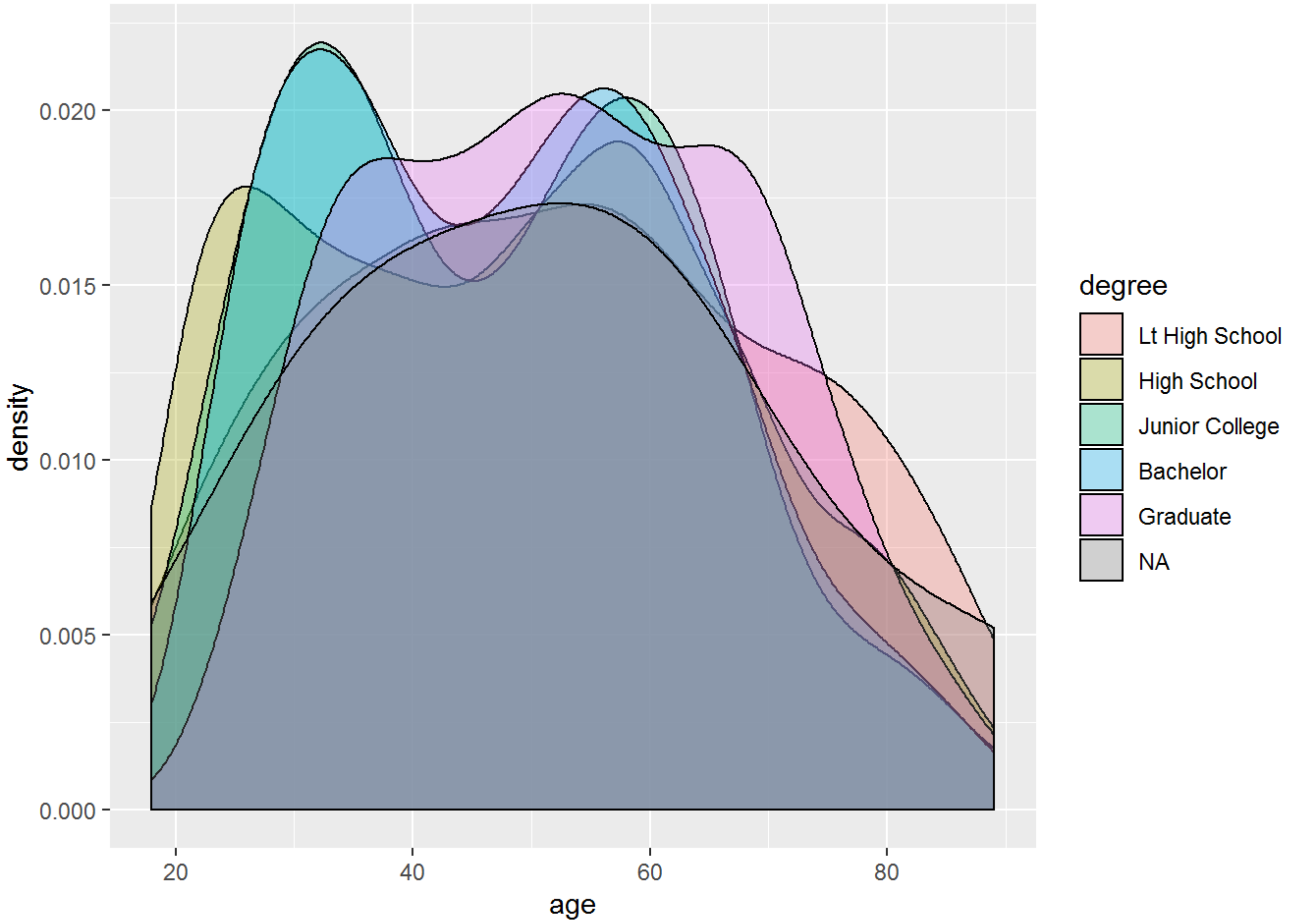


Ans: Variable Agegrp is more appropriate for histogram.becaause agegrp variable is discrete.

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

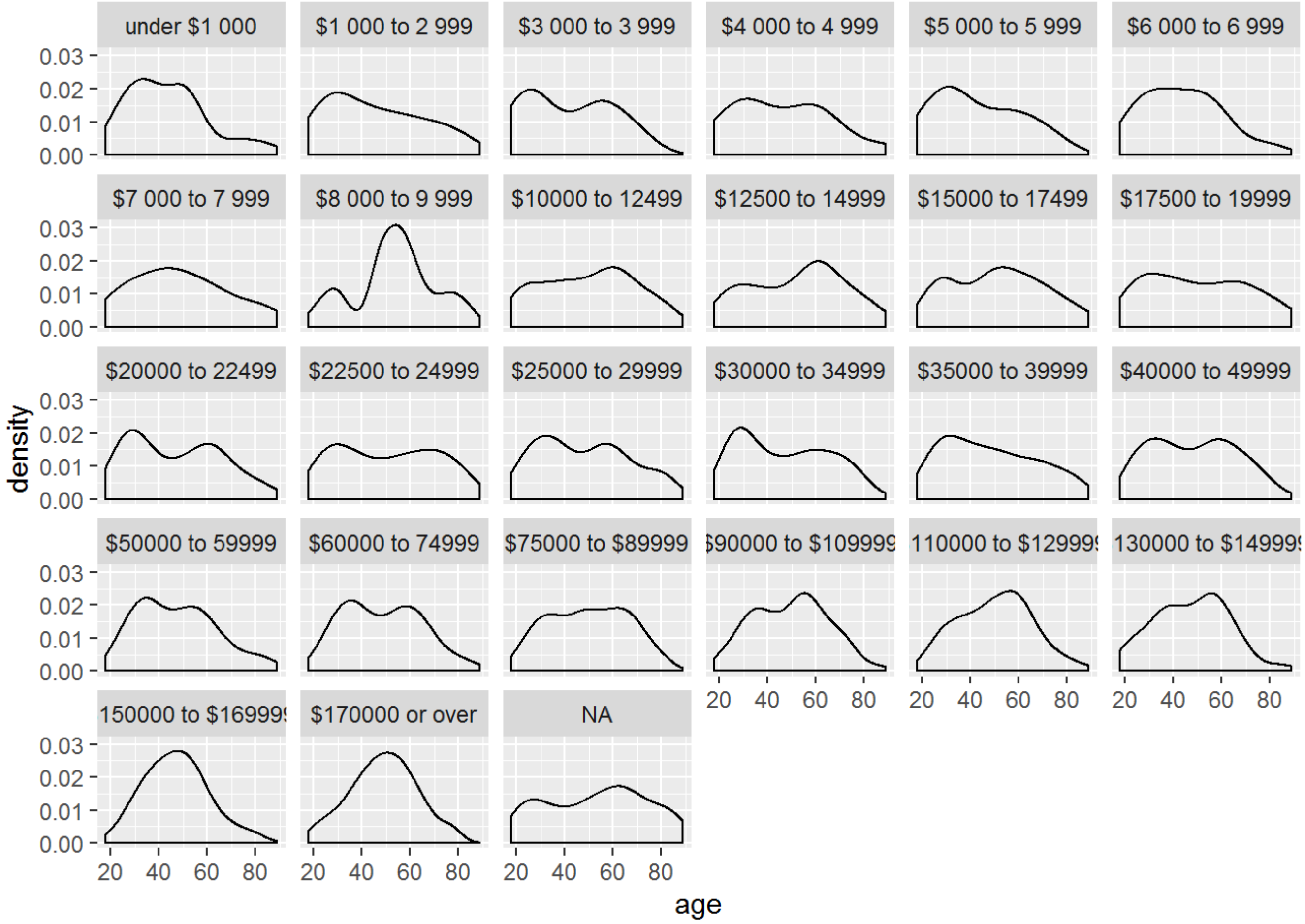
```
p <- ggplot(data=gss_sm,
            mapping = aes(x =age,fill=degree))
p + geom_density(alpha = 0.3)
```

```
## Warning: Removed 10 rows containing non-finite values (stat_density).
```



```
p <- ggplot(data=gss_sm,
            mapping = aes(x =age))
p + geom_density(alpha = 0.3)+facet_wrap(~income16)
```

```
## Warning: Removed 10 rows containing non-finite values (stat_density).
```


under $1 000
$1 000 to 2 999
$3 000 to 3 999
$4 000 to 4 999
$5 000 to 5 999
$6 000 to 6 999
$7 000 to 7 999
$8 000 to 9 999
$10000 to 12499
$12500 to 14999
$15000 to 17499
$17500 to 19999
$20000 to 22499
$22500 to 24999
$25000 to 29999
$30000 to 34999
$35000 to 39999
$40000 to 49999
$50000 to 59999
$60000 to 74999
$75000 to $89999
$90000 to $109999
110000 to $129999
130000 to $149999
150000 to $169999
$170000 or over
NA
density
0.00
0.01
0.02
0.03
age
20
40
60
80