BUDT 730

Data, Models and Decisions

Lecture 02

Descriptive Statistics and Data Exploration

Prof. Sujin Kim

Lecture 2

- Descriptive Statistics (Ch2)
  - Types of Data: Categorical and Numerical Data
  - Summarizing and Visualizing **Categorical Data**: Count, Graphics
  - Summarizing and Visualizing **Numerical Data:** Numerical measures, Graphics

- Finding Relationships among Variables (Ch3)
- Practice:
  - Excel Demonstration – Basic + Pivot Table

Data Exploration and Descriptive Statistics

Ch2-3 from the textbook

Textbook: Business Analytics: Data Analysis and Decision Making by
S. Christian Albright and Wayne L. Winston

Catalog Marketing Data

Data file: Catalog Marketing.xlsx

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Person	Age	Gender	Own Home	Married	Close	Salary	Children	History	Catalogs	Region	State	City	First Purchase	Amount Spent
2	1	1	0	0	0	1	$16,400	1	1	12	South	Florida	Orlando	10/23/2011	$218
3	2	2	0	1	1	0	$108,100	3	3	18	Midwest	Illinois	Chicago	5/25/2009	$2,632
4	3	2	1	1	1	1	$97,300	1	NA	12	South	Florida	Orlando	8/18/2015	$3,048
5	4	3	1	1	1	1	$26,800	0	1	12	East	Ohio	Cleveland	12/26/2012	$435
6	5	1	1	0	0	1	$11,200	0	NA	6	Midwest	Illinois	Chicago	8/4/2015	$106
7	6	2	0	0	0	1	$42,800	0	2	12	West	Arizona	Phoenix	3/4/2013	$759
8	7	2	0	0	0	1	$34,700	0	NA	18	Midwest	Kansas	Kansas City	6/11/2015	$1,615
9	8	3	0	1	1	0	$80,000	0	3	6	West	California	San Francisco	8/17/2009	$1,985
10	9	2	1	1	0	1	$60,300	0	NA	24	Midwest	Illinois	Chicago	5/29/2015	$2,091
11	10	3	1	1	1	0	$62,300	0	3	24	South	Florida	Orlando	6/9/2011	$2,644
12	11	2	1	0	1	1	$94,200	1	3	18	East	New York	Buffalo	4/27/2011	$1,211
13	12	2	1	1	1	0	$73,800	0	3	24	West	Utah	Salt Lake City	8/13/2011	$3,120
14	13	2	1	1	0	1	$45,900	2	1	12	South	Louisiana	New Orleans	6/2/2011	$416

- The Catalog Marketing Excel file contains data on 1000 customers of HyTex marketing company for the current year.
- HyTex wants to extract some useful information about its customers from this data.

Data Sets, Variables, and Observations

- A **data set** is usually a rectangular array of data, with variables in columns and observations in rows.

- A **variable** (or **field**) is a characteristic of members.

- An **observation** is a list of all variable values for a single member.

Catalog Marketing Data

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Person	Age	Gender	Own Home	Married	Close	Salary	Children	History	Catalogs	Region	State	City	First Purchase	Amount Spent
2	1	1	0	0	0	1	$16,400	1	1	12	South	Florida	Orlando	10/23/2011	$218
3	2	2	0	1	1	0	$108,100	3	3	18	Midwest	Illinois	Chicago	5/25/2009	$2,632
4	3	2	1	1	1	1	$97,300	1	NA	12	South	Florida	Orlando	8/18/2015	$3,048
5	4	3	1	1	1	1	$26,800	0	1	12	East	Ohio	Cleveland	12/26/2012	$435
6	5	1	1	0	0	1	$11,200	0	NA	6	Midwest	Illinois	Chicago	8/4/2015	$106
7	6	2	0	0	0	1	$42,800	0	2	12	West	Arizona	Phoenix	3/4/2013	$759
8	7	2	0	0	0	1	$34,700	0	NA	18	Midwest	Kansas	Kansas City	6/11/2015	$1,615
9	8	3	0	1	1	0	$80,000	0	3	6	West	California	San Francisco	8/17/2009	$1,985
10	9	2	1	1	0	1	$60,300	0	NA	24	Midwest	Illinois	Chicago	5/29/2015	$2,091
11	10	3	1	1	1	0	$62,300	0	3	24	South	Florida	Orlando	6/9/2011	$2,644
12	11	2	1	0	1	1	$94,200	1	3	18	East	New York	Buffalo	4/27/2011	$1,211
13	12	2	1	1	1	0	$73,800	0	3	24	West	Utah	Salt Lake City	8/13/2011	$3,120
14	13	2	1	1	0	1	$45,900	2	1	12	South	Louisiana	New Orleans	6/2/2011	$416

Populations and Samples

- A **population** includes all of the entities of interest in a study.

  Examples
  - All potential voters in a presidential election
  - All subscribers to cable television
  - All potential Amazon customers

- A **sample** is a subset of the population, often randomly chosen, and should be representative of the population as a whole.

Strategy for Descriptive Statistics

- When we encounter a set of data, how do you discover the valuable information contained in it?
- First Step: Make sense of data by constructing appropriate summary measures, tables, and graphs

- Important things to think about
  - Types of data: Categorical vs. Numeric
  - What visualizations will help to make sense of the data?
  - What statistical summary measures are relevant?

- This procedure is called descriptive statistics

Types of Data

- Numerical vs. Categorical
  - A variable is **numerical** if meaningful arithmetic can be performed on it.
    - Ex: Salary, Children, Amount Spent
  - Otherwise, the variable is **categorical.**
    - Ex: Gender, Own Home, Married, …

- Categorical variables can be coded numerically using **dummy variable.**
  - Example: Gender can be coded as 1 for males and 0 for females.

Types of Data

- Numerical vs. Categorical
  - A variable is **numerical** if meaningful arithmetic can be performed on it.
    - Ex: Children, Unit Sold, Revenue
  - Otherwise, the variable is **categorical.**
    - Ex: Gender, Own Home, Married, …

- Numerical Variable: Discrete vs. Continuous
  - A numerical variable is **discrete** if it results from a count
    - Ex: Children
  - A **continuous** variable is the result of an essentially continuous measurement
    - Ex: Revenue

Types of Data (Cont'd)

- Categorical variables can be coded numerically.
  - Example: Gender can be coded as 1 for males and 0 for females.
- A **dummy variable** is a 0–1 coded variable for a specific category.
    - It is coded as 1 for all observations in that category and 0 for all observations not in that category.

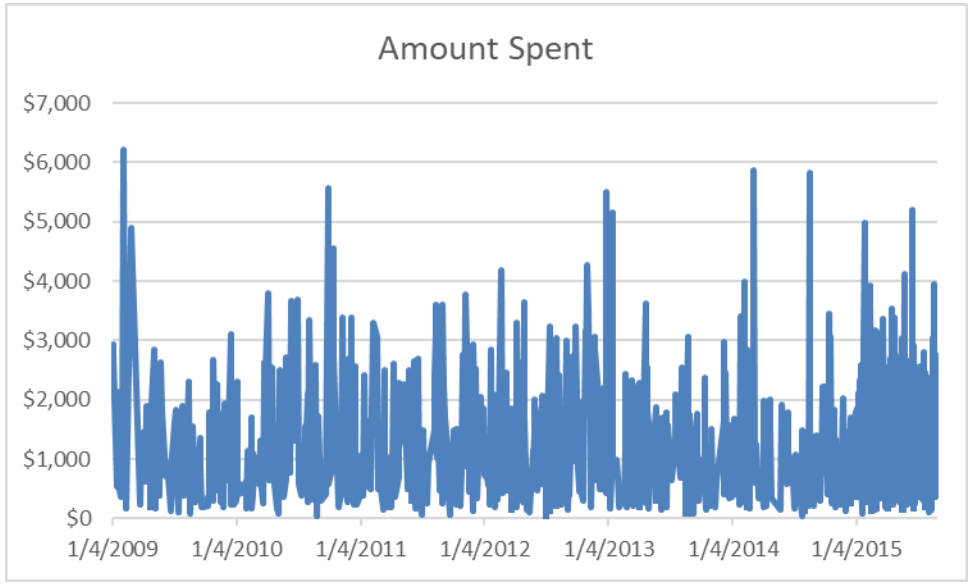
- A **binned** (or **discretized**) **variable** corresponds to a numerical variable that has been categorized into discrete categories.
  - These categories are usually called **bins**.
  - Useful for data visualization -> Histogram

Types of Data (Cont'd)

Cross-Sectional vs. Time Series

- **Cross-sectional:** data are data on a cross-section of a population at a distinct point in time.
- **Time series:** data are data collected over time – We will study time series data later in Ch12

Catalog Marketing – Date vs. Amount Spent



S&P 500

Descriptive Statistics for Categorical Variables & Pivot Table

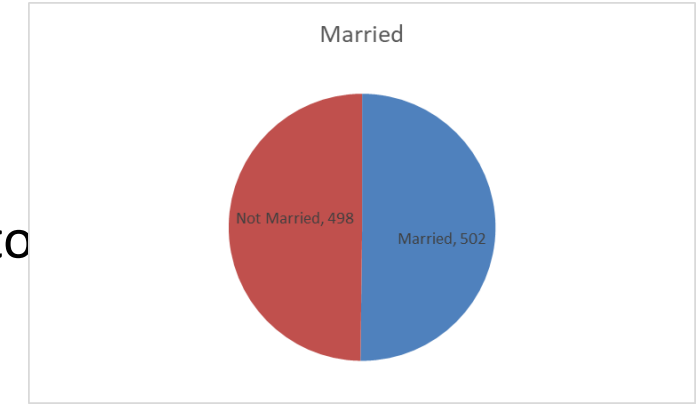
Analyzing Categorical Variables

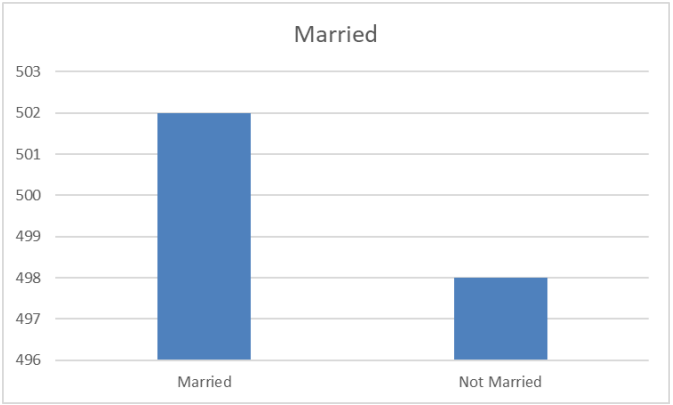
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Person	Age	Gender	Own Home	Married	Close	Salary	Children	History	Catalogs	Region	State	City	First Purchase	Amount Spent
2	1	1	0	0	0	1	$16,400	1	1	12	South	Florida	Orlando	10/23/2011	$218
3	2	2	0	1	1	0	$108,100	3	3	18	Midwest	Illinois	Chicago	5/25/2009	$2,632
4	3	2	1	1	1	1	$97,300	1	NA	12	South	Florida	Orlando	8/18/2015	$3,048
5	4	3	1	1	1	1	$26,800	0	1	12	East	Ohio	Cleveland	12/26/2012	$435
6	5	1	1	0	0	1	$11,200	0	NA	6	Midwest	Illinois	Chicago	8/4/2015	$106
7	6	2	0	0	0	1	$42,800	0	2	12	West	Arizona	Phoenix	3/4/2013	$759
8	7	2	0	0	0	1	$34,700	0	NA	18	Midwest	Kansas	Kansas City	6/11/2015	$1,615
9	8	3	0	1	1	0	$80,000	0	3	6	West	California	San Francisco	8/17/2009	$1,985
10	9	2	1	1	0	1	$60,300	0	NA	24	Midwest	Illinois	Chicago	5/29/2015	$2,091
11	10	3	1	1	1	0	$62,300	0	3	24	South	Florida	Orlando	6/9/2011	$2,644
12	11	2	1	0	1	1	$94,200	1	3	18	East	New York	Buffalo	4/27/2011	$1,211
13	12	2	1	1	1	0	$73,800	0	3	24	West	Utah	Salt Lake City	8/13/2011	$3,120
14	13	2	1	1	0	1	$45,900	2	1	12	South	Louisiana	New Orleans	6/2/2011	$416

Descriptive Statistics for Categorical Variables

- Mostly based on counts and proportions
  - **Counts**: number of observations in each category
  - **Proportions**: proportion of observations in each category, relative to total number of observations, can also convert to percentages (multiply by 100%)

  - Example: 1000 observations with a categorical variable 'Married'
    - Counts: 502 married and 498 not married
    - Proportions: 50.2% married and 49.8% not married
- Once you have the counts, you can display them graphically, usually in a **column** (or **bar**) chart or a **pie** chart.





Relationships between Categorical Variables

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Person	Age	Gender	Own Home	Married	Close	Salary	Children	History	Catalogs	Region	State	City	First Purchase	Amount Spent
2	1	1	0	0	0	1	$16,400	1	1	12	South	Florida	Orlando	10/23/2011	$218
3	2	2	0	1	1	0	$108,100	3	3	18	Midwest	Illinois	Chicago	5/25/2009	$2,632
4	3	2	1	1	1	1	$97,300	1	NA	12	South	Florida	Orlando	8/18/2015	$3,048
5	4	3	1	1	1	1	$26,800	0	1	12	East	Ohio	Cleveland	12/26/2012	$435
6	5	1	1	0	0	1	$11,200	0	NA	6	Midwest	Illinois	Chicago	8/4/2015	$106
7	6	2	0	0	0	1	$42,800	0	2	12	West	Arizona	Phoenix	3/4/2013	$759
8	7	2	0	0	0	1	$34,700	0	NA	18	Midwest	Kansas	Kansas City	6/11/2015	$1,615
9	8	3	0	1	1	0	$80,000	0	3	6	West	California	San Francisco	8/17/2009	$1,985
10	9	2	1	1	0	1	$60,300	0	NA	24	Midwest	Illinois	Chicago	5/29/2015	$2,091
11	10	3	1	1	1	0	$62,300	0	3	24	South	Florida	Orlando	6/9/2011	$2,644
12	11	2	1	0	1	1	$94,200	1	3	18	East	New York	Buffalo	4/27/2011	$1,211
13	12	2	1	1	1	0	$73,800	0	3	24	West	Utah	Salt Lake City	8/13/2011	$3,120
14	13	2	1	1	0	1	$45,900	2	1	12	South	Louisiana	New Orleans	6/2/2011	$416

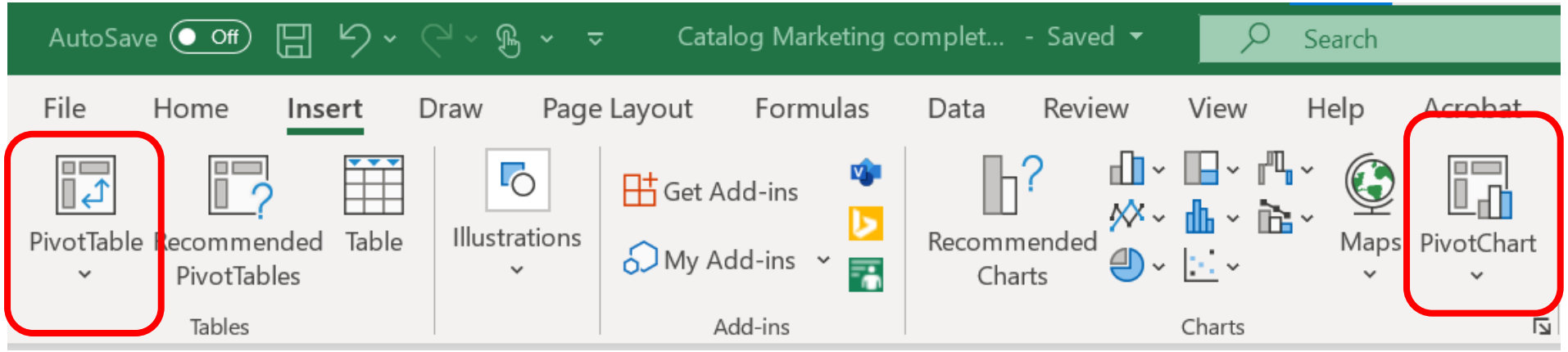
Relationships between Categorical Variables

- Like single categorical variable, the most meaningful way to compare them is with **counts** or **proportions** of the observations that fall into each joint category
- We display these counts or proportions in a ***crosstab*** (short for cross tabulation). This is also sometimes called a contingency table.
- We can easily create crosstab using Pivot Table in Excel.

- **Exercise**: Create crosstabs using Pivot Table in Excel to explore the relationship between Own Home and Married variables.

Pivot Table and Pivot Charts

Under ‘Insert’ tab



Relationship between Own Home and Married

Crosstab of Own House and Married: Frequency

	Married		
Own Home	0	1	Grand Total
0	307	177	484
1	191	325	516
Grand Total	498	502	1000

Relationship between Own Home and Married

Shown as percentage of total

$$\frac{307}{1000} * 100\%$$

	Married		
Own Home	0	1	Grand Total
0	30.7%	17.7%	48.4%
1	19.1%	32.5%	51.6%
Grand Total	49.8%	50.2%	100.0%

Relationship between Own Home and Married

	Married		
Own Home	0	1	Grand Total
0	307	177	484
1	191	325	516
Grand Total	498	502	1000

Shown as percentage of column

Count of Person	Married		
Own Home	0	1	Grand Total
0	61.6%	35.3%	48.4%
1	38.4%	64.7%	51.6%
Grand Total	100.0%	100.0%	100.0%

$\frac{307}{498} * 100\%$

Shown as percentage of row

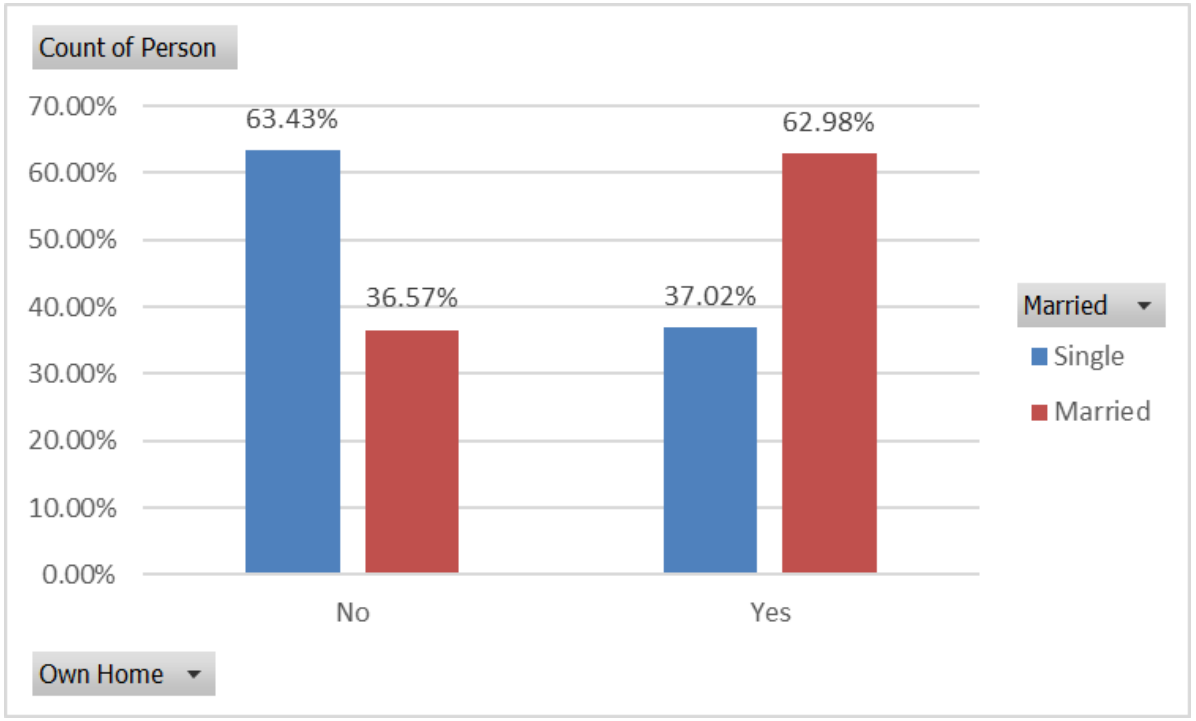
	Married		
Own Home	0	1	Grand Total
0	63.4%	36.6%	100.0%
1	37.0%	63.0%	100.0%
Grand Total	49.8%	50.2%	100.0%

$\frac{307}{484} * 100\%$

Relationship between Own Home and Married – Bar Chart

Use Pivot Chart in Excel

	Married		
Own Home	0	1	Grand Total
0	63.4%	36.6%	100.0%
1	37.0%	63.0%	100.0%
Grand Total	49.8%	50.2%	100.0%



Relationship between Own Home and Married

- What percentage of the customers in the sample is homeowner?
- What percentage of homeowners in the sample is married?
- What percentage of married customers in the sample does not own a home?
- Is there any relationship between Own Home and Married?

Relationship between Own Home and Married

- What percentage of the customers in the sample is homeowner?
  Ans: 51.6%
- What percentage of homeowners in the sample is married?
  Ans: 63.0%
- What percentage of married customers in the sample does not own a home?
  Ans: 35.3%
- Is there any relationship between Own Home and Married?
  Ans: Positive relationship – married customers are more likely to be homeowners

Exercise:
Relationship between Smoking and Drinking

Data: Smoking Drinking.xlsx

Example: Relationship between Smoking and Drinking (Smoking Dringking.xlsx)

- **Objective**: To use a crosstabs to explore the relationship between smoking and drinking.
- **Solution**: Data set lists the smoking and drinking habits of 8761 adults.
- Each variable has three categories: Heavy, Occasional, Non
- Categories have been coded “N,” “O,” “H,” “S,” and “D” for “Non,” “Occasional,” “Heavy,” “Smoker,” and “Drinker.”

	A	B	C
1	**Person**	**Smoking**	**Drinking**
2	1	NS	OD
3	2	NS	HD
4	3	OS	HD
5	4	HS	ND
6	5	NS	OD
7	6	NS	ND
8	7	NS	OD
9	8	NS	ND
10	9	OS	HD
11	10	HS	HD

Exercise: Relationship between Smoking and Drinking

- Create crosstabs that represents the relationship between smoking and drinking using Pivot Table:
  - Frequency/count
  - Percentage of total, column total, and row total

- Create appropriate bar charts to presents the results

Example: Relationship between Smoking and Drinking

Crosstab of Smoking and Drinking: Frequency

Row Labels	HS	NS	OS	Grand Total
HD	733	733	899	2365
ND	163	2118	435	2716
OD	552	2061	1067	3680
Grand Total	1448	4912	2401	8761

Example: Relationship between Smoking and Drinking

$$\frac{2118}{8761} * 100\%$$

Shown as percentage of total

Row Labels	HS	NS	OS	Grand Total
HD	8.37%	8.37%	10.26%	26.99%
ND	1.86%	24.18%	4.97%	31.00%
OD	6.30%	23.52%	12.18%	42.00%
Grand Total	16.53%	56.07%	27.41%	100.00%

Example: Relationship between Smoking and Drinking

Shown as percentage of column

$$\frac{733}{1448} * 100\%$$

Row Labels	HS	NS	OS	Grand Total
HD	50.62%	14.92%	37.44%	26.99%
ND	11.26%	43.12%	18.12%	31.00%
OD	38.12%	41.96%	44.44%	42.00%
Grand Total	100.00%	100.00%	100.00%	100.00%

Shown as percentage of row

$$\frac{1067}{3680} * 100\%$$

Row Labels	HS	NS	OS	Grand Total
HD	30.99%	30.99%	38.01%	100.00%
ND	6.00%	77.98%	16.02%	100.00%
OD	15.00%	56.01%	28.99%	100.00%
Grand Total	16.53%	56.07%	27.41%	100.00%

Relationship between Smoking and Drinking

- What percentage of heavy smokers in the sample is non-drinker?
- What percentage of heavy drinker in the sample is non-smoker?
- What percentage of the adults in the sample is heavy smoker and occasional drinker?

Answers

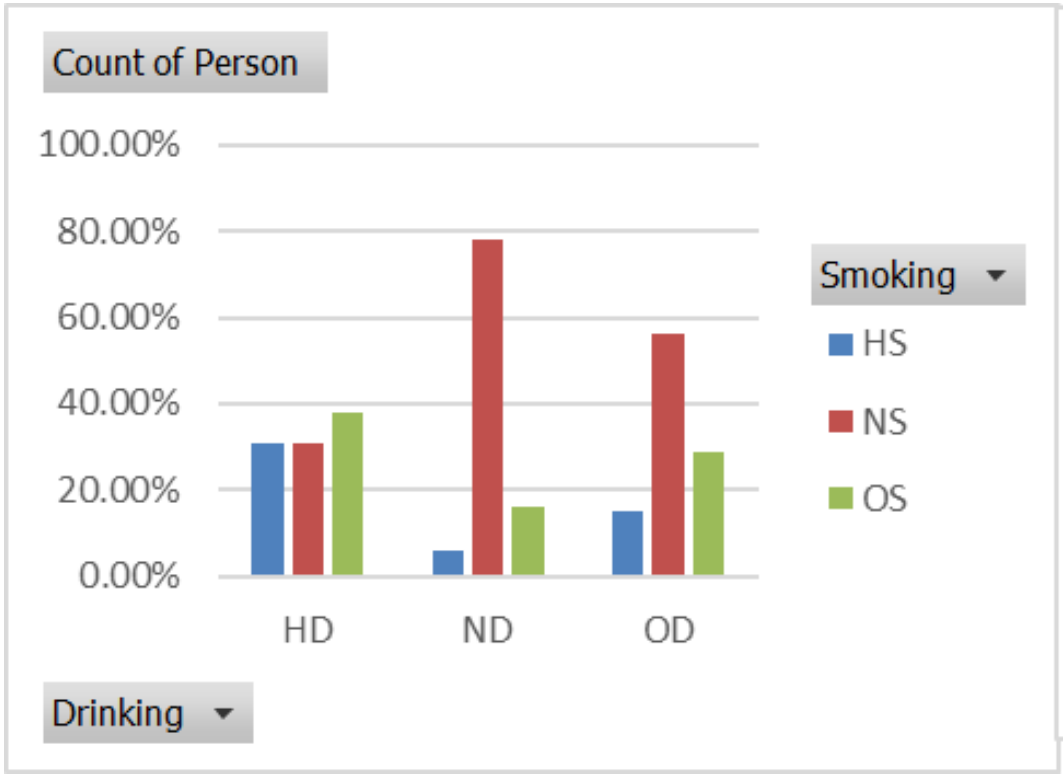
- What percentage of heavy smokers in the sample is non-drinker?
  - 11.26%
- What percentage of heavy drinker in the sample is non-smoker?
  - 30.99%
- What percentage of the adults in the sample is heavy smoker and occasional drinker?
  - 6.30%

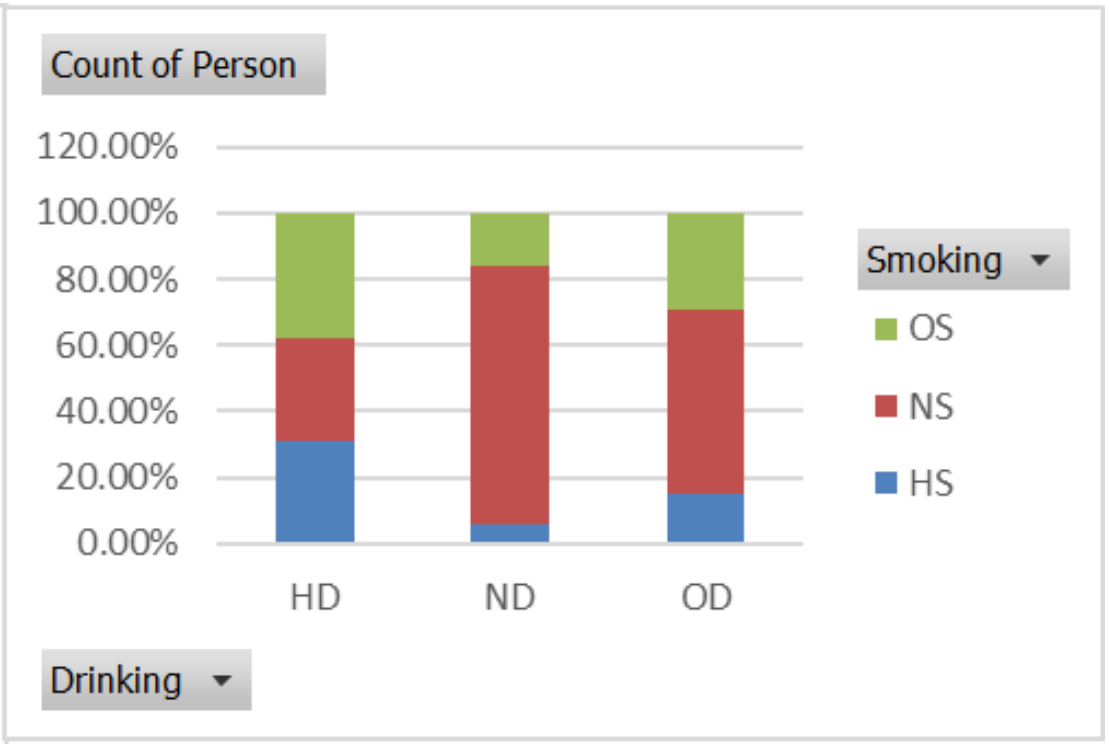
Row Labels	HS	NS	OS	Grand Total
HD	50.62%	14.92%	37.44%	26.99%
ND	11.26%	43.12%	18.12%	31.00%
OD	38.12%	41.96%	44.44%	42.00%
Grand Total	100.00%	100.00%	100.00%	100.00%

Row Labels	HS	NS	OS	Grand Total
HD	30.99%	30.99%	38.01%	100.00%
ND	6.00%	77.98%	16.02%	100.00%
OD	15.00%	56.01%	28.99%	100.00%
Grand Total	16.53%	56.07%	27.41%	100.00%

Row Labels	HS	NS	OS	Grand Total
HD	8.37%	8.37%	10.26%	26.99%
ND	1.86%	24.18%	4.97%	31.00%
OD	6.30%	23.52%	12.18%	42.00%
Grand Total	16.53%	56.07%	27.41%	100.00%

Example: Relationship between Smoking and Drinking





Is there any relationship between smoking and drinking?

Quiz 2

- This quiz is about smoking drinking data. First, create all cosstabs (contingency tables), practice all the questions on page 31, and then take the quiz.
- This quiz is timed: 30 minutes

Descriptive Statistics for Numerical Variables

Descriptive Statistics for Numerical Variables

- Many ways to summarize numerical variables
  - **Numerical descriptive measures**
  - **Visualization - graphs**
- We can ask many questions to learn how the values of a numerical variable are distributed:
  - What are typical values?
  - How spread out are the values?
  - What are the “extreme” values ?
  - Are the data symmetric or skewed in some direction?

Descriptive Statistics for Numerical Variables (Cont'd)

Numerical descriptive measures can be categorized into several groups:

- Measures of central tendency: MEAN, MEDIAN, and MODE
- Relative Standing: MIN, MAX, PERCENTILE and QUARTILE
- Measures of variability: VARIANCE, STANDARD DEVIATION and MAD
- Measures of shape: SKEWNESS

Measures of Central Tendency: **Mean**

- **Mean**: Average of all values of a variable
  - (sample) mean is

$$\bar{X}(or\ \bar{X}_n) = \frac{\sum_{i=1}^{n} X_i}{n}$$

  - n= sample size
  - Excel function is: = AVERAGE (Cell Range)
- Example - Consider the following data set (arranged in ascending order):

11, 12, 13, 14, 14, 16, 16, 16, 18, 20

The (sample) mean is

$$\bar{X} = \frac{11+12+\cdots+18+20}{10} = \frac{150}{10} = 15$$

Measures of Central Tendency: **Mode & Median**

- **Mode**: Value that appears most often
  - Excel function is: = MODE (Cell Range)
  - Example: 11, 12, 13, 14, 14, 16, 16, 16, 18, 20
    The mode is 16

- **Median**: Middle observation when data set has been arranged in ascending order
  - If the sample size n is
    - Even, the median is the average value of the two middle points
    - Odd, the median is the middle point
  - Excel function is: = MEDIAN (Cell Range)
  - Example: 11, 12, 13, 14, 14, 16, 16, 16, 18, 20
    The median is $\frac{14+16}{2} = 15$

Sensitivity of Central Tendency Measures

- Is the median or the mean of a data set more sensitive measure of central tendency?

- Suppose that the last observation is changed to a new value say 100:

  11, 12, 13, 14, 14, 16, 16, 16, 18, 100

  - Mean $= \frac{230}{10} = 23$
  - Median = 15

- Thus, the mean is more sensitive when extreme valued observations are present. In this case, the median may be a better measure for the central tendency

Relative Standing: **MIN and MAX**

- **MIN**: the smallest value of all values
  - Excel function is MIN (Cell Range)
- **MAX**: the largest value of all values
  - Excel function is MAX (Cell Range)
- **RANGE**: The difference between MIN and MAX

- Example: 11, 12, 13, 14, 14, 16, 16, 16, 18, 20
  - The MIN is 11.
  - The MAX is 20.
  - The RAGNE is 20 – 11 = 9.

Relative Standing: **Percentiles**

- For any percentage *p*, the p^{th} **percentile** is the value such that (approximately) *p*% of all values are less than it.
- There are several ways to compute percentiles.
  - Excel function is: = PERCENTILE (Cell Range, p%)

- Example: 11, 12, 13, 14, 14, 16, 16, 16, 18, 20

  25^{th} percentile = PERCENTILE(Cell Range, 25%) (or PERCENTILE(Cell Range, 0.25))

  = 13.25

Relative Standing: **Quartiles**

- **Quartiles** divide the data into four approximately equal-sized groups
  - The 1st, 2nd, and 3rd correspond to the 25th, 50th, and 75th percentiles
  - Excel function is: = QUARTILE (Cell Range, #), # = 0,1,2,3,4

- The **interquartile range (IQR)** is the difference between the 1st and 3rd quartiles

- Example: 11, 12, 13, 14, 14, 16, 16, 16, 18, 20
  - 1st quartile = QUARTILE(Cell, Range, 1)= 13.25
  - 3rd quartile = QUARTILE(Cell, Range, 3)= 16
  - IQR = 16 – 13.25 = 2.75

Measure of Variability

- Why does it matter?
  - In operations and supply chain management, variability could mean less efficient processes or poor quality
  - In finance, variability could mean volatility and risk

Measure of Variability: **Variance and Standard Deviation**

- **Variance:** approximately the average of the squared deviations from the mean
  - (sample) variance

$$S^2 (or\ S_n^2) = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1}$$

  - (sample) standard deviation = $S = \sqrt{S^2}$
- Excel Function
  - Sample Variance = VAR or VAR.S (Cell Range)
  - Sample Standard Deviation = STDEV or STDEV.S (Cell Range)
- Example: 11, 12, 13, 14, 14, 16, 16, 16, 18, 20

$$S^2 = \frac{(11-15)^2 + (12-15)^2 + \cdots (20-15)^2}{9} = \frac{68}{9} = 7.56$$

$$S = \sqrt{7.56} = 2.75$$

Population Mean & Variation

- Notation:
  - Population mean = μ
  - Population variance = σ^2
- In the case of discrete variable,

  $x_1, x_2, \ldots, x_N$: all outcomes

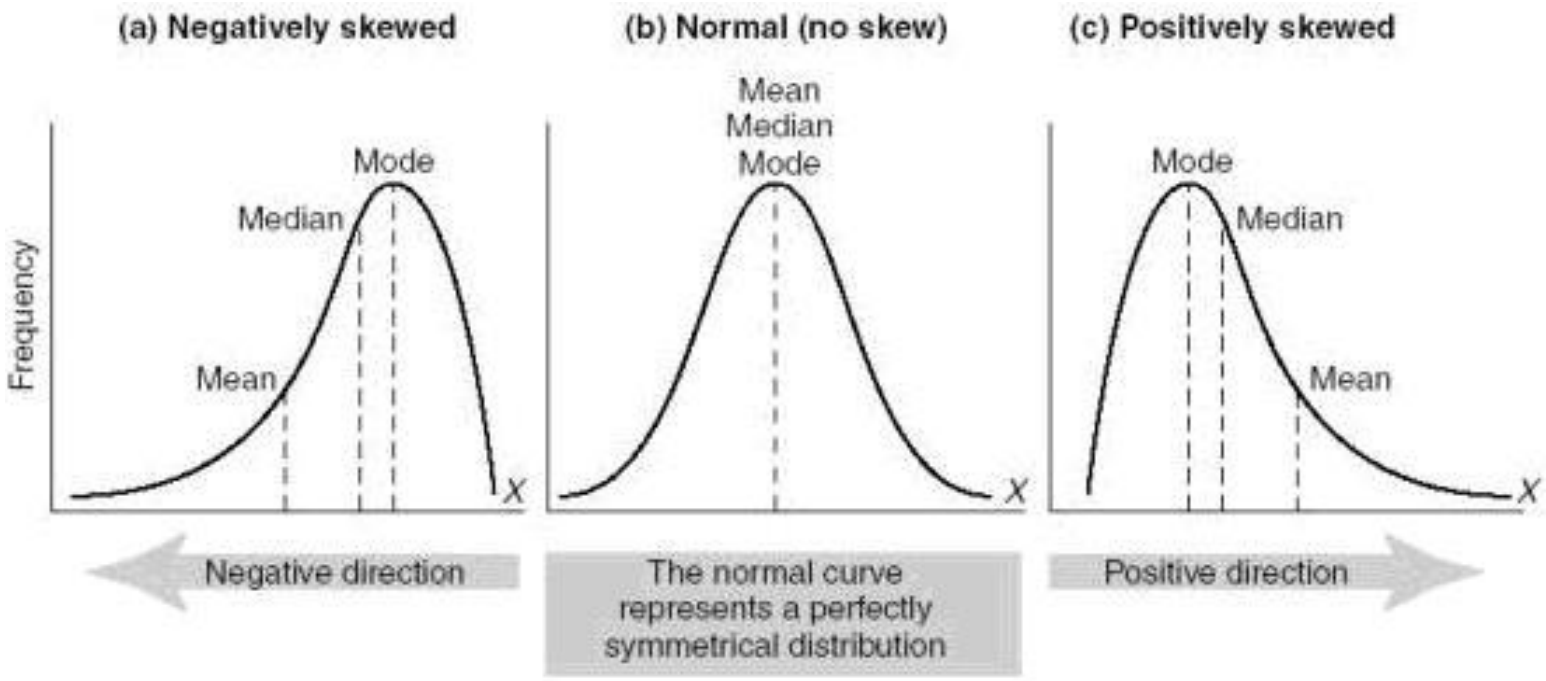
  N = population size
  - $\mu = \frac{\sum_{i=1}^{N} x_i}{N}$
  - $\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$
- Excel functions
  - Population Variance = VAR.P
  - Population Standard Deviation = STDEV.P

Measures of Shape: **Skewness**

- **Skewness** occurs because of a lack of symmetry in the distribution of values
  - Skewness > 0: A variable is skewed to the right (positively skewed) because of some really large values
  - Skewness < 0: A variable is skewed to the left (negatively skewed) because of some really small values

Central Tendency and Skewness

Skewness is easily observed via visualization



Visualizing Numerical Variables

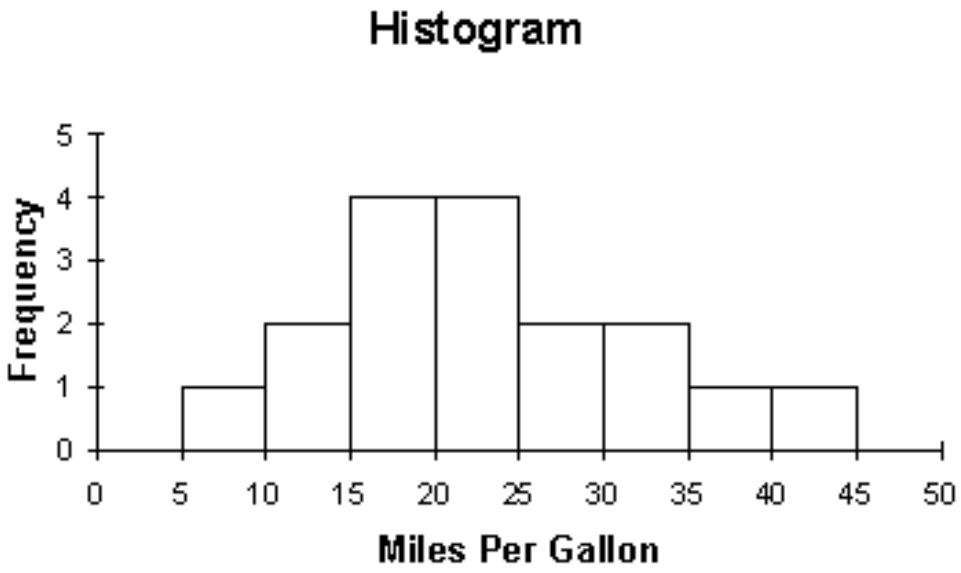
Visualizing Numerical Variables

- There are many graphical ways to indicate the distribution of a numerical variable. The most widely used graphs are:
  - Histogram
  - Box plot (also called box-whisker plot)

Visualizing Numerical Variables: **Histogram**

Histogram

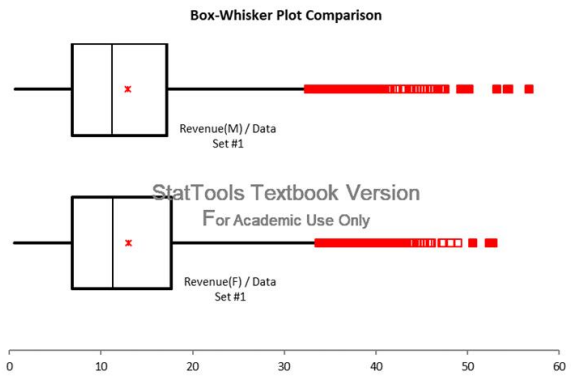
- Based on binning the variable and plotting the frequency or proportion of each bin
- Most common type
- Great for showing the shape of the distribution of a single variable

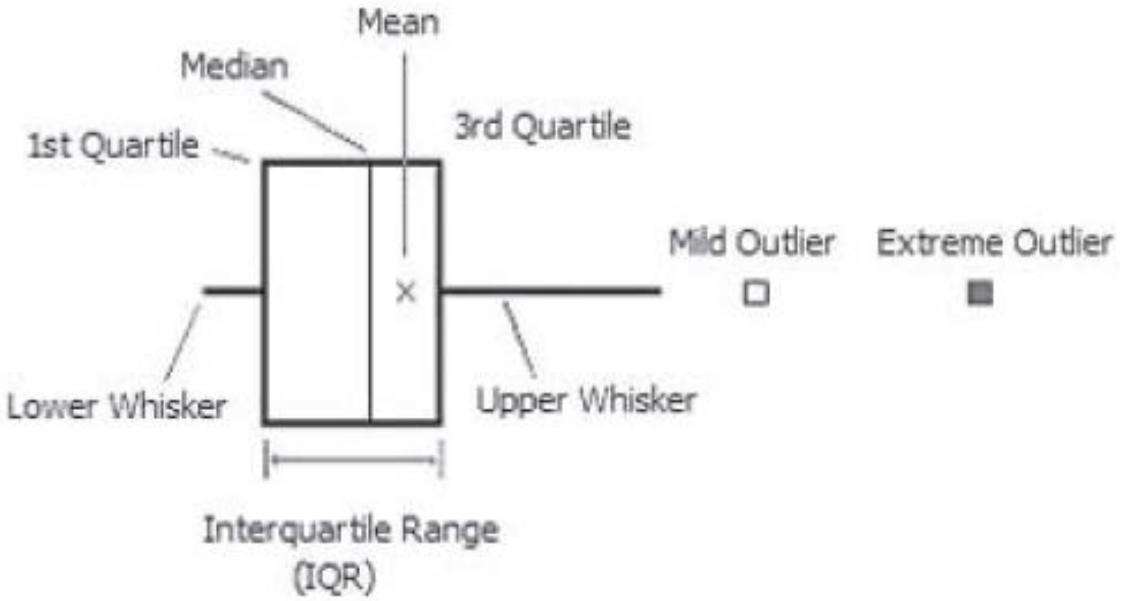


Visualizing Numerical Variables: **Box Plot**

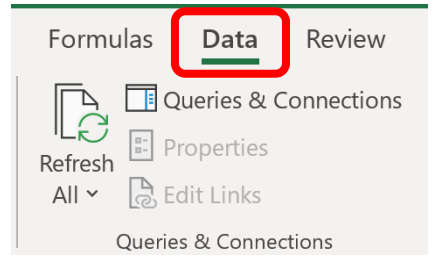
Box Plot

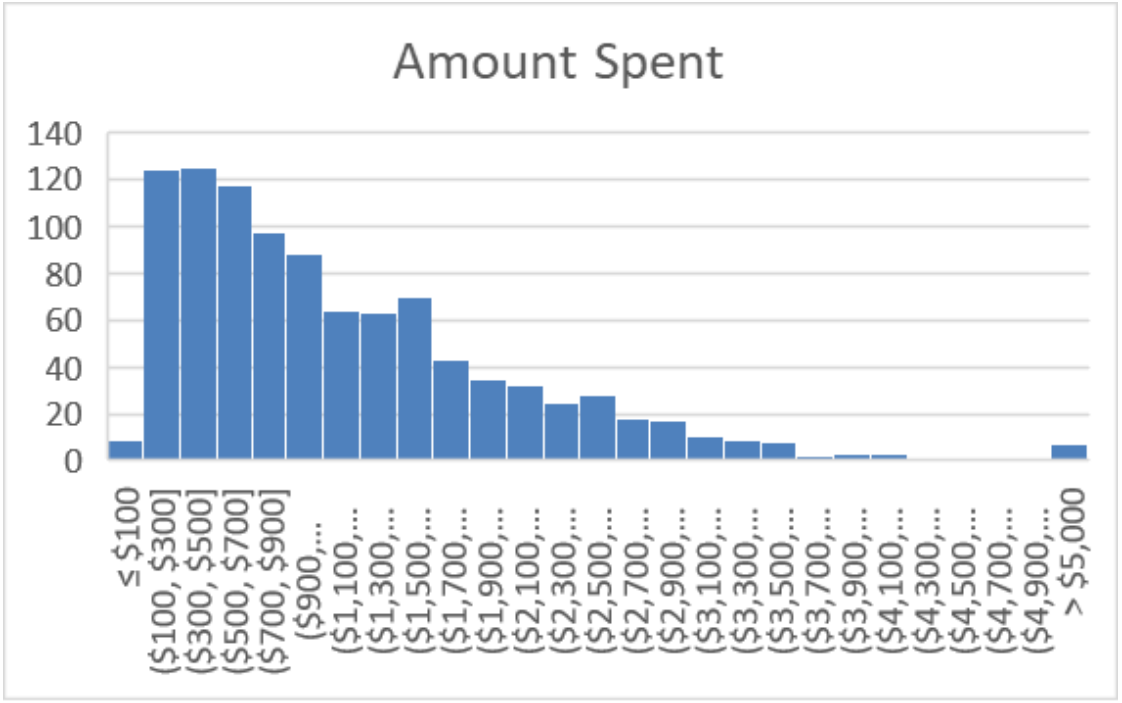
- Percentile-based plot for visualizing the distribution of a variable
- More information dense
- Side-by-side box plots are useful for comparing distributions

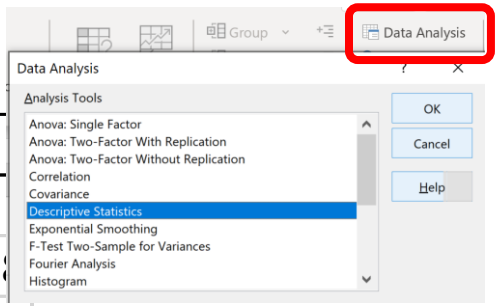




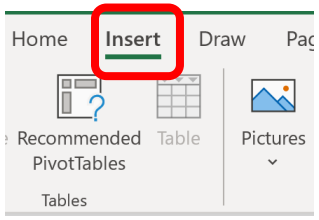
HyTex Catalog Marketing Data

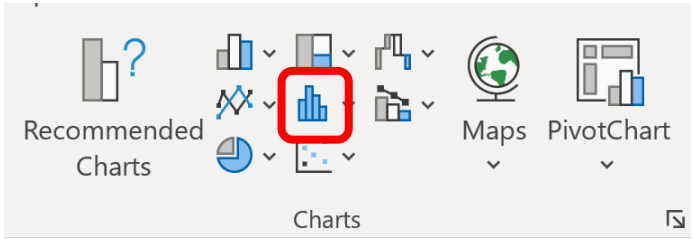


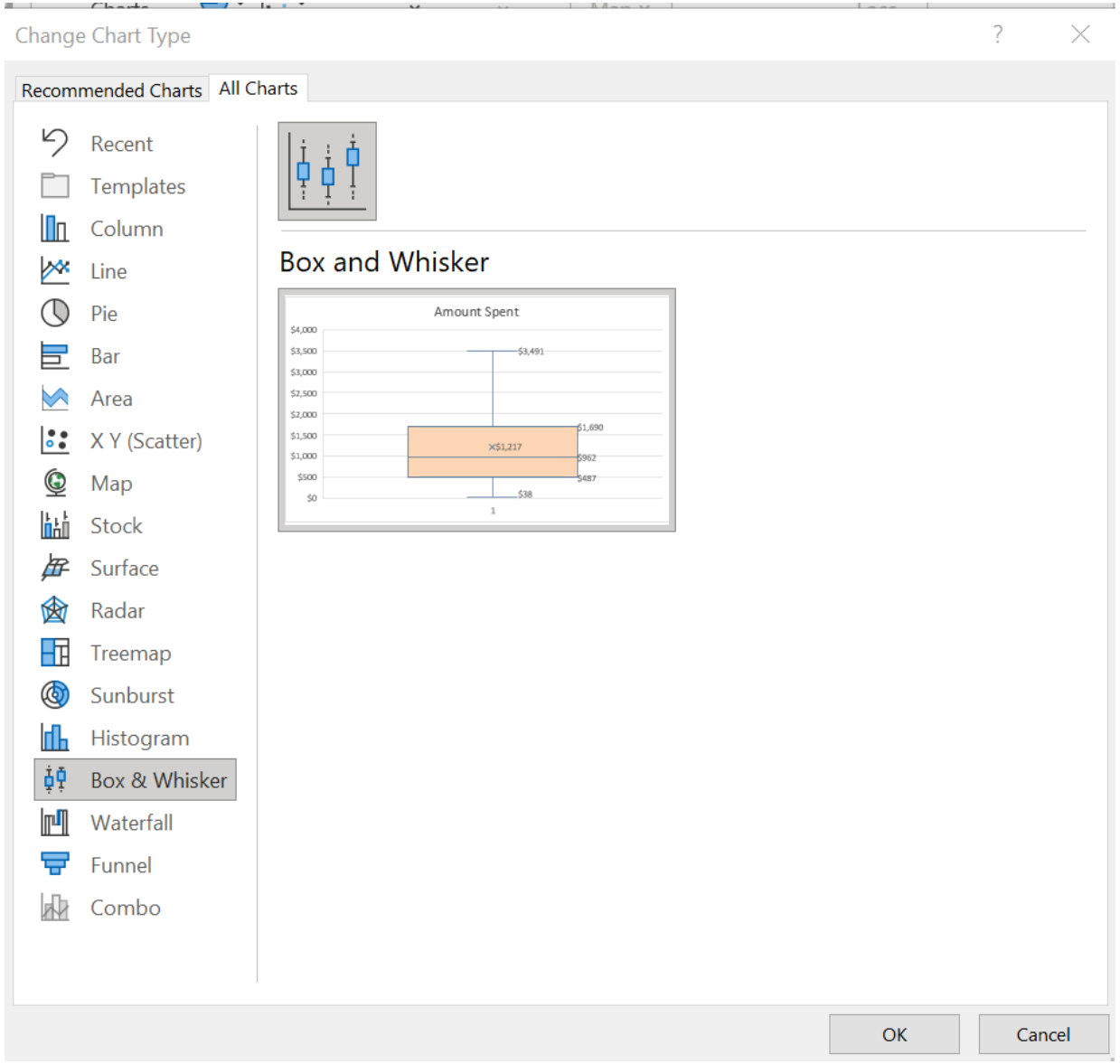


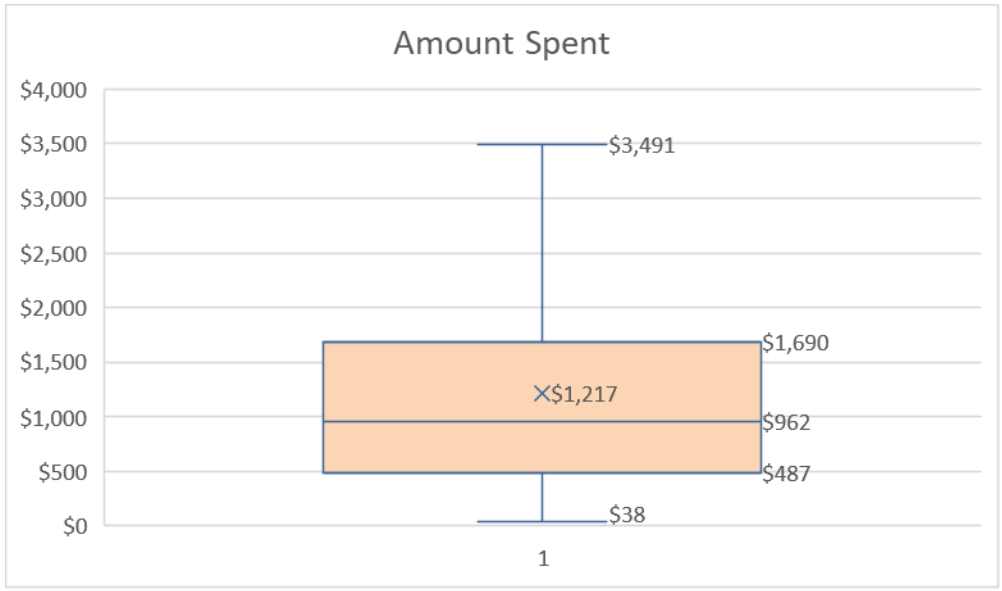


Amount Spent	
Mean	1216.76
Standard Error	30.39187
Median	961.8105
Mode	#N/A
Standard Deviati	961.0754
Sample Variance	923665.9
Kurtosis	2.974078
Skewness	1.469267
Range	6179.536
Minimum	37.807
Maximum	6217.343
Sum	1216768
Count	1000

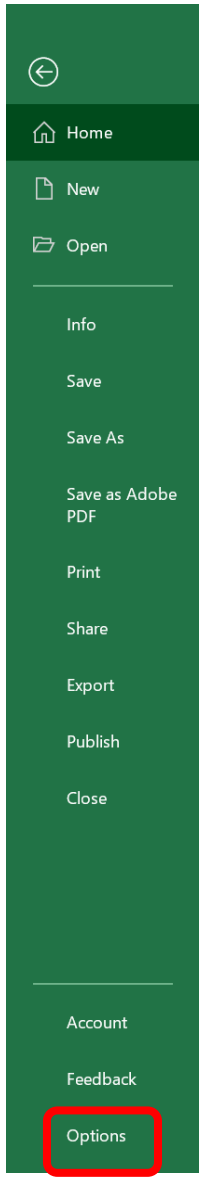


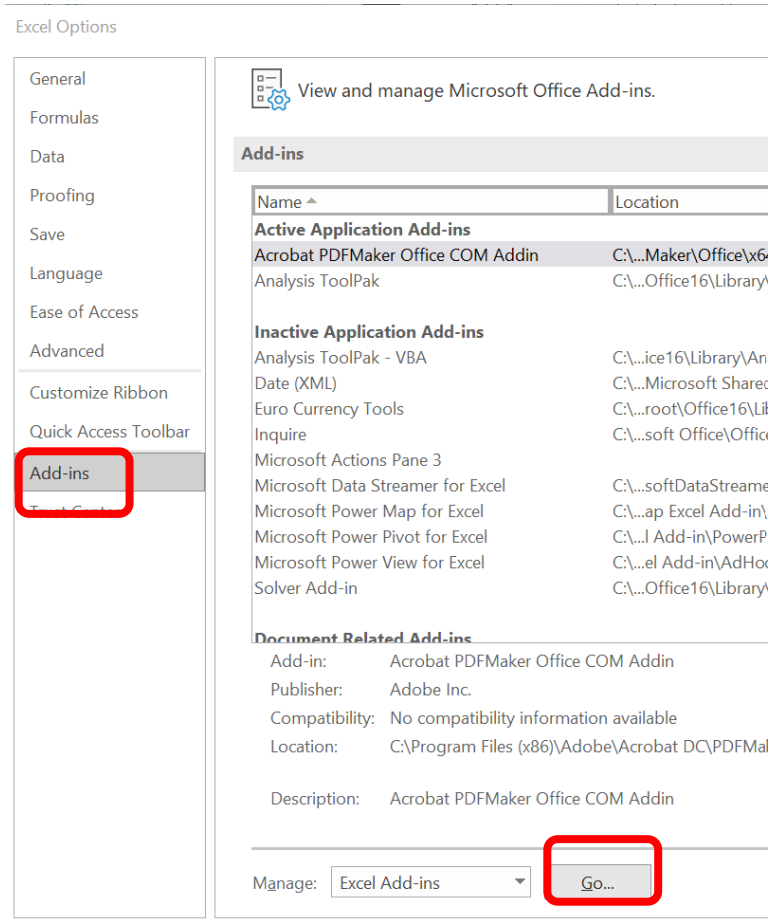


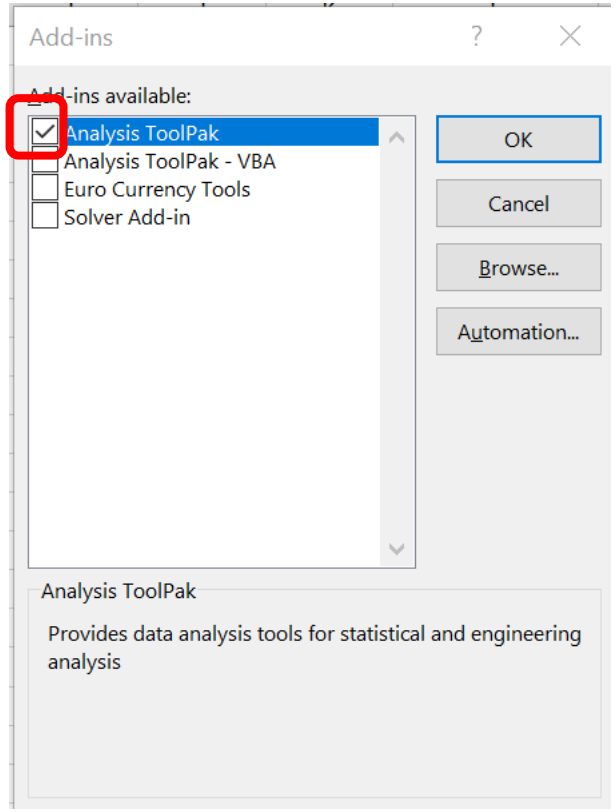

Change Chart Type
Recommended Charts
All Charts
Recent
Templates
Column
Line
Pie
Bar
Area
X Y (Scatter)
Map
Stock
Surface
Radar
Treemap
Sunburst
Histogram
Box & Whisker
Waterfall
Funnel
Combo
Box and Whisker
Amount Spent
$4,000
$3,500
$3,000
$2,500
$2,000
$1,500
$1,000
$500
$0
$3,491
$1,690
$1,217
$962
$487
$38
1
OK
Cancel


Amount Spent
$4,000
$3,500
$3,000
$2,500
$2,000
$1,500
$1,000
$500
$0
$3,491
$1,690
$1,217
$962
$487
$38
1

- How to add Excel Analysis ToolPak?
  - Go to file
  - Click Options on the left menu
  - Click Add ins
  - Go to Excel Add-ins
  - Check Analysis ToolPak







Outliers

- An outlier is an observation that lies well outside of the norm, with respect to one variable or a combination of variables
- General rule of thumb
  - An outlier is any value more than **three standard deviations** from the mean
- Best practice
  - Run analysis two ways: With outliers and without
- Applications – Outlier/anomaly detection
  - Fraud detection, diagnostic medicine, (structural) fault detection, superstar athletes

Missing Values

- As with outliers, we need to know how to detect missing values and what to do about them
- Missing values are coded in many ways (e.g., NA, blank)
  - In Excel, do a Find/Replace to standardize missing values
- More importantly, what to do with missing values?
  1. Ignore them, but you need to know how the software deals with them
  2. Fill in missing values with central measure of existing values
  3. Examine the existing values in the row of a missing value; they may provide information on what a missing value should be

Relationships between Numerical Variables

Relationships Among Numerical Variables

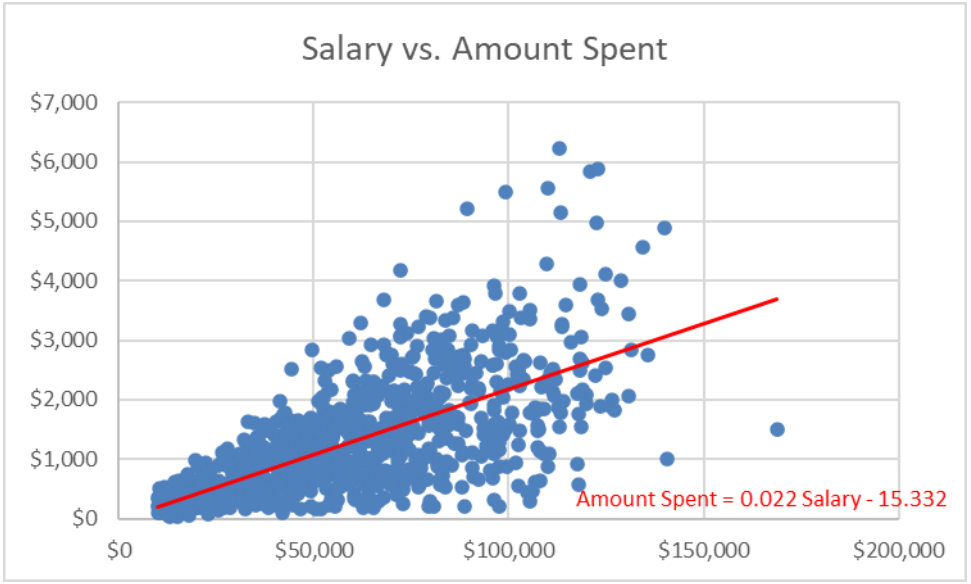
- Visualization technique: **Scatter plots**
- Summary measures
  - Covariance
  - Correlation

Scatterplot

- A **scatterplot** is a scatter of points, where each point denotes the values of an observation for two selected variables.
  - It is a graphical method for detecting relationships between two numerical variables.
  - The purpose of a scatterplot is to make a relationship (or the lack of it) apparent.

Scatter Plot for HyTex Catalog Marketing Data

- We can use scatterplots to visualize the relationship between " Amount Spent" and "Salary"
- What can you say about the relationship between a customer's salary and the amount he/she spends?
- Quantifying the strength of relationship from a scatterplot is hard!
- We can use association measures to quantify the relationship between two variables



Measuring Association: Covariance

- **Covariance** measures the strength and direction of a *linear relationship* between two numerical variables

$$cov(X, Y) = E[(X - E(X)) \cdot (Y - E(Y))]$$

- It is essentially an average of products of deviations from means.
- Excel function: COVAR (array1, array2)

Measuring Association: Correlation

- A standardized measure of association is the **correlation coefficient**:

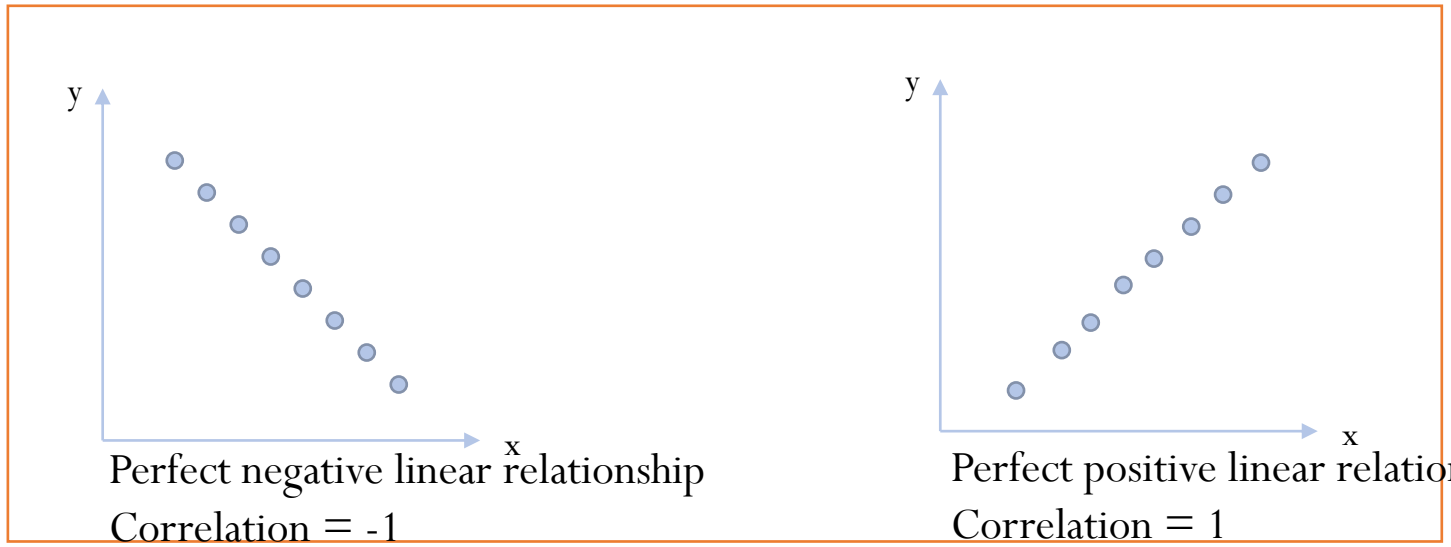
$$corr(X,Y) = \frac{cov(X,Y)}{Stdev(X) \cdot Stdev(Y)}$$

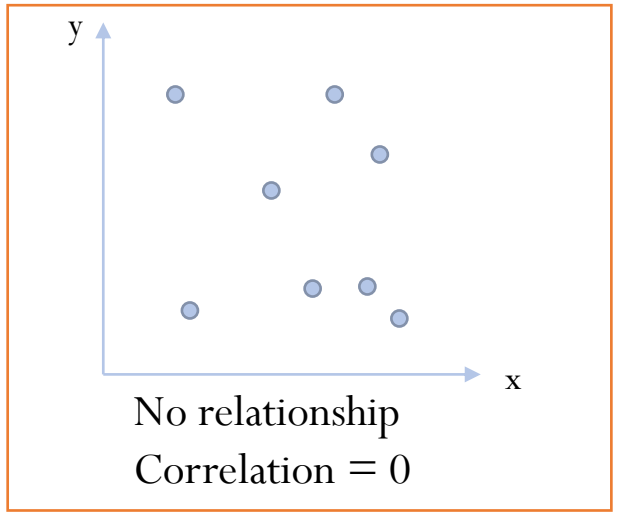
- Excel function: CORREL(array1, array2)
- Correlation is a *unitless* quantity that is unaffected by the measurement scale
- It is easier to interpret because it's *normalized* to values between -1 and 1

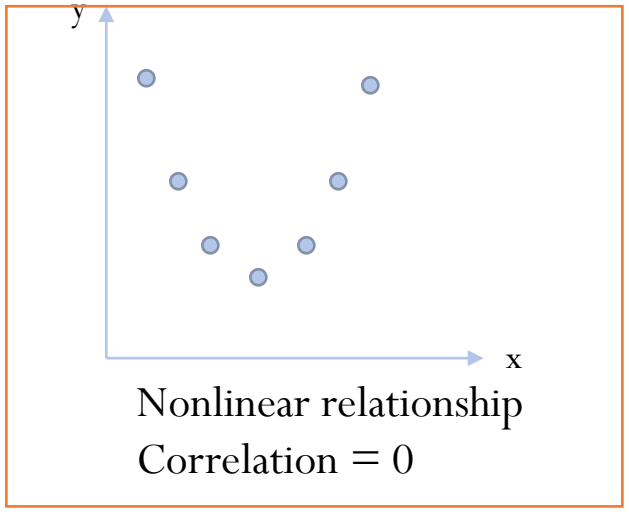
Correlation

- A measure of the linear relationship between variables
  - The range is between -1 and 1.
    - +1 = perfect positive linear relationship
    - 0 = no linear relationship
    - -1 = perfect negative linear relationship

Examples


y
x
Perfect negative linear relationship
Correlation = -1
y
x
Perfect positive linear relationship
Correlation = 1

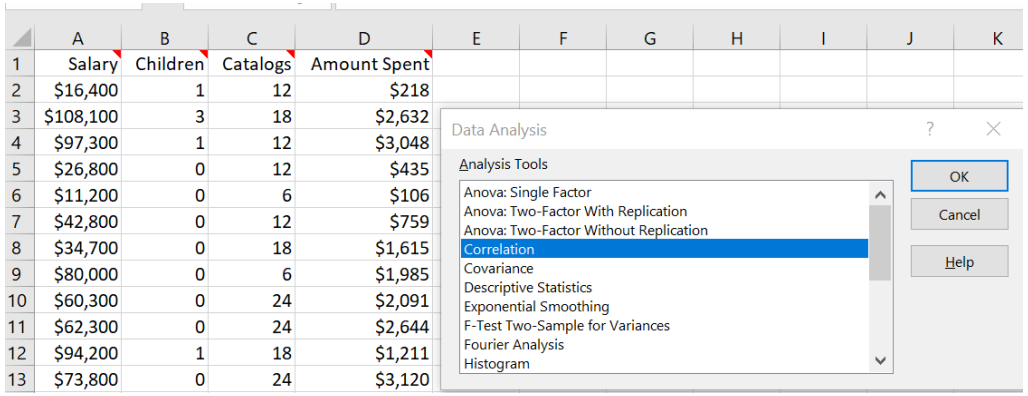

y
x
No relationship
Correlation = 0


y
x
Nonlinear relationship
Correlation = 0

Example: HyTex Catalog Marketing Data

- Which variable, *Salary , Children,* or *Catalogs*, has a stronger relationship with *Amount spent*?

	Salary	*Children*	*Catalogs*	*Amount Spent*
Salary	1			
Children	0.049663	1		
Catalogs	0.183551	-0.11346	1	
Amount S|	0.699598	-0.2223	0.472644	1



	A	B	C	D
1	Salary	Children	Catalogs	Amount Spent
2	$16,400	1	12	$218
3	$108,100	3	18	$2,632
4	$97,300	1	12	$3,048
5	$26,800	0	12	$435
6	$11,200	0	6	$106
7	$42,800	0	12	$759
8	$34,700	0	18	$1,615
9	$80,000	0	6	$1,985
10	$60,300	0	24	$2,091
11	$62,300	0	24	$2,644
12	$94,200	1	18	$1,211
13	$73,800	0	24	$3,120

- We can see from the correlation table that *Salary* has a stronger relationship with *AmountSpent.*

Next …

- Data Visualization
  - Data Visualization with Tableau
  - Complete the tutorial videos posted on Canvas