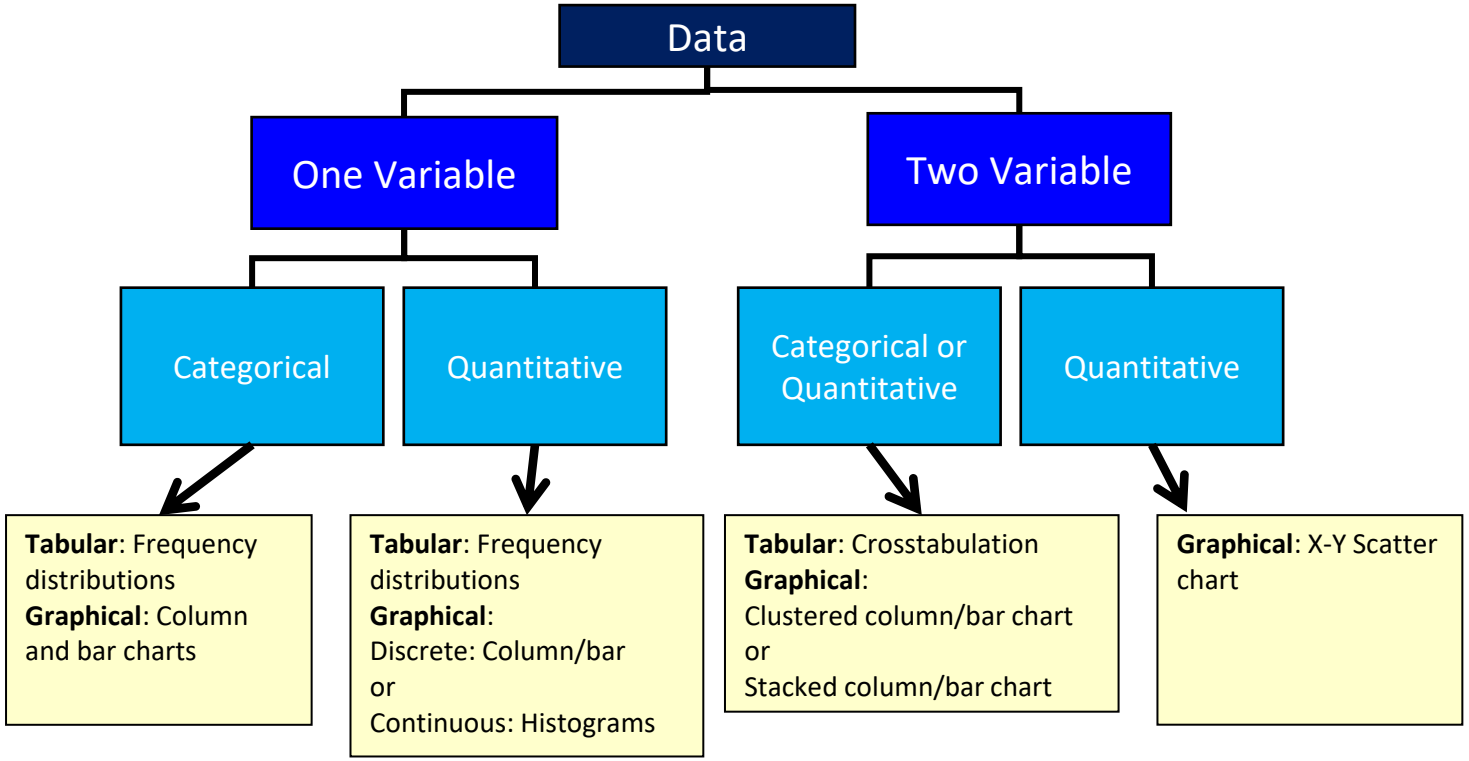
Descriptive Statistics: Tabular, Graphical and Numerical summaries of data.

Section 2: Descriptive Statistics: **Tabular** and **Graphical** summaries of data.

Section 3: Descriptive Statistics: **Numerical** summaries of data.

Why Tabular and Graphical? Because it is hard to see patterns and trends when you are looking at Raw Data! Our Goal is to create useful information from the raw data so that we can see patterns and trends. This will help in Decision Making.



Statistics

Numerical facts like:

USA unemployment rate reported Jan. 2021 was 6.3%

Sioux Radcoolinator student ranked at the 90th percentile for the test

3rd quarter YouTube advertising revenue was $7.20 billion vs. $7.4 billion expected.

Subject of Statistics defined:

Statistics is the art and science of collecting, analyzing, presenting and interpreting data.

Descriptive Statistics:

Data that is summarized and presented

Tabular: table of information

Graphical: charts, graphs, visualizations

Numerical: like an average (mean), median, mode

Inferential Statistics:

The process of using data obtained from a sample to make estimates and test hypotheses about the characteristics/attribute of a population

Take a sample from the population and draw reasonable conclusions that can help to estimate the unknown future.

Define Terms:

Population: The set of all elements of interest in a particular study

(In many situations it is too costly to get data from all the elements in the population)

Example ==>> Census: Collecting data for a population

Sample: A subset of the population

Example ==>> Sample survey: Collecting data for a sample

Categorical data = labels or names to identify categories of similar items	
Quantitative data = number data. Discrete or continuous	
Descriptive statistics to summarize and display data:	
Tabular display= data in a table = summary reports	
Graphical display = charts or other visualizations	
Tabular and graphical can be used for one or more variables	
Tabular and graphical can be used for two variables to show a relationship between the two variables	

Frequency distribution = Tabular summary that shows a unique list of nonoverlapping categories with the counts (frequency) for each category	
The goal of a frequency distribution is to show the distribution of counts across counting categories	
For **categorical** data the categories are a unique list of items from the column of data	
For **quantitative** data you usually have to create lower and upper limits for each **class** or **counting category**. The counting categories must be **collectively exhaustive** (enough categories so nothing is left out) and **mutually exclusive** (no item can fit into more than one category). Details below.	
Five types of columns in a frequency distribution table:	
Frequency = count for each category/class **Relative frequency** = (frequency of class) / (number of observations in data set). Used to build a probability distribution based on past data (ch 4). **% Frequency** = relative frequencies with percent number format applied. We will not use method of multiplying by 100.	
Cumulative Frequency = for each counting category in a grouped PivotTable, the count is made for "less than" the upper limit of class. The last class will be equal to the count of all items in the data set.	
% Frequency Distribution = cumulative relative frequencies with percent number format applied. We will not use method of multiplying by 100.	
Creating classes (counting categories) for quantitative variables in a frequency distribution and histogram:	
The goal is to reveal the natural distribution or shape or variation of the data. This is the "art side of statistics". It takes practice to get the hang of it.	
Step 1	**Determine the number of nonoverlapping classes**. Goal is to have enough to show natural shape of data. One general guideline is: 2^k > n, where n = count and k = number of classes.
Step 2	**Determine the width of each class** with something like: approx. width = (max-min)/(number of classes). Trial and error is usually required.
Step 3	**Determine the class limits**, which are the lower and upper limits used in an AND Logical test to count how many values occur between the lower and upper limit. The key is to not create classes that would double count. Once you create the class limits, you list the counting categories from smallest to biggest before you calculate the counts and make a histogram chart. Trial and error is usually required.
If you have a discrete variable (or a continuous variable that is shown as a whole number) it is just a matter of getting the lower and upper limit, like: 0-9, 10-19...	
If you have a continuous variable and you choose to use the upper limit from the previous class as the lower limit for the current class, be sure to include the equal sign on only one side, either the lower or upper, but not both. Create classes like: 0 <= Sales < 20, 20 <= Sales < 40... or 0 up to 20, 20 up to 40...	
When you create a set of classes, you are creating a type of category for your continuous quantitative variable	
Charts are more easily interpreted if the class width is the same for all classes.	
Sometimes if there are a few large values or small values, it may be efficient to create an open ended class	

Visualizing Data
Why? To get a quick impression of the data. Recognize patterns, trends and "the shape of the data"
Overriding rule: Remove all elements in visual that do not help deliver the message. "No Chart Junk".

Column and Bar Excel charts = graphical display that compares relative differences across categories/classes.
Column chart: Height of column conveys number **Bar chart**: Length of bar conveys number.
The difference between the two charts is that the bar chart, as compared to the column chart, can more forcefully emphasize differences across categories and can accommodate longer category names.
In some statistics and math textbooks, authors will refer to column charts as bar charts. However, in Excel, column charts use vertical rectangles and bar charts use horizonal rectangles.
Both charts are good for displaying frequency, relative frequency or % frequency
For **categorical data**: 1) Columns do not touch (to indicate "gap" between categories) and 2) Order of categories conveys no information.
For **discrete quantitative data**: 1) Columns do not touch (to indicate "gap" between categories) and 2) Order of categories is smallest to biggest help to show the distribution or variation or pattern in data.
For **continuous quantitative data**: 1) Columns must touch to indicate that there is no gap between counting categories. 2) Order of categories is smallest to biggest help to show the distribution or variation or pattern in data.
Specific types of column/bar charts:
Pareto chart = quality control categorical data plotted in a column chart with columns sorted by frequencies left to right from biggest to smallest
Used in quality control to show highest to lowest frequency of problems from left to right. Often a cumulative line is added to chart.

Histogram chart created from a column/bar chart = continuous quantitative data plotted in a column/bar chart using counting categories with a lower and upper limit, where counting categories are sorted left to right from smallest to biggest and there is no gap between columns to indicate that no data can occur between the successive lower and upper limits. This chart is used to visualize the frequencies from a frequency distribution for a continuous variable. Do not use built-in Excel Histogram Chart: it assumes a normal distribution and does not allow you to provide the lower limit for the first class.
Correct graphical display for revealing the distribution or variation or pattern in how frequencies occur in the data set. This chart shows the shape of the data or the skew in the data.
Histogram Notes:
Column or bar charts where columns are touching to indicate that the variable is continuous
Columns touch to indicate that no numbers can fit between classes. "No numbers can fit between columns - no gaps"
Height of columns convey count
Order of classes is important to help reveal shape of data, or distribution of data
Skew of Histograms:
What does the distribution of histogram columns look like?
Skew left or negative means a few short histogram columns are on the low end (pull mean down)
Skew right or positive means a few short histogram columns are on the high end (pull mean up)
No skew means the distribution is bell shaped or nearly bell shaped (mean = median = mode)

Summarize data for one categorical variable
Tabular: Frequency distribution **Graphical:** Column and bar Excel charts
Summarize data for one quantitative variable
Tabular: Frequency distribution **Graphical:** Histogram

Tabular and Graphical Displays For Two Variables
Crosstabulation = tabular summary for counting with two conditions / criteria / variables.
One variable is the row header and the other is the column header. On the inside of the table, the intersection of the row and column provides the count based on an AND Logical test. Variables can be categorical and/or quantitative.
Allows you to compare two variables and see relationship based on numerical counts.
Usually, for quantitative data, you must group the data into counting categories.
PivotTables are perfect for creating crosstabulations, including the ability to show percentages with by right-clicking the Values area and point to "Show Values as" and then choose "% of Grand Total", "% of Row Total", "% of Column Total"
Row totals show the frequency distribution for row variable and column totals show the frequency distribution for column variable, but these "bonus" elements in the crosstabulation do not say anything about the relationship between the two variables.
Simpson's Paradox for a cross tabulated report = a conclusion from an aggregation of two or more crosstabulations may be reversed when the data is unaggregated into two new cross tabulations that show a hidden variable. Said a different way: In cross tabulated reports, watch out for hidden variables.
Clustered column chart (Excel name) = graphical display for a crosstabulation. Emphasis is on comparing the categories listed in the legend. Also known as: "Side-by-side bar chart"
Stacked column (Excel name) = graphical display for a crosstabulation. Emphasis is on comparing the categories listed in the horizontal axis. . Also known as: "Stacked bar chart"
Scatter Chart for X-Y Data = graphical display to show the relationship between two categorical variables.
Horizontal Axis = Independent Variable = x. Vertical Axis = Dependent Variable = f(x) = y
To plot point: 1) Move along x axis, then 2) move along y axis, record point.
To use the Excel X-Y Scatter chart, the source table of data should have X values on the left of the y values, and have field names at top of each column
Use X-Y Scatter Plot Chart, not Line Chart (common mistake).
Type of relationship:
A **direct** or **positive relationship** indicates that as the x value increases, the y value tends to increase. An **indirect** or **negative relationship** indicates that as the x value increases, the y value decrease. **No relationship** indicates that as x increases, it is hard to predict where the y value will be.

Summarize data for two variables
Tabular: Crosstabulation
Graphical: Clustered column/bar or stacked column/bar Excel charts
Showing relationship between two quatitative variables
Graphical: X-Y Scatter chart.

Prevent PivotTable from Auto fitting column width:
Right-click PivotTable, click PivotTable options, on the Display & Format tab, uncheck the check box for "Autofit Column widths on update"
Set Default Settings for PivotTable:
File, Options, on right-side of Excel Options dialog box select Data, click Edit Default Layout button to change the settings to match your goals. I set mine to: 1) Report Layout = Show in Tabular Form (will show Field Names in Report rather than generic "Row Labels"), 2) PivotTable Options = uncheck the check box for "Autofit Column widths on update".

Date Variable:
Number (Quantitative)
Discrete Variable (Sometimes Grouped)
But sometimes treated as Continuous, as with a Time Series Line Chart

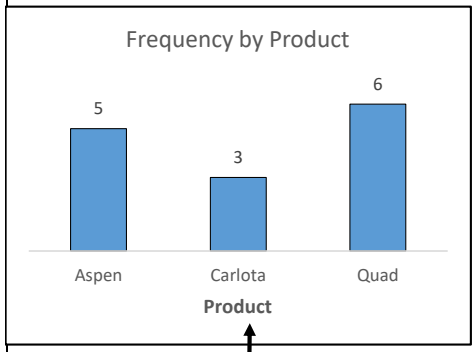
Product	Frequency
Aspen	5
Carlota	3
Quad	6
Grand Total	**14**

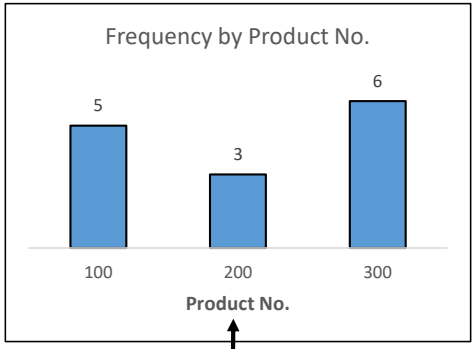
Product No.	Frequency
100	5
200	3
300	6
Grand Total	**14**

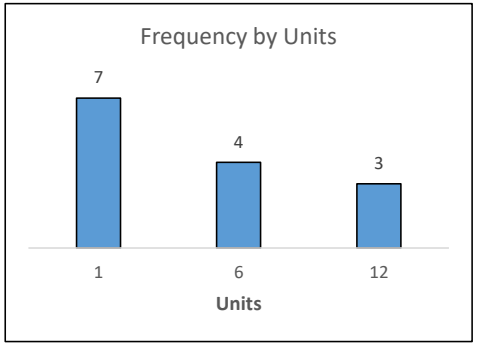
Units	Count of Units
1	7
6	4
12	3
Grand Total	**14**

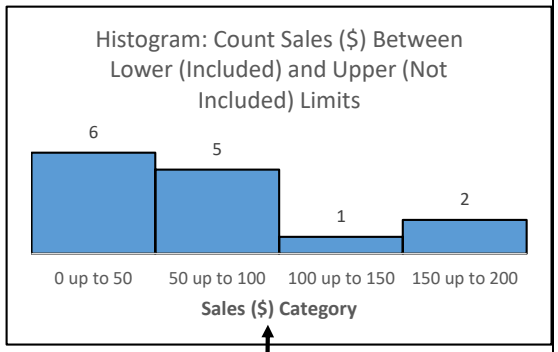
Sales ($) Category	Frequency
0 up to 50	6
50 up to 100	5
100 up to 150	1
150 up to 200	2
Grand Total	**14**

Column Charts:









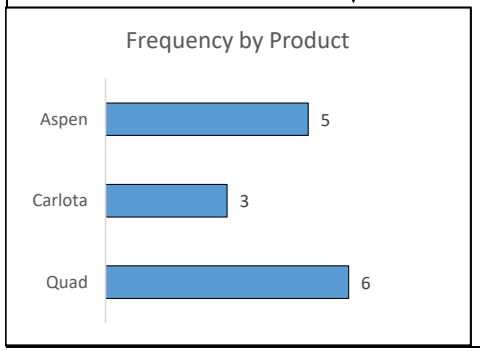
1. Categorical variable as condition for calculation = **Gaps** between columns/bars

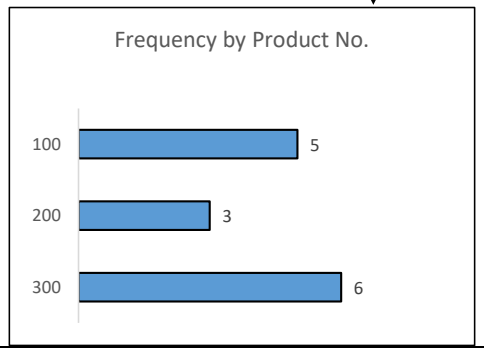
2. Categorical number variable as condition for calculation = **Gaps** between columns/bars

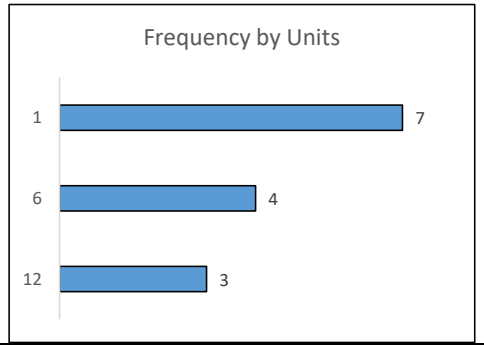
3. Discrete number variable as condition for calculation = **Gaps** between columns/bars

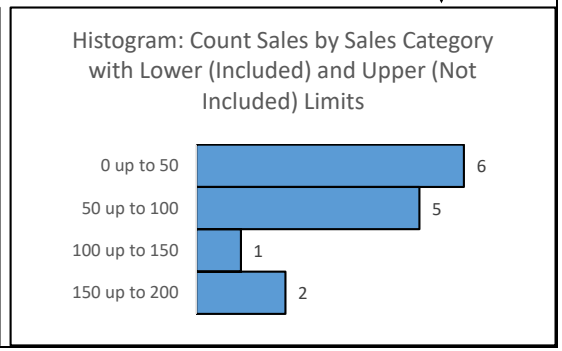
4. Continuous quantitative variable as condition for calculation = **NO Gaps** between columns/bars

Bar Charts:



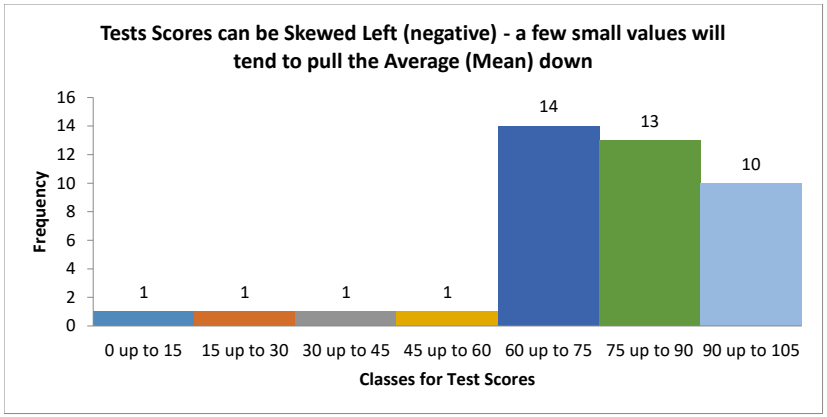


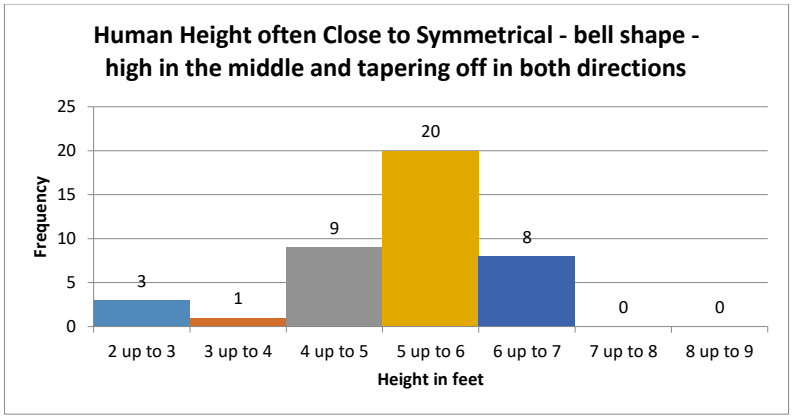


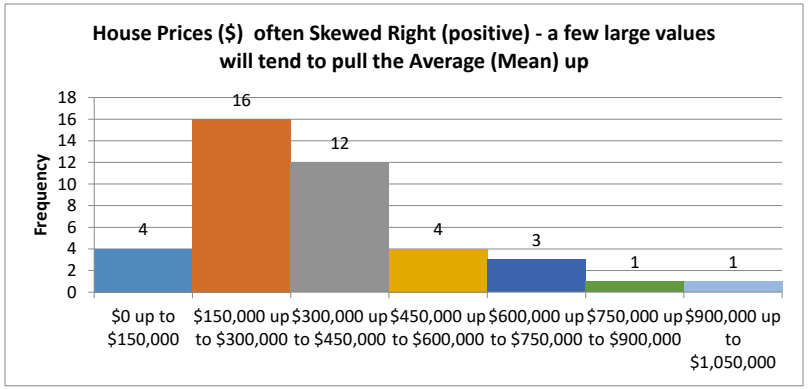


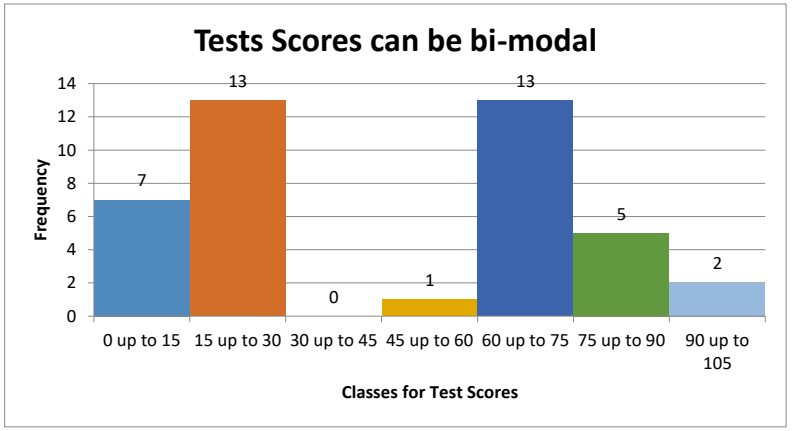
Bar Charts vs. Column Charts:

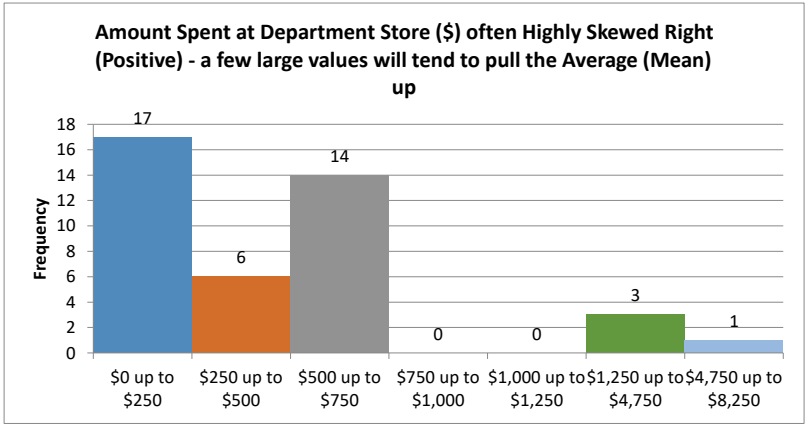
1) Bar emphasizes differences across categories more forcefully that Column because of horizontal dimensions, 2) Bar more easily accommodates longer category labels.











Skew of Histograms:

What does the distribution of histogram columns look like?

Skew left or negative means a few short histogram columns are on the low end (pull mean down)

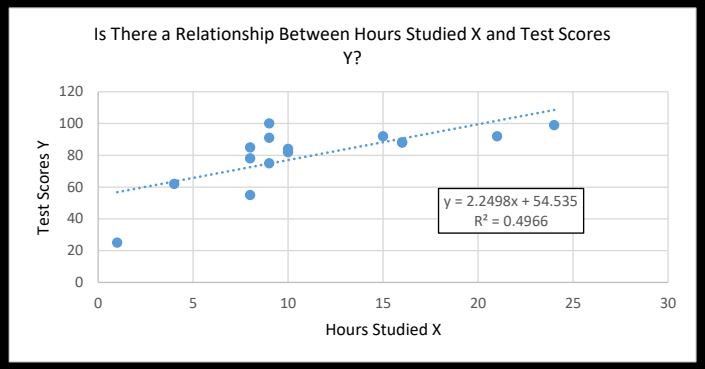
Skew right or positive means a few short histogram columns are on the high end (pull mean up)

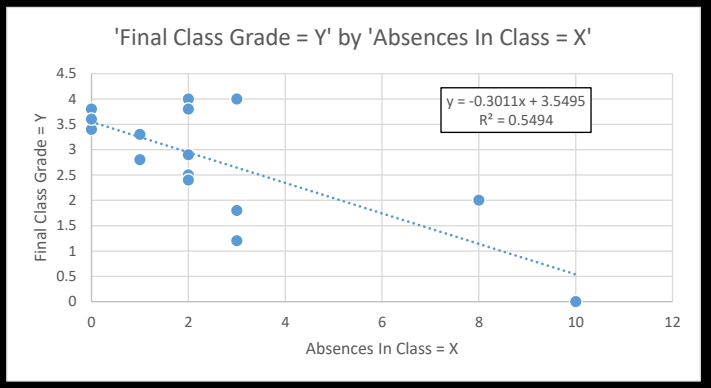
No skew means the distribution is bell shaped or nearly bell shaped (mean = median = mode)

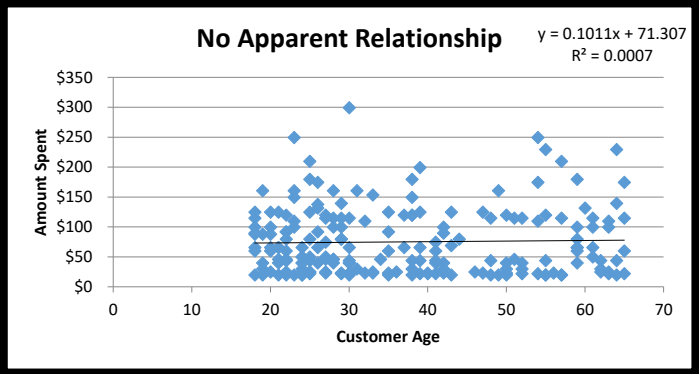
Numbers
10
11
12
11

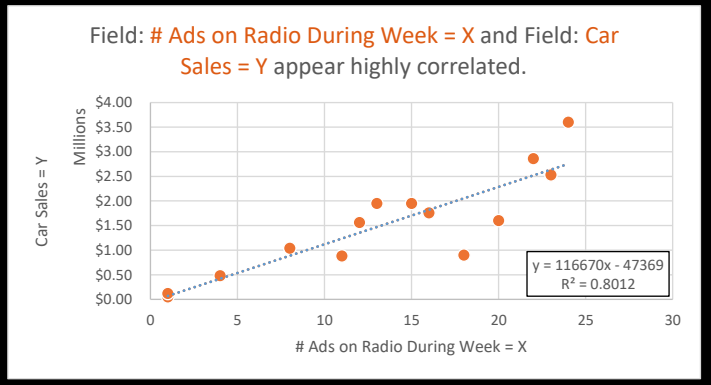
Average = Mean

11


Is There a Relationship Between Hours Studied X and Test Scores Y?
Test Scores Y
Hours Studied X
y = 2.2498x + 54.535
R² = 0.4966


'Final Class Grade = Y' by 'Absences In Class = X'
Final Class Grade = Y
Absences In Class = X
y = -0.3011x + 3.5495
R² = 0.5494


No Apparent Relationship
y = 0.1011x + 71.307
R² = 0.0007
Amount Spent
Customer Age


Field: # Ads on Radio During Week = X and Field: Car Sales = Y appear highly correlated.
Car Sales = Y
Millions
# Ads on Radio During Week = X
y = 116670x - 47369
R² = 0.8012