Visualizing Data Using R

Fall 2017

mtcars Data Frame

- R has a built-in data frame called mtcars
- Useful R functions
  - length(object) # number of variables
  - str(object) # structure of an object
  - class(object) # class or type of an object
  - names(object) # names
  - dim(object) # number of observations and variables
- In the console, call each function using **mtcars** as the object

mtcars Data Frame

The data was extracted from the 1974 *Motor Trend* US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models).

[1]	mpg	Miles/(US) gallon
[2]	cyl	Number of cylinders
[3]	disp	Displacement (cu.in.)
[4]	hp	Gross horsepower
[5]	drat	Rear axle ratio
[6]	wt	Weight (1000 lbs)
[7]	qsec	1/4 mile time
[8]	vs	V/S (vshape or straight line engine)
[9]	am	Transmission (0 = automatic, 1 = manual)
[10]	gear	Number of forward gears
[11]	carb	Number of carburetors

Recoding Variables

- Copy mtcars to tempMtcars to protect mtcars data
  > tempMtcars = mtcars

- Recode am variable as amCategorical
  > tempMtcars$amCategorical = as.factor (mtcars$am)

  > tempMtcars$amLabels = factor (mtcars$am,
  levels=c('0','1'), labels=c("auto", "manual"))

  > tempMtcars$amOrdered = factor (mtcars$am,
  levels=c('1‘,’0’), labels=c("manual“, “auto”),
  ordered=TRUE)

  > barplot(summary(tempMtcars$amOrdered))

  > barplot(summary(tempMtcars$amLabels))

table function

- The table function will return a vector of table counts
- For instance, transmission=table(tempMtcars$am) will return a count of the number of automatic (value is 0) and manual (value is 1) transmission types

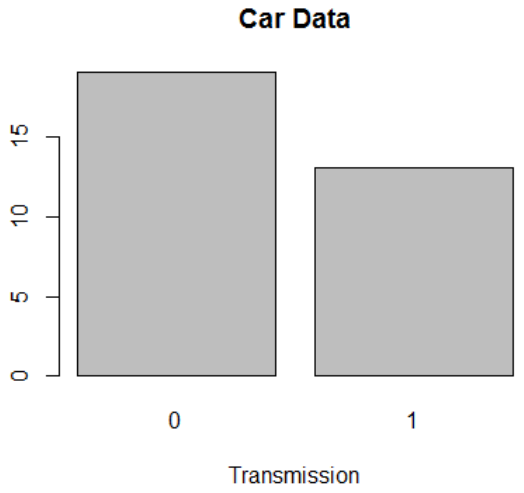
```
> transmission=table(tempMtcars$am)
> transmission

 0  1
19 13
```

Bar Chart

http://statmethods.net/graphs/bar.html

- A **bar chart** or **bar graph** is a chart that presents grouped data using rectangular bars with lengths proportional to the values that they represent.
- function table returns a vector of frequency data

```
> barplot(table(tempMtcars$amCategorical),
main = "Car Data",
xlab = "Transmission")
```



Recoding Variables

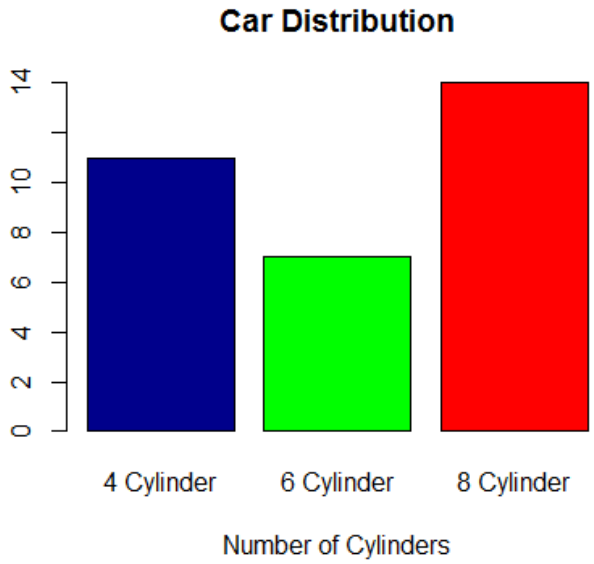
- Create a new variable mpgClass where mpg<=25 is “low”, mpg>25 is “high”

```
> tempMtcars$mpgClass[tempMtcars$mpg <= 25] = "low"
> tempMtcars$mpgClass[tempMtcars$mpg > 25] = "high"
> tempMtcars$mpgClass
[1] "low" "low" "low" "low" "low" "low" "low" "low"
[9] "low" "low" "low" "low" "low" "low" "low" "low"
[17] "low" "high" "high" "high" "low" "low" "low" "low"
[25] "low" "high" "high" "high" "low" "low" "low" "low"
> typeof(tempMtcars$mpgClass)
[1] "character"
```

```
barplot(table(tempMtcars$mpgClass), main = "Car Data",
xlab="MPG")
```

Bar Chart

```
> barplot (table(mtcars$cyl),
main = "Car Distribution",
xlab = "Number of Cylinders",
col = c("darkblue", "green", "red"),
names.arg = c("4 Cylinder", "6 Cylinder", "8 Cylinder"))
```

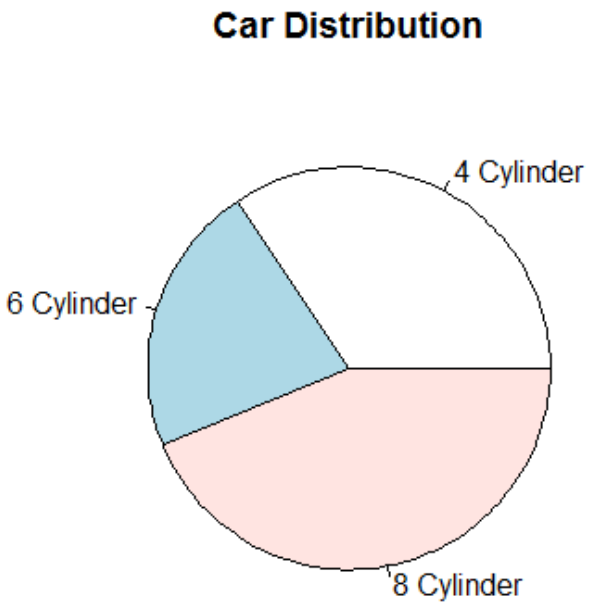


Pie Chart

http://statmethods.net/graphs/pie.html

- A pie chart is a circular graphical representation of data that illustrates a numerical proportion
- A pie chart gives a better visualization of the frequency of occurrence as a percent

```
> pie(table (mtcars$cyl),
labels = c("4 Cylinder", "6 Cylinder", "8 Cylinder"),
main="Car Distribution")
```

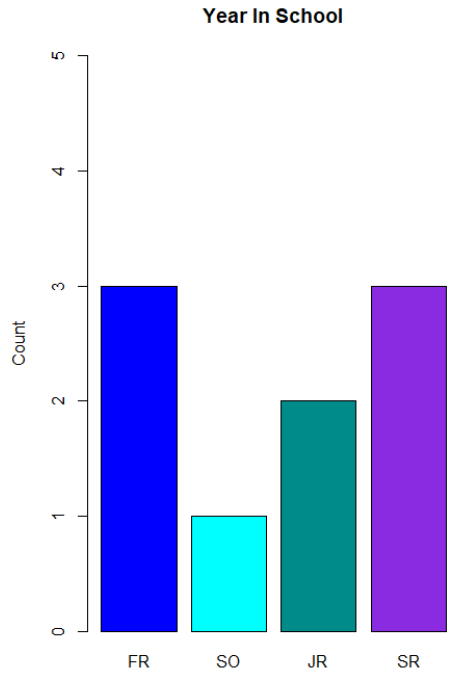


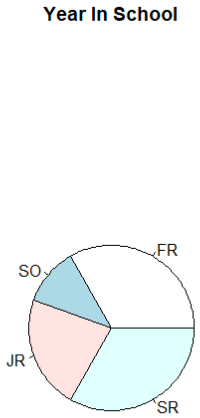
CS100 Problem

- For the given CS100 class information, create a data frame, cs100DataFrame that displays pie and bar chart representations of the Year data properly labeled. See next slide.

ID	Year	Age
0001	FR	18
0002	FR	18
0003	SR	22
0004	JR	22
0005	SO	19
0006	FR	19
0007	SR	23
0008	SO	19
0009	SR	22

CS100 Problem Continued



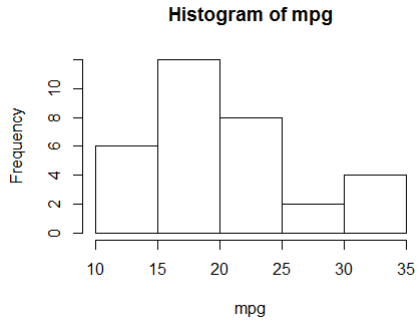


Histogram

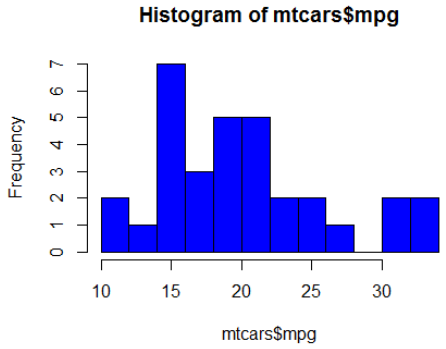
http://statmethods.net/graphs/density.html

- A histogram is a graphical representation of the distribution of numerical data
- Bin – are adjacent intervals usually of equal size
- Notice: breaks <> number of bins and breaks is just a suggestion and not guaranteed

>hist (mtcars$mpg)



> hist (mtcars$mpg, breaks=10, col="blue")



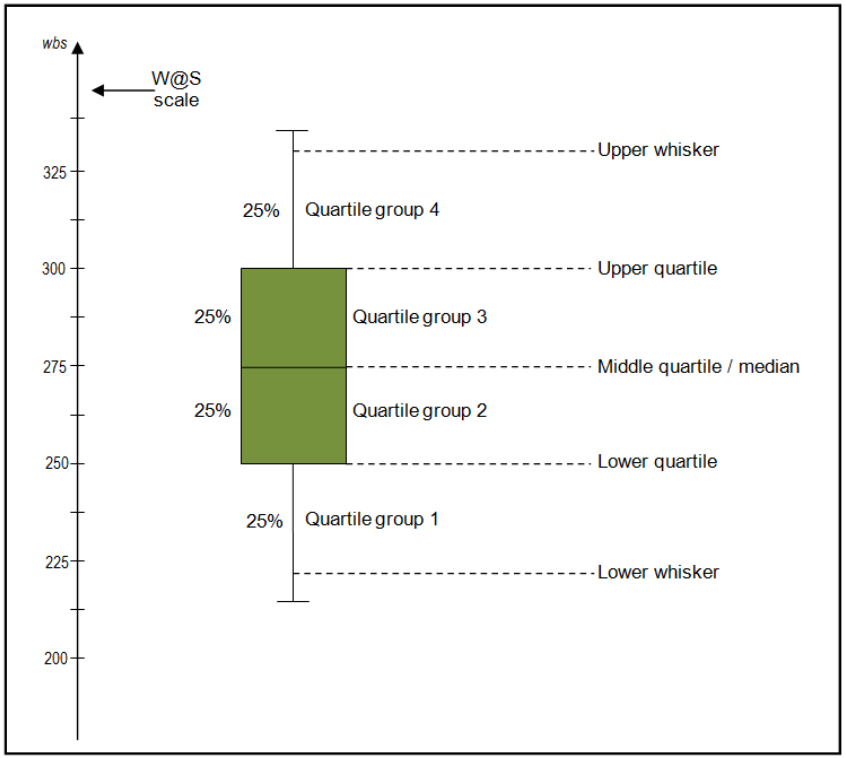
Boxplots

http://statmethods.net/graphs/boxplot.html

- A boxplot is a way of graphically showing numerical data through quartiles
- A box-and-whisker plot is a boxplot that shows variability outside the upper and lower quartiles
- Quartile – the three points that divide the ranked data values into 4 equal sized groups

Box-and-Whisker

https://www.wellbeingatschool.org.nz/information-sheet/understanding-and-interpreting-box-plots



FYI - Quartile Definitions

https://en.wikipedia.org/wiki/Quartile

https://www.mathsisfun.com/data/quartiles.html

http://dsearls.org/other/CalculatingQuartiles/CalculatingQuartiles.htm

- **first quartile/lower quartile/25th percentile/** Q_1
  - splits off the lowest 25% of data from the highest 75%
- **second quartile /median/50th percentile /** Q_2
  - cuts data set in half
- **third quartile/upper quartile/75th percentile /** Q_3
  - splits off the highest 25% of data from the lowest 75%
- **interquartile range /** IQR
  - $IQR = Q_3 - Q_1$

Problem Continued

- Using R, show the box-and-whisker plot and quantiles for the following values
  - 1, 2, 3, 4, 5, 6, 7, 8

Paint Problem

- Let's put everything together
- A paint manufacturer tested two experimental brands of paint over a period of months to determine how long they would last without fading. Here are the results:

BrandA	**BrandB**	Report on the following
10	25	-Mean
20	35	-Median
60	40	-Mode
40	45	-Std Deviation
50	35	-Minimum
30	30	-Maximum

Paint Problem

1. Using Rstudio, create an R script on your desktop called paintDataFrame.R that creates a data frame paintData for the paint data.
2. Enter the data
3. Output the data frame
4. Save and run the script. Show me.

Paint Problem Continued

5. Compute and output the mean, median, std deviation, minimum, and maximum for each brand of paint

```
[1] "Brand A Mean = 35"
[1] "Brand A Median = 35"
[1] "Brand A Std Dev = 18.7082869338697"
[1] "Brand A Minimum = 10"
[1] "Brand A Maximum = 60"
[1] ""
[1] "Brand B Mean = 35"
[1] "Brand B Median = 35"
[1] "Brand B Std Dev = 7.07106781186548"
[1] "Brand B Minimum = 25"
[1] "Brand B Maximum = 45"
```

Paint Problem Continued

5 Output a Box-and-Whisker Plot for each brand of paint as follows. Get as close as possible. This isn't easy but give it a try.

6. What do the descriptive statistics tell us?

7. Which paint would you buy? Justify your answer