July 17, 2020 8:39 AM

Summarizing Data – Graphical Methods

Topics

Introduction – Descriptive versus Inferential Statistics

- **What is/are statistics?**

<u>Statistics</u> is the branch of mathematics concerned with *making sense of data*: collecting, summarizing, organizing, presenting, and analyzing numerical information.

<u>Descriptive statistics</u> covers methods for organizing and summarizing possibly rather large collections of data.

Example – student survey in R (built-in dataset).

survey

Filter

	Sex	Wr.Hnd	NW.Hnd	W.Hnd	Fold	Pulse	Clap	Exer	Smoke	Height	M.I	Age
1	Female	18.5	18.0	Right	R on L	92	Left	Some	Never	173.00	Metric	18.250
2	Male	19.5	20.5	Left	R on L	104	Left	None	Regul	177.80	Imperial	17.583
3	Male	18.0	13.3	Right	L on R	87	Neither	None	Occas	NA	NA	16.917
4	Male	18.8	18.9	Right	R on L	NA	Neither	None	Never	160.00	Metric	20.333
5	Male	20.0	20.0	Right	Neither	35	Right	Some	Never	165.00	Metric	23.667
6	Female	18.0	17.7	Right	L on R	64	Right	Some	Never	172.72	Imperial	21.000
7	Male	17.7	17.7	Right	L on R	83	Right	Freq	Never	182.88	Imperial	18.833
8	Female	17.0	17.3	Right	R on L	74	Right	Freq	Never	157.00	Metric	35.833
9	Male	20.0	19.5	Right	R on L	72	Right	Some	Never	175.00	Metric	19.000
10	Male	18.5	18.5	Right	R on L	90	Right	Some	Never	167.00	Metric	22.333
11	Female	17.0	17.2	Right	L on R	80	Right	Freq	Never	156.20	Imperial	28.500
12	Male	21.0	21.0	Right	R on L	68	Left	Freq	Never	NA	NA	18.250
13	Female	16.0	16.0	Right	L on R	NA	Right	Some	Never	155.00	Metric	18.750
14	Female	19.5	20.2	Right	L on R	66	Neither	Some	Never	155.00	Metric	17.500
15	Male	16.0	15.5	Right	R on L	60	Right	Some	Never	NA	NA	17.167
16	Female	17.5	17.0	Right	R on L	NA	Right	Freq	Never	156.00	Metric	17.167
17	Female	18.0	18.0	Right	L on R	89	Neither	Freq	Never	157.00	Metric	19.333
18	Male	19.4	19.2	Left	R on L	74	Right	Some	Never	182.88	Imperial	18.333

Showing 1 to 21 of 237 entries, 12 total columns

Inferential statistics (or statistical inference) which provides methods for (1) estimating values of population parameters (statistical estimation) or (2) evaluating the degree of support for a statement or claim about some characteristic of a population (hypothesis testing) using data obtained from a random sample of that population.

(1) "What is the average height of a woman?"

(2) "Claim: the average woman is 165 cm tall."

Descriptive Statistics

- **Graphical Methods**

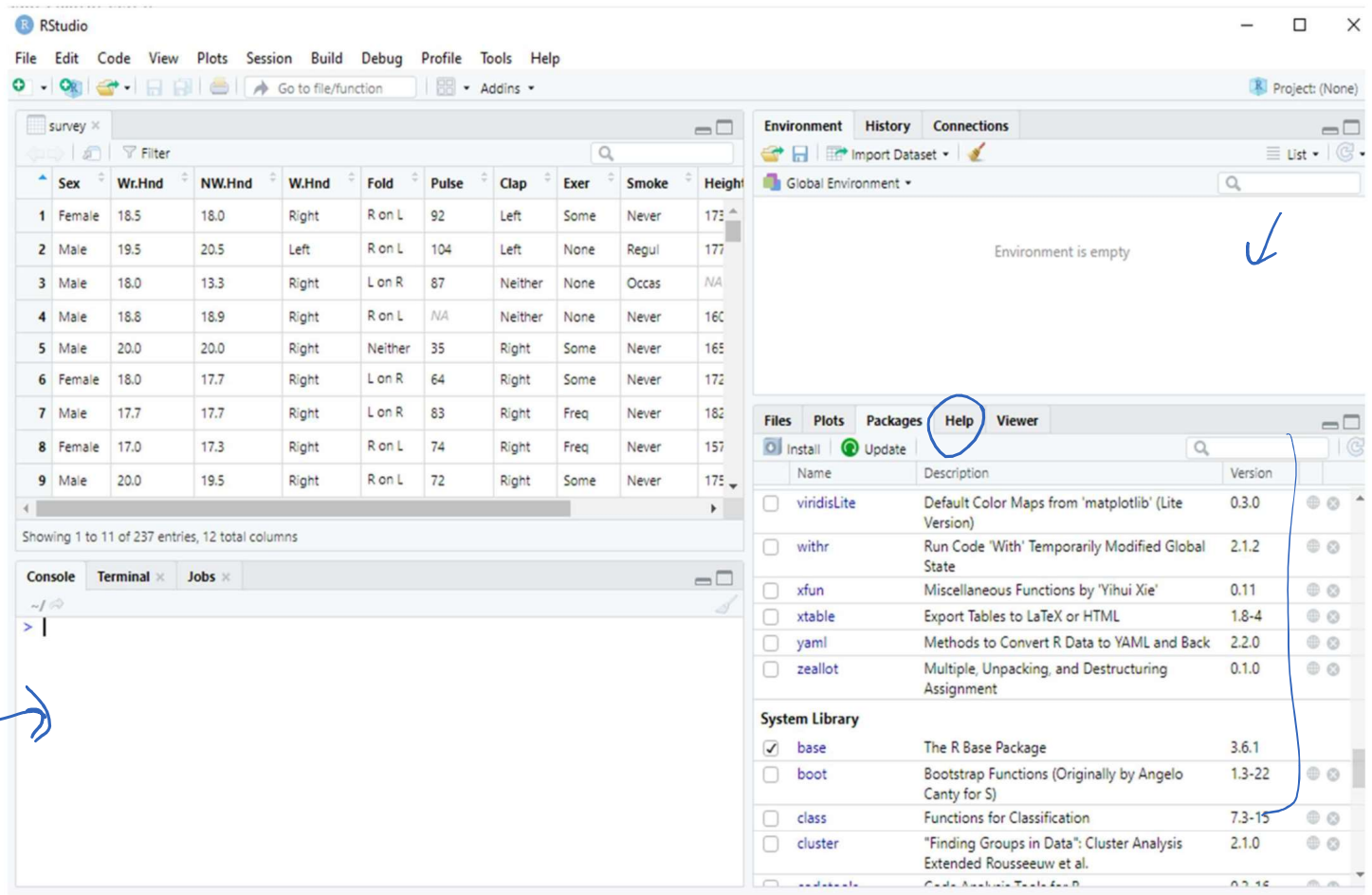
- **Numerical Methods (Numerical measurements of position and variation)**

↳ average (mean, median, mode)
↳ position, eg "in the top 10% of data"
↳ variation, eg - standard deviation
is the data spread apart or clustered together?

Graphical ways of summarizing data

Pie Chart
Stem-and-Leaf Plots
Frequency Histograms
Ogive
Scatter Diagrams

We will be using the software R to create graphs and charts. R is a programming language and free, open-source software environment for statistical computing and graphics. The R language is widely used among statisticians and data miners for developing statistical software and data analysis.



Qualitative data:

Pie chart

Pie charts are generally a poor way to represent quantitative data. However, they provide a simple way to visualize qualitative data.

The built-in dataset *HairEyeColor* in R gives the hair and eye colour of 592 students.

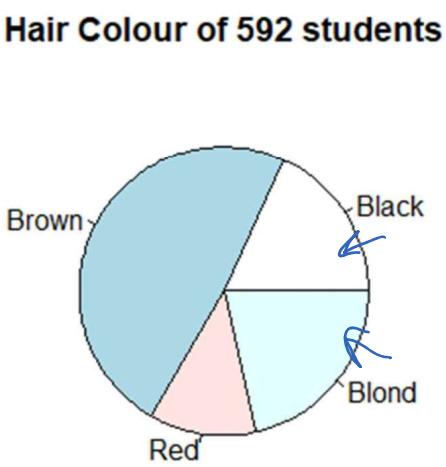
survey × HairEyeColor ×

Filter

	Hair	Eye	Sex	Freq
1	Black	Brown	Male	32
2	Brown	Brown	Male	53
3	Red	Brown	Male	10
4	Blond	Brown	Male	3
5	Black	Blue	Male	11
6	Brown	Blue	Male	50
7	Red	Blue	Male	10
8	Blond	Blue	Male	30
9	Black	Hazel	Male	10
10	Brown	Hazel	Male	25
11	Red	Hazel	Male	7

Showing 1 to 12 of 32 entries, 4 total columns

If we are interested in visualizing the proportion of students who have each different hair colour, we can easily represent that information in a pie chart.



MATH 3042: Applied Probability and Statistics for CST

Quantitative data

We will be focusing on quantitative data in this course.

When we have a raw data set, it is often difficult to get information from it, especially if the data set is large. Consider, for example, the built-in dataset *faithful*, which gives the duration of eruptions as well as the time between eruptions (both in minutes) from the geyser Old Faithful.

faithful

Filter

	eruptions	waiting
1	3.600	79
2	1.800	54
3	3.333	74
4	2.283	62
5	4.533	85
6	2.883	55
7	4.700	88
8	3.600	85
9	1.950	51
10	4.350	85
11	1.833	54

Showing 1 to 12 of 272 entries, 2 total columns

It is hard to obtain any meaningful information from this dataset in its current form.

How can we organize the data?

- list eruptions in order (low → high)
- put eruption lengths in frequency table

eruption (min)	frequency
1.6–1.8	#
1.8–2.0	

The data comes from the built-in dataset *faithful* in R. We can view the data in R with the command *View(faithful)*.

It is clear that we are mostly interested in roughly how long the eruptions were – "between 1.8 and 2 minutes" is sufficient information. This will allow us to group the data.

A *stem-leaf plot* allows us to group the data without losing any information. To create one, we divide the data into stems – based on the leftmost significant figures in the data – and the leaves. Here is the stem-leaf plot produced by the command *stem(faithful$eruptions)*:

```
  The decimal point is 1 digit(s) to the left of the |

  16 | 070355555588
  18 | 000022233333335577777777888822335777888
  20 | 00002223378800035778
  22 | 0002335578023578
  24 | 00228
  26 | 23
  28 | 080
  30 | 7
  32 | 2337
  34 | 250077
  36 | 0000823577
  38 | 2333335582225577
  40 | 000000335778888800233555577778
  42 | 033355557788002333335555777778
  44 | 022223355577800000000233333577788888
  46 | 000023335770000002357 8
  48 | 00000022335800333
  50 | 0370
```

R helpfully tells us how to interpret the plot. The first row, for instance, gives the following times for eruptions (in minutes): 1.60, 1.67, 1.70, 1.73, 1.75, 1.75, 1.75, 1.75, 1.75, 1.75, 1.78, 1.78.

What can we say about the eruptions?

There are many "short" eruptions and many "long" eruptions, but not many "medium-length" eruptions.

We can customize our stem plot by adjusting parameters.

eg- stems 1.6, 1.7, 1.8, 1.9 ...
or 1.5, 2.0, 2.5, ...

Frequency distribution

For large data sets, a stem and leaf plot is ridiculous to construct. Even the Old Faithful stem and leaf plot is a bit unwieldy and gives us more information than we need. Instead we can form intervals (classes) and tally the number of values within each interval. We can do this by sorting the data into *classes* with a frequency distribution.

Tips: **The classes should all be the same width.**
The classes should capture all the data.
The classes should not overlap.

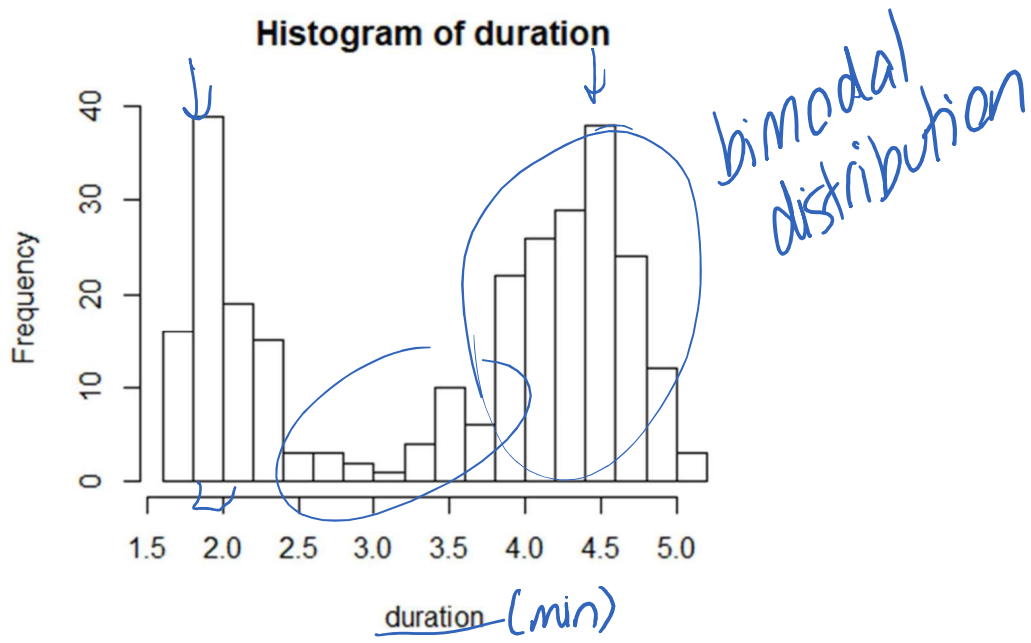
If we use the same classes of 0.2 minutes apiece that we used for the stem plot, R can give us the following frequency table:

[1.6,1.8)	12
[1.8,2)	39
[2,2.2)	20
[2.2,2.4)	18
[2.4,2.6)	3
[2.6,2.8)	3
[2.8,3)	2
[3,3.2)	1
[3.2,3.4)	4
[3.4,3.6)	6
[3.6,3.8)	10
[3.8,4)	16
[4,4.2)	31
[4.2,4.4)	29
[4.4,4.6)	35
[4.6,4.8)	28
[4.8,5)	11
[5,5.2)	4

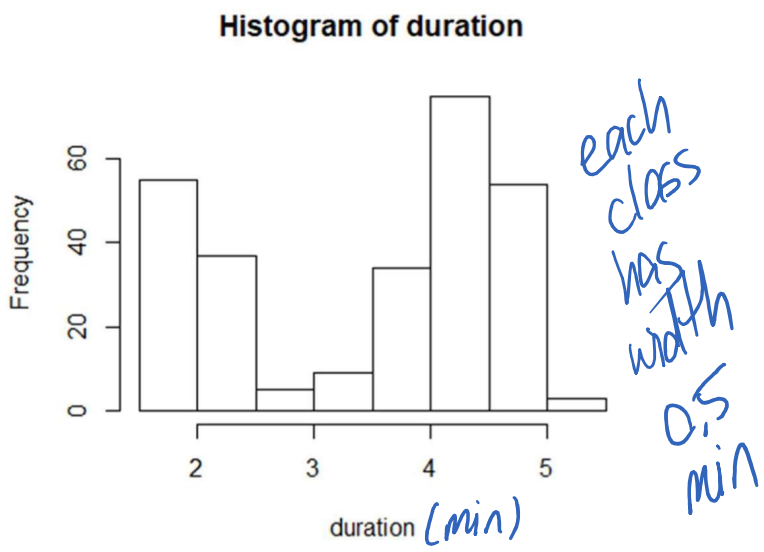
includes eruptions of 4.00 min but not eruptions of 4.20 min

For example, there were 12 eruptions out of 272 that were between 1.6min and 1.8min in length. Note the brackets: [1.6,1.8) denotes a class that includes 1.6 minute eruptions but excludes 1.8 minute ones. This is so that there is no overlap between classes.

It is still hard to make sense of this data in the above form, so we can represent it graphically with a histogram.



This histogram still looks a bit cluttered. We can change the number of classes as well. We want a number of classes so that we can obtain meaningful information about the data at a glance. Between 5 and 20 classes is a standard guideline, with the classes chosen so that the numbers are "nice". (eg, a class of [1.6,1.8) is nicer than a class of [1.56,1.87).)



Cumulative Frequency Table

Often we are interested in knowing how much of our data is *less* than a certain value. For instance, we might be interested in knowing how many of Old Faithful's eruptions were less than 4 minutes. We can create a *cumulative frequency* table with that information.

Using the intervals of 0.2 minutes again, R gives us the following cumulative frequency table:

[1.6,1.8)	12
[1.8,2)	51
[2,2.2)	71
[2.2,2.4)	89
[2.4,2.6)	92
[2.6,2.8)	95
[2.8,3)	97
[3,3.2)	98
[3.2,3.4)	102
[3.4,3.6)	108
[3.6,3.8)	118
[3.8,4)	134
[4,4.2)	165
[4.2,4.4)	194
[4.4,4.6)	229
[4.6,4.8)	257
[4.8,5)	268
[5,5.2)	272

39+12

89 eruptions were of duration less than 2.4 min

total number of eruptions

This table tells us that 134 of the eruptions were less than 4 minutes in duration. Note that the cumulative frequencies increase.

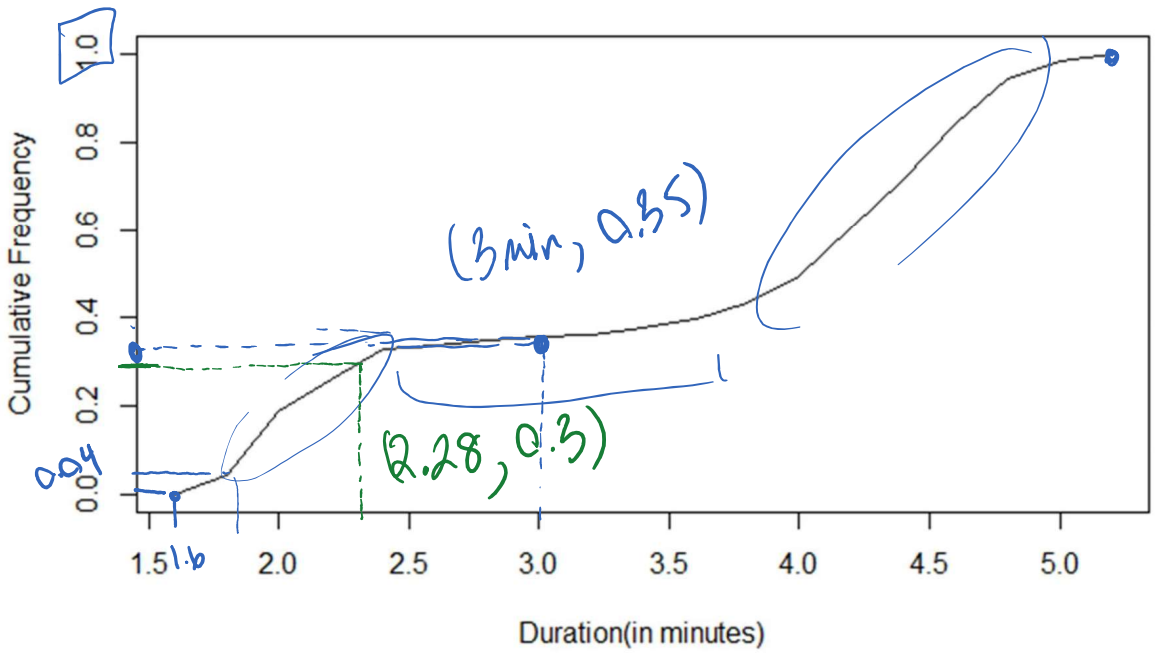
By dividing each of these cumulative frequencies by the total number of eruptions (which is equal to the last of these entries, 272), we can obtain a *cumulative relative frequency table*:

[1.6,1.8)	0.04
[1.8,2)	0.19
[2,2.2)	0.26
[2.2,2.4)	0.33
[2.4,2.6)	0.34
[2.6,2.8)	0.35
[2.8,3)	0.36
[3,3.2)	0.36
[3.2,3.4)	0.38
[3.4,3.6)	0.40
[3.6,3.8)	0.43
[3.8,4)	0.49
[4,4.2)	0.61
[4.2,4.4)	0.71
[4.4,4.6)	0.84
[4.6,4.8)	0.94
[4.8,5)	0.99
[5,5.2)	1.00

Here we can see that 49% of eruptions were less than 4 minutes in duration. (Note that these numbers are rounded to two decimal places. By default R displays 8 digits, which is excessive.)

We can use the table above to form an ogive (“oh-jive”). An ogive is a line graph that depicts cumulative frequencies or cumulative percent frequencies.

July 17, 2020 8:40 AM



Approximately what percentage of eruptions were less than 3 minutes?

≈35%, 38%, 34%, 36%

Approximately 30% of eruptions were less than how long?

≈2.28 min

Note: from table/ogive, 30% of eruptions were less than some amount of time between 2.2 min & 2.4 min, but we can't get a more precise number directly from the table/ogive.

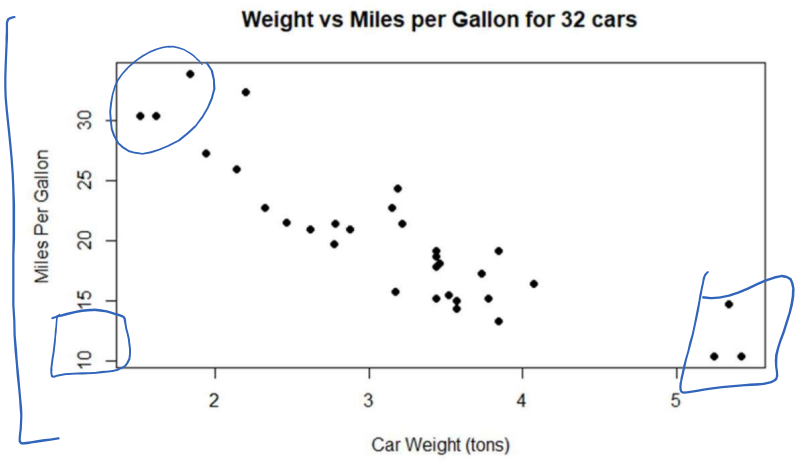
Scatter Diagrams

Often we are interested in knowing how two variables in a dataset are related. The file *01 – cars.csv*, based on R's dataframe *mtcars*, contains information about 32 makes of cars.

mtcars

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3
Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3
Valiant	18.1	6	225.0	105	2.76	3.460	20.22	1	0	3
Duster 360	14.3	8	360.0	245	3.21	3.570	15.84	0	0	3
Merc 240D	24.4	4	146.7	62	3.69	3.190	20.00	1	0	4
Merc 230	22.8	4	140.8	95	3.92	3.150	22.90	1	0	4

Showing 1 to 11 of 32 entries, 11 total columns



We can see how the cars' weights and their efficiencies (in miles per gallon) by creating a *scatter diagram*, or *scatterplot*.

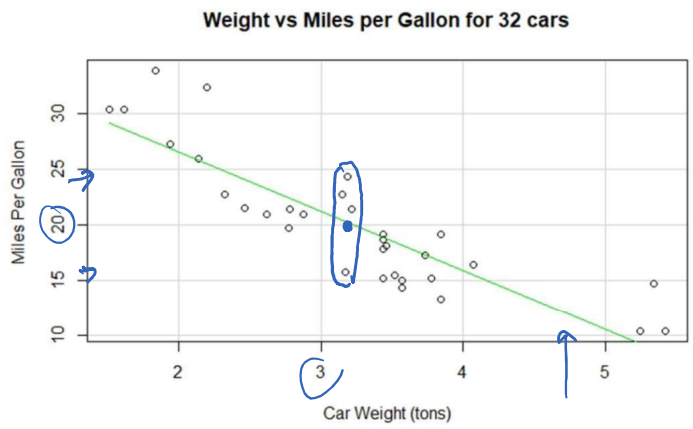
What trend can we see from the scatterplot?

In general, heavy cars are less efficient (fewer miles per gallon)
ie, as weight ↗, mpg ↘

negative correlation

July 17, 2020 8:40 AM

We can get R to create a *line of best fit.*



Here, R tells us that the equation of the line of best fit is

y=-5x+37

where x is the car's weight in tons, and y is the number of miles per gallon. Since the data doesn't fall exactly on the line, there is some error associated with this equation. Later, we will see when it is appropriate to construct a line of best fit, and when other methods are necessary.