Stat 201:
Introduction to Statistics

Standard 4: Graphical Summaries

Chapter Two

Summaries

Graphical Displays

Variable Type	Graphical Display	Numerical Summary
Categorical	Pie chart or bar graph	Frequency table
Quantitative	Histogram or box plot – can also try dotplot or stem & leaf	Quantitative Summary
1-Categorical and 1-Quantitative	Side by Side boxplots	Quantitative Summary for groups
2-Categorical	Side by side pie charts or bar graphs best: stacked bar chart	Contingency Table or side by side frequency tables
2-Quantitative	Scatter plot	Side by side Quantitative Summaries

Misrepresentation of Data

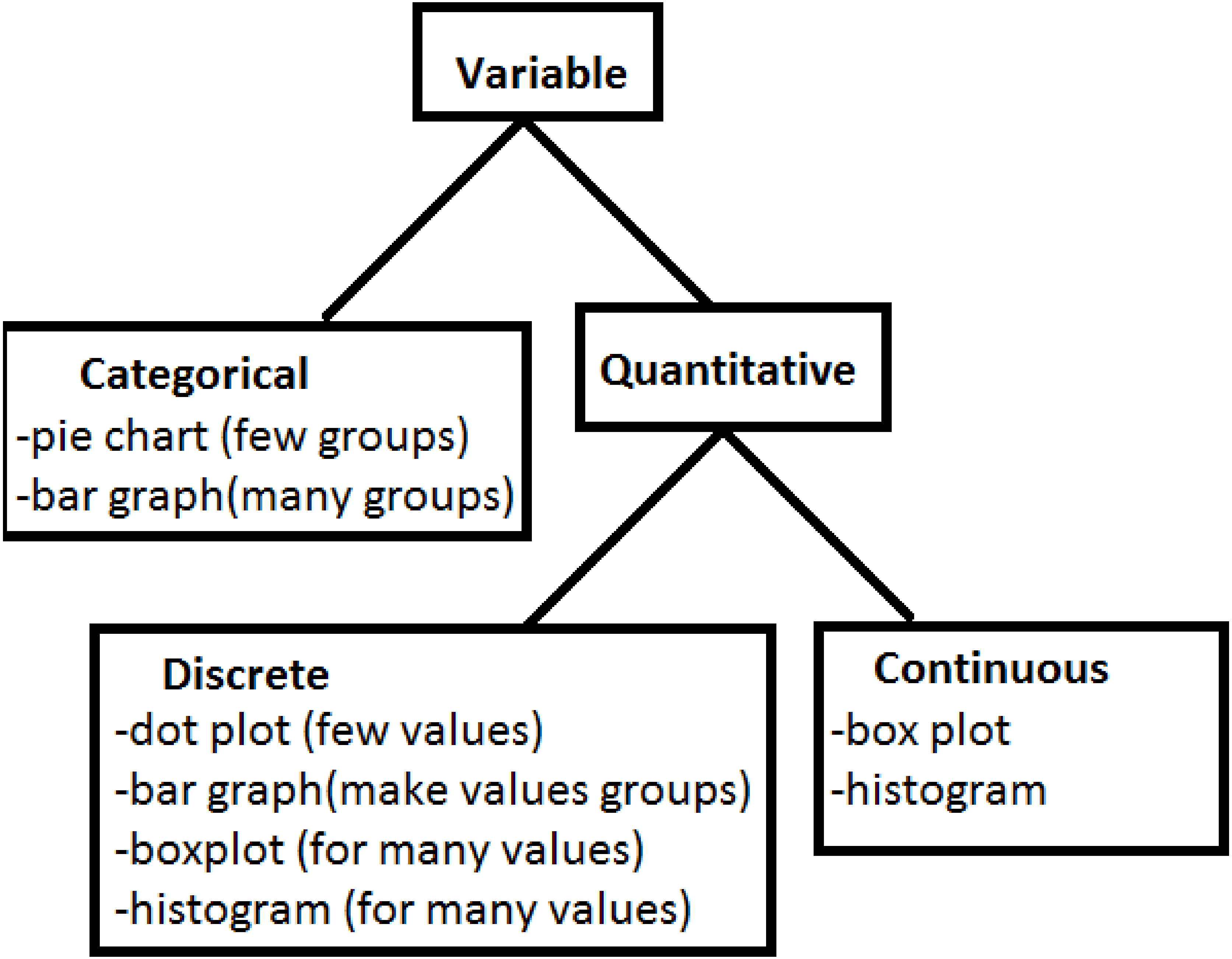
- You should be able to look at your graphs and realize when you’ve made a mistake

  -The percentages of all relative frequency graphs should add to 1 or 100%

  -The scale should be understandable and constant

  -Consider whether or not you need to start your y axis at zero or caution against misreading the graph

  -Graphs should be simple and easy to interpret correctly in just a few moments.



Variable

Categorical

-pie chart (few groups)

-bar graph(many groups)

Quantitative

Discrete

-dot plot (few values)

-bar graph(make values groups)

-boxplot (for many values)

-histogram (for many values)

Continuous

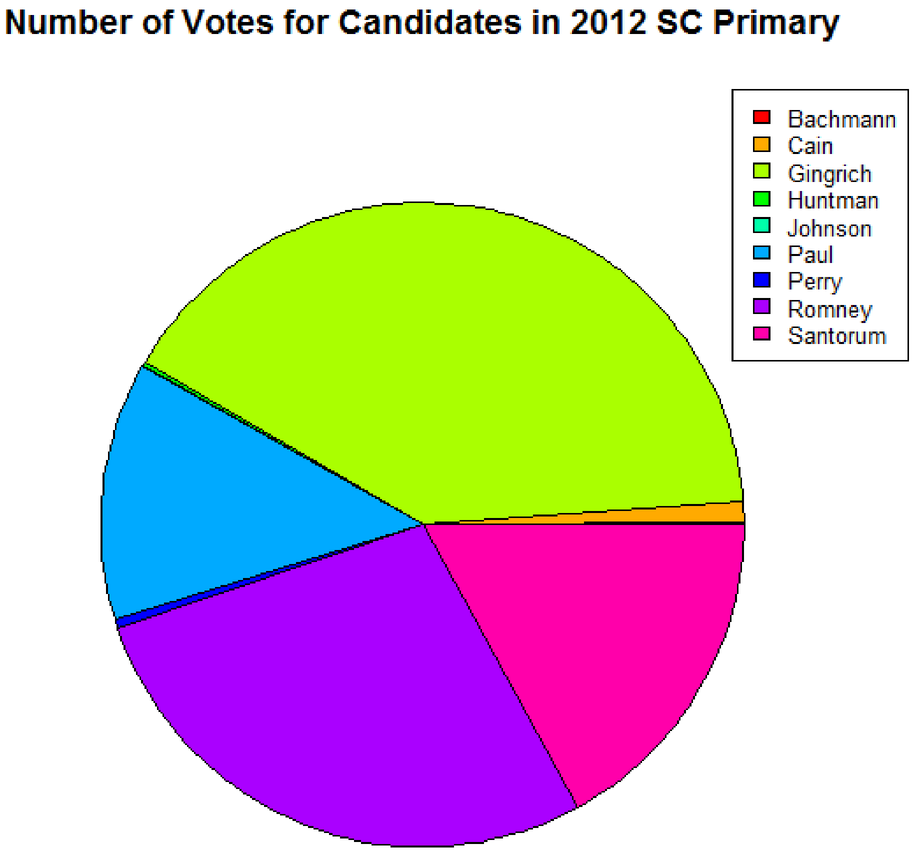
-box plot

-histogram

Walkthrough

Summarizing Qualitative Data: Pie Chart

- Useful when there are a small number of categories



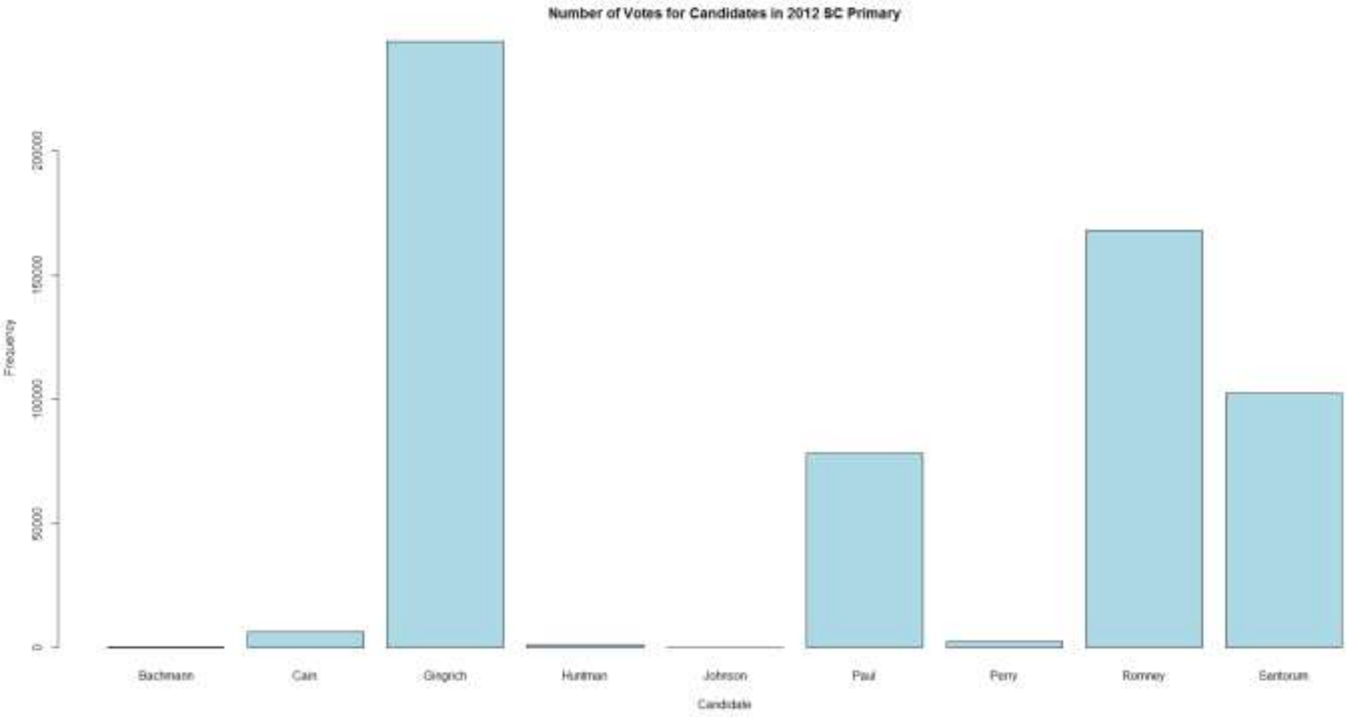
Data: Graphical Summary

- **StatCrunch Command:**

Graph→Pie Chart→w/data→ Select your variable(s)→Compute

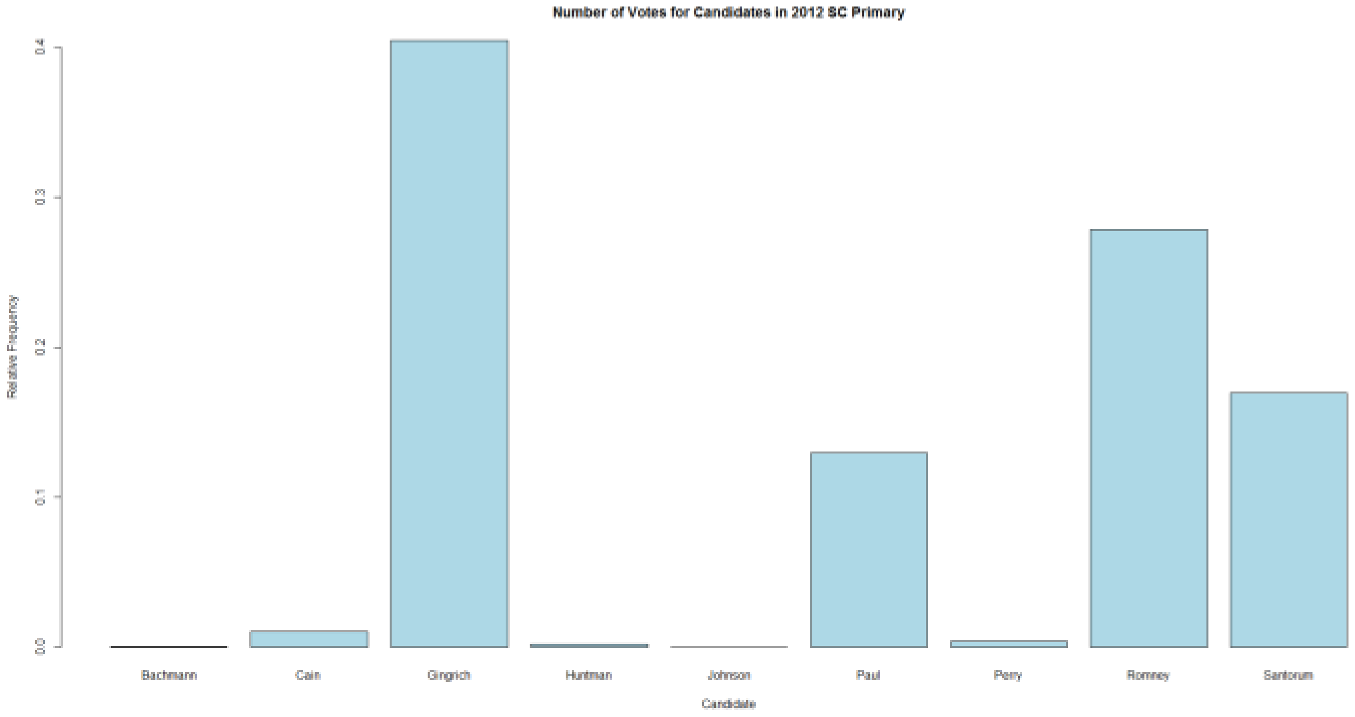
Summarizing Qualitative Data: Bar Graph

- Useful when there are many categories of the variable
- Useful to compare groups



Summarizing Qualitative Data: Bar Graph

- **Note:** the relative frequency chart has the same shape but a different y-axis



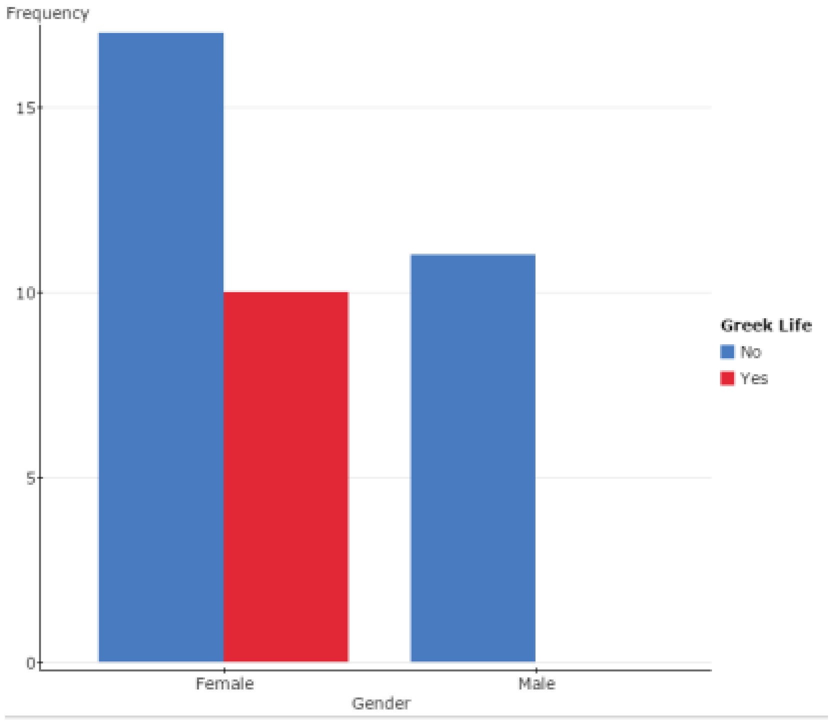
Data: Graphical Summary

- **StatCrunch Command:**

Graph→Bar Plot→w/data→ Select your variable(s)→Compute

Categorical Summary: Side by Side Bar Graph

- We could draw bar graphs side by side to compare the data for the two different groups.



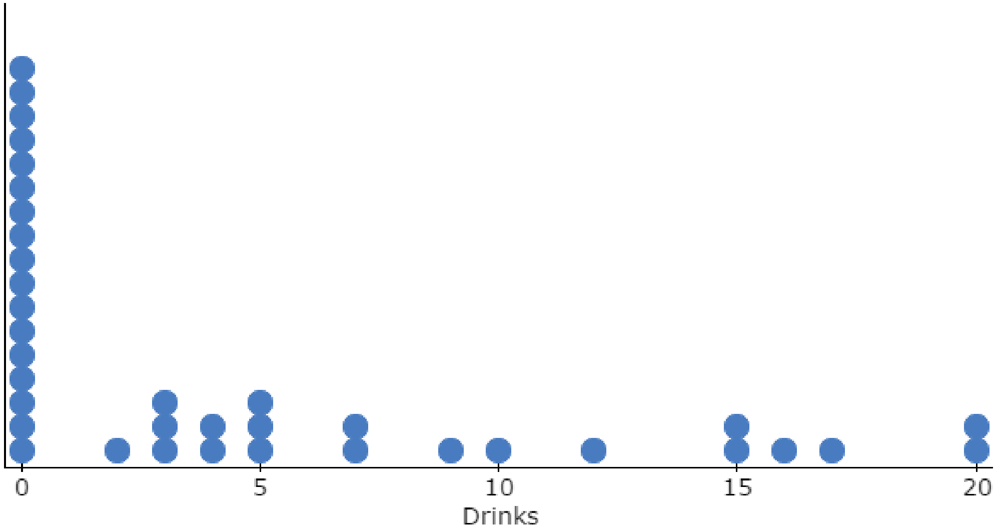
Data: Graphical Summary

- **StatCrunch Command:**

Graph→Bar Plot→w/data→ Select the variable you'd like on the x-axis→Group by the variable you would like the bars to be split by→Compute

Quantitative Summary: Dot Plot

- Useful for smaller datasets
- Useful for finding outliers
- I don’t like these – histograms are **almost** always better



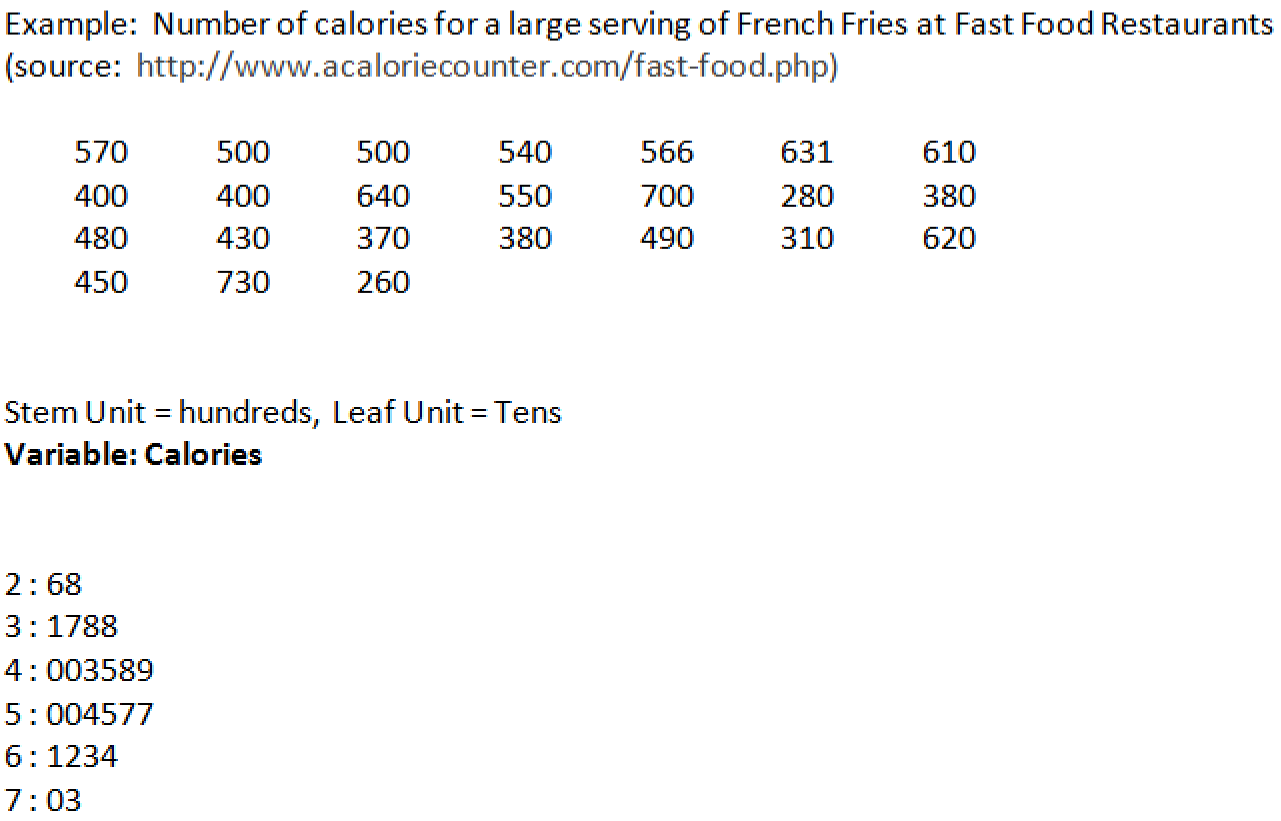
Data: Graphical Summary

- **StatCrunch Command:**

Graph→Dot Plot→w/data→ Select the variable(s)→Compute

Quantitative Summary: Stem and Leaf

- Retains actual data values



Example: Number of calories for a large serving of French Fries at Fast Food Restaurants
(source: http://www.acaloriecounter.com/fast-food.php)

570	500	500	540	566	631	610
400	400	640	550	700	280	380
480	430	370	380	490	310	620
450	730	260				

Stem Unit = hundreds, Leaf Unit = Tens
Variable: Calories

2 : 68
3 : 1788
4 : 003589
5 : 004577
6 : 1234
7 : 03

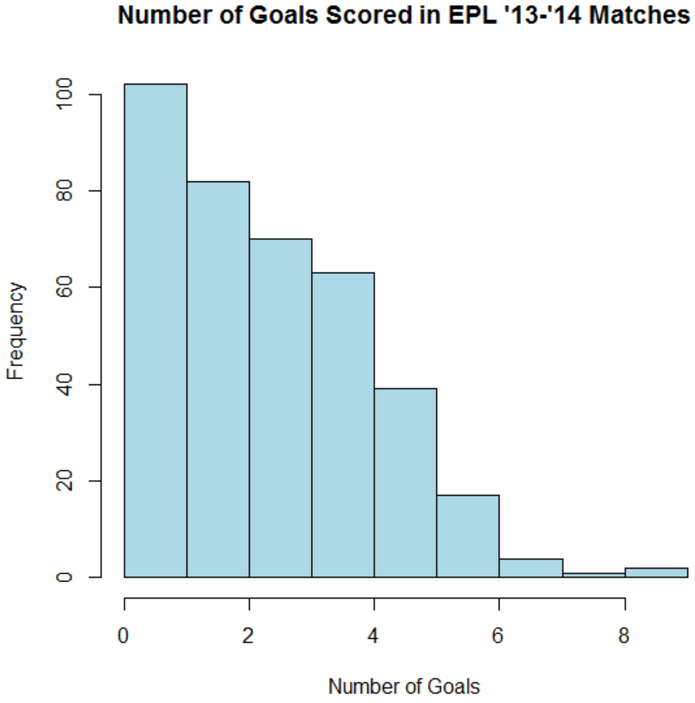
Data: Graphical Summary

- **StatCrunch Command:**

Graph→Stem and Leaf→ Select the variable(s)→Compute

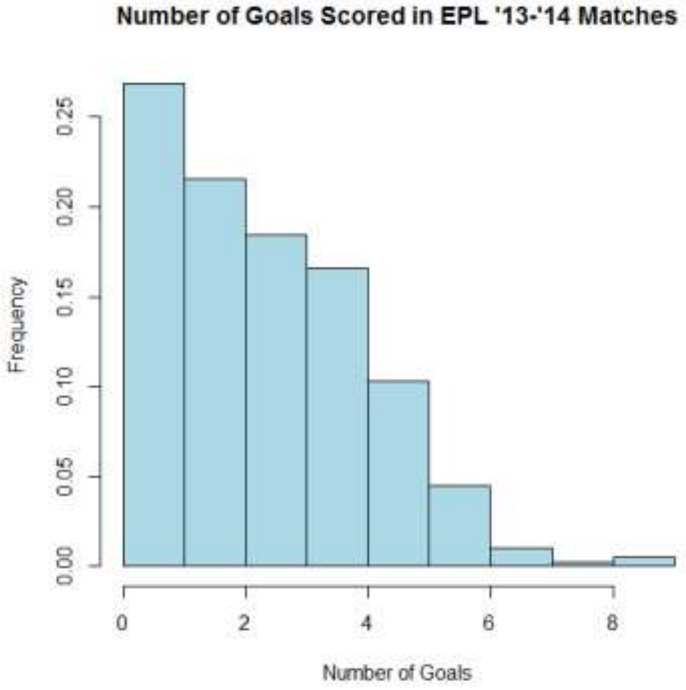
Summarizing Quantitative Data: Histogram

- Histograms are used to summarize quantitative data and will be our main tool for continuous data



Summarizing Quantitative Data: Histogram

- **Note:** the relative frequency chart has the same shape but a different y-axis



Data: Graphical Summary

- **StatCrunch Command:**

Graph→Histogram→ Select the variable(s)→Compute

Histograms Vs. Bar Charts

- With bar charts, each column represents a group defined by a categorical variable

- With histograms, each column represents a group defined by a quantitative variable.

Histograms Vs. Bar Charts

- With bar charts, each column represents a group defined by a class of a qualitative (categorical) variable

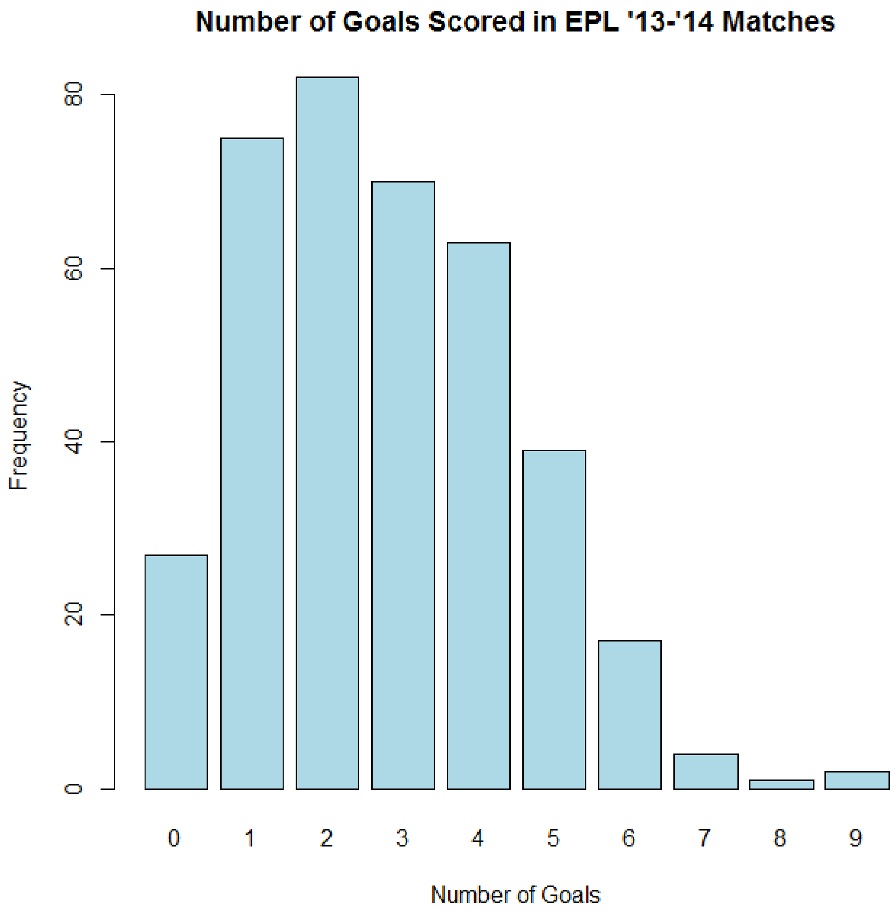
- With histograms, each column represents a group defined by a quantitative variable. R will automatically generate classes for the quantitative data

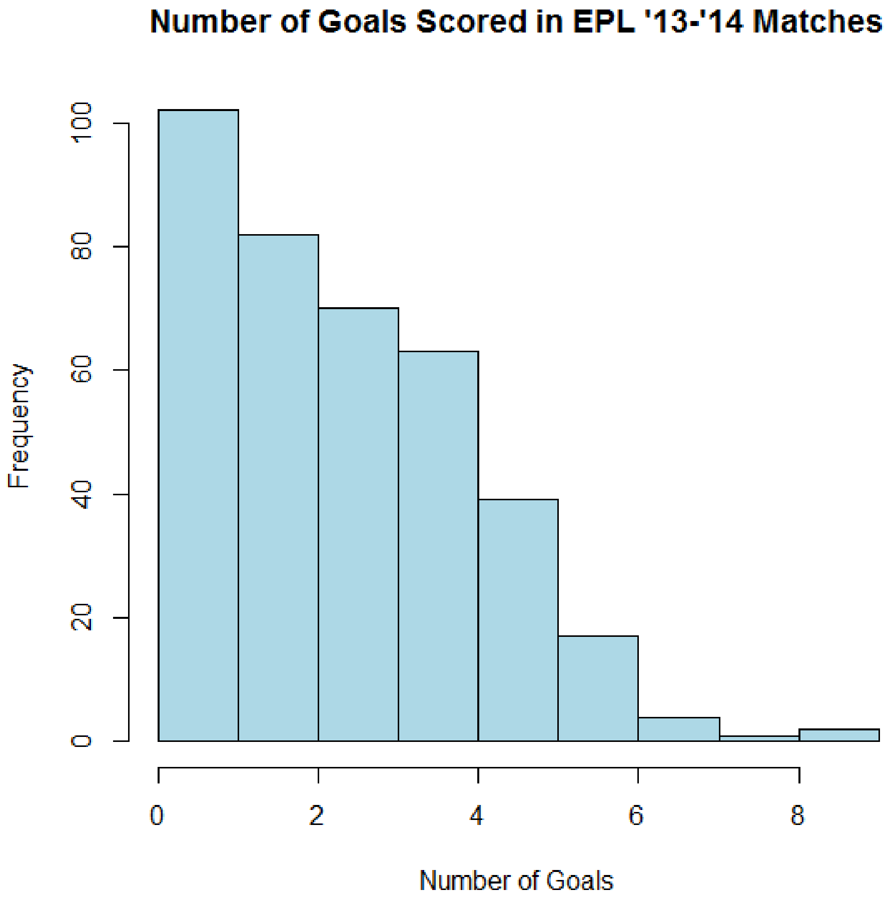
Histograms Vs. Bar Charts

- In our example of EPL goals over the ’13-’14 season the groups that R creates for the histogram are as follow

[0,1]	**102**
(1,2]	**82**
(2,3]	**70**
(3,4]	**63**
(4,5]	**39**
(5,6]	**17**
(6,7]	**4**
(7,8]	**1**
(8,9]	**2**

Histograms Vs. Bar Charts





Histograms Vs. Bar Charts

- In this case, because there are so few observable values the histogram is actually a little misleading – it just combines the bars at 0 and 1 and the rest is the same as the bar plot

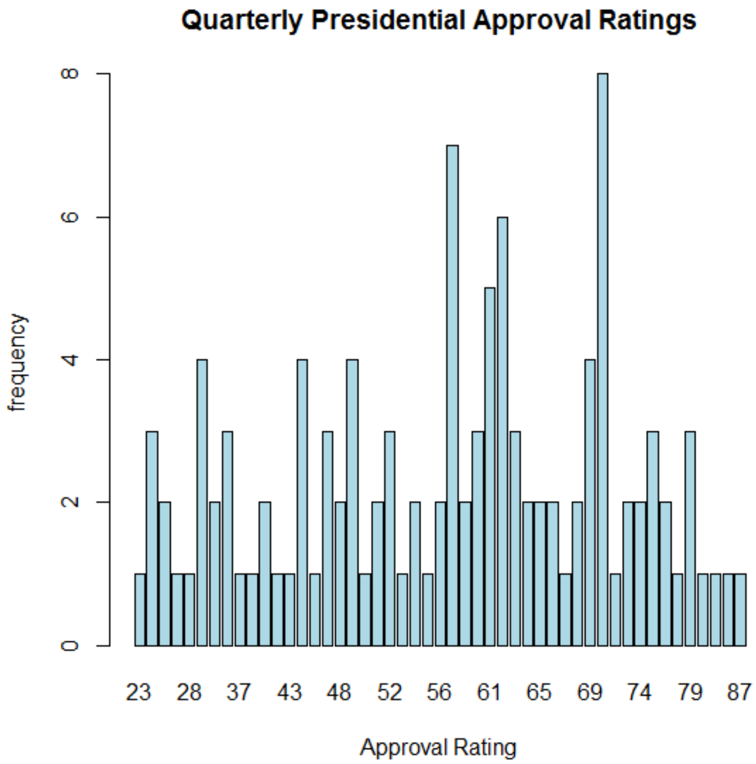
Summarizing Quantitative Data: Histograms

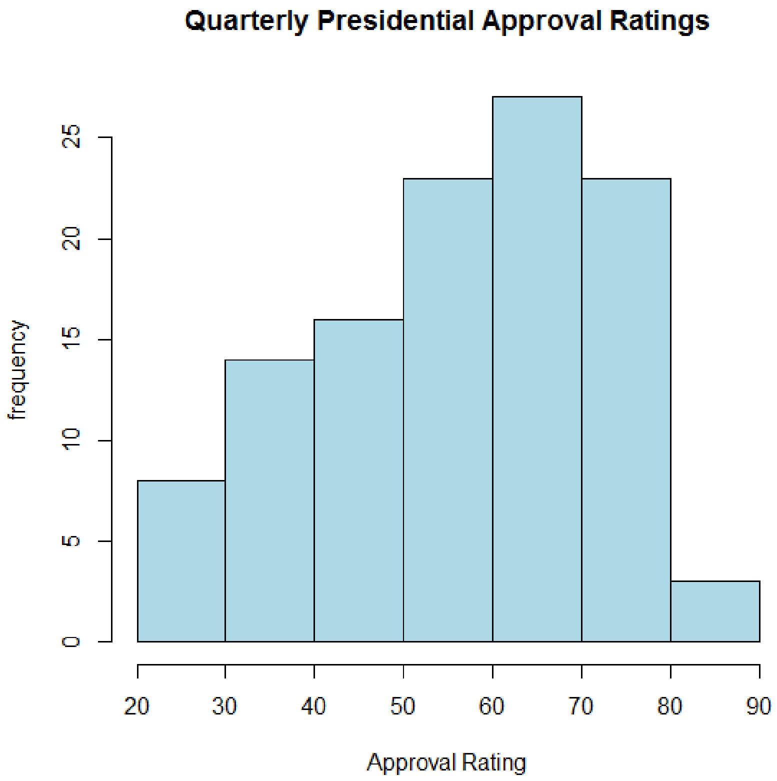
- Let’s consider a different dataset – as we mentioned earlier, the small number of observable values allows us to use the qualitative(categorical) approach with this EPL data
- We will continue looking at histograms by considering the discrete quantitative data considering the quarterly presidential approval ratings from ‘54 to ‘74

Summarizing Quantitative Data: Histograms

- Among the quarterly presidential approval ratings there are 49 observable values ranging from 23 (Truman in ‘51) to 87(Truman in ‘45)
- Here, if we followed what we did for qualitative (categorical data) we would find a frequency table with 49 rows and a bar graph with 49 bars
- Here a histogram is easily a better visual

Summarizing Quantitative Data: Histograms





Histograms Vs. Bar Charts

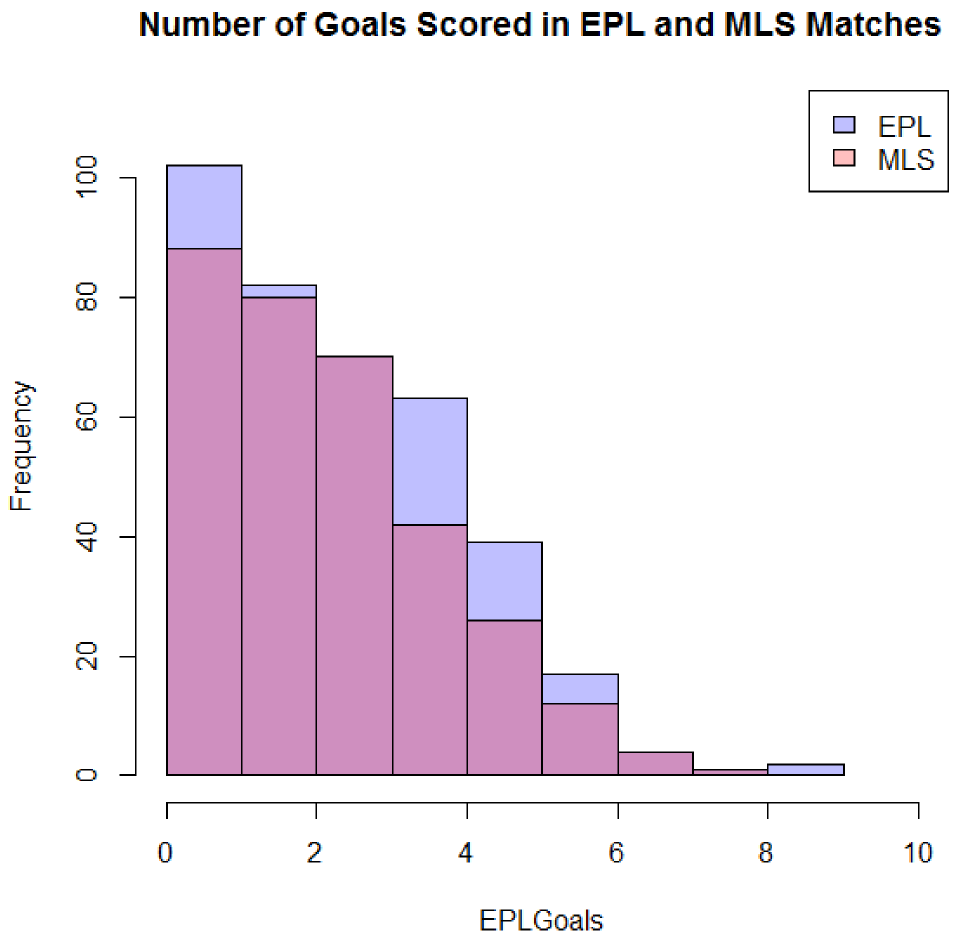
- In our example of Presidential approval ratings the groups that R creates for the histogram are as follow:

[20,30]	**8**
(30,40]	**14**
(40,50]	**16**
(50,60]	**23**
(60,70]	**27**
(70,80]	**23**
(80,90]	**43**

Talking about Two Things at Once

- In many cases we’re looking at two groups and comparing them.
- Here we consider the EPL goals data and compare it to another league to see if teams score more or less over their season
- The following graphs compare goals in the EPL ’13-’14 season and goals in the MLS ’13 season

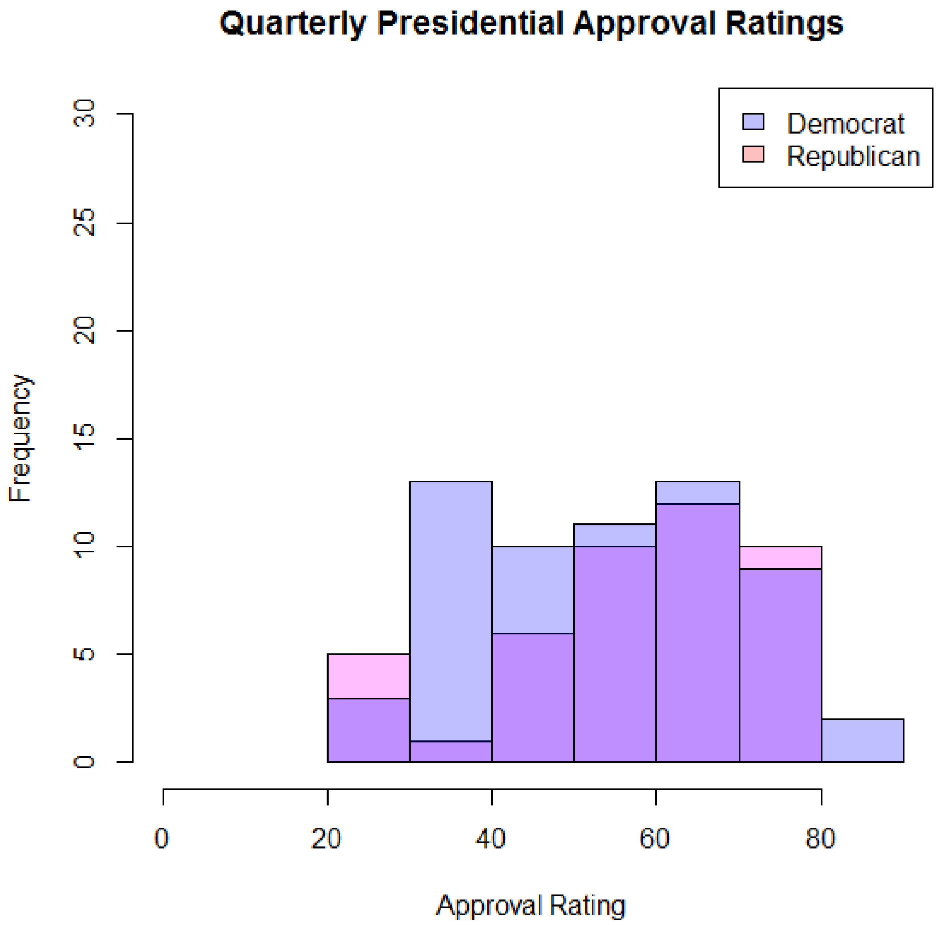
Talking about Two Things at Once



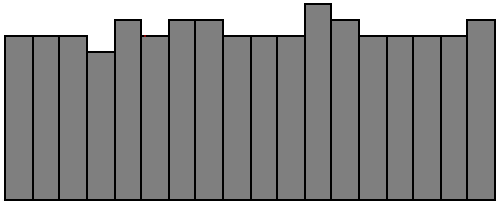
Talking about Two Things at Once

- Here. we consider the presidential approval data and split it into democratic and republican presidents to compare the two parties ratings
- The following graphs compare quarterly ratings of republican and democrat presidents

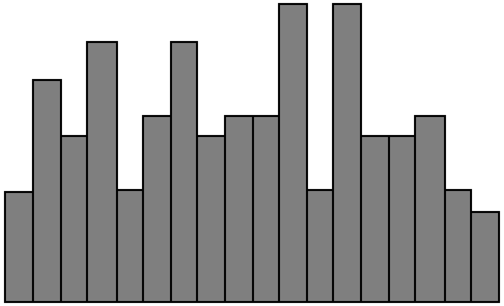
Talking about Two Things at Once



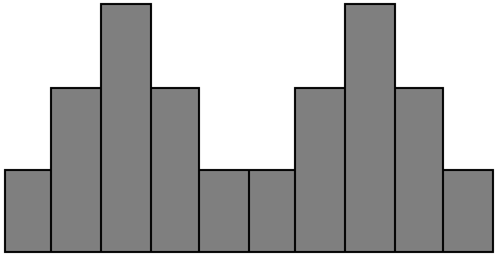
Quantitative Summary: Histogram Shapes



Uniform

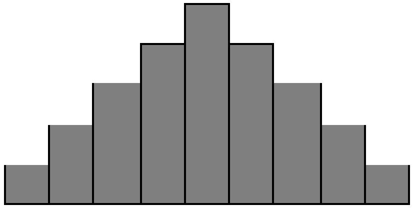


Random



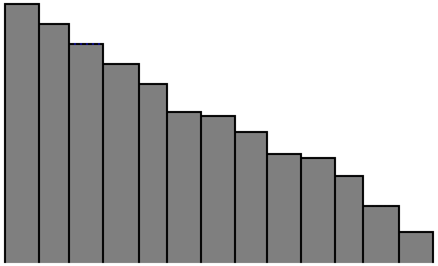
Bimodal

Quantitative Summary: Histogram Shapes



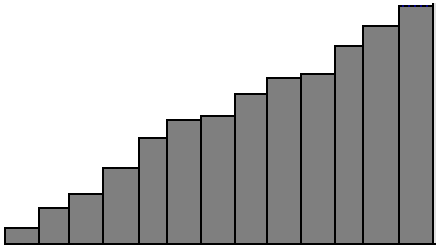
Bell-shaped - Unimodal

$mean \approx median$



Skewed Right

$mean > median$

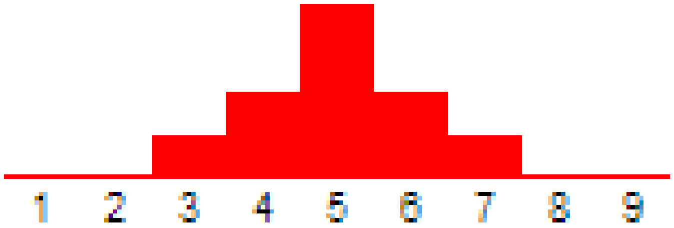


Skewed Left

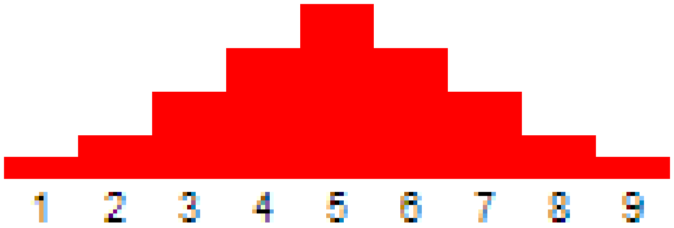
$mean < median$

Histogram

- Spread:



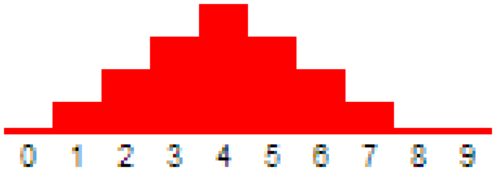
Less spread



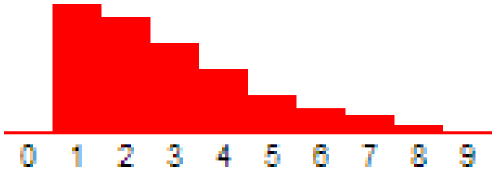
More spread

Histogram

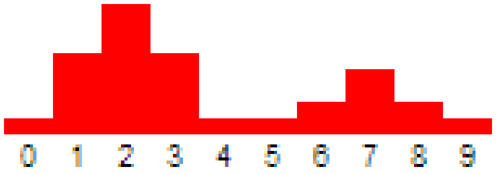
- Shape:



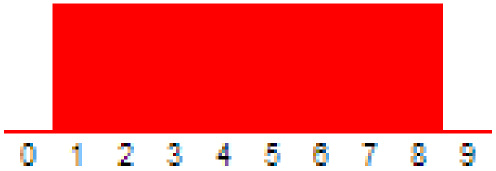
Symmetric, unimodal, bell-shaped



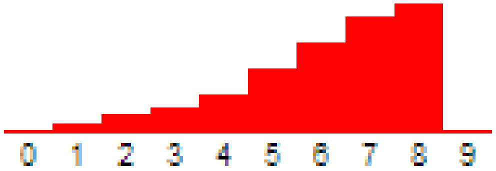
Skewed right



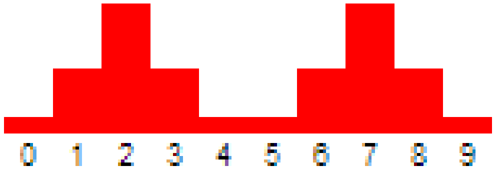
Non-symmetric, bimodal



Uniform



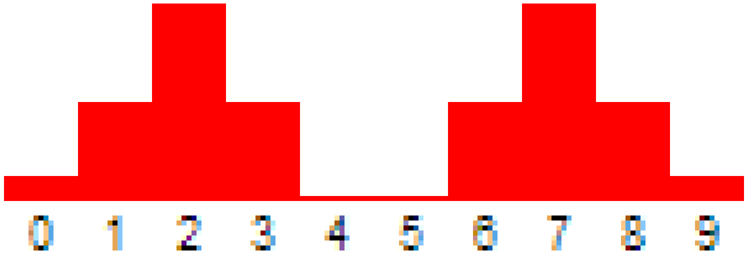
Skewed left



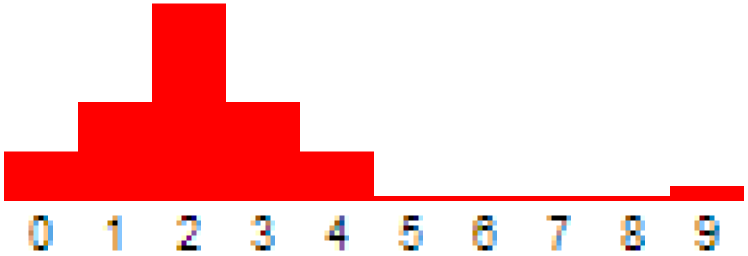
Symmetric, bimodal

Histogram

- Gap vs. Outlier:



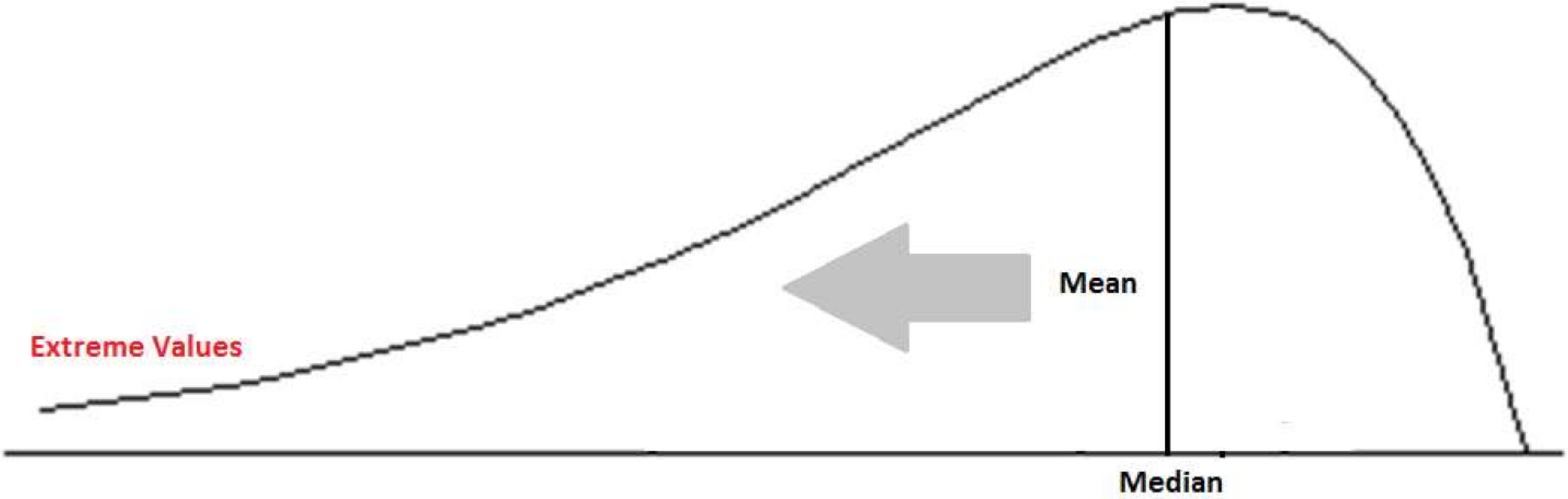
Gap



Outlier

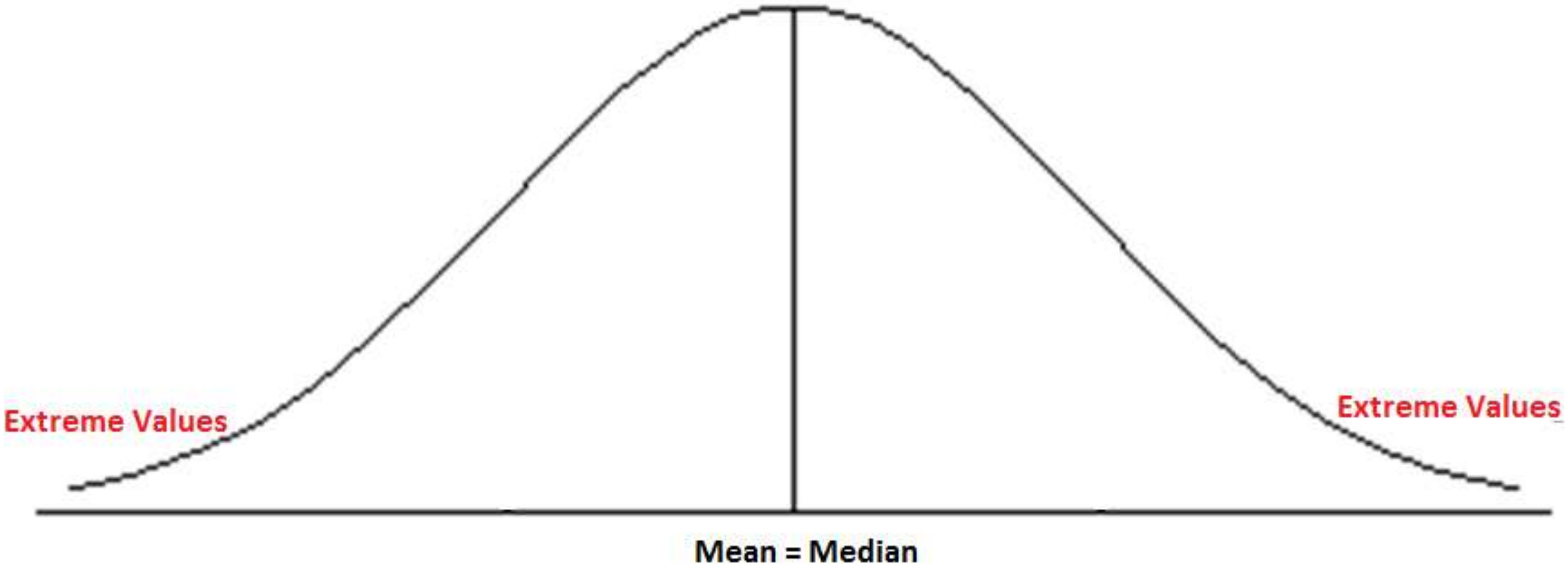
Quantitative Summary: Histograms – Left Skewed

- Here we see a left skewed graph – the extreme values on the left drag the mean to the left tail causing Mean<Median



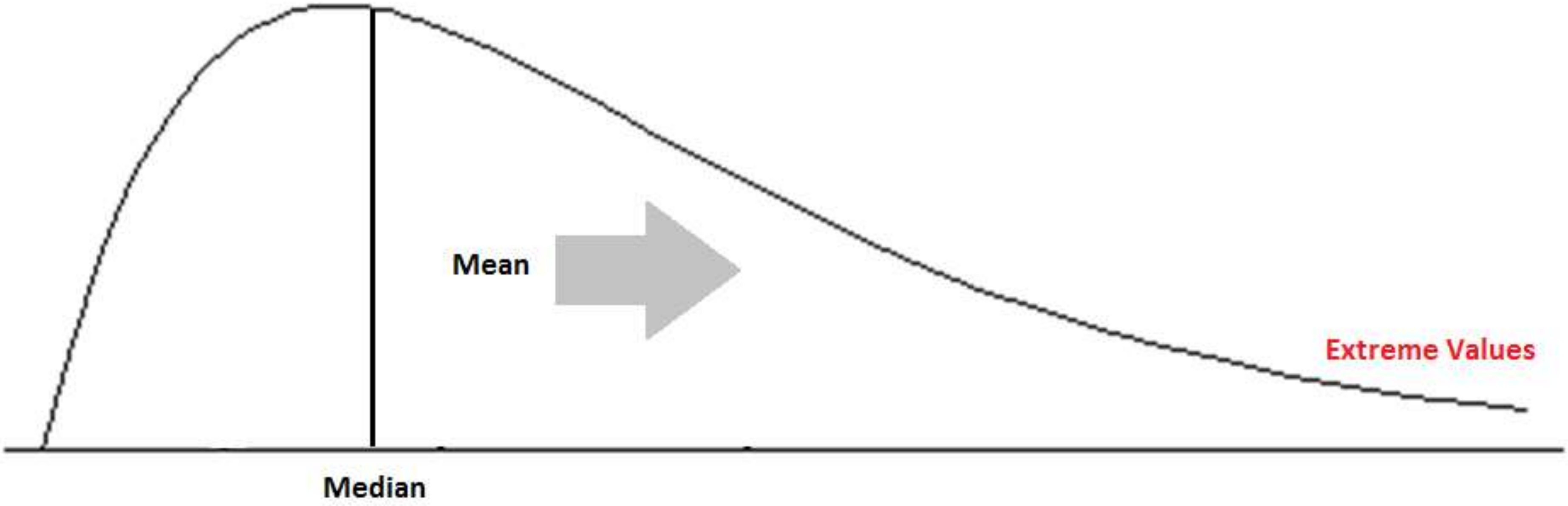
Quantitative Summary: Histograms – Bell Shaped

- Here there is no skew – the extreme values on both side cancel any outlying effect on the mean

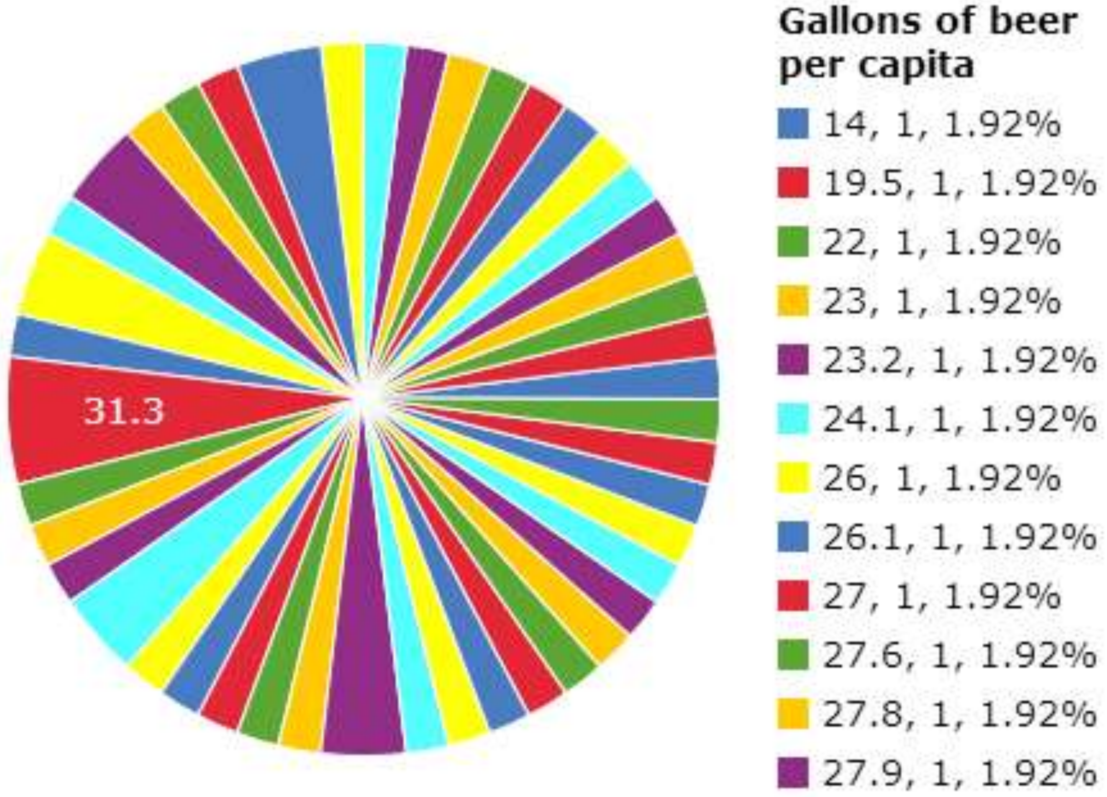


Quantitative Summary: Histograms – Left Skewed

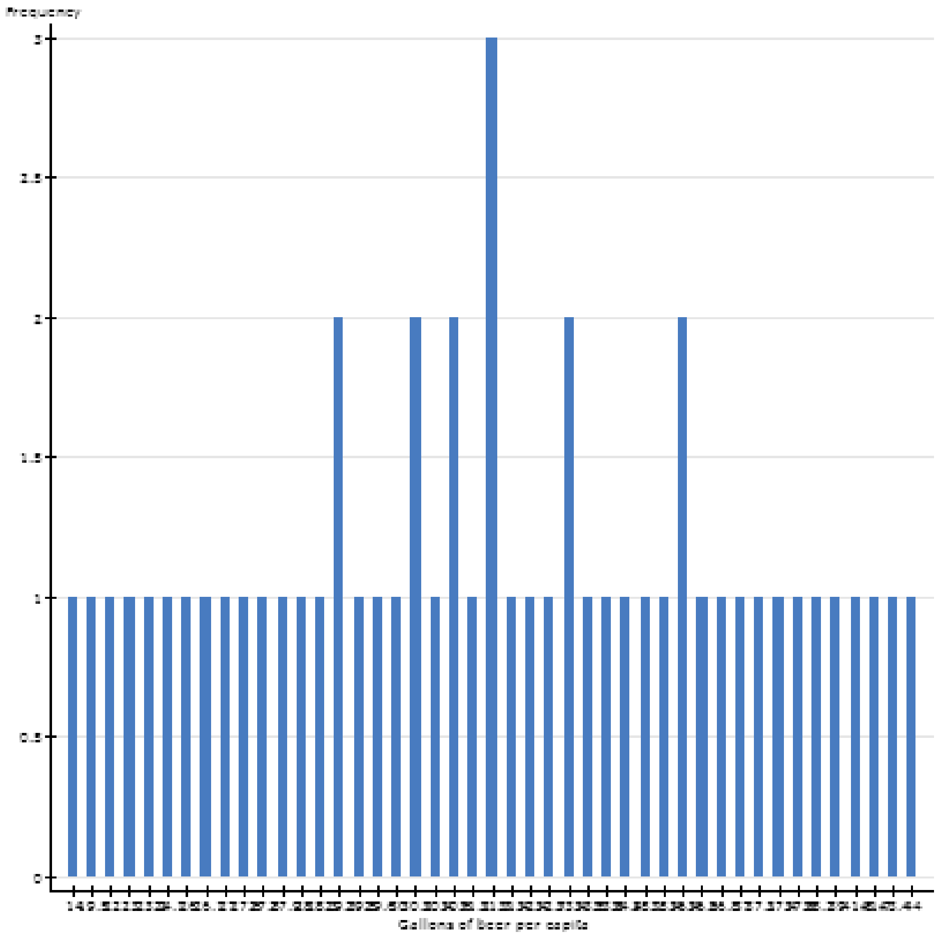
- Here we see a right skewed graph – the extreme values on the right drag the mean to the right tail causing Mean>Median



Remember: With graphs, if it’s ugly it’s probably not right.



Remember: With graphs, if it’s ugly it’s probably not right.



Misrepresentation of Data

- You should be able to look at your graphs and realize when you’ve made a mistake

  -The percentages of all relative frequency graphs should add to 1 or 100%

  -The scale should be understandable and constant

  -Consider whether or not you need to start your y axis at zero or caution against misreading the graph

  -Graphs should be simple and easy to interpret correctly in just a few moments.