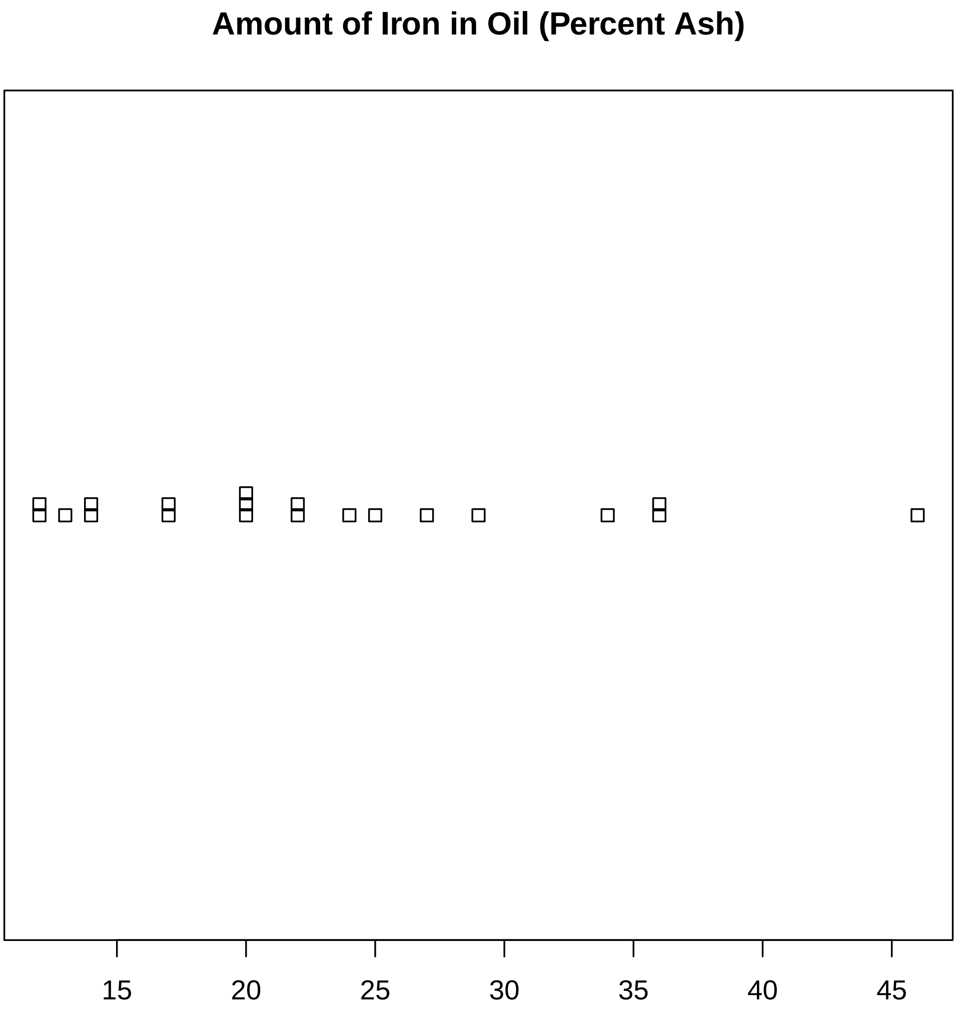
1. A federal government study of the oil reserves in Elk Hills, CA, included data on the amount of iron present in the oil. The data on 20 reserves is below:

   20, 14, 22, 12, 34, 20, 22, 13, 12, 36, 25, 14, 17, 27, 17, 20, 29, 24, 36, 46

   (a) Create a dot plot of the data. Identify any key features of the data that can be identified in the graph.
   The plot is below. The value of 46 is largest and seperated from the other values - maybe an outlier?. There looks to be slight right skew. In particular, the distribution of smaller values seem to be denser. So, it is not sampled uniformly.

   Outliers are the data that do not fit in other data. Generally, it is important to check if a point is outlier in linear regression in order to check the assumptions. Statistical models are always built based on appropriate assumptions, so it is important to check these assumptions. If there are many outliers, we can say the model's assumptions are not satisfied. Therefore, we can conclude the model is not appropriate for the data. Furthermore, we often need to run statistical tests to prove the data is not uniformly distributed.



However, this figure seems to be too complicated, which will make us hard to make some statements. There are too many groups. Can we use a simpler plot to illustrate the distribution of points so that we can make some solid statements?

(b) Create a stem-and-leaf plot of the data. Why did you choose the stems that you did? Based on the plot, would you say the data is skewed, or roughly symmetric? *Two possible plots are below. The one with two stems per tens place seems a bit better. The first one might have too few stems. The right-skew seems more evident in these graphs.*

There are two barplots in the following. Barplots will count the amount of data in every stem. We will lose information in this step. But if this losing information is useless to us, it is reasonable to make the plot simpler because we can extract what we want more easily.

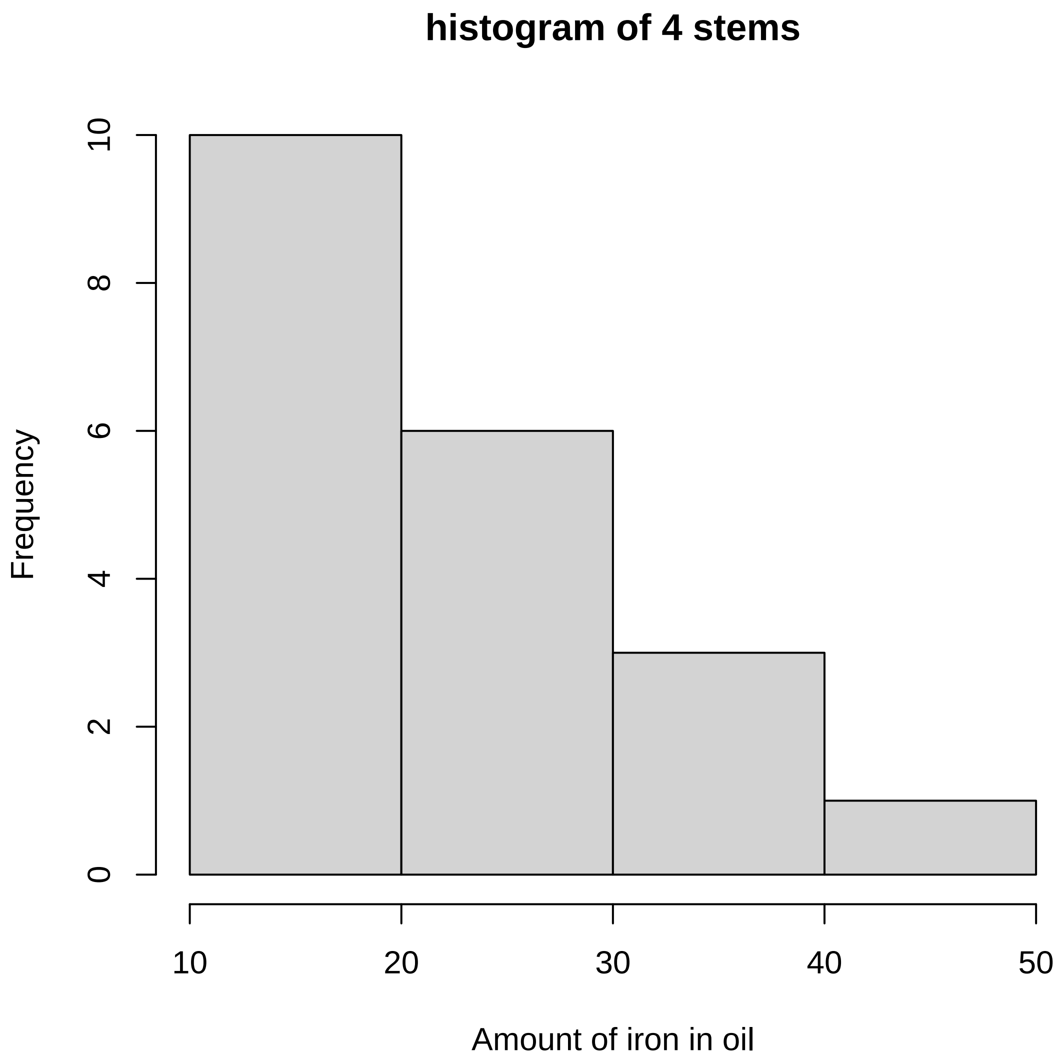
We say losing information means we cannot recover the former data by the transfered information. In this example, we can use 20, 14, 22, 12, 34, 20, 22, 13, 12, 36, 25, 14, 17, 27, 17, 20, 29, 24, 36, 46 to get the barplots, but we cannot use barplot to recover these data.

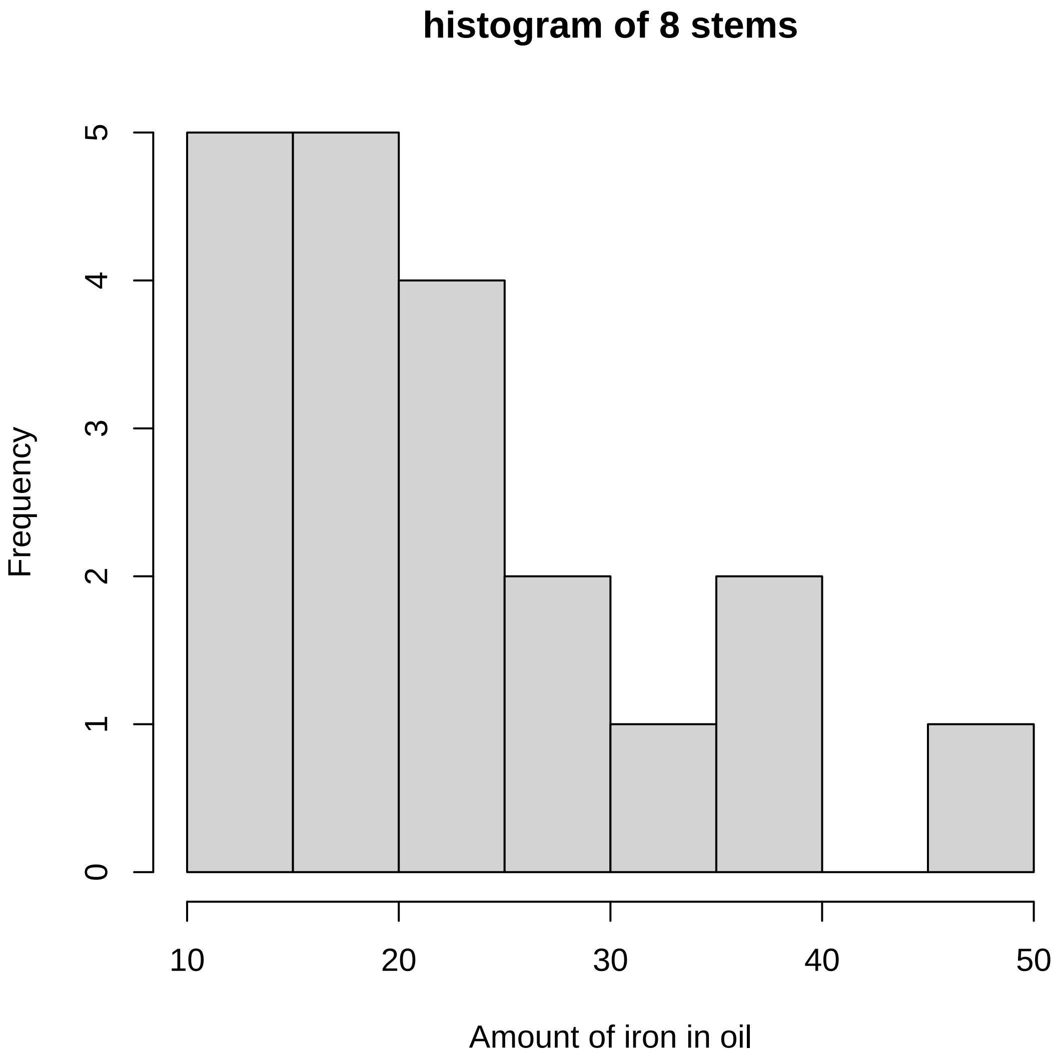
```
Key: 1|2 means 12

1 | 2234477
2 | 000224579
3 | 466
4 | 6
```

```
Key: 1|2 means 12

1 | 22344
1 | 77
2 | 000224
2 | 579
3 | 4
3 | 66
4 |
4 | 6
```


histogram of 4 stems
Frequency
0
2
4
6
8
10
10
20
30
40
50
Amount of iron in oil



Although we want to simplify the figures, we still do not want the plot to be too simple. We should balance this simplification to eliminate some information but not too much. Notice that the default histogram plot will have slight difference from stem-and-leaf what we have defined before.

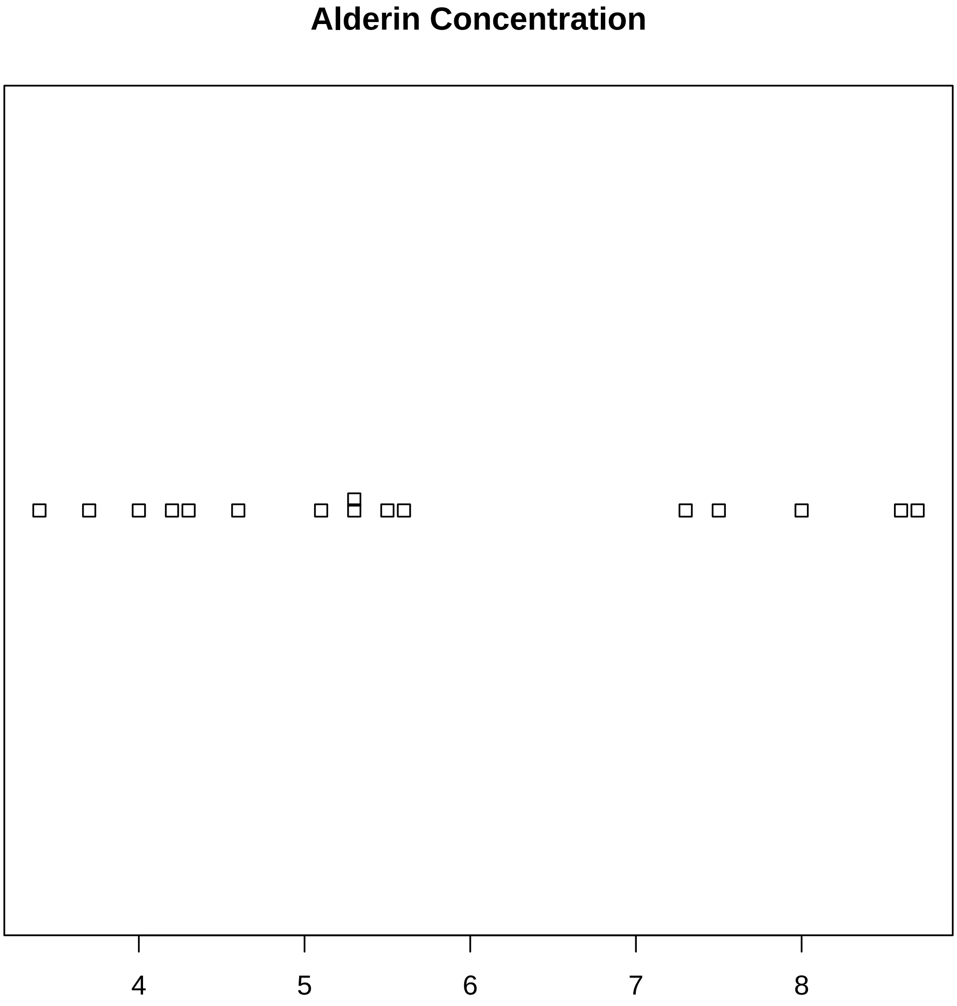
(c) Which graphical summary, the dot plot or stem-and-leaf plot, do you think was better for this data? Why? *The stem-and-leaf graph seems better. You get a better sense for the shape of the data with this summary. Also, it will give you an appropriate simplification of the plot because we are not interested in specific data.*

2. As a part of the United States Department of Agriculture's Super Dump cleanup efforts in the early 1990s, various sites in the country were targeted for cleanup. River X had become contaminated with pesticides because it was located near an abandoned pesticide dump. Measurements of the concentration of aldrin (a commonly used pesticide from the 1970's, which is now banned in many countries) were taken at sixteen

randomly selected locations along the river and are given below:

3.4, 4.0, 5.6, 3.7, 8.0, 5.5, 5.3, 4.2, 4.3, 7.3, 8.6, 5.1, 8.7, 4.6, 7.5, 5.3

(a) Create a dot plot of the data. Do you see anything unusual? *The plot is below. It seems strange that there are no measurements between about 5.6 and 7.2.*

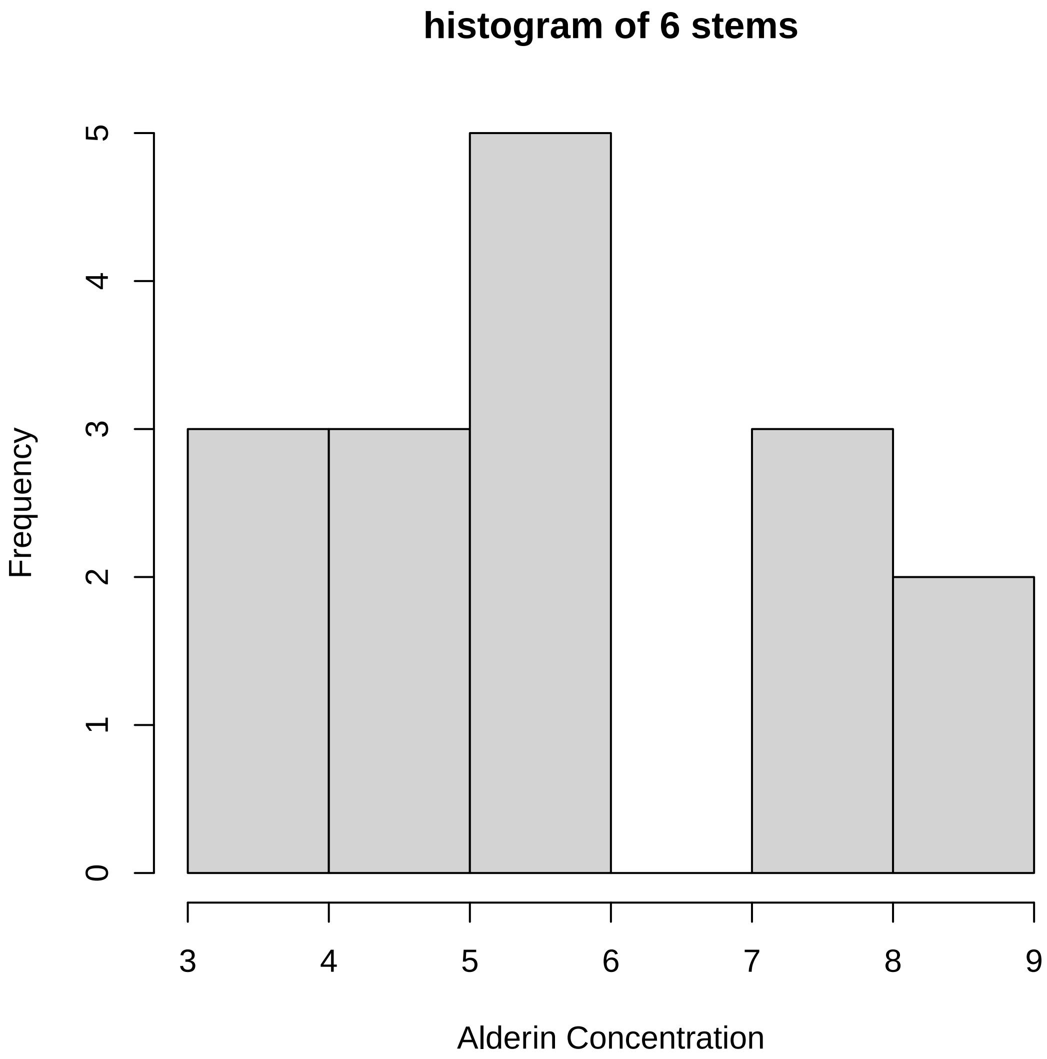


(b) Create a stem-and-leaf plot of the data. Why did you choose the stem values that you did? *The plot is below. Using individual ones places seemed to capture the shape of the data well (show the two clumps). Also, we want each groups to have enough data, and we have enough number of groups.*

```
Key: 3|4 means 3.4

3 | 47
```

```
4 | 0236
5 | 13356
6 |
7 | 35
8 | 067
```



3. What is true about the type of data in both examples that make dot plots and stem and leaf graphs appropriate?

   Both sets of data have numerical measurements - it is quantitative data.

   Qualitative data is defined as the data that approximates and characterizes. Examples of Qualitative Data: The cake is orange, blue, and black in color (qualitative). The property of color is qualitative attribute.

   Examples of Quantitative Data: the cake is 12 inches high. The height of cake should be quantitative attribute.

Examples of Quantitative Data: the cake was baked for 1 hour. The baking time should be quantitative attribute.

Notice that we can say the cake belongs to group 2. This belonging attribute should be qualitative.

4. A professor at UW gives a survey to their class to find out the breakdown of majors in their class. They produce the following graph:

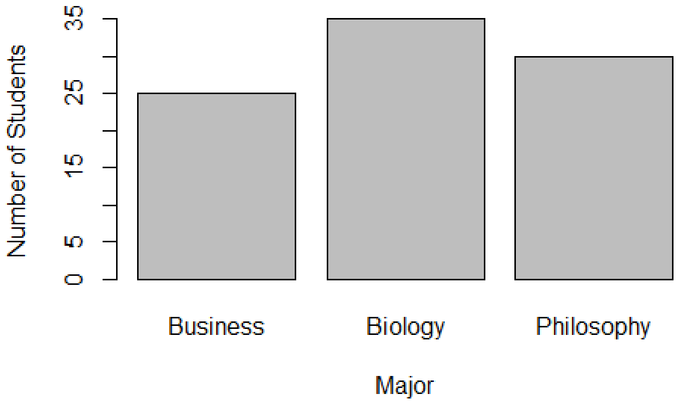


Figure 1: College Majors

What is the variable being measured?

(a) A Single Student

(b) Number of Students

(c) *Major*

(d) Business, Biology and Philosophy

(e) None of the above

The sample would be students. For every measured data in the sample, the variable is major.

The measured subject is students, and the variable is students' feature. Their major is a qualitative variable and for each student, this variable can be Business, Biology and Philosophy.

5. A principal has record of all her 580 students' test scores. She calculates that the average score was a 76% and 45% of the scores were above proficient. She was surprised when she asked a sample of 20 students how their test went and 90% reported scoring proficient. Identify any parameter(s) or statistic(s) in the scenerio described.

The sample is the 20 students that the professor asked. 76% and 45% are both parameters as they are summaries of the performance of all 580 test scores because these data are based on all units. The 90% is a statistic since it is calculated on a sample of students. Statistics should be the function of a sample.