3/24/2022

Frequency Tables
Statistical Graphs – which graph types go with which data types
2.1 Stem, Line, Bar
2.2 Histograms, Polygons, Time Series
2.3 Measures of Location
2.4 Box Plots
Other kinds of statistical graphs: Pareto, Pie,…

Excel Sheet 1:
Frequency Table

| Row Labels | Count of Region |
|---|---|
| East | 252 |
| Midwest | 261 |
| South | 253 |
| West | 234 |
| **Grand Total** | **1000** |

Relative Frequency Table

| Row Labels | Count of History |
|---|---|
| 1 | 23.00% |
| 2 | 21.20% |
| 3 | 25.50% |
| NA | 30.30% |
| **Grand Total** | **100.00%** |

From Excel Sheet 2:
Two-way Table

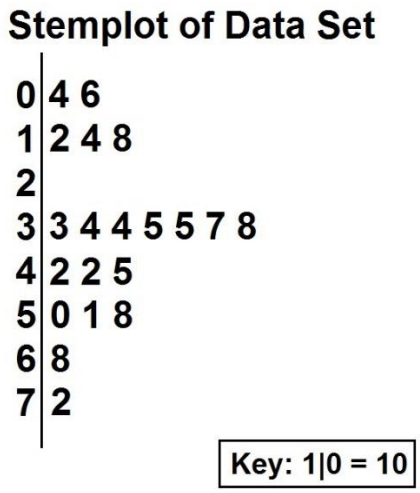
| Count of Region | Column Labels | | | | |
|---|---|---|---|---|---|
| **Row Labels** | **1** | **2** | **3** | **NA** | **Grand Total** |
| East | 58 | 52 | 63 | 79 | 252 |
| Midwest | 57 | 63 | 59 | 82 | 261 |
| South | 63 | 47 | 61 | 82 | 253 |
| West | 52 | 50 | 72 | 60 | 234 |
| **Grand Total** | **230** | **212** | **255** | **303** | **1000** |

Statistical Graphs.

Stemplot (Stem-and-Leaf Plot) – this is a graph type we can't do in Excel.
Line Graph – connects ordered values on a graph (the horizontal axis is the order, often this variable is time)

Bar Graph – for categorical/qualitative data, it graphs the frequency table.

Stemplot example:




The number to the left of the bar is called the stem. And the number to the right of the bar are the leaves.

The advantage of a stemplot is that it contains all the original data.
4, 6, 12, 14, 18, 33, 34, 34, 35, 35, 37, 38, 42, 42, 45, 50, 51, 58, 68, 72.

The leading is often treated as the stem, but if I were to add 100 to all these values, so in that case the first two digits could be the stem.

If the data is clustered very tightly, you can break up the stems into two categories, so that the leaves 0 – 4 go in the first stem group, and 5-9 go into the second.

```
epresents 0.12
it: 0.01
    n: 164
  2* | 12
  2. | 677
  3* | 12223344
  3. | 5666778889
  4* | 00011222333334444
  4. | 55556666678888889999
  5* | 01111113333344
  5. | 555666777788888999999
  6* | 11112233444441010101010
  6. | 55667777899
  7* | 11222333101010
  7. | 556677888889
  8* | 011111222
  8. | 55689
  9* | 013
```

(I cut off the key!)

Back-to-back stemplots for comparison:

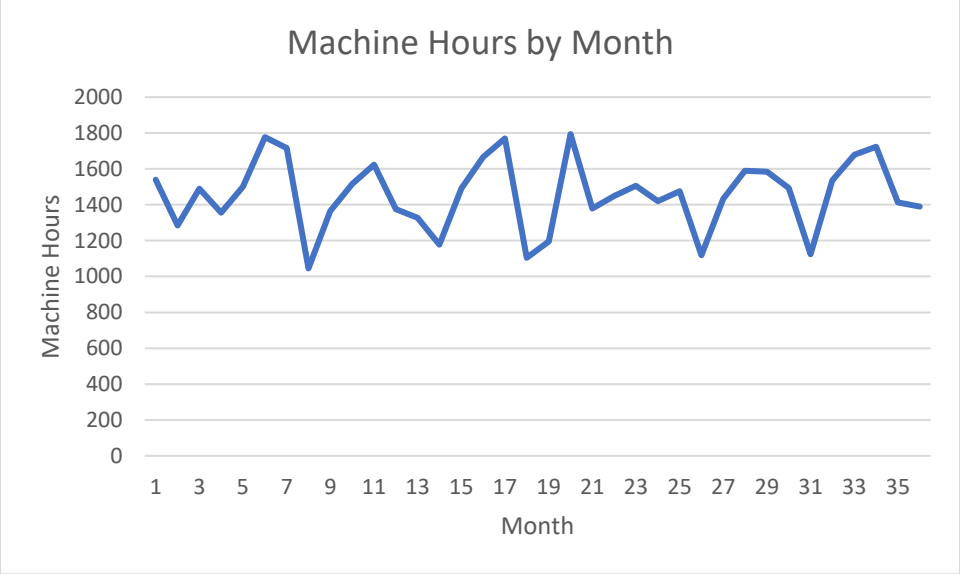
| Monday section | | | | | | | | | | | Wednesday section | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | 9 | 4 | | | | | | | | | | | | |
| | | | | | | 8 | 6 | 1 | 5 | | | | | | | | | | | |
| | | | | | 8 | 3 | 0 | 0 | 6 | | 2 | 3 | 9 | | | | | | | |
| | | | | 9 | 8 | 6 | 6 | 2 | 1 | 7 | | 2 | 6 | | | | | | | |
| 9 | 9 | 8 | 8 | 8 | 8 | 7 | 6 | 6 | 4 | 1 | 8 | | 1 | 1 | 2 | 3 | 5 | 6 | 8 | 8 | 9 |
| | | | | | | 5 | 5 | 1 | 0 | 9 | | 0 | 2 | 2 | 2 | 5 | 5 | 5 | 5 | | |

Stemplots are for numerical/quantitative data. They allow you to see the way the data is distributed without losing the original values.

Stemplots are best for relatively small datasets.

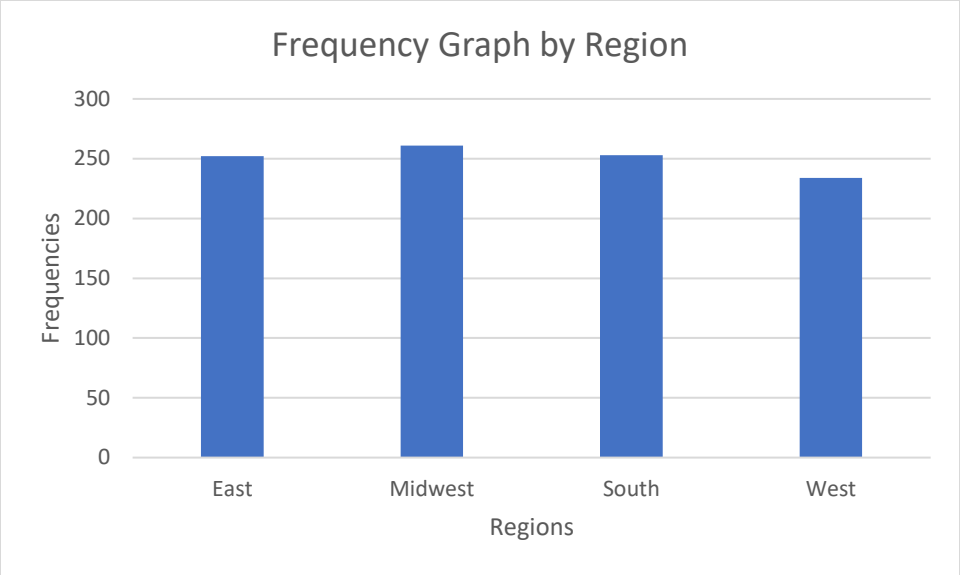
Line graph – we will build this in Excel. Generally, these are based ordered observations, often using time to provide the sequence (for the horizontal axis). The vertical axis is a numerical/quantitative variable.

Excel Sheet 3
Make sure you have a good graph.
Your graph should have a title (a descriptive title).
And you have axis labels.



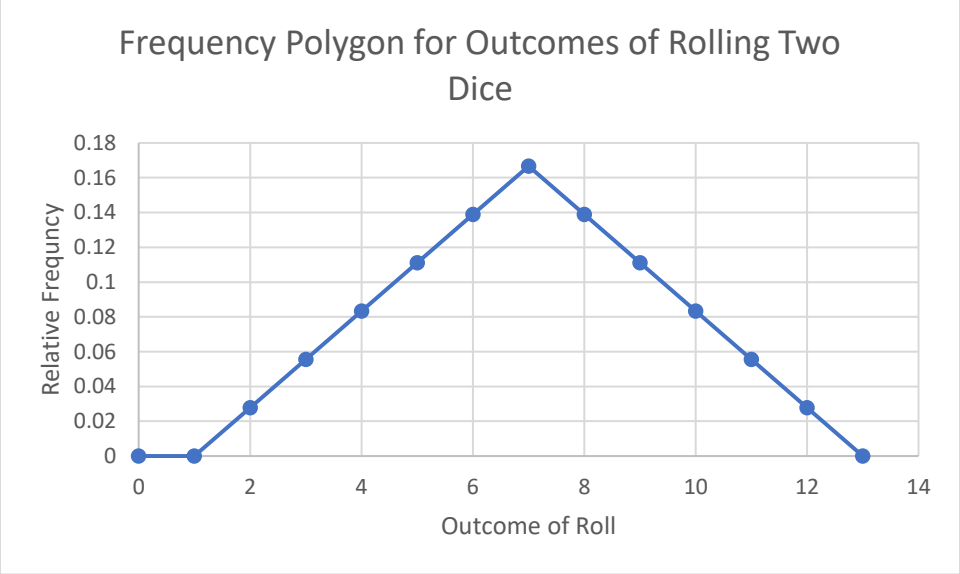

Bar Graph – used to graph the frequencies or relative frequencies of categorical data.
Want to start with a frequency table.

The Regional data from Sheet 1 is a good candidate for a bar graph.
Do not include the total column.



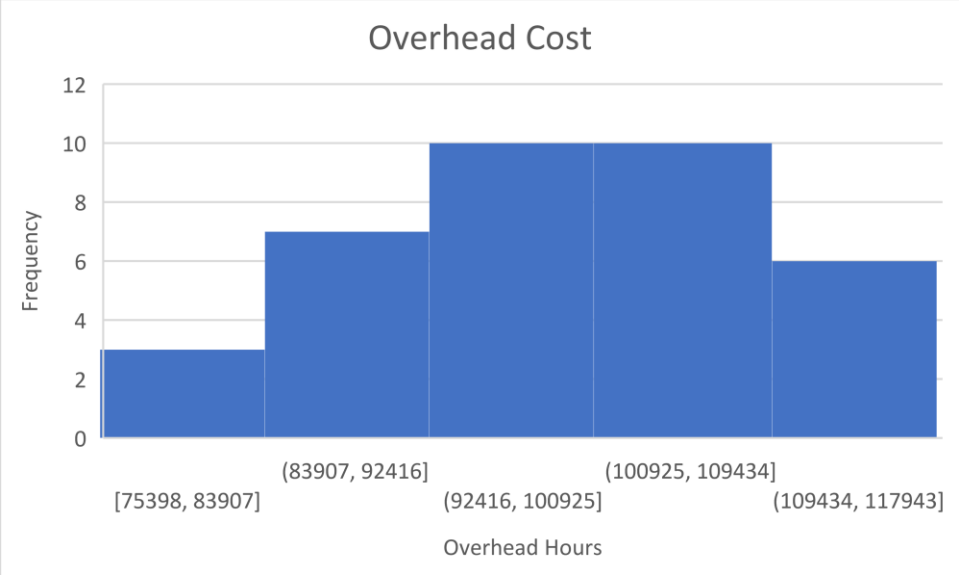

Make sure, as before, you have a descriptive title, and axis labels. And for bar graphs, it's important to start at 0 otherwise the differences in the heights of the bars can send the wrong message (they can be misleading).

Time Series graphs are basically just line graphs.
Polygons: are frequency polygons: line graphs that graph relative frequency for ordered variables.

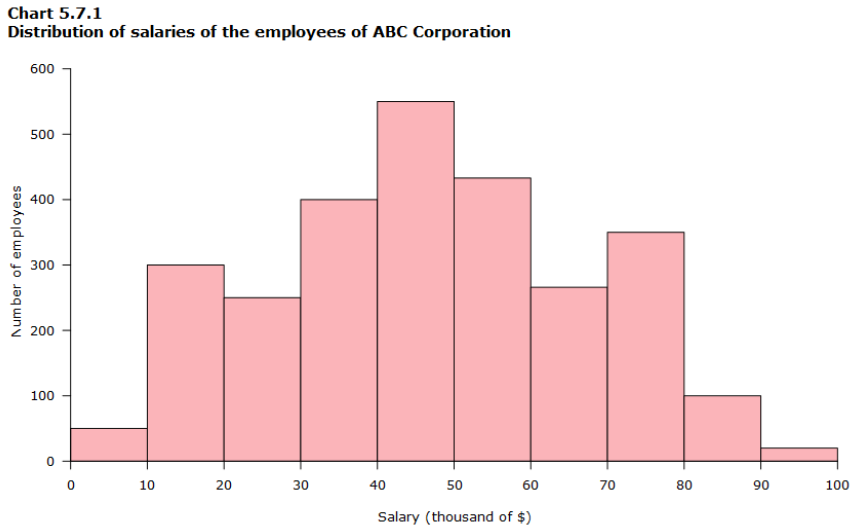



Histogram is a bar graph for quantitative/numerical data. The data is grouped into "bins" of a common width. The number of observations in each bin is counted, and then the result is a frequency table, which can then be graphed like a bar graph. Only the bins are not qualitative categories, they are ranges of values.

The rule of thumb on the bins is between 5 and 20. More bins with more observations.




Experiment with the number of bins. Too many will not be meaningful, nor will too few.




In Excel, the horizontal axis is given as intervals, but on this version, just the breaks are plotted.

Measures of Location (we will revisit these on Tuesday as part the descriptive section).

Boxplots (Box-and-whisker plots) depend on these measures of location: they depend mostly on the 5-number summary.

5-number summary: Minimum, 1st Quartile, Median, 3rd Quartile, Maximum.

Median: This is the value where half the data is below this value and half the data is above this value (in an ordered list). If the list has an odd number of values, then the middle value is the median. If the list has an even number of values, then the median is the average of the two values closest to the middle.

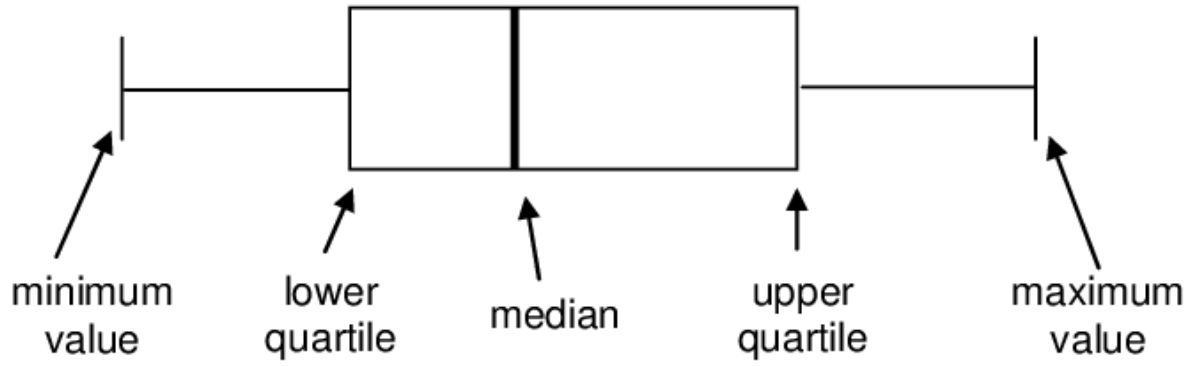
4, 6, 12, 14, 18, 33, 34, 34, 35, 35, 37, 38, 42, 42, 45, 50, 51, 58, 68, 72.

There are two values left in the middle, the average is the median: 36

Quartiles: they are the median of each half of the data.
The first quartile is the median of the first half (bottom half)
The third quartile is the median of the second half (top half)

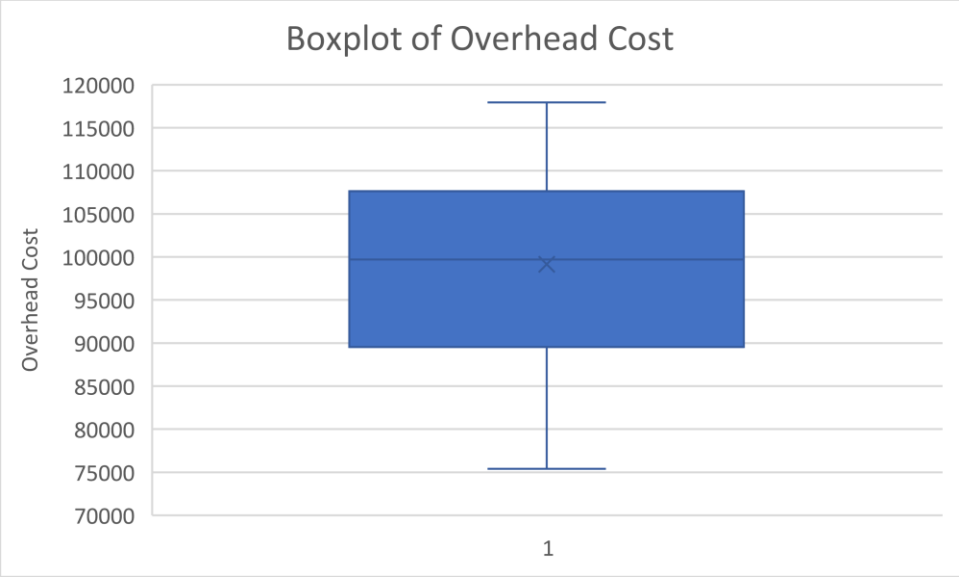
Bottom half: 4, 6, 12, 14, 18, 33, 34, 34, 35, 35
Top Half: 37, 38, 42, 42, 45, 50, 51, 58, 68, 72

1st quartile: the average of 18 and 33 is 25.5
3rd quartile: the average of 45 and 50 is 47.5

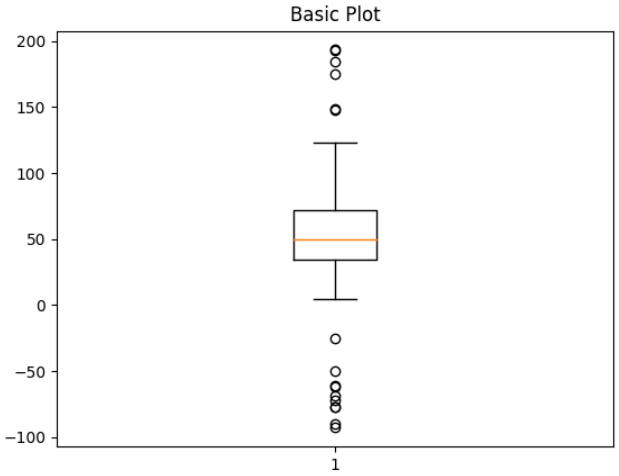



More sophisticated versions of the boxplot will calculate the "fences". These determine if any of the values are considered "extreme".

Boxplots are for numerical/quantitative data (just like the histogram).

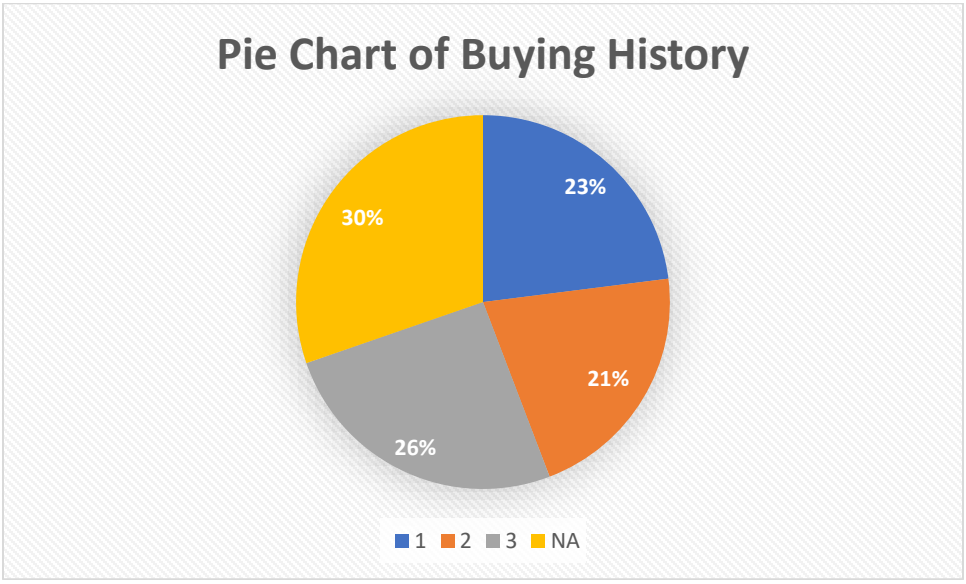


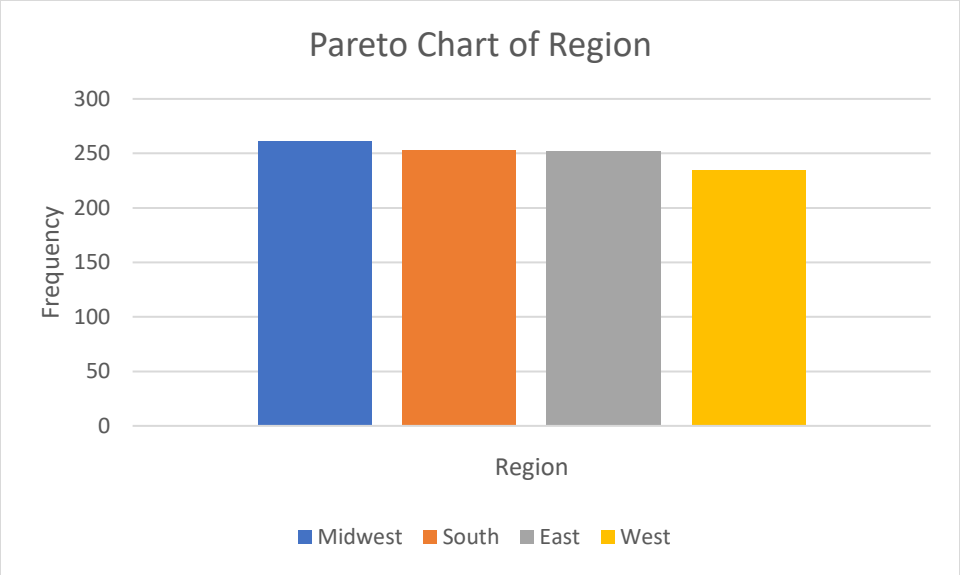

A version with extreme values:




Pie chart and a Pareto chart.
These are both graphs for categorical data. And specifically only for one variable.
Pie chart is sometimes also called a circle graph (a doughnut graph is similar it just has a hole in the middle). It graphs relative frequencies. The percentages must add up to 100%.
The Pareto chart is basically a bar graph where the bars are ordered by height (either smallest to largest, or largest to smallest).

Make both from the frequency table.






Good graphs vs. Bad graphs.
Good graphs have descriptive titles (you need to know what the graph is about without having the original data).
They have axis labels (for pie charts this is the percentage on the slices).
They may need a legend if the graph is color-coded or otherwise complex (eg. A graph of a two-way table).
Some types of graphs (like bar graphs) should start the vertical axis at 0.

Elements of a bad graph:
3D effects can make a graph more difficult to read, and distort perspective
Used "stacked" options only when it makes sense to do so (the top of the bar or top line is the total).

The orientation of the graph can be confusing (bars vs. columns)

You may want to be mindful of colorblindness (don't contrast red and green).

You could absolutely be tested on good vs. bad graphs.

Examples of Bad Graphs:
https://www.statisticshowto.com/probability-and-statistics/descriptive-statistics/misleading-graphs/

https://towardsdatascience.com/misleading-graphs-e86c8df8c5de

https://www.businessinsider.com/the-27-worst-charts-of-all-time-2013-6

Next Time: descriptive statistics.