DSCI 610: Introduction to EDA

What is an Exploratory Data Analysis (EDA)?

John W. Tukey, a renown statistician and an early visionary of the field of data science, coined the phrase *exploratory data analysis (EDA)*. EDA is one of his remarkable contributions to statistics and modern day data science. He reenergized descriptive statistics through EDA and changed the language and paradigm of statitsics in doing so. Tukey introduced EDA by describing its characteristics and creating novel tools. Some of his descriptions include the following:

- "Three of the main strategies of data analysis are: 1. graphical presentation. 2. provision of flexibility in viewpoint and in facilities, 3. intensive search for parsimony and simplicity."
- "In exploratory data analysis there can be no substitute for flexibility; for adapting what is calculated—and what we hope plotted—both to the needs of the situation and the clues that the data have already provided."
- "'Exploratory data analysis' is an attitude, a state of flexibility, a willingness to look for those things that we believe are not there, as well as those we believe to be there."
- "Exploratory data analysis isolates patterns and features of the data and reveals these forcefully to the analyst".

More recently Hadley Wickham, the creator of the R package **tidyverse** and others, said that EDA is an iterative cycle of

- generating questions about data,
- searching for answers by visualising, transforming, and modelling data,
- refining questions and/or generating new questions using what is leared from exploring data.

EDA is an important part of data analysis and should be carried out even if we know the study obejctives in terms of hypothesis testing or prediction.

The primary objective of doing an EDA is to develop an understanding of data at hand. Setting a series of relevant questions and investigating accordingly is right approach to understand the data at hand. Quoting John Tukey again,"**Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise**."

In practice, two types of questions are useful for making discoveries during an exploratory data analysis.

1. What type of variation occurs within the variables?
2. What type of covariation occurs between the variables?

A Case Study on EDA: National Health and Nutrition Examination Study (NHANES)

The **NHANES** package contains data from the US National Health and Nutritional Examiation study collected by the US National Center for Health Statistics (NCHS). Use `?NHANES` to get more information after you load the package into R workspace.

The package includes two versions of NHANES data : i) NHANES and ii) NHANESraw, each having 75 variables for the 2009-2010 and 2011-2012 sample years. For this case study we will consider NHANES data that contains 10000 rows of data resampled from NHANESraw.

The following code chunk loads the packages `NHANES` and `tidyverse` with `library()` function. Then the `as_tibble()` function coerce the dataframe `NHANES` into a tibble that has some desirable properties including printing only first 10 rows and limited number of columns to fit the screen.

```
library(NHANES)
library(tidyverse)
```

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

```
## v ggplot2 3.2.1     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.4
## v tidyr   1.0.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

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

```
as_tibble(NHANES)
```

```
## # A tibble: 10,000 x 76
##       ID SurveyYr Gender   Age AgeDecade AgeMonths Race1 Race3 Education
##    <int> <fct>    <fct>  <int> <fct>         <int> <fct> <fct> <fct>
##  1 51624 2009_10  male      34 " 30-39"        409 White <NA>  High Sch~
##  2 51624 2009_10  male      34 " 30-39"        409 White <NA>  High Sch~
##  3 51624 2009_10  male      34 " 30-39"        409 White <NA>  High Sch~
##  4 51625 2009_10  male       4 " 0-9"           49 Other <NA>  <NA>
##  5 51630 2009_10  female    49 " 40-49"        596 White <NA>  Some Col~
##  6 51638 2009_10  male       9 " 0-9"          115 White <NA>  <NA>
##  7 51646 2009_10  male       8 " 0-9"          101 White <NA>  <NA>
##  8 51647 2009_10  female    45 " 40-49"        541 White <NA>  College ~
##  9 51647 2009_10  female    45 " 40-49"        541 White <NA>  College ~
## 10 51647 2009_10  female    45 " 40-49"        541 White <NA>  College ~
## # ... with 9,990 more rows, and 67 more variables: MaritalStatus <fct>,
## #   HHIncome <fct>, HHIncomeMid <int>, Poverty <dbl>, HomeRooms <int>,
## #   HomeOwn <fct>, Work <fct>, Weight <dbl>, Length <dbl>, HeadCirc <dbl>,
## #   Height <dbl>, BMI <dbl>, BMICatUnder20yrs <fct>, BMI_WHO <fct>,
## #   Pulse <int>, BPSysAve <int>, BPDiaAve <int>, BPSys1 <int>, BPDia1 <int>,
## #   BPSys2 <int>, BPDia2 <int>, BPSys3 <int>, BPDia3 <int>, Testosterone <dbl>,
## #   DirectChol <dbl>, TotChol <dbl>, UrineVol1 <int>, UrineFlow1 <dbl>,
## #   UrineVol2 <int>, UrineFlow2 <dbl>, Diabetes <fct>, DiabetesAge <int>,
## #   HealthGen <fct>, DaysPhysHlthBad <int>, DaysMentHlthBad <int>,
## #   LittleInterest <fct>, Depressed <fct>, nPregnancies <int>, nBabies <int>,
## #   Age1stBaby <int>, SleepHrsNight <int>, SleepTrouble <fct>,
```

```
## #   PhysActive <fct>, PhysActiveDays <int>, TVHrsDay <fct>, CompHrsDay <fct>,
## #   TVHrsDayChild <int>, CompHrsDayChild <int>, Alcohol12PlusYr <fct>,
## #   AlcoholDay <int>, AlcoholYear <int>, SmokeNow <fct>, Smoke100 <fct>,
## #   Smoke100n <fct>, SmokeAge <int>, Marijuana <fct>, AgeFirstMarij <int>,
## #   RegularMarij <fct>, AgeRegMarij <int>, HardDrugs <fct>, SexEver <fct>,
## #   SexAge <int>, SexNumPartnLife <int>, SexNumPartYear <int>, SameSex <fct>,
## #   SexOrientation <fct>, PregnantNow <fct>
```

We will create an analysis dataset by selecting only the variables we are interesd in. Our exploratory data analysis will be based on the new data frame `df_eda`.

```
df_eda <- select(NHANES, ID, SurveyYr, Gender, Age, Race1, Poverty, HomeOwn, Weight, Height, BMI, BPSys
df_eda
```

```
## # A tibble: 10,000 x 15
##       ID SurveyYr Gender   Age Race1 Poverty HomeOwn Weight Height   BMI
##    <int> <fct>    <fct>  <int> <fct>   <dbl> <fct>    <dbl>  <dbl> <dbl>
##  1 51624 2009_10  male      34 White    1.36 Own       87.4   165.  32.2
##  2 51624 2009_10  male      34 White    1.36 Own       87.4   165.  32.2
##  3 51624 2009_10  male      34 White    1.36 Own       87.4   165.  32.2
##  4 51625 2009_10  male       4 Other    1.07 Own       17     105.  15.3
##  5 51630 2009_10  female    49 White    1.91 Rent      86.7   168.  30.6
##  6 51638 2009_10  male       9 White    1.84 Rent      29.8   133.  16.8
##  7 51646 2009_10  male       8 White    2.33 Own       35.2   131.  20.6
##  8 51647 2009_10  female    45 White    5    Own       75.7   167.  27.2
##  9 51647 2009_10  female    45 White    5    Own       75.7   167.  27.2
## 10 51647 2009_10  female    45 White    5    Own       75.7   167.  27.2
## # ... with 9,990 more rows, and 5 more variables: BPSysAve <int>,
## #   BPDiaAve <int>, TotChol <dbl>, Diabetes <fct>, SmokeNow <fct>
```

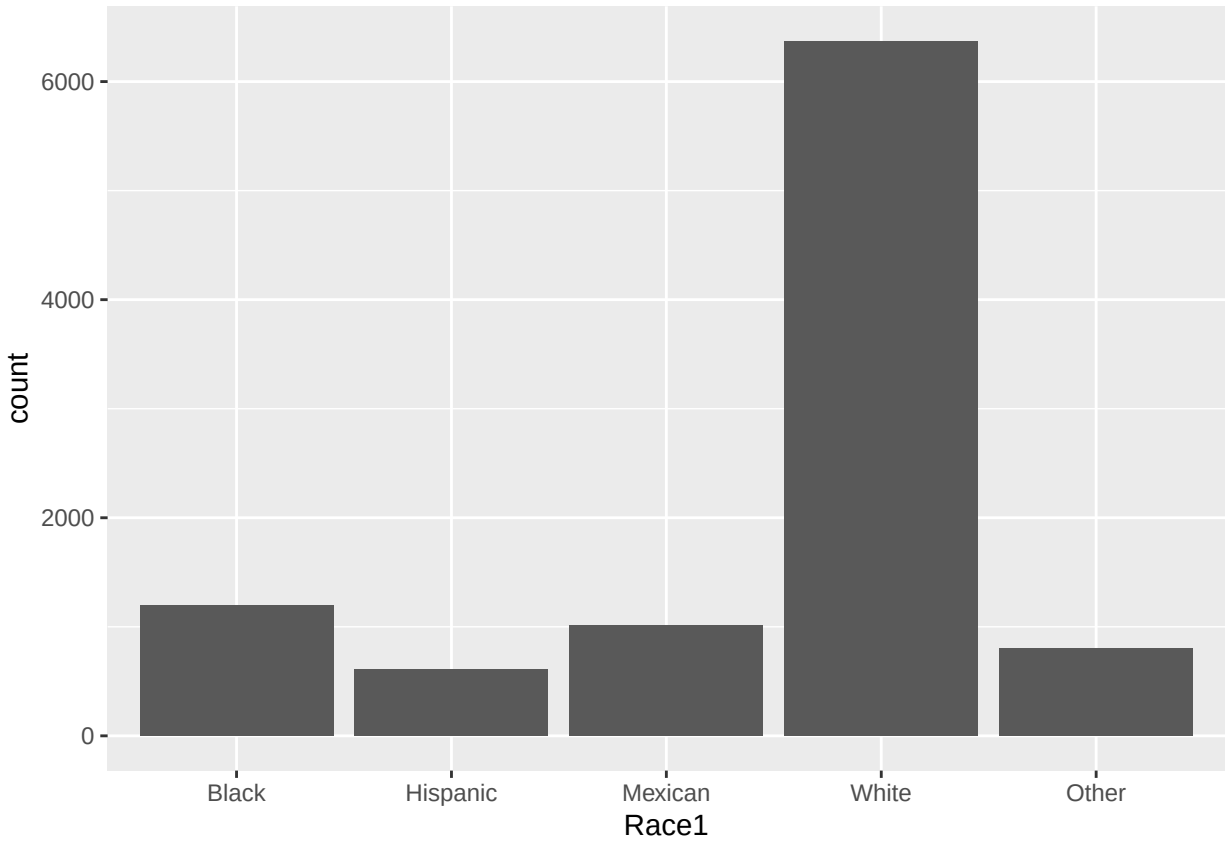
Visualising variation within categorical variables

To visualize variation within a single variable, you can inspect the distribution of the variable graphically. However, visualising the distribution of a variable depends on whether the variable is categorical or continuous. A variable is categorical if it can only take one of a small set of values.

Visualising variation within a single categorical variable

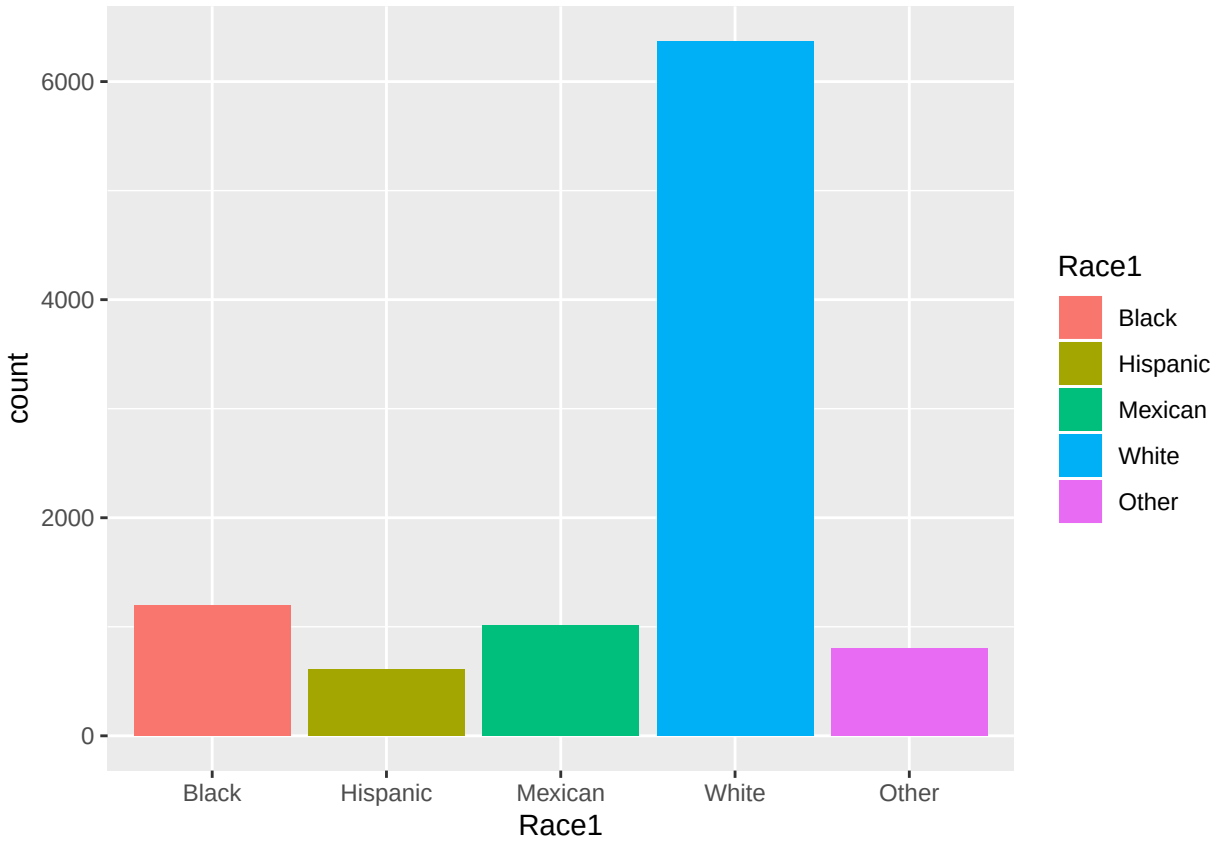
In R, categorical variables are usually saved as factors or character vectors. To examine the distribution of a categorical variable, use a bar chart. A basic bar chart can be created with `geom_bar()` function from the package `ggplot2` which is integrated into the `tidyverse` package.

```
ggplot(data=df_eda)+
  geom_bar(mapping = aes(x=Race1))
```



The bar chart shows the total number of participants in the `df_eda` dataset grouped by their race. As expected, white is the dominant race among the participants. You can color the bars using another aesthetic `fill` as follows.

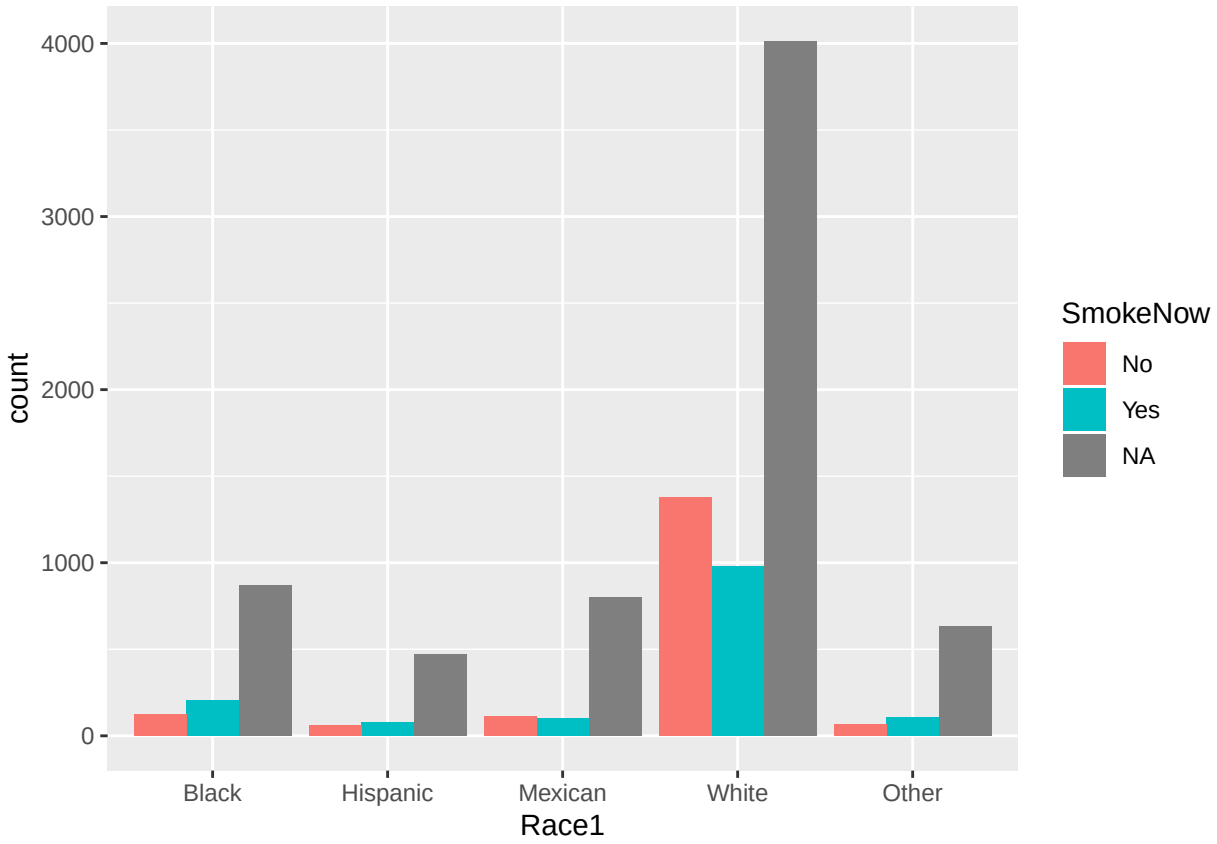
```
ggplot(data=df_eda)+
  geom_bar(mapping = aes(x=Race1, fill=Race1))
```



Visualising covariation between two categorical variables

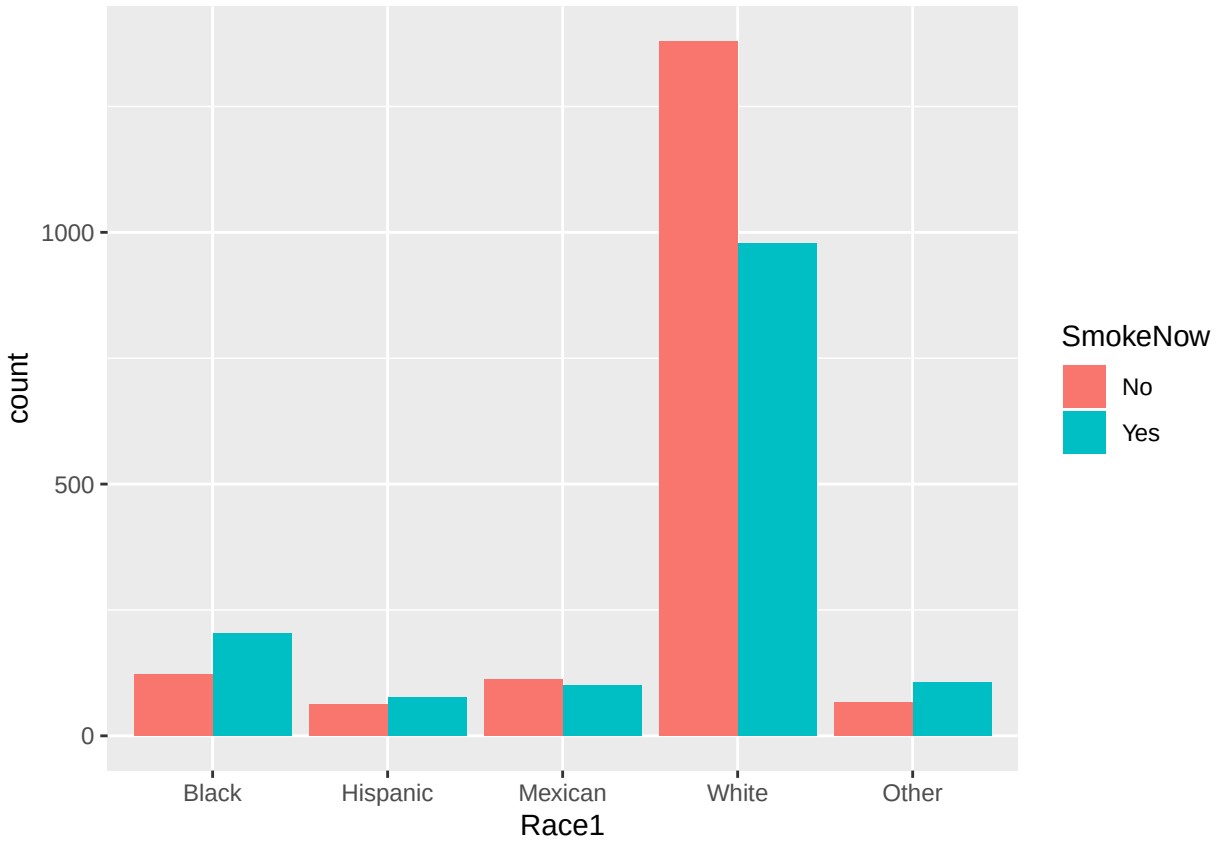
To investigate covariation between two categorical variables, you can use the `fill` aesthetic with `position` adjustment. The following code chunk creates a bar chart for race by the smoking status of the participants.

```
ggplot(data=df_eda)+
  geom_bar(mapping = aes(x=Race1, fill=SmokeNow), position="dodge")
```



Note the variable `SmokeNow` has many missing values labeled as `NA (not available)`. We can use the `filter()` function to remove the missing values from the `SmokeNow` variable.

```
df_eda <- filter(df_eda, SmokeNow !="NA")
#table(df_eda$SmokeNow)
ggplot(data=df_eda)+
  geom_bar(mapping = aes(x=Race1, fill=SmokeNow), position="dodge")
```



You can see that the number of participants not smoking is higher among the `whites` followed by the `Mexicans`.

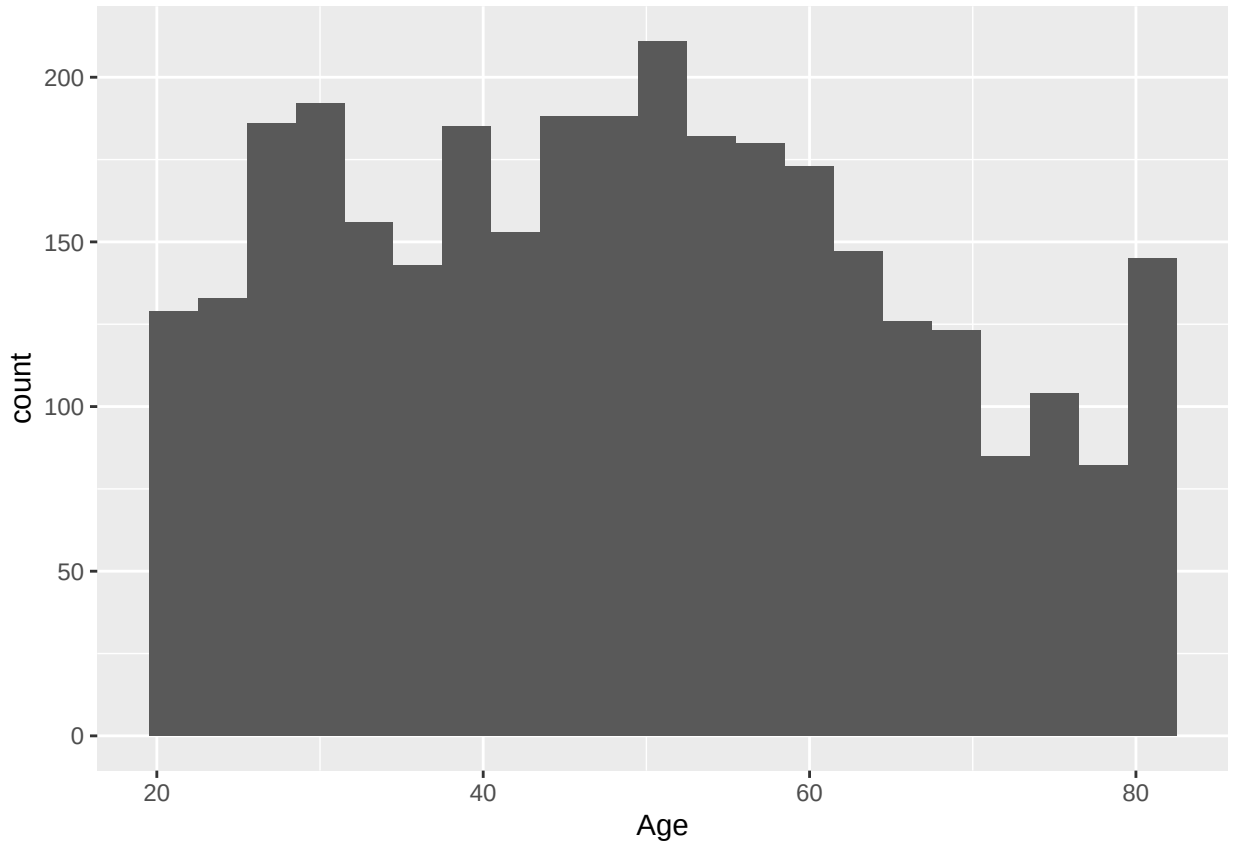
Visualising variation within continuous variables

A variable is continuous if it can take any of an infinite set of ordered values. In the `df_eda` data set there are a number of continuous variables including `Weight`, `Hieght`, `Age`, `BPSysAve`, `BPDiaAve`, `TotChol` etc.

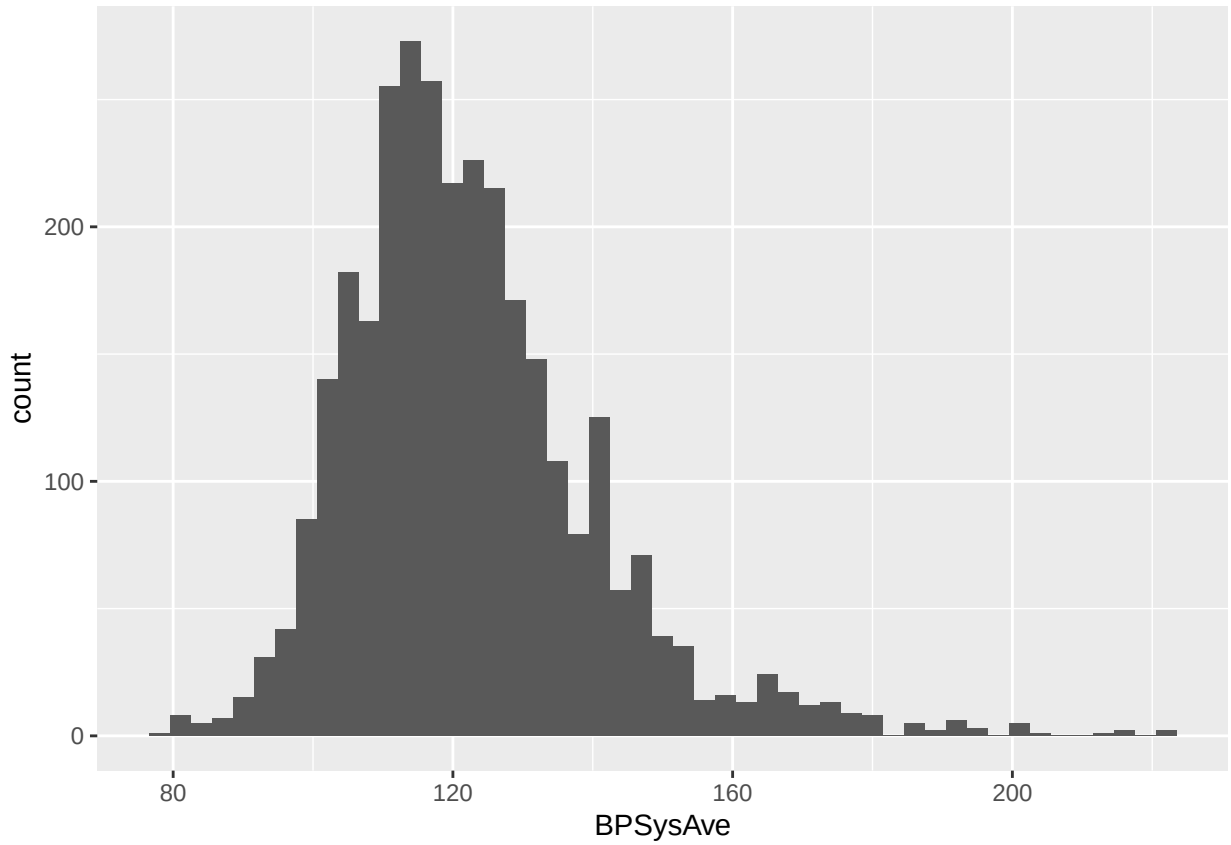
Visualising variation within a single contiuous variable

To examine the distribution of a continuous variable, you can use a histogram, a box-plot or a dotplot. A histogram divides the x-axis into equally spaced bins and then uses the height of a bar to display the number of observations that fall in each bin. You can set the width of the intervals in a histogram with `binwidth` argument, which is measured in the units of the x variable.

```
ggplot(data = df_eda) +
  geom_histogram(mapping = aes(x = Age), binwidth = 3 )
```



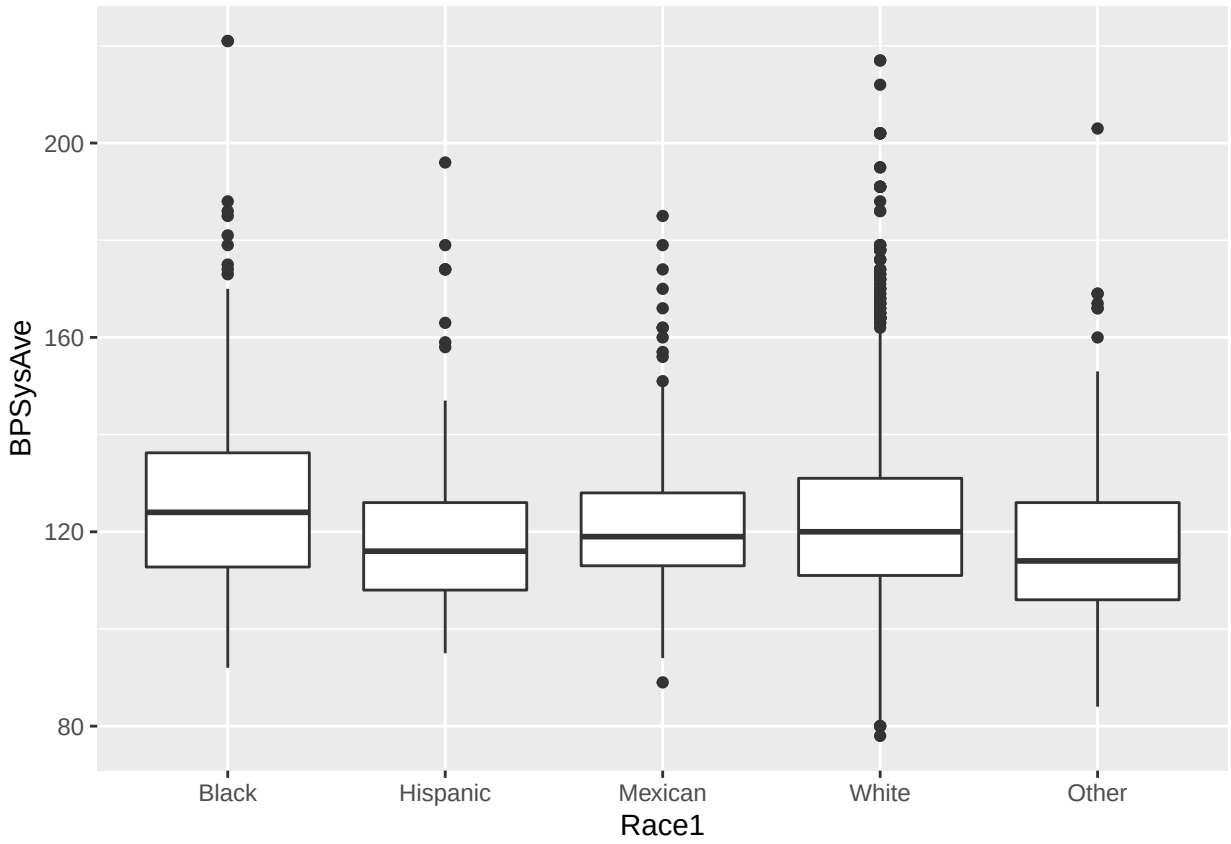

```
df_eda <- filter(df_eda, BPSysAve !="NA")
ggplot(data = df_eda) +
  geom_histogram(mapping = aes(x = BPSysAve), binwidth = 3 )
```



Visualising a categorical and a continuous variable

It is common to explore the distribution of a continuous variable broken down by a categorical variable. For example, we can explore how the average systolic blood pressure varies with participants' race. To display the distribution of a continuous variable broken down by a categorical variable you can use the boxplot.

```
ggplot(data = df_eda, mapping = aes(x = Race1, y = BPSysAve)) + geom_boxplot()
```

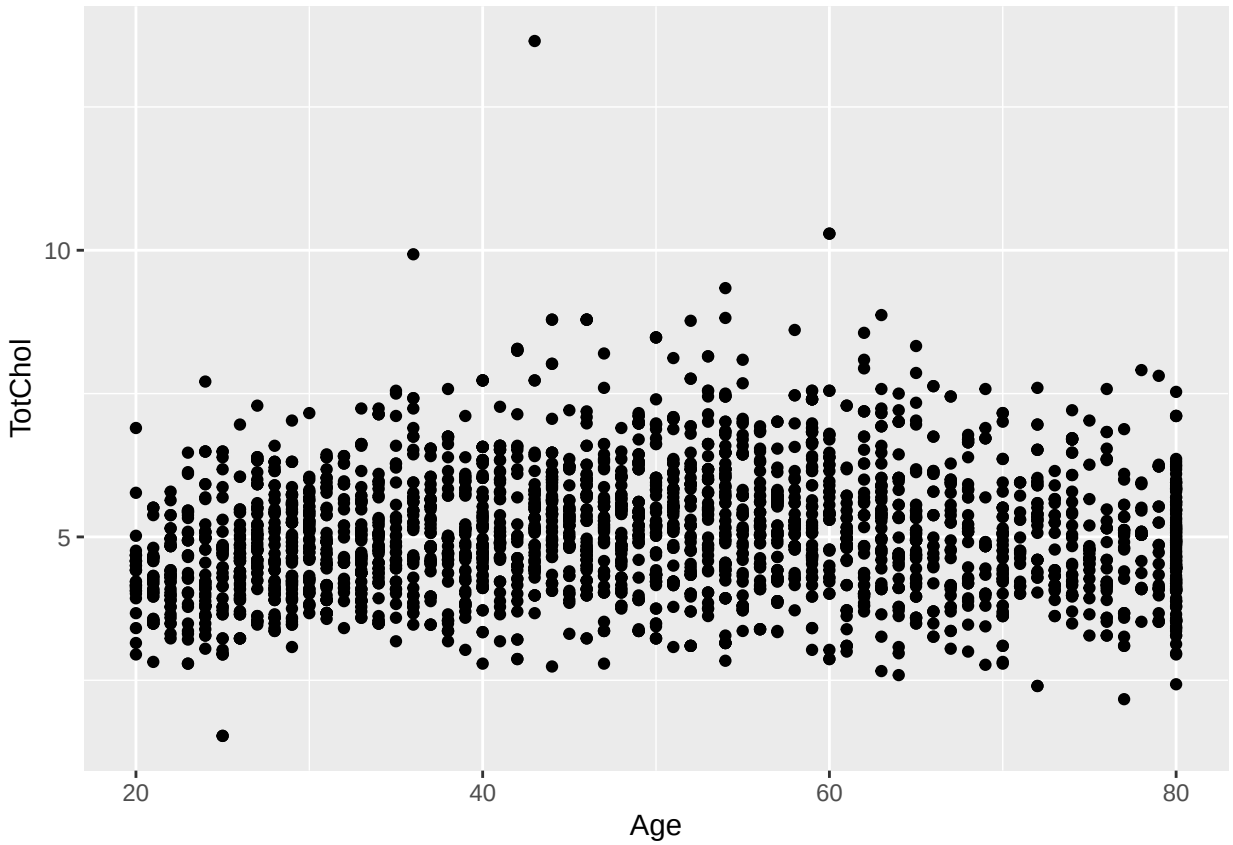


You can see on average black participants have higher systolic blood pressure than the rest of the races.

Visualising two continuous variables

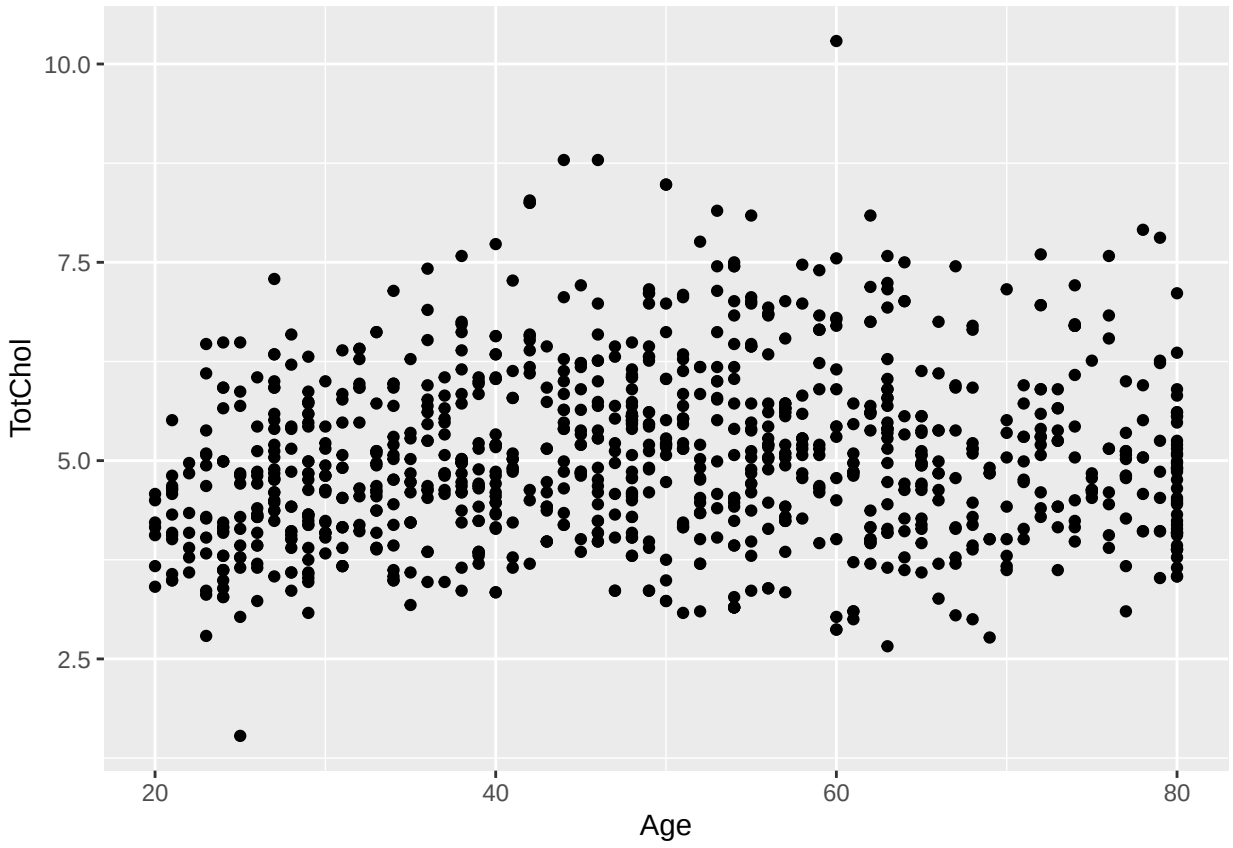
To visualise the covariation between two continuous variables you can draw a scatterplot with the function geom_point(). For example you may want to know how total cholesterol levels of the participants vary with respect to their age.

```
df_eda <- filter(df_eda, TotChol !="NA")
ggplot(data = df_eda) +
  geom_point(mapping = aes(x = Age, y = TotChol))
```



The scatterplot is nott very interesting as most of the values are within somewhat similar range across the ages. You can take a sample of 1000 rows from the data and check for pattern if any.

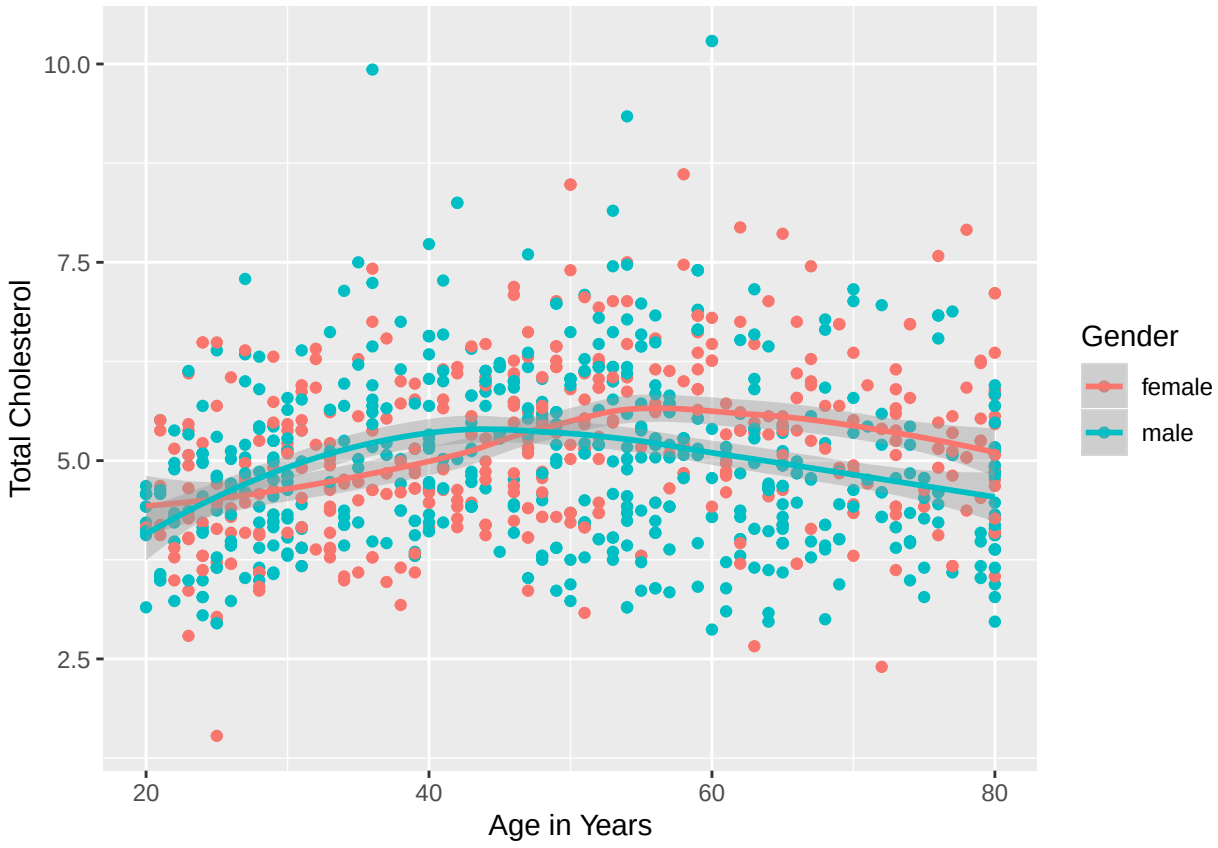
```
ggplot(data = sample_n(df_eda, size=1000), aes(x = Age, y = TotChol))+
             geom_point()
```



The pattern in cholesterol distribution from 1000 random participants across age is similar to that of all 10,000 in the `df_eda` dataset. You can investigate if there is any differences in the total cholesterol patterns over age for men and women.

```
ggplot(data = sample_n(df_eda, size=1000), aes(x = Age, y = TotChol, color=Gender))+
              geom_point() + geom_smooth() + xlab("Age in Years") + ylab("Total Cholesterol")
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



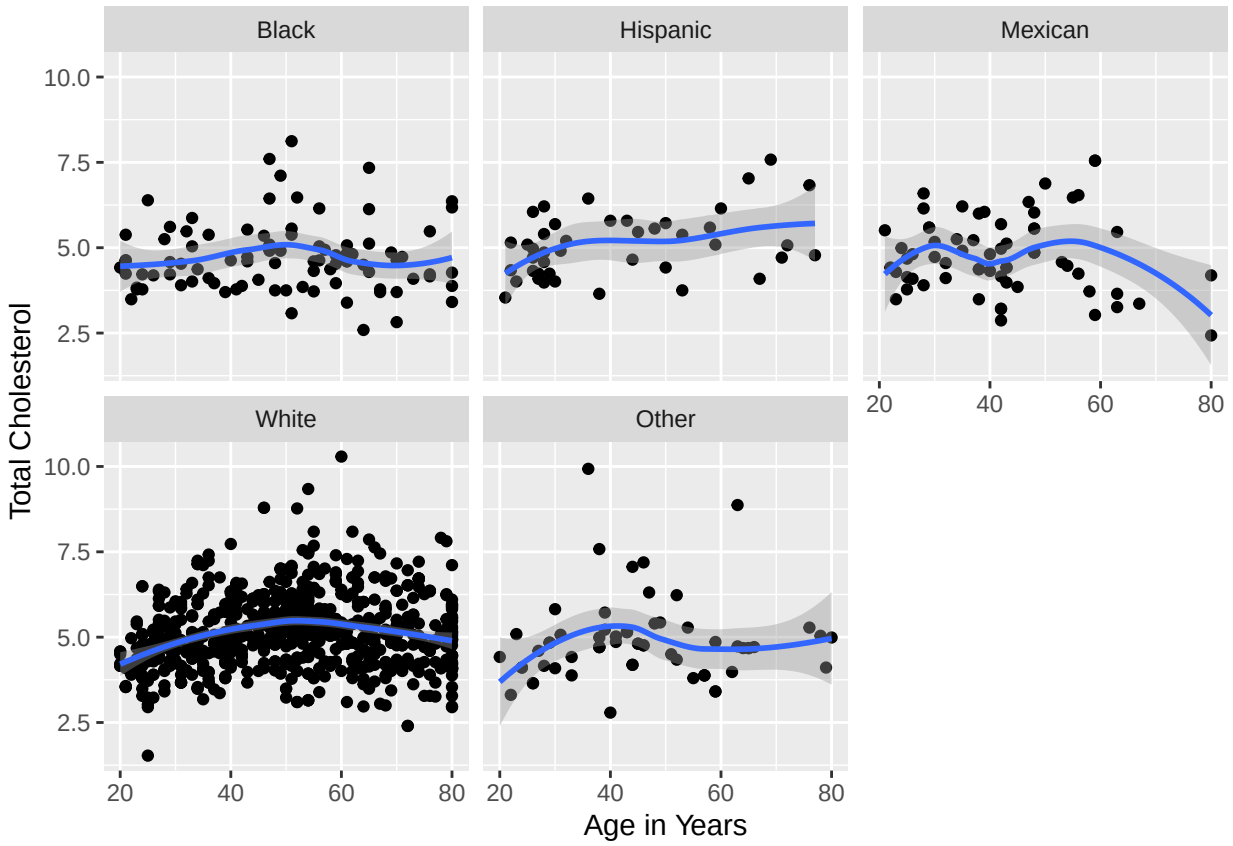
The above plot is created using `geom_point()` and `geom_smooth()` functions. A `geom` is the geometrical object that a plot uses to represent the data. `geom_point` generated the scatterplot and `geom_smooth` fitted a smooth line to the data.

There is not much striking difference in the cholesterol distribution across age for men and women. However, on avergae, younger men of age 25 or less have lower cholestrol than their female counterparts in that age range which flips for the age range of 30-50. After age 50, female participants have higher total cholesterol than their male counterparts.

Note the above plot presents the visualization of two continuous variables, age and total cholesterol, grouped by one categorical variable, gender. Another way to bring a categorical variable into this picture is by splitting the plot into facets.

```
ggplot(data = sample_n(df_eda, size=1000), aes(x = Age, y = TotChol))+
              geom_point() + geom_smooth() + facet_wrap(~Race1, nrow=2)+ xlab("Age in Years") + ylab("T
```

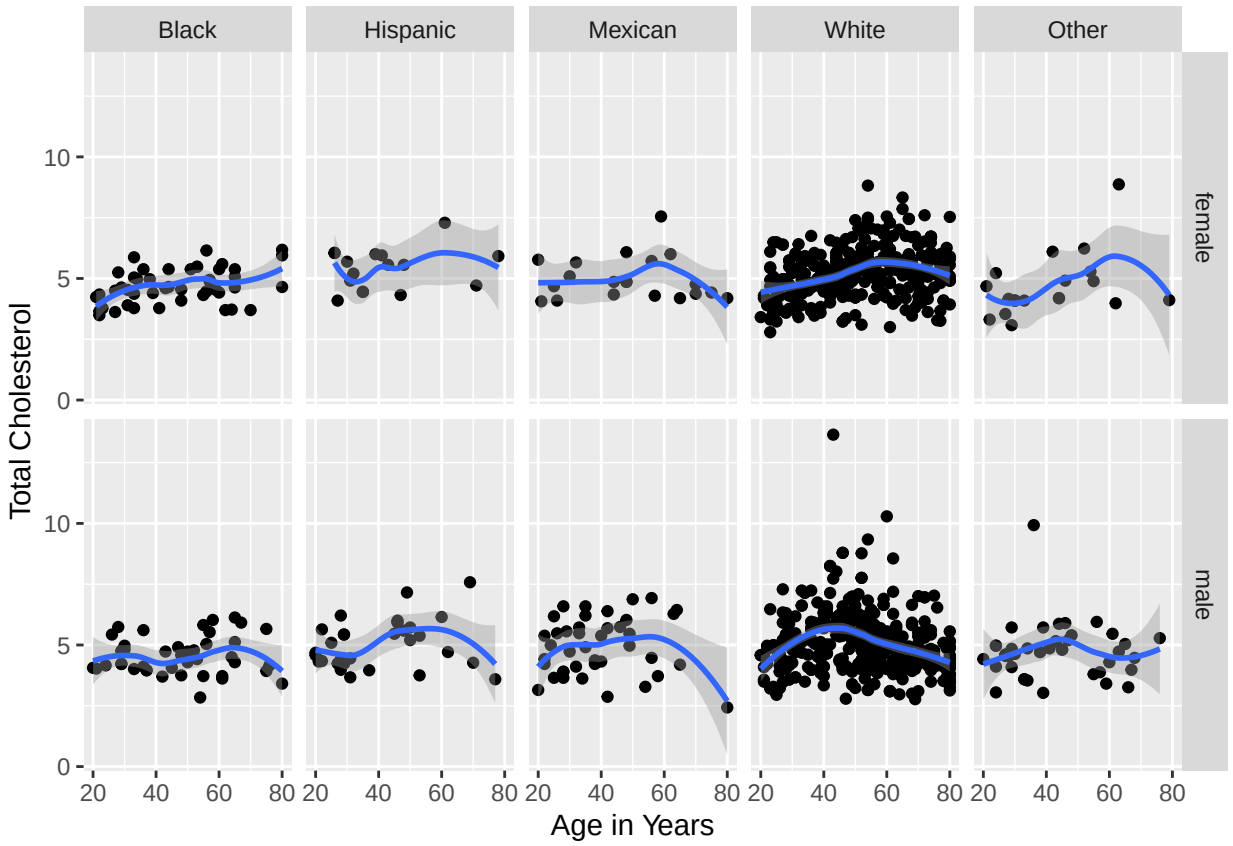
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



You can facet your plot on the combination of two variables by adding facet_grid() to your plot call. For example the following code chuck creates scatterplots of age and total cholesterol by race and gender of participants.

```
ggplot(data = sample_n(df_eda, size=1000), aes(x = Age, y = TotChol))+
              geom_point() + geom_smooth() + facet_grid(Gender~Race1)+ xlab("Age in Years") + ylab("Tota
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



The above plot demonstrates the cholesterol patterns over age at granular levels of race and gender combinations. There is wide variation in the pattern for various races for males and females.

References

1. Hadley Wickham and Garrett Grolemund. R for Data Sicence https://r4ds.had.co.nz/
2. David R. Brillinger, University of California, Berkeley. Data Analysis, Exploratory.