Statistics

A PDF version of Excel exercise solutions that don't visualize properly

Categorical variables – Visualization techniques
(appears in '2.3.Categorical-variables.Visualization-techniques-exercise-solution.xlsx')

Categorical variables. Visualization techniques

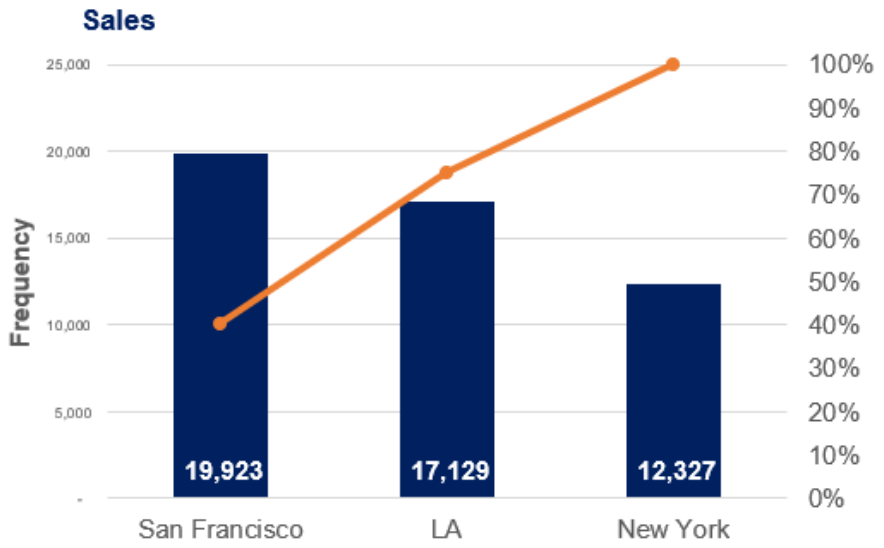
Ice cream shop

Background	You have a frequency distribution table with all the sales. You also have the relative frequency from the pie chart problem.
Task 1	Order the table by frequency.
Task 2	Create a bar (column) chart representing the ordered data.
Task 3	In a new column, calculate the cumulative frequency of the data.
Task 4	On a second axis in the same chart, represent the cumulative frequency of the data.

Solution:

Ordered	Frequency	Relative frequency	Cumulative frequency
San Francisco	19,923	40%	40%
LA	17,129	35%	75%
New York	12,327	25%	100%
Total	49,379	100%	

Adding a second axis is not so straightforward in Excel.
This may be done in various ways. Here is a link to Microsoft's article on the topic. Link
For the purposes of statistics, you need to understand the application of the cumulative frequency line.
Drawing it in Excel is not top priority for this course.



Categorical variables – Visualization techniques
(appears in '2.3.Categorical-variables.Visualization-techniques-exercise-solution.xlsx')

Categorical variables. Visualization techniques

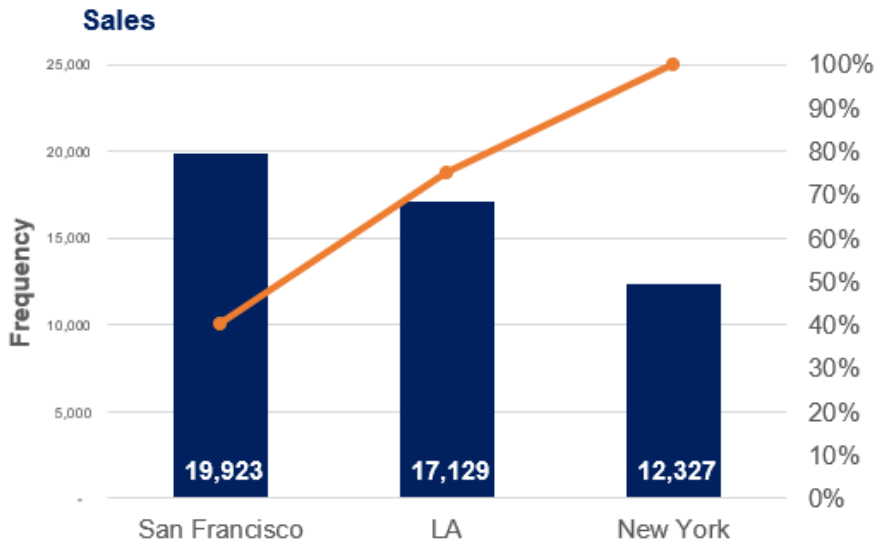
Ice cream shop

Background	You have a frequency distribution table with all the sales. You also have the relative frequency from the pie chart problem.
Task 1	Order the table by frequency.
Task 2	Create a bar (column) chart representing the ordered data.
Task 3	In a new column, calculate the cumulative frequency of the data.
Task 4	On a second axis in the same chart, represent the cumulative frequency of the data.

Solution:

Ordered	Frequency	Relative frequency	Cumulative frequency
San Francisco	19,923	40%	40%
LA	17,129	35%	75%
New York	12,327	25%	100%
Total	49,379	100%	

Adding a second axis is not so straightforward in Excel.
This may be done in various ways. Here is a link to Microsoft's article on the topic. Link
For the purposes of statistics, you need to understand the application of the cumulative frequency line.
Drawing it in Excel is not top priority for this course.



The Histogram (Part I)
(appears in '2.5.The-Histogram-exercise-solution.xlsx')

The histogram

Background You are given a dataset.
Task 1 Construct a frequency distribution table.
Note: Go to the next sheet if you wish to skip this part.
Task 2 Create a histogram with 10 intervals, based on your dataset.

Solution:

Dataset
13
68
165
193
216
228
361
470
500
529
544
602
647
692
696
699
809
892
899
936

Frequency distribution table. Exact width

Intervals	10
Interval width	92.3

Interval start	Interval end	Absolute frequency	Relative frequency
13	105.3	2	0.10
105.3	197.6	2	0.10
197.6	289.9	2	0.10
289.9	382.2	1	0.05
382.2	474.5	1	0.05
474.5	566.8	3	0.15
566.8	659.1	2	0.10
659.1	751.4	3	0.15
751.4	843.7	1	0.05
843.7	936	3	0.15
		20	1.00

2

Frequency distribution table. Rounded up width

Intervals	10
Interval width	93

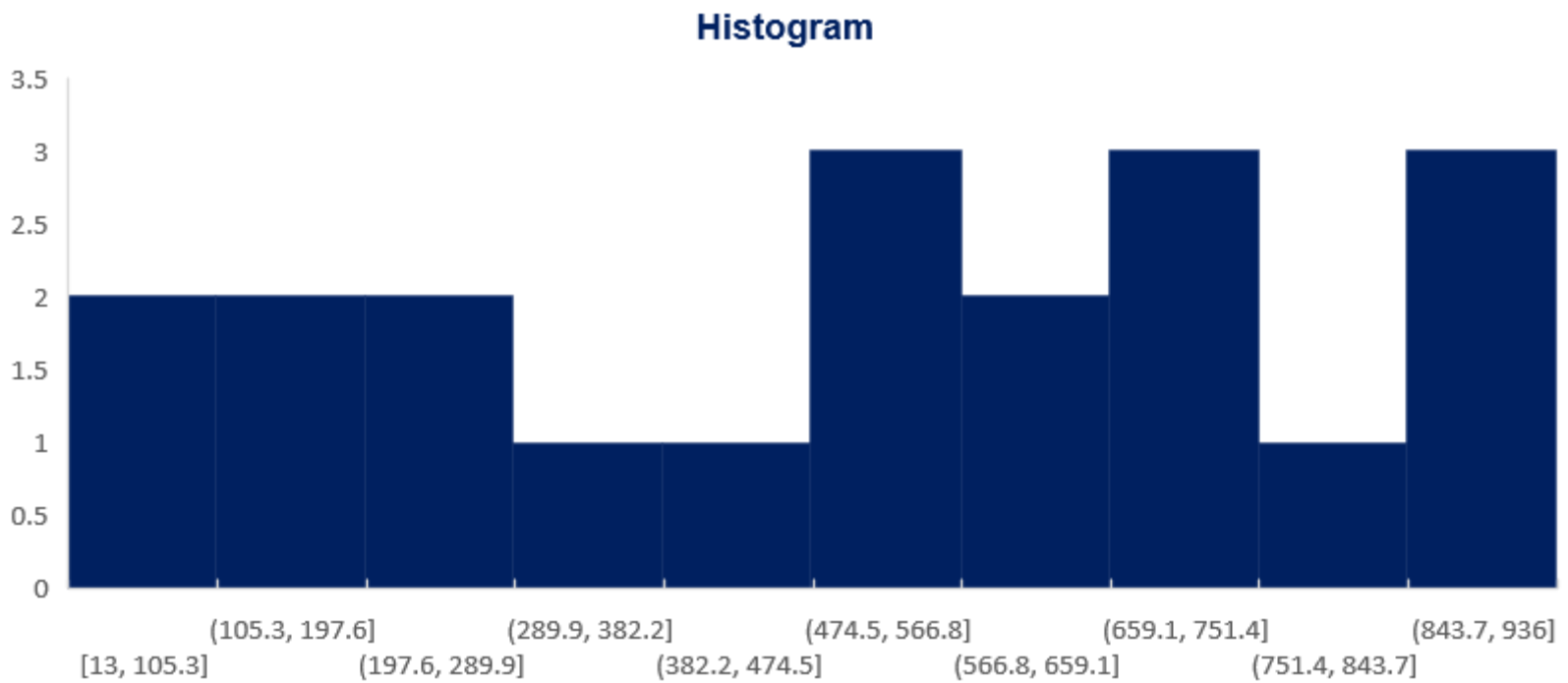
Interval start	Interval end	Absolute frequency	Relative frequency
13	106	2	0.10
106	199	2	0.10
199	292	2	0.10
292	385	1	0.05
385	478	1	0.05
478	571	3	0.15
571	664	2	0.10
664	757	3	0.15
757	850	1	0.05
850	943	3	0.15
		20	1.00

In Excel, the histogram is a special type of chart. In the latest versions of Excel, you should only select your dataset and insert a type of chart, called Histogram.
Alternatively, there is a special histogram tool in the Analysis Toolpak that Excel supports.
Other software will only let you create a histogram, once you have created a frequency distribution table. This is also the skill that you should acquire.

The intervals in the excel histogram are called 'bins'. You can specify the number of bins or the width of the bins.

The Histogram (Part II)

(appears in '2.5.The-Histogram-exercise-solution.xlsx')



Skewness
(appears in '2.8.Skewness-exercise-solution.xlsx')

Skewness

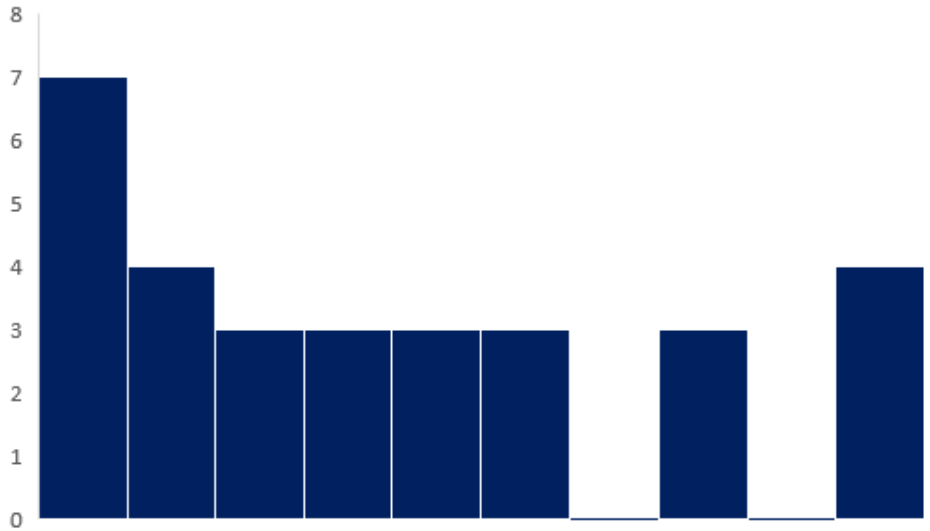
Background You are given two datasets
Task 1 Identify the skewness of dataset 1. You may use the formula from the lesson, the skewness formula in excel (=SKEW) or you can plot it on a graph
Task 2 Identify the skewness of dataset 2. You may use the formula from the lesson, the skewness formula in excel (=SKEW) or you can plot it on a graph

Solution:

Dataset 1
212
869
220
654
511
624
420
121
428
865
799
405
230
670
870
366
99
55
489
312
493
163
221
84
144
48
375
86
168
100

Task 1:

Skewness 0.63



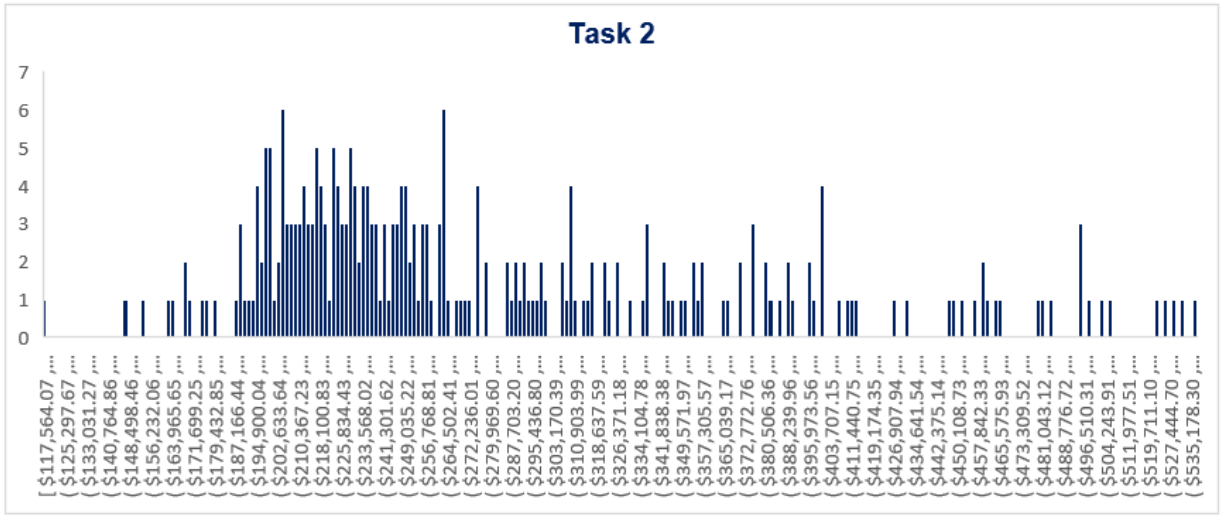
The skew of this dataset is positive.

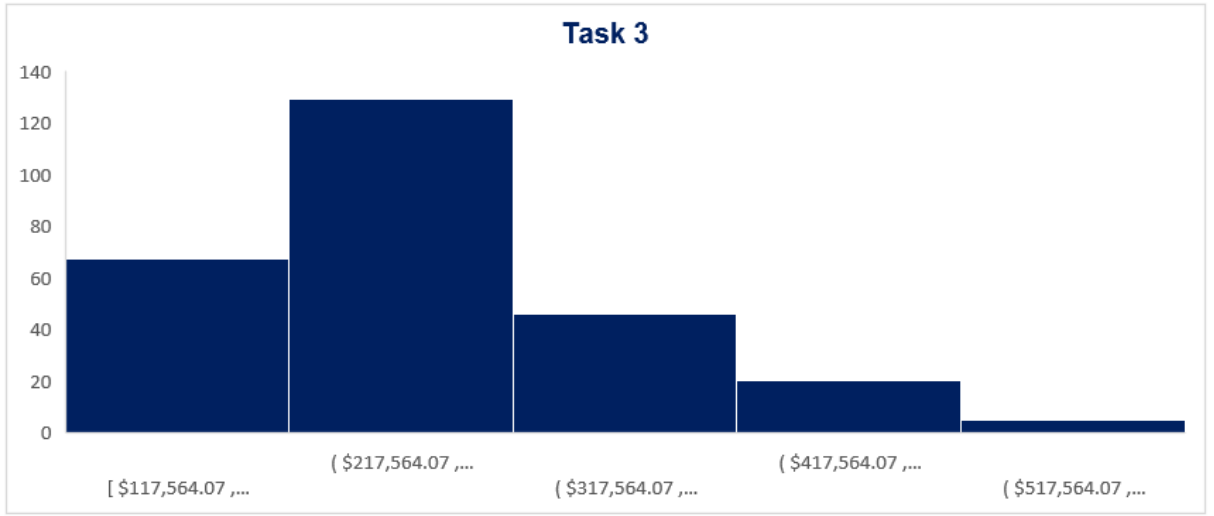
Practical Example – Descriptive Statistics
(appears in '2.13.Practical-example.Descriptive-statistics-exercise-solution.xlsx')

365 DataScience RE California Database
Histograms. Graphing numerical data

Task 2: Create a frequency distribution graph (that is a histogram with the highest possible number of bins - 272). Use data on all properties, no matter if sold or not.
Task 3: Create a histogram which represents the Price variable. Choose interval width (bins) of length $100,000. If you don't know how to do that, refer to the Course notes on descriptive statistics provided with the first lecture in this section.
Use the data on all properties, no matter if sold or not.
Task 4: Interpret the results.

Solution:





Task 4: The histograms point to similar insights - most of the properties' prices are concentrated in the interval ($217,564.07 to 317,564.07)

Standard Normal Distribution
(appears in '3.4.Standard-normal-distribution-exercise-solution.xlsx')

Standard normal distribution

Background	You are given an approximately normally distributed dataset
Task 1	Calculate the mean and standard deviation of the dataset
Task 2	Standardize the dataset
Task 3	Plot the data on a graph to see the change

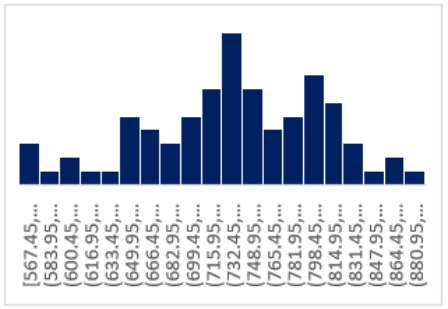
Solution:

Original dataset
567.45
572.45
572.45
589.12
613.87
615.78
628.45
644.87
650.45
652.20
656.87
661.45
666.45
667.70
668.95
675.28
675.78
685.53
694.28
697.62
705.78
705.87

Task 1

Mean	743.03
St. dev	73.95

Task 3.1

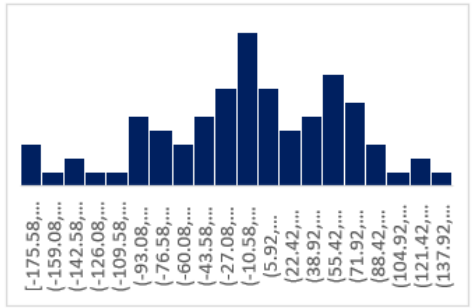


Subtract mean
-175.58
-170.58
-170.58
-153.91
-129.16
-127.24
-114.58
-98.16
-92.58
-90.83
-86.16
-81.58
-76.58
-75.33
-74.08
-67.74
-67.24
-57.49
-48.74
-45.41
-37.24
-37.16

Task 2

Mean	0.00
St. dev	73.95

Task 3.2

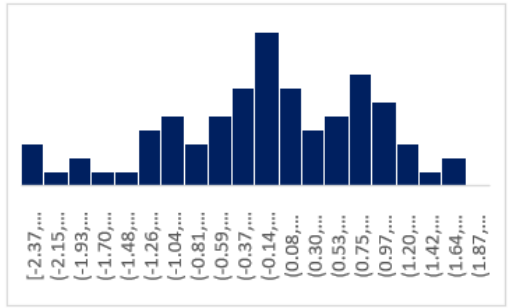


Standardized
-2.37
-2.31
-2.31
-2.08
-1.75
-1.72
-1.55
-1.33
-1.25
-1.23
-1.17
-1.10
-1.04
-1.02
-1.00
-0.92
-0.91
-0.78
-0.66
-0.61
-0.50
-0.50

Task 2

Mean	0.00
St. dev	1.00

Task 3.3



You can see that the difference in the graphs is almost unnoticeable
However, the mean (center of the graph) and the standard deviation (the spread) of the graph are completely different.