Data Visualization with ggplot2 (Single Quantitative Variable)

Learning Objectives

1. 5 number summary of a quantitative variable
2. Summarize a quantitative variable using boxplots
3. Summarize a quantitative variable across categories using side by side boxplots and violin plots
4. Summarize a quantitative variable using histograms and density plots

We will use the dataset ClassDataPrevious.csv as a working example. Download the dataset from Collab and read it into R. Also load the `tidyverse` package (which automatically loads the `ggplot2` package).

```
library(tidyverse)
Data<-read.csv("ClassDataPrevious.csv", header=TRUE)
```

1. 5 number summary of a quantitative variable

The `summary()` function, when applied to a quantitative variable, produces the 5 number summary: the minimum, the first quartile (25th percentile), the median (50th percentile), the third quartile (75th percentile), and the maximum, as well as the mean.

```
summary(Data$Age)
```

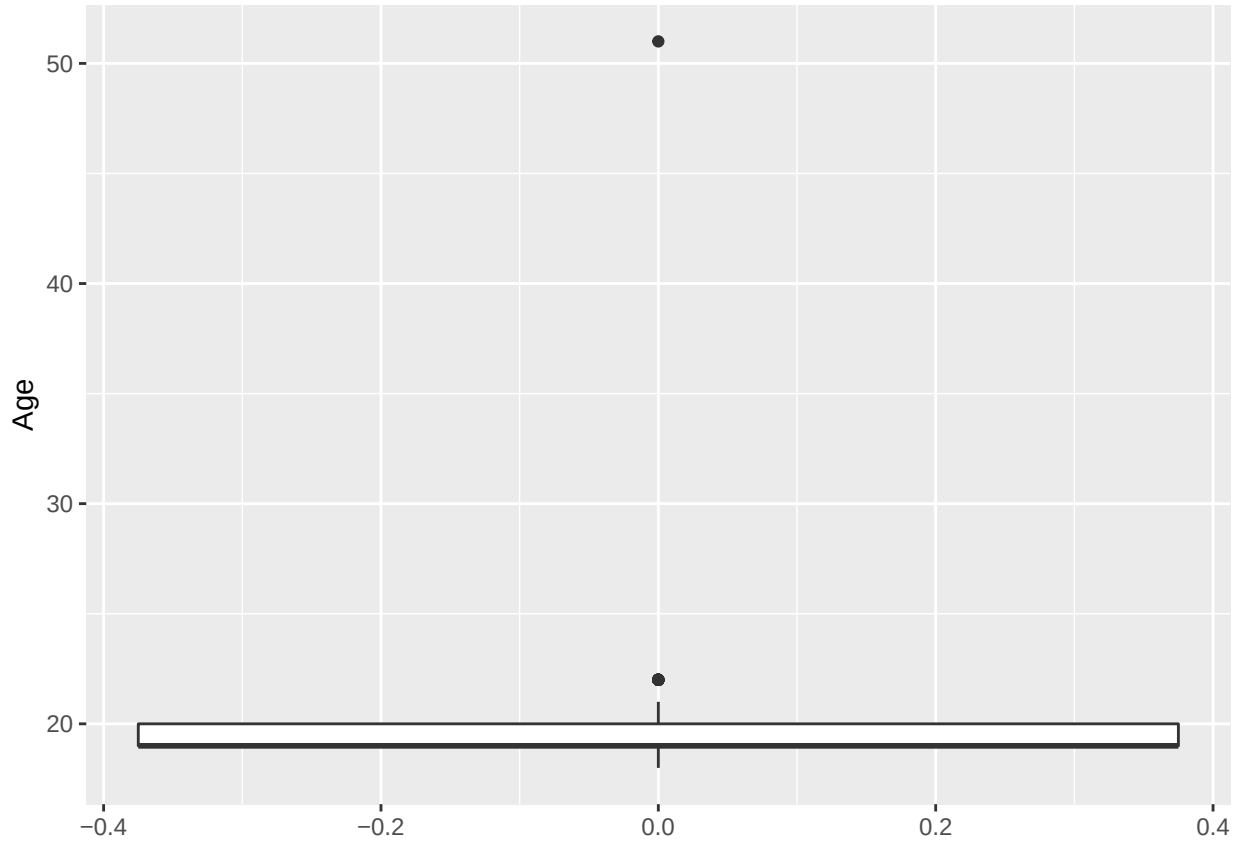
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.00   19.00   19.00   19.57   20.00   51.00
```

The average age of the observations in this dataset is 19.57 years old. Notice the first quartile and the median are both 19 years old, that means at least a quarter of the observations are 19 years old. Also note the maximum of 51 years old, so we have a student who is quite a lot older than the rest.

2. Summarize a quantitative variable using boxplots

A boxplot is a graphical representation of the 5 number summary. To create a generic boxplot, we have the following two lines of code when using the `ggplot()` function

```
ggplot(Data, aes(y=Age))+
  geom_boxplot()
```



Note we are still using the same structure when creating data visualizations with `ggplot()`

1. Use the `ggplot()` function, and supply the name of the data frame, and the x- and/or y- variables via the aes() function. End this line with a `+` operator, and then press enter.
2. In the next line, specify the type of graph we want to create (called `geoms`). For a boxplot, type `geom_boxplot`.

Notice there are outliers that are denoted by the dots. One is the 51 year old, and 22 year olds are deemed to be outliers. The rule being used is the $1.5 \times IQR$ rule.

Similar to bar charts, we can change the orientation of boxplots by adding an additional layer as before

```
ggplot(Data, aes(y=Age))+
  geom_boxplot()+
  coord_flip()
```

We can change the color of the box and the outliers similarly

```
ggplot(Data, aes(y=Age))+
  geom_boxplot(color="blue", outlier.color = "orange" )
```

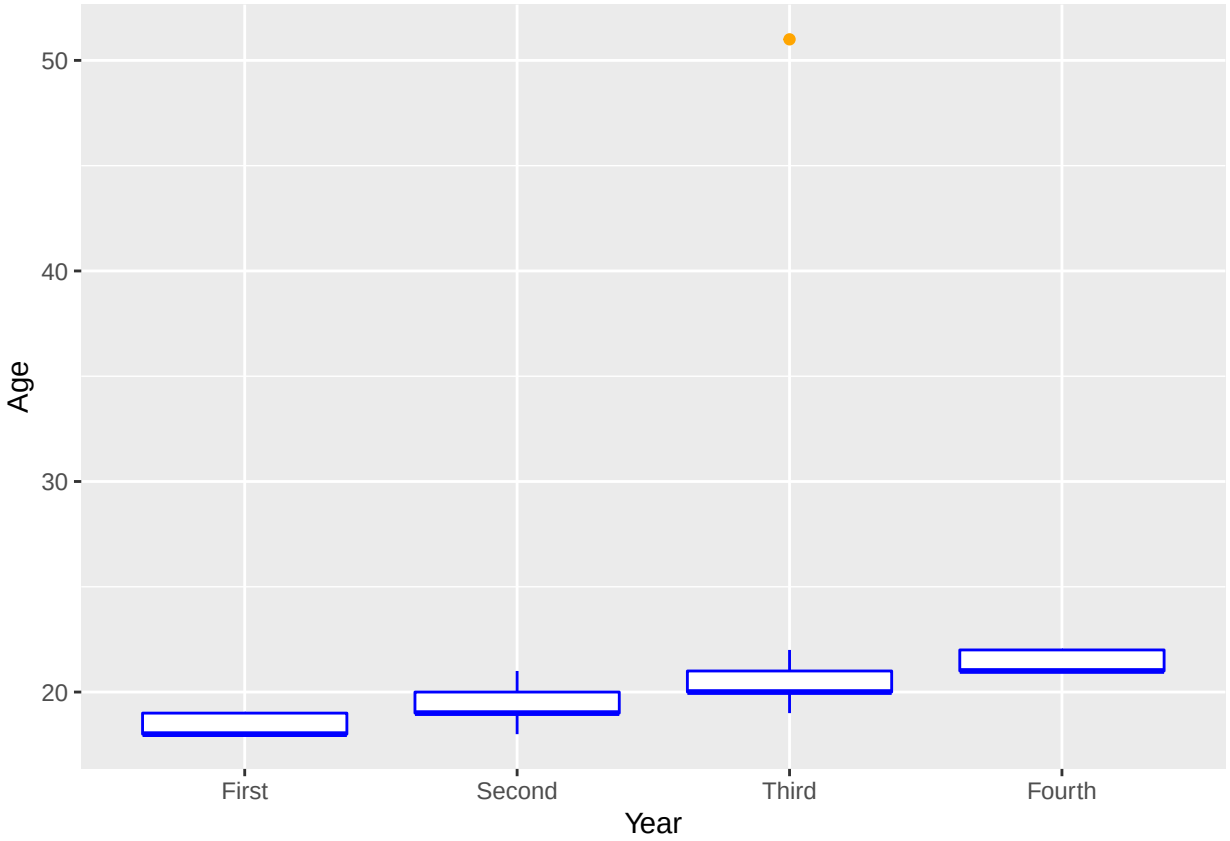
3. Summarize a quantitative variable across categories using side by side boxplots and violin plots

We can also create boxplots for `Age` across `Years`. But first, we need to reorder `Years` so the output makes sense

```
Data$Year<-factor(Data$Year,
                  levels=c("First","Second","Third","Fourth"))
levels(Data$Year)
```

```
## [1] "First"  "Second" "Third"  "Fourth"
```

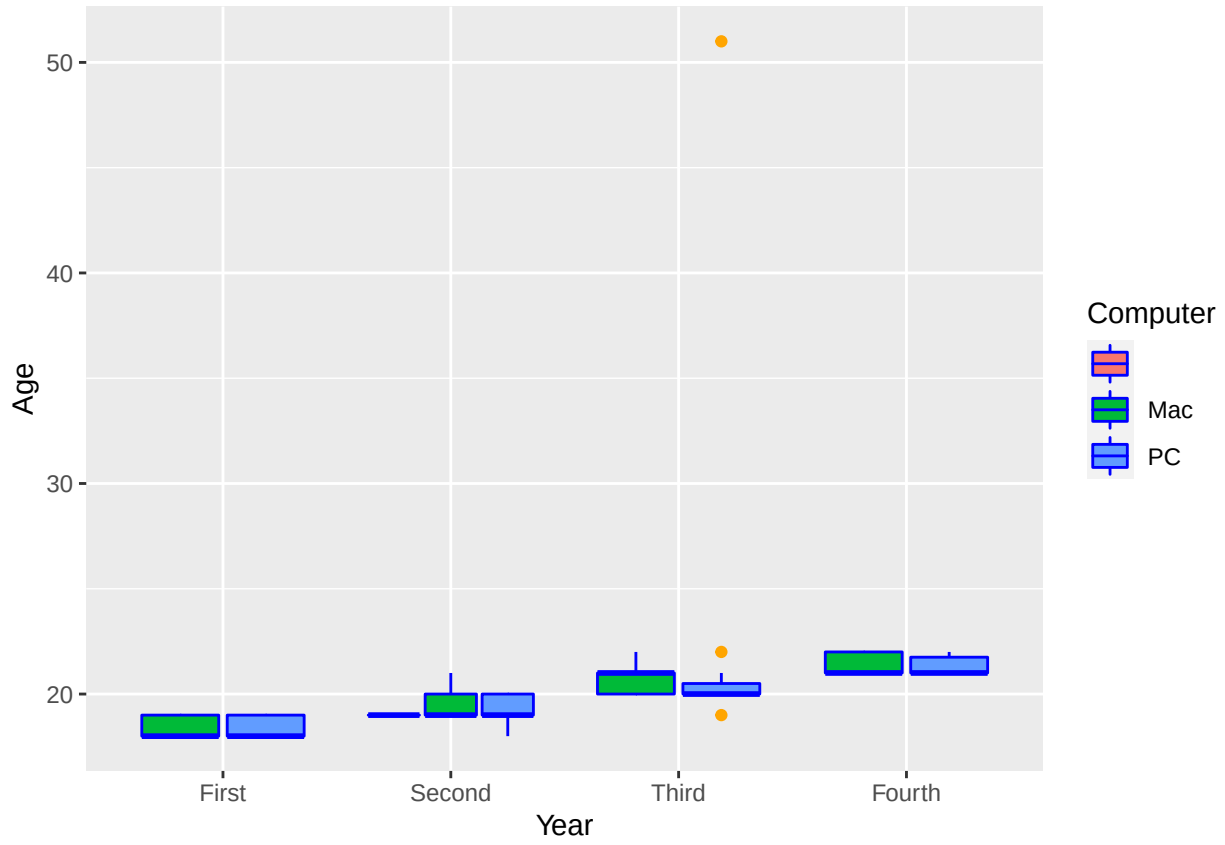
```
ggplot(Data, aes(x=Year, y=Age))+
  geom_boxplot(color="blue", outlier.color = "orange" )
```



These are called side by side boxplots, which we use to compare a quantitative variable across levels. Not surprising to see that as `Year` increases, the age of the students also increase. Note the 51 year old student is technically a third year student.

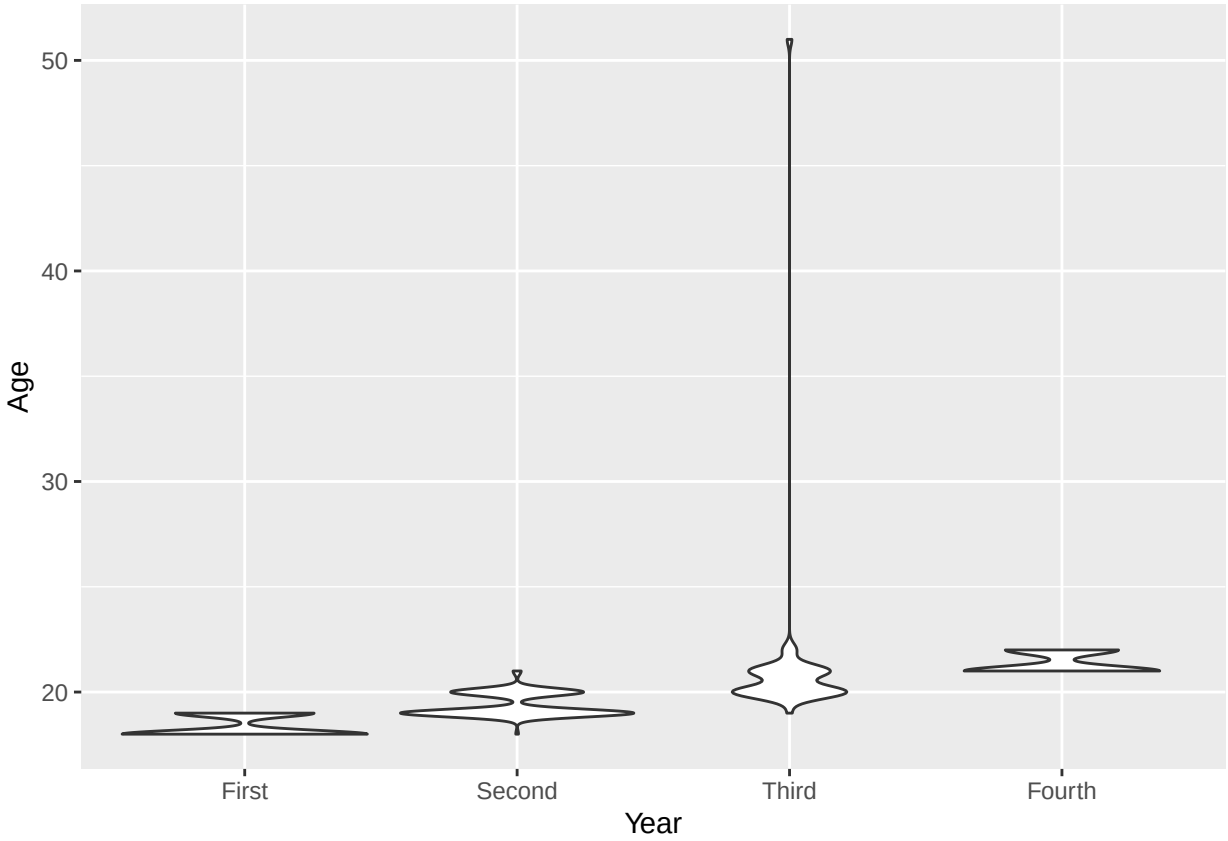
We can also compare boxplots of ages across two categorical variables by adding an additional variable in `aes()` using `fill`, i.e. how do ages vary across years and type of computer used?

```
ggplot(Data, aes(x=Year, y=Age, fill=Computer))+
  geom_boxplot(color="blue", outlier.color = "orange" )
```



A violin plot is a bit different from side by side boxplots. Instead of informing us the values of the 5 number summary, violin plots inform us how often each value occur.

```
ggplot(Data, aes(x=Year, y=Age))+
  geom_violin()
```



The width of the violin informs us how often a value occurs, relative to other values. Looking at the violin plot for first years, we can see that 18 year olds are more common than 19 year olds, since the violin is wider at age 18 than at age 19.

Similar to boxplots, we can add an additional categorical variable to compare age across, as well as change the color of the plots

```
ggplot(Data, aes(x=Year, y=Age, fill=Computer))+
  geom_violin(color="blue")
```

4. Summarize a quantitative variable using histograms and density plots

A histogram displays the number of observations within each bin on the x-axis.

```
ggplot(Data,aes(x=Age))+
  geom_histogram()
```

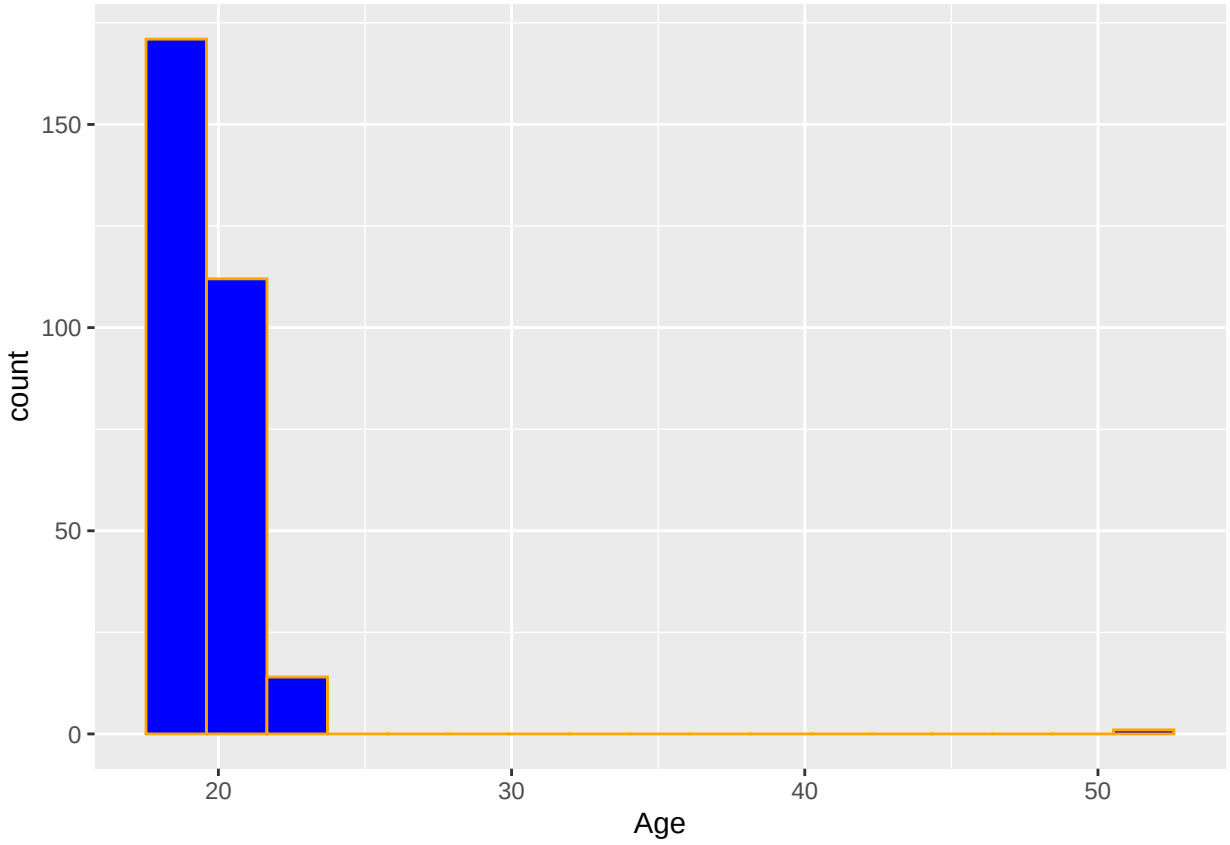
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

Notice a warning message is displayed when creating a basic histogram. To fix this, we use the squareroot of the number of rows (whole number) and use the argument `bins` inside `geom_histogram()`.

```
sqrt(nrow(Data))
```

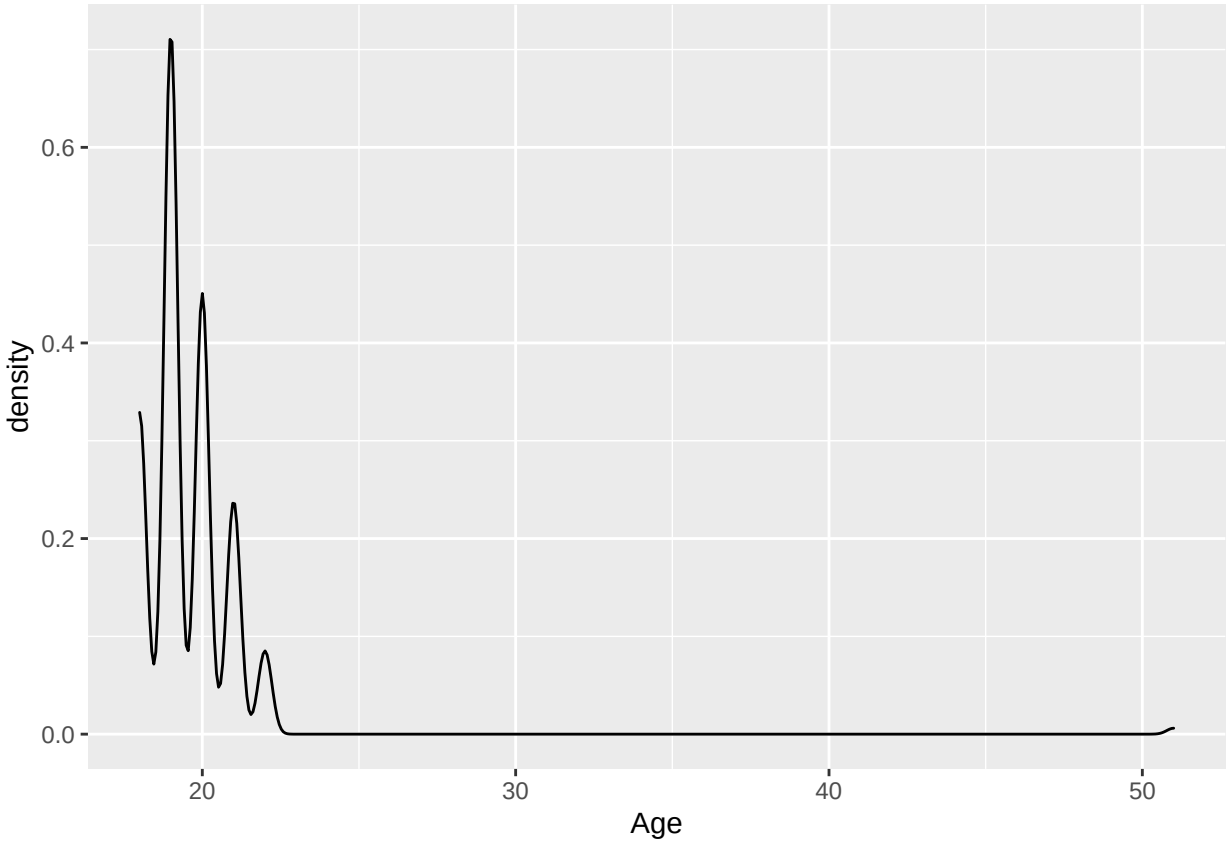
```
## [1] 17.26268
```

```
ggplot(Data,aes(x=Age))+
  geom_histogram(bins = 17,fill="blue",color="orange")
```



A similar graph is a density plot. The heights inform us how frequently each age occurs in the data.

```
ggplot(Data,aes(x=Age))+
  geom_density()
```


density
0.6
0.4
0.2
0.0
20
30
40
50
Age