



Lecture 23

Produced by Dr. Worldwide

Welcome to the 305

Descriptive Statistics



- A **statistic** is a quantity computed from a sample AKA a **function** of the data
- **Descriptive statistics** refers to the analysis of data that helps to describe, visualize, or summarize features in a sample
- Purposes for descriptive statistics
 - Understand the data without seeing the entire dataset
 - Quantify the center of the data
 - Quantify the spread of the data
 - Capture patterns in the data
- A **frequency table** is an array that gives that lists all possible values with counts of how often they occur in the sample (shows the sample distribution)
- For a **numeric** variable it is helpful to organize possible values in bins

Ex: Starting Salaries



- Suppose we have starting salaries of 50 recently graduated students with B.S. degrees in data analytics, recorded the in thousands of dollars per year

- Download [Salaries.xlsx](#) from link [Sheet 1](#) on course website

- Focus on sheet named [Frequency](#)

- Raw data contained in cells A4:A53

- Frequency table contained in cells D4:E17

- To create, use the following Excel function


`=FREQUENCY(data array, bins array)`

- For more help, see [Link 1](#) on course website

Salaries		Bins	Frequency
3		80	2
4	84.8181	82	3
5	90.4642	84	6
6	82.7153	86	6
7	83.319	88	2
8	89.9589	90	7
9	99.1958	92	8
10	85.382	94	3
11	92.2283	96	5
12	88.6465	98	2
13	96.7041	100	4
14	83.4656	102	0
		104	0
			2

Ex: Starting Salaries



- Bins need to be specified beforehand
- The first bin counts the number of values smaller than the bin value and all other bins count the number of values larger than the previous bin but smaller than the current one
- To use FREQUENCY, you need to select the array where the results will be displayed and press
 - CTRL+SHIFT+ENTER (Windows)
 - CTRL+U or +RETURN (Mac)
- Current frequency table breaks up the range of values into bins of size 2
- Go to sheet named **Practice** to create frequency table with bins of size 3

-



- 

Ex: Starting Salaries



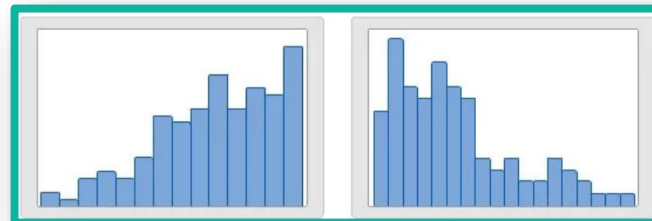
- Frequencies need to be added to the table
 - Select D4 (or D4:D13) and input the formula
`=FREQUENCY(A4:A53, C4:C12)`
 - You may need to use shortcuts to create vector of frequencies
- Q: Why is the frequency table better than the original data?

Bins	Frequency
80	2
83	4
86	11
89	7
92	10
95	6
98	4
101	4
104	0
	2

Descriptive Statistics



- A **histogram** is a graph based off a frequency table for a numeric variable
- The histogram is the most common graph used to understand the variability in numeric data
- The histogram is built off **rectangles** whose heights represent the number of observations within each bin
- Histogram gives us an **estimate** of the distribution in the population
- Three things we learn from a histogram
 - Shape of the distribution (left skewed or right skewed)
 - Range of values
 - Outlier identification



Ex: Starting Salaries



- Various ways to build a histogram in Excel
- Created from frequency table
 - Copy and paste frequency **values** (not formulas)

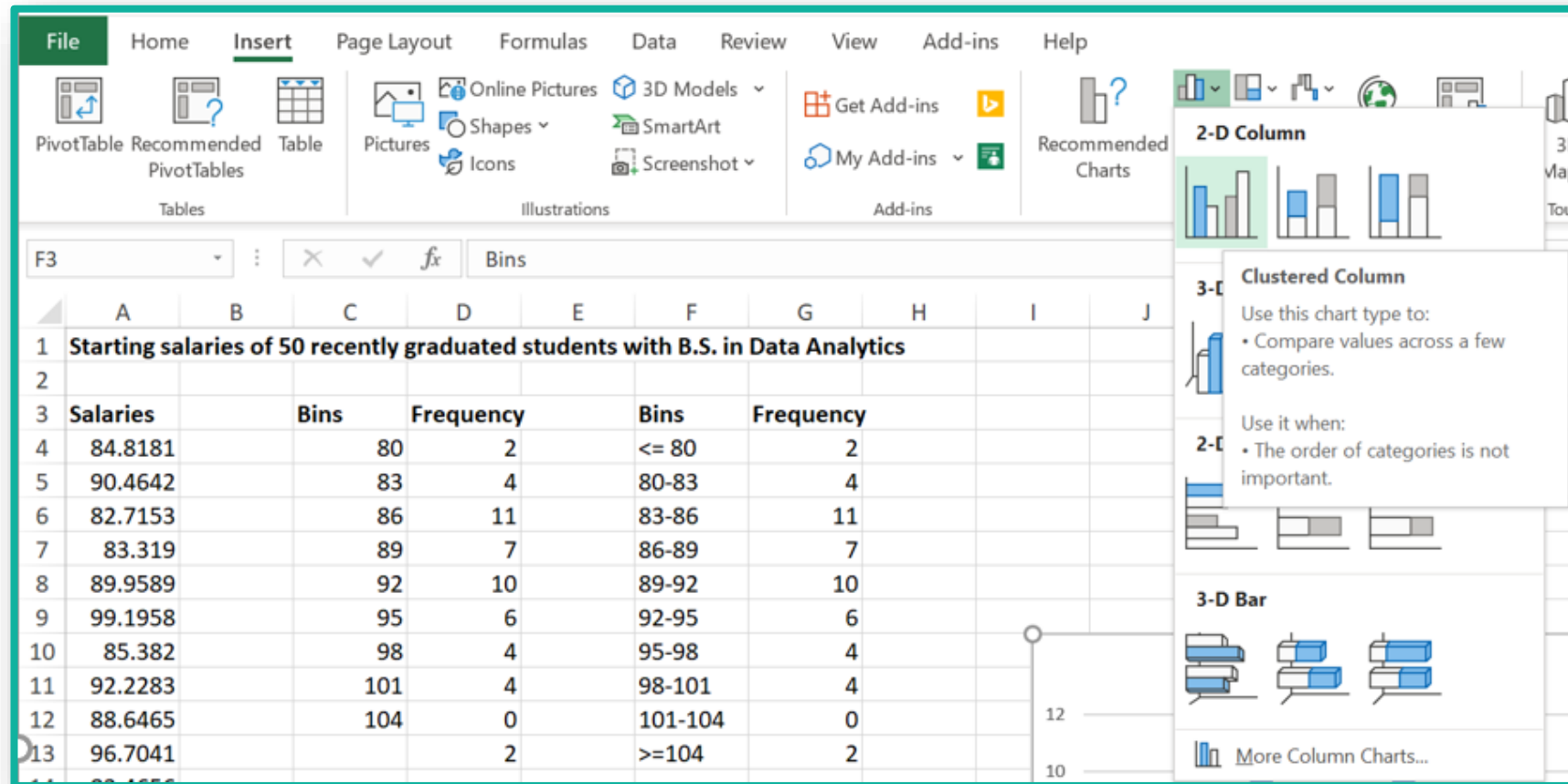
Bins	Frequency	Bins	Frequency
80	2	<= 80	2
83	4	80-83	4
86	11	83-86	11
89	7	86-89	7
92	10	89-92	10
95	6	92-95	6
98	4	95-98	4
101	4	98-101	4
104	0	101-104	0
	2	>=104	2

- Highlight the two columns of bin names and frequencies

Ex: Starting Salaries



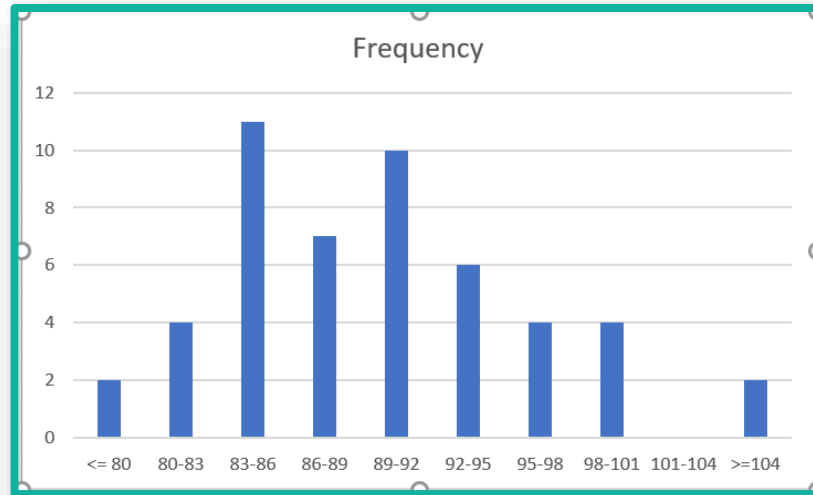
- Created from frequency table
 - Select appropriate chart (**clustered column**) through **Insert** menu



Ex: Starting Salaries



- Created from frequency table
 - Observe the default plot which needs to be fixed

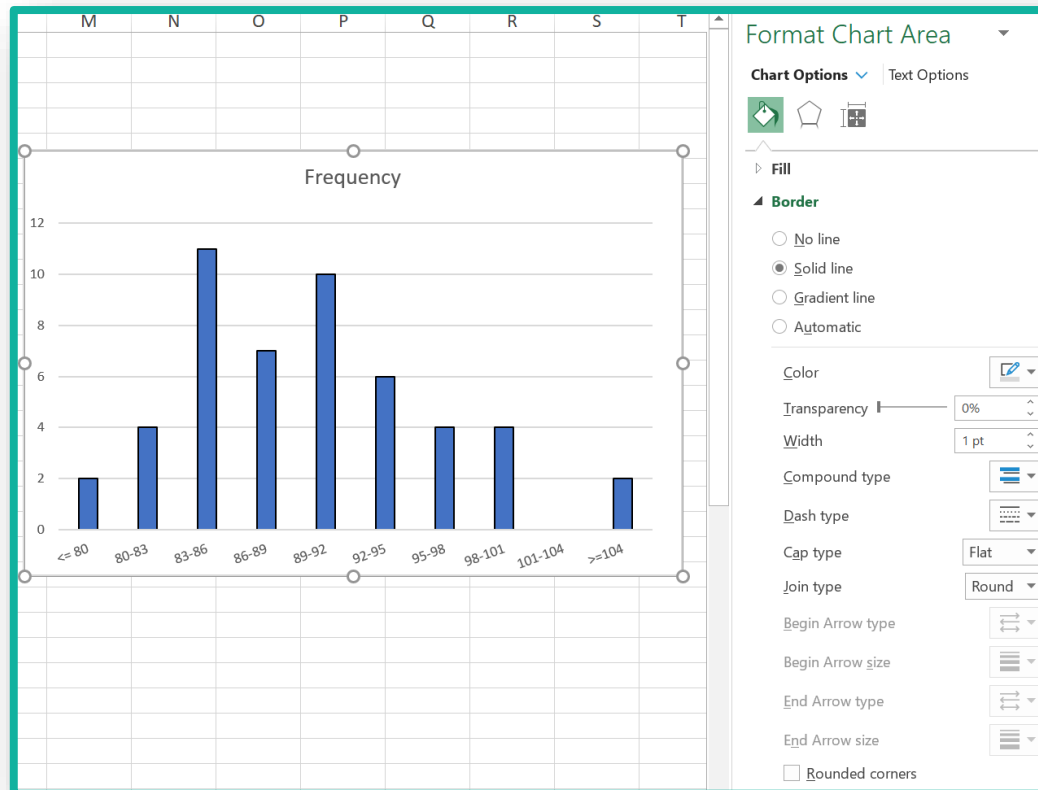


- Double click on plot to get sidebar menu for easy editing of elements
- Select the x-axis labels for modification and try to figure out how to rotate the text to -21 degrees

Ex: Starting Salaries



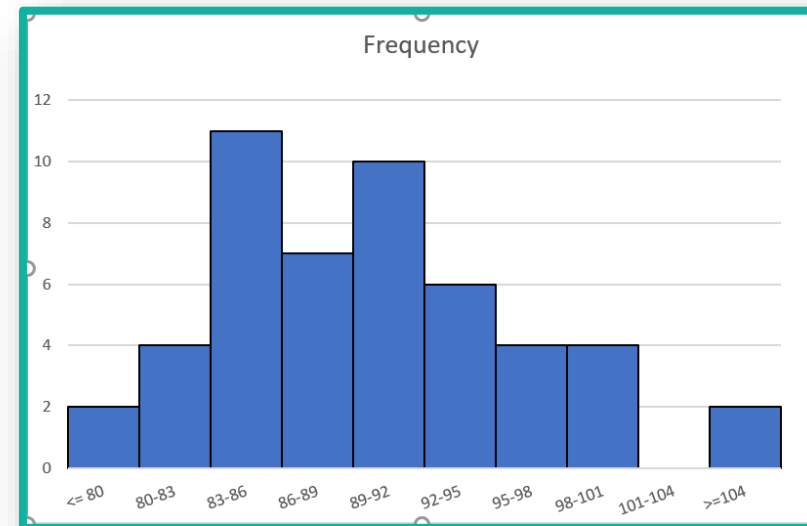
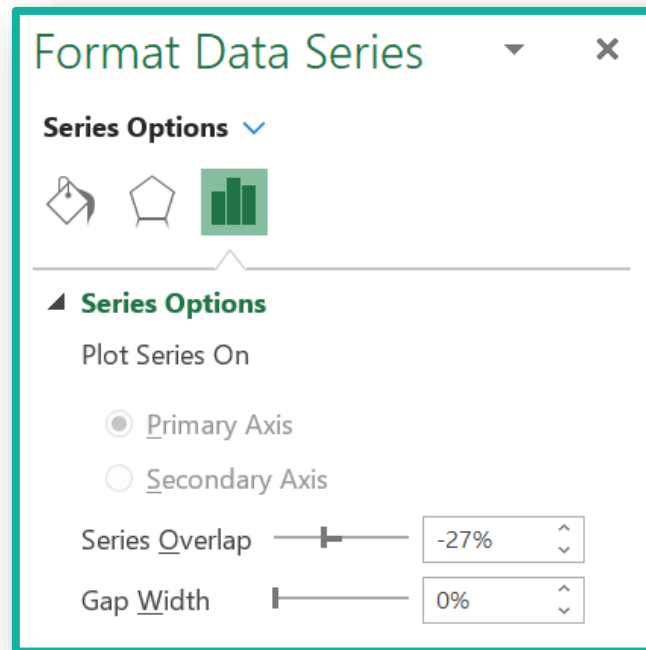
- Created from frequency table
 - Select **text options**, then **textbox**, then make **custom angle** -21 degrees
 - Select the rectangles and give boxes a 1pt black border



Ex: Starting Salaries



- Created from frequency table
 - When rectangles are selected, select **Series Options**, and make **Gap Width=0**

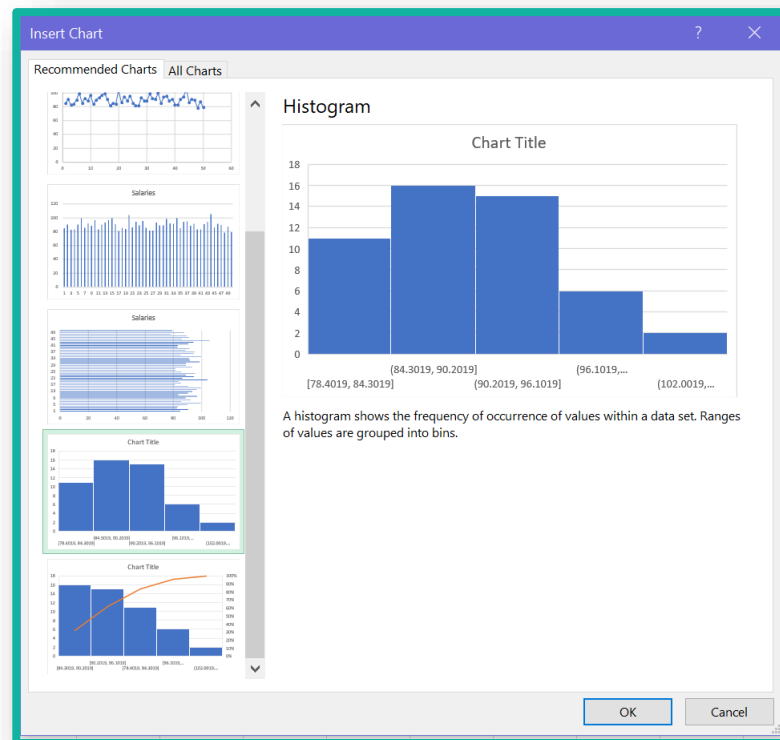


- Select the title **Frequency** and change to **Starting Salaries**

Ex: Starting Salaries



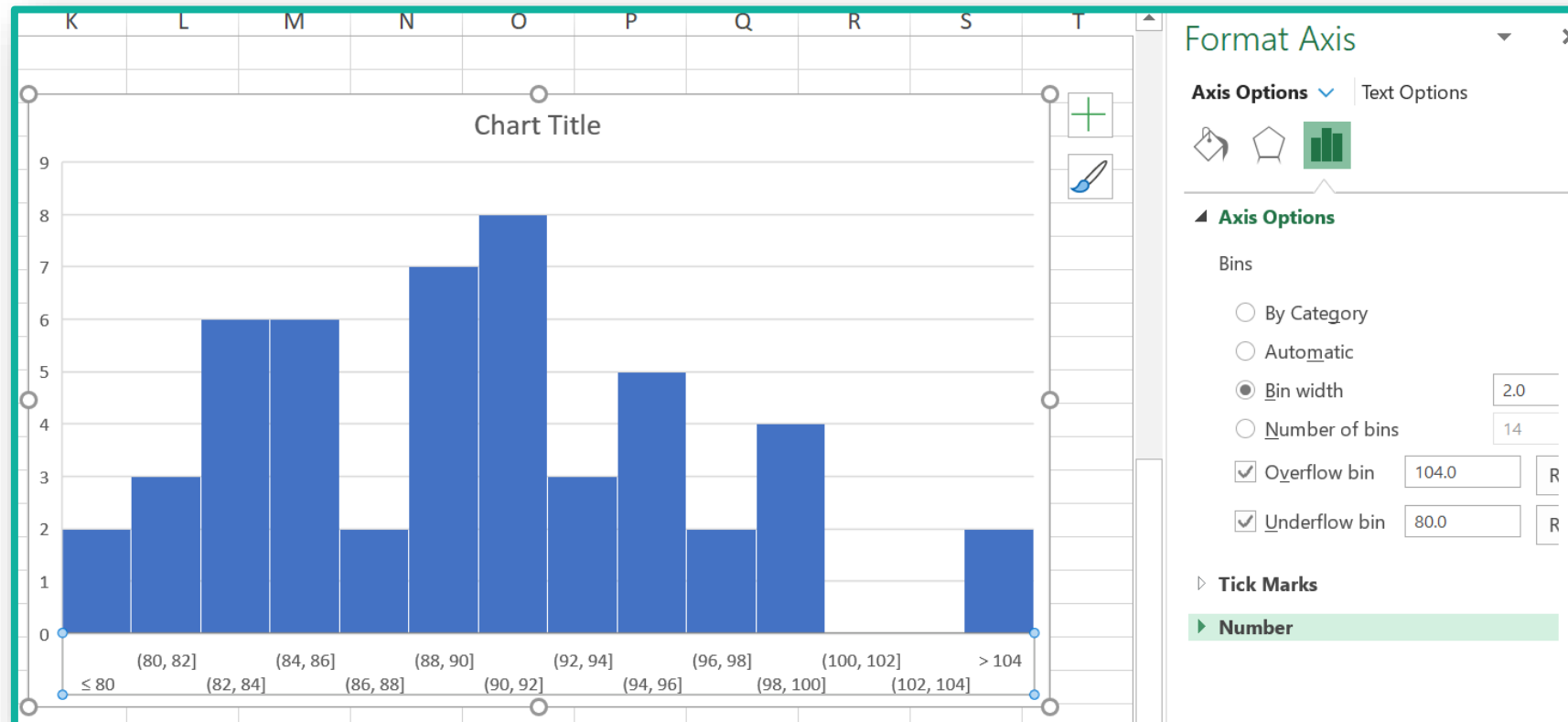
- Created from raw data
 - Highlight raw data and variable name in A3:A53
 - Select **Recommended Charts** in **Insert** menu
 - Select Histogram and select OK



Ex: Starting Salaries



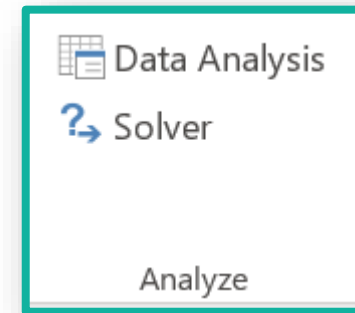
- Created from raw data
 - Select x-axis and modify **Axis Options**
 - We can build a histogram similar to previous histogram



Ex: Starting Salaries



- Created using the Analysis ToolPak add-in
 - Look at [Link 2](#) for help with loading the Analysis ToolPak
 - This may be automatically loaded on your computer
 - Above Excel Solver you will find **Data Analysis**
 - Select **Histogram** and fill out fields

A screenshot of the "Histogram" dialog box in Excel. The dialog box has a title bar with a question mark and a close button. It contains several sections: "Input" with "Input Range" set to "\$A\$3:\$A\$53" and "Bin Range" set to "\$C\$3:\$C\$12"; "Labels" with a checked checkbox; "Output options" with "New Worksheet Ply:" selected and "Plot Example" entered; and checkboxes for "Pareto (sorted histogram)", "Cumulative Percentage", and "Chart Output" (which is checked). There are "OK", "Cancel", and "Help" buttons on the right.

Histogram

Input

Input Range:

Bin Range:

☒ Labels

Output options

☐ Output Range:

☒ New Worksheet Ply:

☐ New Workbook

☐ Pareto (sorted histogram)

☐ Cumulative Percentage

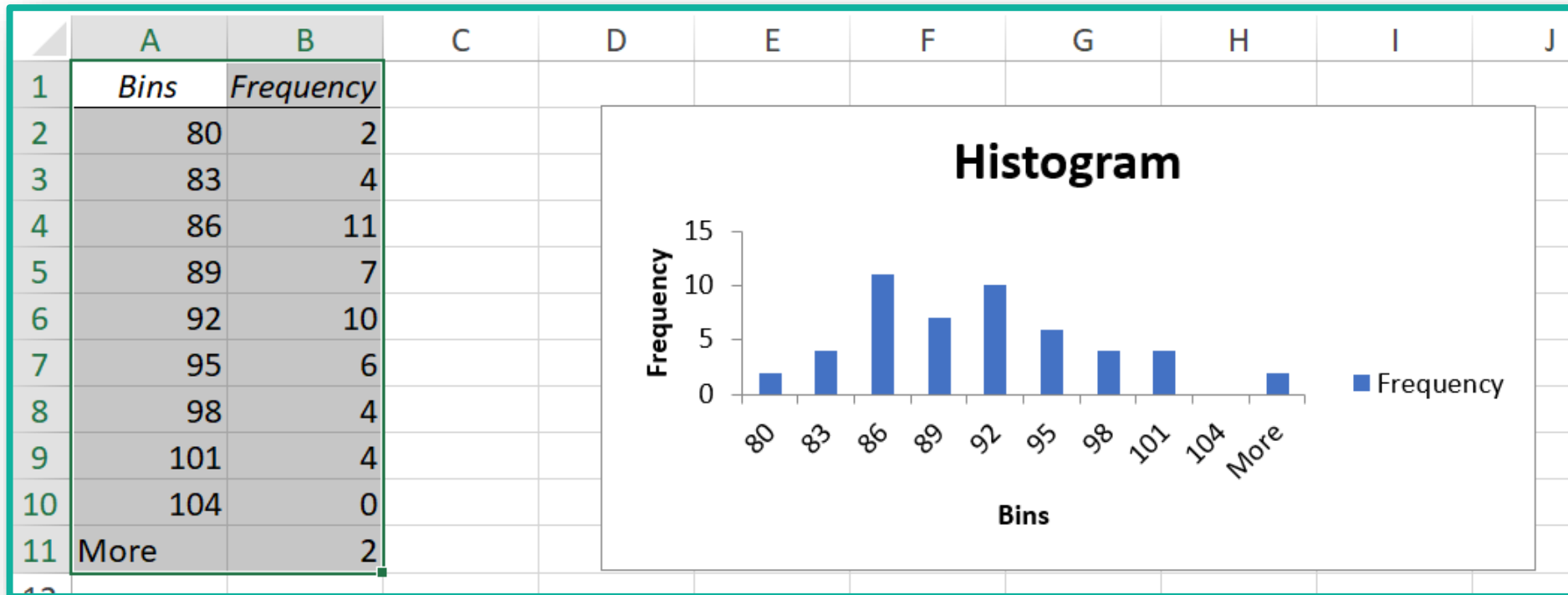
☒ Chart Output

OK Cancel Help

Ex: Starting Salaries



- Created using the Analysis ToolPak add-in
 - Result found in new sheet named **Plot Example**



- Requires more modification

Descriptive Statistics



- Collection of data points is called a **sample**, and we interpret it as a subset of observations from some underlying random phenomenon
- We denote the i th point in the sample as X_i
- We denote the whole set of observations as $\{X_1, X_2, \dots, X_n\}$
- To measure the center of the data, we compute three quantities
 - **Sample mean**: the average value of our observations

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

- **Sample median**: the value that divides the bottom 50% by the top 50%
- **Mode**: the most frequently occurring value (**discrete** or **categorical** only)

Descriptive Statistics



- Q: Why calculate the sample mean and sample median?
- To measure the spread of the data, we compute three quantities
 - **Sample variance**: the average squared distance between an observation and the sample mean

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

- **Sample standard deviation**: more convenient than the sample variance

$$S_X(n) = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}(n))^2}$$

- **Range**: the difference between the largest value and smallest value

Descriptive Statistics



- Percentiles are also helpful
 - The *k*th percentile is a value that divides the bottom $k\%$ from the top $(1-k)\%$
 - The median is the 50th percentile
 - The 25th and 75th percentiles (Q_1 and Q_3) are useful for understanding the variability in the middle of the distribution
 - The *interquartile range* (IQR) is the difference between Q_3 and Q_1

Ex: Starting Salaries



- Analyze the formulas for these statistics in the **Frequency** sheet

Sample Mean		89.772812
Sample Median		89.73195
Sample Variance		38.9904019
Sample SD		6.24422949
Min		78.4019
Max		105.5129
Range		27.111
Q1		85.2254
Q3		93.963575
IQR		8.738175

Ex: Starting Salaries



- More information can be gathered using **Data Analysis** in the **Data** menu

Descriptive Statistics

Input

Input Range:

Grouped By: ☒ Columns ☐ Rows

☒ Labels in first row

Output options

☐ Output Range:

☒ New Worksheet Ply:

☐ New Workbook

☒ Summary statistics

☒ Confidence Level for Mean: %

☐ Kth Largest:

☐ Kth Smallest:

OK Cancel Help

Ex: Starting Salaries



- Observe the results in the created sheet named **Descriptive**

	A	B
1	<i>Salaries</i>	
2		
3	Mean	89.772812
4	Standard Error	0.883067403
5	Median	89.73195
6	Mode	#N/A
7	Standard Deviation	6.244229491
8	Sample Variance	38.99040194
9	Kurtosis	-0.155889833
10	Skewness	0.417303487
11	Range	27.111
12	Minimum	78.4019
13	Maximum	105.5129
14	Sum	4488.6406
15	Count	50
16	Confidence Level(90.0%)	1.480507443



The End



Dale

