

Summarizing Data with Statistics

Webster dictionary defines data as factual information (such as measurements or statistics) used as a basis for reasoning, discussion, or calculation'. Factual information can take many forms. We will discuss these forms of data and what they mean.

Data can be broadly classified into two categories viz categorical and quantitative data. As the name suggests qualitative data is non-numeric in nature whereas quantitative data is numeric. Categorical data can be further classified as ordinal and nominal, whereas quantitative data can be subcategorized as discrete and continuous.

Categorical Data

A categorical variable measures something and identifies a group to which the thing belongs. They describe a quality or characteristic of a data unit like what type or which category. They tend to be represented by a non-numeric value and fall into mutually exclusive and exhaustive categories. Sometimes a categorical variable is stored as a string. Ex: Blood group of a person

Categorical data are classified as:

- Nominal data there is no natural order between the categories (eg eye color, Gender i.e Male or Female),
 - Ordinal data an ordering exists (eg. exam results, socio-economic status, Educational Level - Kindergarten, Primary, Secondary, Higher Secondary, Graduation, etc.).
- Quantitative Data

Quantitative or numerical data arise when the observations are counts or measurements. The data are said to be discrete if the measurements are integers (ex. number of people in a household, number of cigarettes smoked per day) and continuous if the measurements can take on any value, usually within some range (ex. weight).

- Interval data is numeric unlike nominal and ordinal data and the difference between the successive numbers is constant. Such variables do not have a zero point. This means that a value zero does not mean the absence of this variable. eg. (IQ score of individuals, temperature).
- Ratio is quite the same as interval variables with the addition that they have a meaningful zero point. This means a ratio value zero indicates the absence of the given variable. eg. (height, age, the weight of individuals, dimensions of a room).

We should be aware of these different types of data, as we need to apply different processes to the same. We will look at these processes during the data pre-processing step.

Let us now take a look at a real dataset and try to identify the type of variables the dataset contains.

We will be using a rich dataset on housing prices from Ames, Iowa. Each row in the dataset describes the properties of a single house as well as the amount it was sold for. The original data set contains 82 features and 2930 data points. You can read more about this dataset [here](#) and download the dataset from [here](#).

However, for the purpose of learning descriptive statistics, we will specifically, work on the following features of the house

- Lot Area
- Number of Bedrooms
- Garage Area
- Condition of the house
- Sale Price

Brief explanation of the dataset & features

- Lot Area (Continuous) : Lot size in square feet.
- Bedroom (Discrete) : Bedrooms above grade (does NOT include basement bedrooms).
- Garage Area (Continuous) : Size of garage in square feet.
- Overall Cond (Ordinal) : Rates the overall condition of the house. 10 Very Excellent 9 Excellent 8 Very Good 7 Good 6 Above Average 5 Average 4 Below Average 3 Fair 2 Poor 1 Very Poor
- SalePrice: Sale price of the house

Structured vs. Unstructured Data

Structured data is generally organized in rows and columns and stored in some kind of database. Some examples of structured data are

- Airline reservation system
- Student database at a university
- Comparison chart for automobiles

On the other hand, unstructured data is not organized and can be stored in different formats. Some examples of unstructured data are

- Images
- PDF documents of reports
- X-ray reports
- Text messages

Following table gives us a comparison and overview of structured and unstructured data

	Structured Data	Unstructured Data
Characteristics	<ul style="list-style-type: none"> • Pre-defined data models • Usually text only • Easy to search 	<ul style="list-style-type: none"> • No pre-defined data model • May be text, images, sound, video or other formats • Difficult to search
Resides in	<ul style="list-style-type: none"> • Relational databases • Data warehouses 	<ul style="list-style-type: none"> • Applications • NoSQL databases • Data warehouses • Data lakes
Generated by	Humans or machines	Humans or machines
Typical applications	<ul style="list-style-type: none"> • Airline reservation systems • Inventory control • CRM systems • ERP systems 	<ul style="list-style-type: none"> • Word processing • Presentation software • Email clients • Tools for viewing or editing media
Examples	<ul style="list-style-type: none"> • Dates • Phone numbers • Social security numbers • Credit card numbers • Customer names • Addresses • Product names and numbers • Transaction information 	<ul style="list-style-type: none"> • Text files • Reports • Email messages • Audio files • Video files • Images • Surveillance imagery

Histogram

While dealing with data, often you will want to get a sense of how the variables are distributed. The histogram is a visual method to plot frequency distribution for a continuous variable. The histogram helps us identify different characteristics of data that we will be learning later in this chapter.

Let us now go through a simple example to see how a histogram is constructed. The following table represents SalePrice of 20 houses from our dataset.

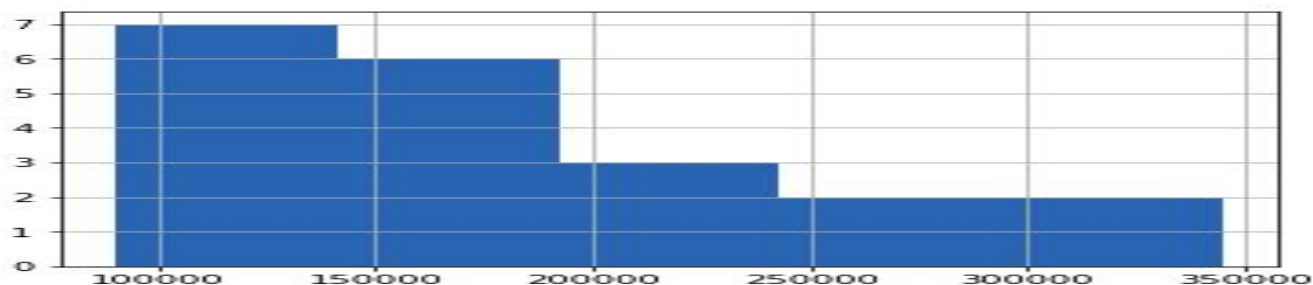
208500	181500	223500	140000	250000
143000	307000	200000	129900	118000
129500	345000	144000	279500	157000
132000	149000	90000	159000	139000

We will build a frequency distribution table for these observations now.

Class	Frequency
90000 - 141000	7
141000 - 192000	6

192000 - 243000	3
243000 - 294000	2
294000 - 345000	2

Now we can plot a histogram as follows. The X-axis represents the SalePrice of the house and the Y-axis represents the frequency of the class.



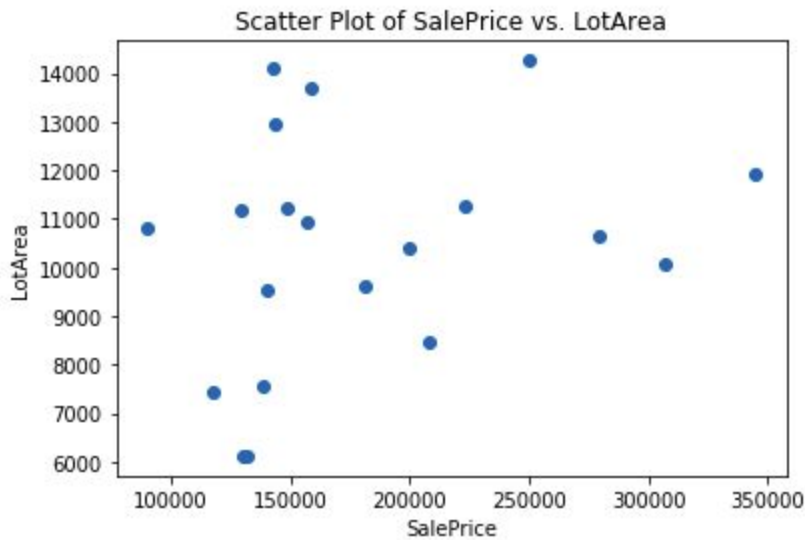
Scatter Plot

Scatter plots are used to depict the relationship between two variables. The data is represented by a collection of points where each point represents a value of a set of variables given on the X and Y-axis. Scatter plots are used when we want to define relationships between quantitative variables.

Let us now see how to draw a scatter plot. We will work with 20 observations from our dataset and draw a scatter plot for variables SalePrice and LotArea.

	LotArea	SalePrice
0	8450	208500
1	9600	181500
2	11250	223500
3	9550	140000
4	14260	250000
5	14115	143000
6	10084	307000
7	10382	200000
8	6120	129900
9	7420	118000
10	11200	129500
11	11924	345000
12	12968	144000
13	10652	279500
14	10920	157000
15	6120	132000
16	11241	149000
17	10791	90000
18	13695	159000
19	7560	139000

Now based on these observations we will draw a scatter plot where the X-axis represents the SalePrice of the house and the Y-axis represents the LotArea of the house. Each observation is represented using a point on the graph.



Here we can see that there is no specific relationship between the variables.

Pie charts

Pie charts are used to display the total number of observations of different types in a dataset in terms of percentage.

Let us look at an example to build intuition. In our dataset, we have information about how many FullBaths each house has. We want to summarize the data and find out the proportion of houses having 0,1,2 and 3 FullBaths. We can first count the number of houses and build a small table as follows

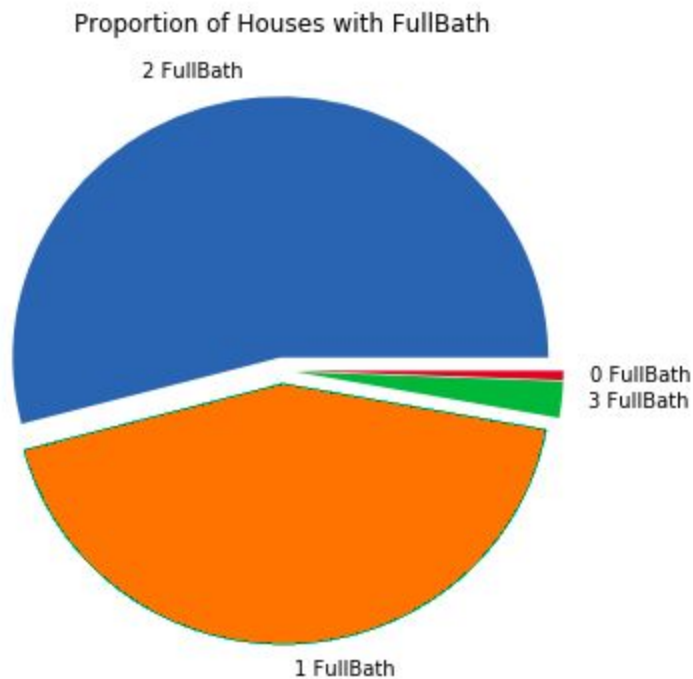
No. of FullBath	No. of houses
0	8
1	594
2	746
3	31

In order to draw a pie chart, we need proportions for all the readings. These are calculated and summarized in following tables

No. of FullBath	No. of houses	Proportion

0	8	1%
1	594	43%
2	746	54%
3	31	2%

Now we can draw a pie chart as follows for the variable FullBath



We can clearly get intuition on the proportion of houses having a different number of FullBaths.

Mathematical Average

In the previous chapter, we looked at different types of data and a way to summarize the same visually. In this chapter, we will learn about more characteristics of data. The first such characteristic is the central tendency of the data. The central tendency of the data can be defined as a numerical value around which most of the values from data tend to cluster.

The simplest central tendency that most of us might know is a simple average. Let us see what are the different measures of central tendency and look into details of the same one by one.

The central tendency can be broadly divided into two categories viz Mathematical Average and Positional Average. These can be further divided as follows

Mathematical Average	Position Average
Arithmetic	Median
Geometric	Mode
Harmonics	-

In this section, we will learn about difference mathematical averages.

Arithmetic Mean

The arithmetic mean is the most popular and simple measure to calculate the central tendency of the data. We can calculate the arithmetic mean of a variable by adding all the values and dividing the sum by a number of values in the data. Mathematically it can be represented as follows

$$arithmetic\ mean = \frac{x_1 + x_2 + x_3 + ... + x_n}{n} = \frac{1}{n} \sum_{i=1}^n x_i$$

where x_1x_n represents **n** values present in the data.

represents n values present in the data.

Example: Let's take a data table consisting of ten records of heights (in cm) and weights (in kg) and calculate the mean value of each column

Height (x)	Weight (y)
173	80
176	86
167	65
178	82
180	115

167	62
156	36
177	83
172	77
175	81

The mean value of Height

The mean value of Height

$$\bar{x} = \frac{1}{10} \sum_{i=1}^{10} (Height)_i$$

$$= \frac{173 + 176 + 167 + 178 + 180 + 167 + 156 + 177 + 172 + 175}{10} = 172.1$$

The mean value of Weight

$$(\bar{y}) = \frac{1}{10} \sum_{i=1}^{10} (Weight)_i$$

$$= \frac{80 + 86 + 65 + 82 + 115 + 62 + 36 + 83 + 77 + 81}{10} = 76.7$$

Geometric Mean

In many business and economics problems, such as calculation of compound interest and inflation, variables change over a period of time. In such cases, we need to know the average percentage change in the variable's value over a period of time. We use geometric mean to calculate this. Mathematically it can be represented as follows

$$geometric\ mean = \sqrt[n]{(x_1 \times x_2 \times x_3 \times \dots \times x_n)}$$

where x_1, \dots, x_n represents n values present in the data.

Example : Consider a set of values - 2, 4, 6, 10, 15

Since there are 5 numbers, $n=5$

With the above formula, the geometric mean of these numbers =

$$\sqrt[5]{(2 \times 4 \times 6 \times 10 \times 15)} = \sqrt[5]{7200} = 5.90$$

Harmonic Mean

Harmonics mean is a very specific kind of mathematical average. The harmonic mean is generally useful when we want to calculate the average of ratios or rates. Mathematically it can be represented as follows

$$\text{harmonic mean} = \frac{n}{\sum_{i=1}^n \left(\frac{1}{x_i} \right)}$$

Example : Consider the same set of values - 2, 4, 6, 10, 15

The harmonic mean for the above values is -

$$\frac{5}{\frac{1}{2} + \frac{1}{4} + \frac{1}{6} + \frac{1}{10} + \frac{1}{15}} = 1.084$$

Positional Average

Mathematical averages can be calculated for numerical data as seen in the earlier topic. However, they are not suitable measures when the data is categorical or non-numerical in nature. Also, some of these measures are prone to distortion based on extreme values in data. As the name suggests positional averages are calculated based on the position of particular value in the dataset. Let us explore these positional average one by one.

Median

Median can be defined as the value at the center of the dataset when the dataset is arranged in an order of increasing or decreasing magnitude. Position of median value can be mathematically calculated as follows

For odd number of observations

For odd number of observations

$$median = value\ of\ \left(\frac{n+1}{2}\right)^{th}\ element$$

For even number of observations

$$median = \frac{value\ of\ \left(\frac{n}{2}\right)^{th}\ element + value\ of\ \left(\frac{n}{2} + 1\right)^{th}\ element}{2}$$

Considering our previous example, let's calculate the median values for both Height and Weight.

Height (x)	Weight (y)
173	80
176	86
167	65
178	82
180	115
167	62
156	36
177	83
172	77
175	81

Arranging the above table in ascending order:

Height (x)	Weight (y)
------------	------------

156	36
167	62
167	65
172	77
173	80
175	81
176	82
177	83
178	86
180	115

Both the columns have 10 observations. So the median values for Height and Weight can be found using the second formula :

Both the columns have 10 observations. So the median values for Height and Weight can be found using the second formula :

$$\text{Median} = \frac{\text{value of } \left(\frac{n}{2}\right)^{\text{th}} \text{ element} + \text{value of } \left(\frac{n}{2} + 1\right)^{\text{th}} \text{ element}}{2}$$

$$\text{Median} = \frac{\text{value of } \left(\frac{10}{2}\right)^{\text{th}} \text{ element} + \text{value of } \left(\frac{10}{2} + 1\right)^{\text{th}} \text{ element}}{2} =$$

$$\text{Median}(\text{Height}) = \frac{173 + 175}{2} = \frac{348}{2} = 174$$

$$\text{Median}(\text{Weight}) = \frac{80 + 81}{2} = \frac{161}{2} = 80.5$$

Partition Values - Quartiles, Deciles and Percentiles

The value of observations in a data set, when arranged in an ordered sequence, can be divided into four equal parts using three quartiles namely Q1, Q2, and Q3.

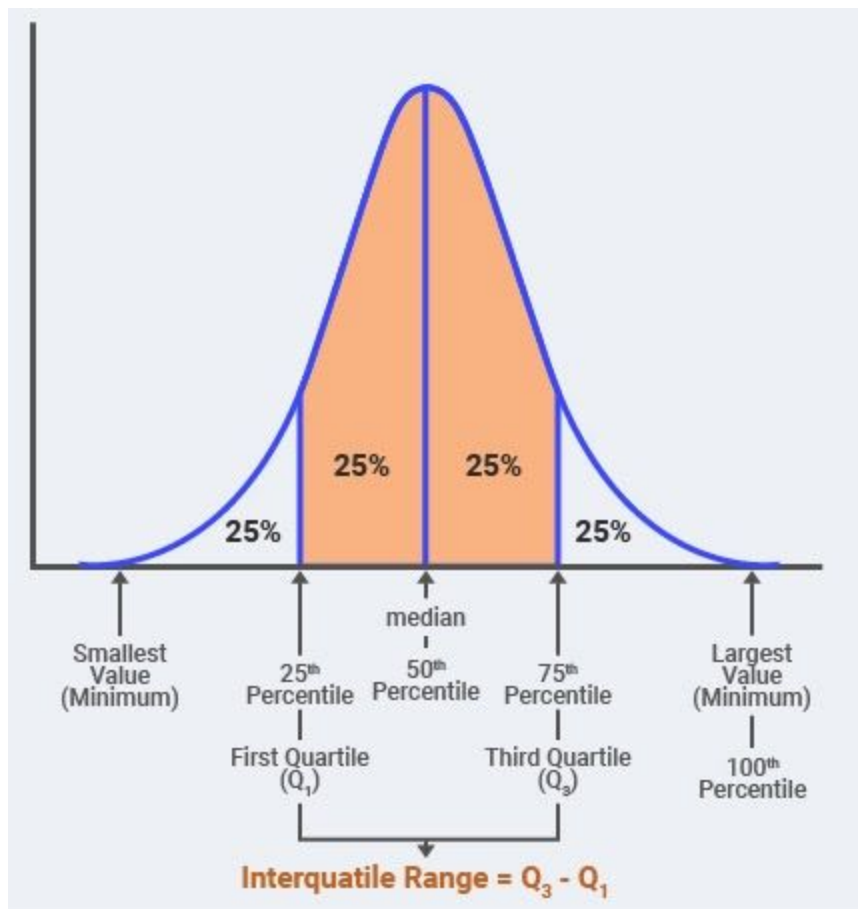
- The first quartile Q1 divides a distribution in such a way that 25% of observations have a value less than Q1 and 75% have a value more than Q1.
- The second quartile Q2 divided a distribution into two equal halves and hence it is median
- The third quartile Q3 divides the data set in such a way that 75% of the observations have a value less than Q3 and 25% have a value greater than Q3

In a similar fashion, Deciles divide the data set in 10 equal parts and percentiles divide the data set on 100 equal parts.

Mode

Mode is the value of observation which occurs the most number of times in a dataset.

The following visual summarizes all the measures of central tendency that we have learned.



Calculate arithmetic mean, median, mode for SalePrice and plot the same on the histogram

In this task, we will calculate the central tendency measures for the column of `SalePrice` from our data

Instructions

- The dataset is stored in DataFrame `df`
- Calculate arithmetic mean for `SalePrice` using `mean()` method of pandas and store the value in a variable names `mean`
- Calculate mode for `SalePrice` using `mode()` method of pandas and store the value in a variable names `mode`
- Calculate the median for `SalePrice` using `median()` method of pandas and store the value in a variable names `median`
- Use `hist()` method from matplotlib to plot the histogram for `SalePrice`, keep `bins=40` as a parameter
- Use the following block of code to plot the mean, median and mode on the histogram.

```
plt.plot([mode]*300, range(300), label='mode')
plt.plot([median]*300, range(300), label='median')
plt.plot([mean]*300, range(300), label='mean')
```

- Plot the legend for the chart

```
# Code starts here
mean = df['SalePrice'].mean()
mode = df['SalePrice'].mode()
median = df['SalePrice'].median()
df.hist(column = 'SalePrice',bins = 40)
plt.plot([mode]*300, range(300), label='mode')
plt.plot([median]*300, range(300), label='median')
plt.plot([mean]*300, range(300), label='mean')
plt.show()
# Code ends here
```

Use of correct measure

Now that we have gone through various measures of central tendency it is time to know which type of measure is to be used based on the data.

When is the mean the best measure of central tendency?

- The mean is usually the best measure of central tendency to use when your data distribution is continuous and symmetrical, such as when your data is normally distributed.
- However, it all depends on what you are trying to show from your data.

- You will learn about Normal Distributions during your session in Inferential Statistics

When is the mode the best measure of central tendency?

- The mode is the least used of the measures of central tendency
- The mode will be the best measure of central tendency (as it is the only one appropriate to use) when dealing with nominal data.
- The mean and/or median are usually preferred when dealing with all other types of data, but this does not mean it is never used with these data types

When is the median the best measure of central tendency?

- The median is usually preferred to other measures of central tendency when your data set is skewed (i.e., forms a skewed distribution) or you are dealing with ordinal data. We will learn about skewness in upcoming chapters.
- However, the mode can also be appropriate in these situations, but IS NOT as commonly used as the median.

What is the most appropriate measure of central tendency when the data has outliers?

- The median is usually preferred in these situations because the value of the mean can be distorted by the outliers.
- However, it will depend on how influential the outliers are. If they do not significantly distort the mean, using the mean as the measure of central tendency will usually be preferred.

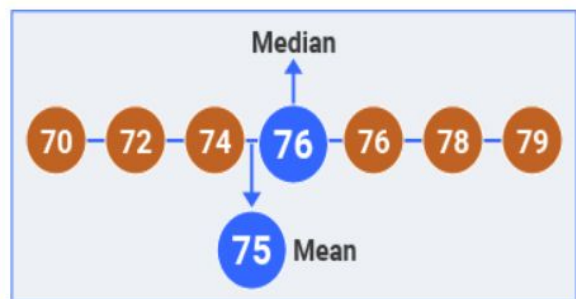
eg. Let's consider the below numbers :

70, 72, 74, 76, 76, 78, 79

The mean for the above numbers =

$$\frac{70 + 72 + 76 + 74 + 78 + 76 + 79}{7} = 75$$

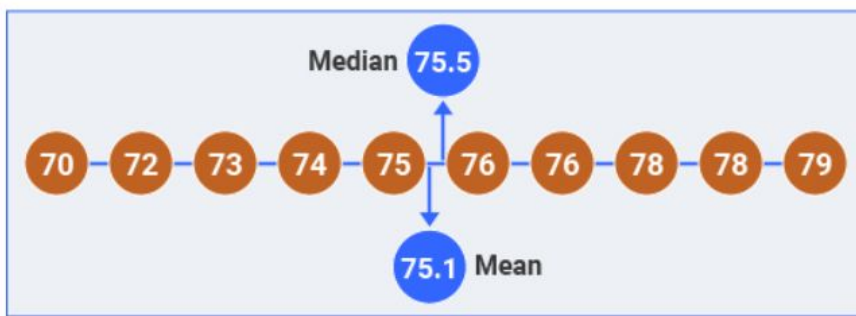
and the median is the 4th value which is 76



Let's add 3 additional values 78, 73, 75 that belong to the same range (between 70-80) as the original numbers.

70, 72, 76, 74, 78, 76, 79, 78, 73, 75

The mean is now = $\frac{70+72+76+74+78+76+79+78+73+75}{10} = 75.1$ while the median will be the average of 75 and 76 = 75.5



You can see that the mean and median don't differ much from the previous values.

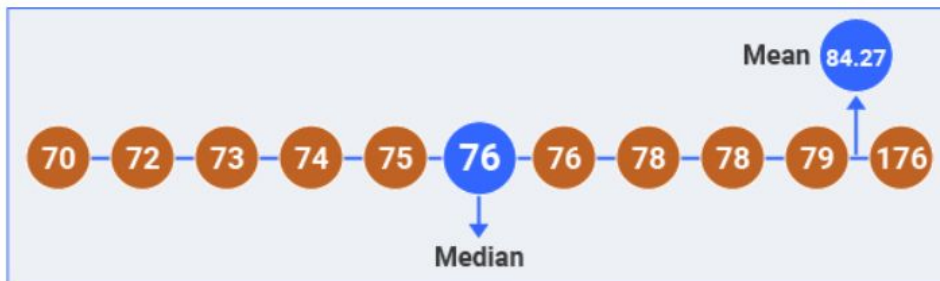
Now, let's introduce a value 176 that is quite further away from the given range of numbers. We can call this value an outlier.

70, 72, 76, 74, 78, 76, 79, 78, 73, 75, 176

The mean for the above numbers is =

$$\frac{70 + 72 + 76 + 74 + 78 + 76 + 79 + 78 + 73 + 75 + 176}{11} = 84.27$$

Since there are 11 numbers, which is an odd value, the median will be $(\frac{11+1}{2})^{th}$ value i.e. $\frac{12}{2} = 6^{th}$ value = 76



We can see that the mean has jumped from 75.1 to 84.27 but the median is around the same values as before.

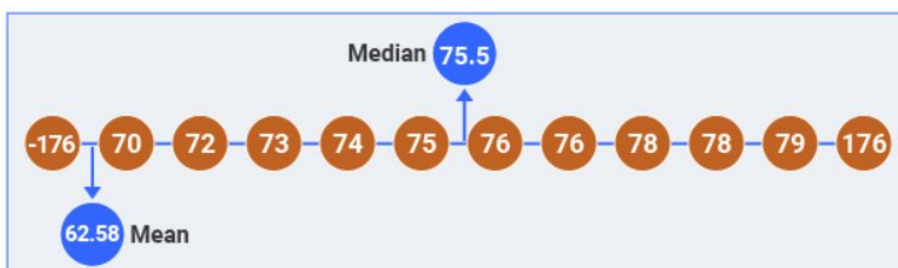
176 was an outlier that belonged to the positive extreme. Let's check what happens when we introduce a value -176 to our set of numbers.

-176, 70, 72, 76, 74, 78, 76, 79, 78, 73, 75, 176

The mean is =

$$\frac{-176 + 70 + 72 + 76 + 74 + 78 + 76 + 79 + 78 + 73 + 75 + 176}{12} = 62.58$$

while the median is 75.5



The mean has again swung towards the outlier -176 while the median remains fixed around its position.

From the above example, we can see that the mean is heavily influenced by outliers while the median stays around the same value. Hence in case of outliers that are far away from the majority of the numbers in the group, the median is a better measure around which the data is centered since it is not influenced by outliers.

Range

Central tendency describes the central value around which most of the data is clustered in a dataset, but it does not describe by what magnitude is this data spread on either side of central value. We need another measure to look at this.

The range is one of the simplest measures to describe the dispersion of data. The range is defined as the difference between the highest and the lowest observation in a dataset.

Range = highest observation – lowest observation

Let's consider the heights and weights of different individuals.

Height (in cm)	Weight (in kg)
173	80
176	86
167	65
178	82
180	115
167	62
156	36
177	83
172	77
175	81

The maximum value for height is 180 cm

The minimum value for height is 156 cm

So the range covered by all the data = $180 - 156 = 24$ cm

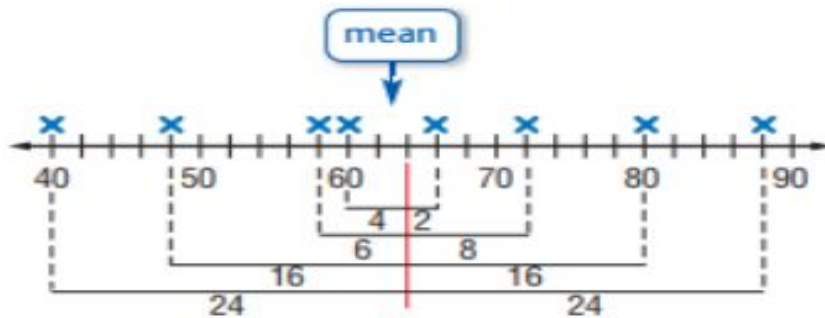
Mean Absolute Deviation

Although range is a good measure of dispersion for some kind of variables, it does not describe how data is scattered around the central value or dispersed across the range, we need some other measure to calculate this.

One way to calculate the spread of data is to calculate the distance of each data point in the data set from its central value. We can calculate the average of these distances and understand the average spread of data, but there is one problem, by definition of mean the sum of these distances would be always zero. One way to get around this problem is to take absolute values of these distances. Therefore, mathematically Mean Absolute Deviation can be calculated as follows

$$MAD = \frac{1}{n} \sum_{i=1}^n |x_i - \mu|$$

where μ is the mean of observations $x_1 \dots x_n$



Taking the previous example of heights and weights, let's calculate the mean absolute deviation for Weight (in kg)

Height (in cm)	Weight (in kg)
173	80
176	86
167	65
178	82

180	115
167	62
156	36
177	83
172	77
175	81

The mean value of Weight =

The mean value of Weight =

$$\frac{1}{10} \sum_{i=1}^{10} (Weight)_i = \frac{80 + 86 + 65 + 82 + 115 + 62 + 36 + 83 + 77 + 81}{10}$$

= 76.7 kg

Using the mean, we can now calculate the deviation of every weight from the mean.

Height (in cm)	Weight (in kg)	Deviation of Weights from the mean (Weight - Mean Weight)
173	80	3.3
176	86	9.3
167	65	-11.7
178	82	5.3
180	115	38.3
167	62	-14.7

156	36	-40.7
177	83	6.3
172	77	0.3
175	81	4.3

Mean absolute deviation for Weight =

$$\frac{1}{10} \sum_{i=1}^{10} |(Weight)_i - Mean\ Weight|$$

Calculate the MAD for SalePrice

In this task, we will calculate the Mean Absolute Deviation for the features of `SalePrice`

- Calculate the arithmetic mean of the variable `SalePrice` using `mean()` method and store it in a variable `mean`
- Calculate the absolute distances from mean for each observation of `SalePrice` and store it in variable `distance`
- Calculate the MAD using the above calculated values and store in `mad`

```
mean = df.SalePrice.mean()

distance = abs(df.SalePrice - mean)

mad = distance.sum()/len(distance)

print('Mean : ', mean)

print('Mean Absolute Deviation : ', mad)
```

Standard Deviation and Variance

One major disadvantage of MAD is that it violates the algebraic principle by ignoring the + and – signs while calculating the deviations of the different items from the central value of a dataset. To overcome this, we can calculate a measure by squaring the deviations from mean, this measure is called variance. Square root of this variance is known as standard deviation. Standard deviation can be mathematically calculated as follows

For quantitative data, we can use MAD, Standard Deviation or variance as measures of dispersion. The standard deviation is usually preferable.

$$\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2}$$

where μ is the mean of observations $x_1 \dots x_n$

Let's check the standard deviation for both the columns Height and Weight.

Height (x)	Weight (y)	Deviation of Height from the mean $(x-\bar{x})$	Deviation of Weight from the mean $(y-\bar{y})$
173	80	0.9	3.3
176	86	3.9	9.3
167	65	-5.1	-11.7
178	82	5.9	5.3
180	115	7.9	38.3
167	62	-5.1	-14.7
156	36	-16.1	-40.7
177	83	4.9	6.3
172	77	-0.1	0.3
175	81	2.9	4.3

Using the equation for standard deviation, we can calculate the standard deviation for Height as

$$\sigma_{Height} = \sqrt{\frac{1}{10} \sum_{i=1}^{10} (Height_i - Mean\ Height)^2}$$

=

$$\sqrt{\frac{0.9^2 + 3.9^2 + (-5.1)^2 + 5.9^2 + 7.9^2 + (-5.1)^2 + (-16.1)^2 + 4.9^2 + (-1.9)^2 + (-1.9)^2}{10}}$$

Applying the same for Weight, we can calculate as

$$\sigma_{Weight} = \sqrt{\frac{1}{10} \sum_{i=1}^{10} (Weight_i - Mean\ Weight)^2}$$

=

$$\sqrt{\frac{3.3^2 + 9.3^2 + (-11.7)^2 + 5.3^2 + 38.3^2 + (-14.7)^2 + (-40.7)^2 + 6.3^2 + (-1.7)^2 + (-1.7)^2}{10}}$$

However, the standard deviation (or variance) isn't appropriate when there are extreme scores and/or skewness in your data set. In this situation, the interquartile range (to be covered in the next topics) is usually preferable.

In this task, we will calculate the Standard Deviation for the feature of `SalePrice`

- Calculate the arithmetic mean of the variable `SalePrice` using `mean()` method and store it in a variable
- Calculate the squared distances from mean for each observation of `SalePrice`
- Calculate the Standard Deviation using the above calculated values and store the same in `sd`

```
mean = df.SalePrice.mean()

distance = (df.SalePrice - mean)**2

sd = (distance.sum()/len(distance))**(1/2)

print('Mean : ', mean)

print('Standard Deviation : ', sd)
```

Coefficient of Variation

The coefficient of variation is a relative measure of dispersion. It can be used while comparing

- two or more data sets expressed in different units
- data sets with the same unit, but different mean values

For example: Suppose we have a data set with variables as `height` and `weight` of people of a particular city. We want to know whether the `height` of the people has more variation or `weight` of the people? We will not be able to compare their standard deviations as `height` is measured in cms and `weight` is measured in kilograms. In this scenario, we need a measure with no units, and hence we can use CV to compare the variation of these two variables in the data set.

The coefficient of variation measures the standard deviation relative to the mean in percentage terms, mathematically it can be represented as

$$CV = \frac{\text{standard deviation}}{\text{mean}} \times 100$$

For our example, we have already calculated the standard deviation and mean for weights and heights.

So the coefficient of variation for Weights =

$$\frac{\text{Standard Deviation of Weights}}{\text{Mean Of Weights}} = \frac{19.131}{76.7} * 100 = 24.9\%$$

And for Heights =

$$\frac{\text{Standard Deviation of Heights}}{\text{Mean Of Heights}} = \frac{6.75}{172.1} * 100 = 3.9\%$$

Example:

LESSON PRACTICE

Calculate CV for GarageArea and LotArea and compare the same

In this task, we will calculate the coefficient of variation of features `GarageArea` and `LotArea` and compare which feature shows greater variation

Instructions

- Calculate mean and standard deviation for `GarageArea` (using `df.mean()` and `df.std()`) and store the same in variables `garage_mean` and `garage_std` respectively
- Calculate mean and standard deviation for `LotArea` (using `df.mean()` and `df.std()`) and store the same in variables `lot_mean` and `lot_std` respectively
- Calculate the CV for `GarageArea` and `LotArea`, save the same in variables `garage_cv` and `lot_cv`.
- Print the values of `garage_cv` and `lot_cv`

CODE DATA

RESET CODE Code Saved

```
1 # code starts here
2 garage_mean = df['GarageArea'].mean()
3 garage_std = df['GarageArea'].std()
4 print(garage_mean)
5 print(garage_std)
6
7 lot_mean = df['LotArea'].mean()
8 lot_std = df['LotArea'].std()
9 print(lot_mean)
10 print(lot_std)
11
12 garage_cv = (garage_std/garage_mean)*100
13 lot_cv = (lot_std/lot_mean)*100
14
15 print(garage_cv)
16 print(lot_cv)
17 # Code ends here
```

Previous Code

TRY IT SUBMIT

Inter - Quartile Range (IQR)

We have seen that range as a measure of dispersion is prone to error due to extreme values. A better estimate for measuring dispersion is hence Inter - Quartile Range or IQR.

As we have learned in the earlier chapter that quartiles divide the data into four equal parts, the IQR is calculated by taking the difference between Q3 and Q1. IQR hence takes into consideration only 50% of the data present at the center and less prone to variation due to extreme values.



Suppose we have a dataset and we want to calculate the IQR(). So in IQR as we can see that in the image we are going to arrange the data in ascending order(min to max). We take the Median of that. The Median splits the data into two regions - the lower 50% (Min to Median) and the upper 50% (Median to Max). We then take the median for the lower half and it is given as Q1. Q1 further divides the lower half into two regions. Similarly, we take the median of the upper half (Q3) which again divides the region into two parts. So Q1, Median and Q3 divide the entire data into 4 quartiles each containing 25% of the data. The IQR is then the middle 50% from Q1 to Q3 and can be given mathematically as

$$\text{IQR} = Q3 - Q1$$

Let's walk through an example and calculate the IQR for the column Weight (in kg)

Height (in cm)	Weight (in kg)
173	80
176	86
167	65
178	82
180	115

167	62
156	36
177	83
172	77
175	81

Arranging the Weight (in kg) column in ascending order, we get -

Weight (in kg)
36
62
65
77
80
81
82
83
86
115

Since there are even number of observations in Weight, the median

Since there are even number of observations in Weight, the median(Q_2) for the above column is the average of the 5th and 6th observations =



$$\frac{80 + 81}{2} = 80.5$$

(Q_1) for Weight is the median of the first five values = Median of 36,62,65,77,80 = 65 and upper half (from observations 6 to 10).

(Q_3) is the median of the last five values = Median of 81,82,83,86,115 = 83

(Q_3) is the median of the last five values = Median of 81,82,83,86,115 = 83

$$\therefore IQR = Q_3 - Q_1 \\ = 83 - 65 = 18$$

Example:

Inter - Quartile Range (IQR)

In this task, we will be calculating the interquartile range for the feature of

SalePrice

Instructions

- Calculate the value of Q_1 for SalePrice using quantile() method and store the value in the variable named q1
- Calculate the value of Q_3 for SalePrice using quantile() method and store the value in the variable named q3
- Calculate the IQR and store it in variable iqr

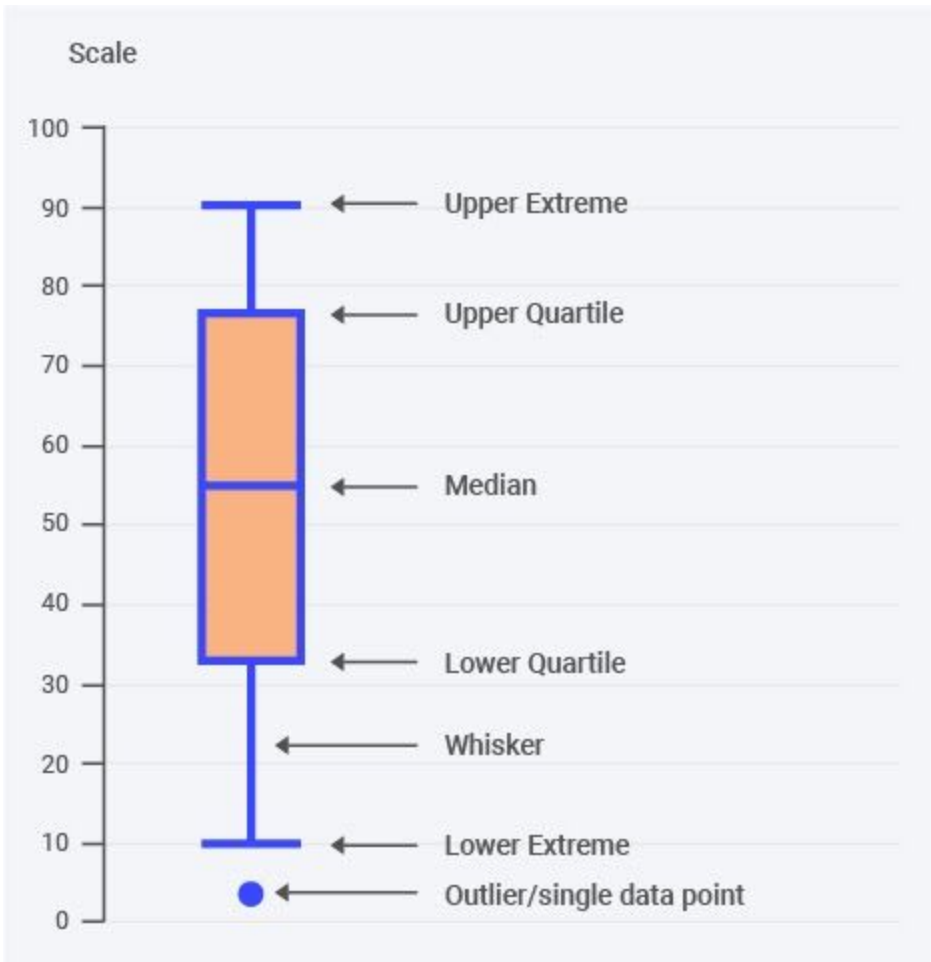


```
1 # Code starts here
2 q1 = df['SalePrice'].quantile(q=0.25)
3 q3 = df['SalePrice'].quantile(q=0.75)
4 iqr = q3 - q1
5 print(iqr)
6
7 # Code ends here
```

Box Plots

One of the intuitive ways to visualize the summary statistics of the data is by using a box plot.

Following is a diagram of the box plot explaining all its components. The lines extending parallel from the boxes are known as the “whiskers”, which are used to indicate variability outside the upper and lower quartiles. Outliers are sometimes plotted as individual dots that are in-line with whiskers. Box Plots can be drawn either vertically or horizontally.

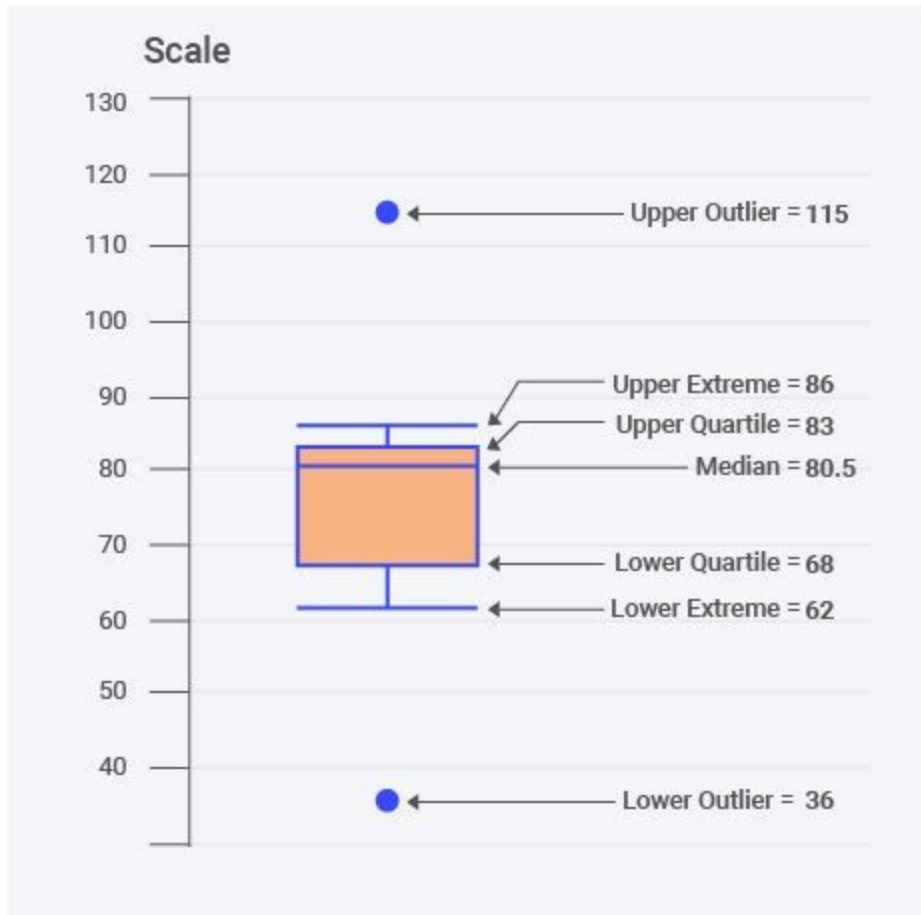


For our data of height and weight, suppose we want to plot a boxplot for the Weight column:

Height (in cm)	Weight (in kg)
173	80
176	86
167	65
178	82
180	115
167	62
156	36
177	83

172	77
175	81

We can use either seaborn or matplotlib libraries in python for plotting a boxplot. On plotting a boxplot using any of the two libraries, we will get an output as shown below for the Weight column :



box plot

In this task we will plot a Box Plot for the feature of `SalePrice`

Instructions

- The dataset has been stored in the DataFrame `df`
- Use `boxplot()` method from pandas to plot the boxplot for column `SalePrice`

Label the plot appropriately

Skills Covered:

Data Wrangling

Reference Solution

```
1 df.boxplot(column='SalePrice')
```

```
2 import seaborn as sns
3 sns.boxplot(df['SalePrice'])
4
5 # Code ends here
```

- We can see from the boxplot that there are many outliers on the higher side for `SalePrice`, this property of a data distribution is called as skewness which we will learn in the next chapter.

CONTINUE

> TRY IT

SU

Covariance

Covariance helps us quantify how much two variables vary with respect to one another and how two variables are related. A positive covariance indicates that they move in the same direction, on the other hand, negative correlation indicates that they move in opposite directions. Mathematically it can be represented as

$$Cov(x, y) = \sigma_{xy} = \frac{1}{n} \sum_{i=1}^n (x - \bar{x})(y - \bar{y})$$

How much does weight vary with height? Let's find out by calculating the covariance between height and weight

Height (x)	Weight (y)
173	80
176	86
167	65
178	82
180	115
167	62
156	36
177	83
172	77
175	81

The mean value for height is given as =

$$\frac{1}{10} \sum_{i=1}^{10} (Height)_i$$

=

$$\frac{173 + 176 + 167 + 178 + 180 + 167 + 156 + 177 + 172 + 175}{10} = 172.1cm$$

The mean value for weight is given as =

$$\frac{1}{10} \sum_{i=1}^{10} (Weight)_i = \frac{80 + 86 + 65 + 82 + 115 + 62 + 36 + 83 + 77 + 81}{10} = 76$$

The deviation of every value from the mean is as shown in the table below

Height (x)	Weight (y)	Deviation of Height from the mean $(x - \bar{x})$	Deviation of Weight from the mean $(y - \bar{y})$
173	80	0.9	3.3
176	86	3.9	9.3
167	65	-5.1	-11.7
178	82	5.9	5.3
180	115	7.9	38.3
167	62	-5.1	-14.7
156	36	-16.1	-40.7
177	83	4.9	6.3
172	77	-0.1	0.3
175	81	2.9	4.3

The deviation of every value from the mean is as shown in the table below

Using the formula for Covariance(x,y)

$$Cov(Height, Weight) = \sigma_{Height, Weight}$$

=

$$\leftarrow \frac{1}{10} \sum_{i=1}^{10} (Height_i - Mean\ Height)(Weight_i - Mean\ Weight)$$

=

$$\frac{3.3 * 0.9 + 9.3 * 3.9 + (-11.7) * (-5.1) + 5.3 * 5.9 + 38.3 * 7.9 + (-14.7) * (-1.7) + (-1.7) * (-1.7) + (-1.7) * (-1.7) + (-1.7) * (-1.7) + (-1.7) * (-1.7)}{10}$$

The mean value for height is given as =

$$\frac{1}{10} \sum_{i=1}^{10} (Height)_i$$

=

$$\frac{173 + 176 + 167 + 178 + 180 + 167 + 156 + 177 + 172 + 175}{10} = 172.1cm$$

The mean value for weight is given as =



$$\frac{1}{10} \sum_{i=1}^{10} (Weight)_i$$

=

$$\frac{80 + 86 + 65 + 82 + 115 + 62 + 36 + 83 + 77 + 81}{10} = 76.7kg$$

Height (x)	Weight (y)	Deviation of Height from the mean $(x - \bar{x})$	Deviation of Weight from the mean $(y - \bar{y})$
173	80	0.9	3.3
176	86	3.9	9.3
167	65	-5.1	-11.7
178	82	5.9	5.3
180	115	7.9	38.3

167	62	-5.1	-14.7
156	36	-16.1	-40.7
177	83	4.9	6.3
172	77	-0.1	0.3
175	81	2.9	4.3

Using the formula for Covariance(x,y)

$$Cov(Height, Weight) = \sigma_{Height, Weight}$$

= 

$$\frac{1}{10} \sum_{i=1}^{10} (Height_i - Mean Height)(Weight_i - Mean Weight)$$

=

$$\frac{3.3 * 0.9 + 9.3 * 3.9 + (-11.7) * (-5.1) + 5.3 * 5.9 + 38.3 * 7.9 + (-14.7) * (-1.7)}{10}$$

[Here is a link to a good video on understanding covariance.](#)

For example: Let us calculate the covariance between `SalePrice` and `LotArea` of the house to understand how these variables are related to each other. For simplicity, we consider the first 10 observations from our dataset.

	LotArea	SalePrice
0	8450	208500
1	9600	181500
2	11250	223500
3	9550	140000
4	14260	250000
5	14115	143000
6	10084	307000
7	10382	200000
8	6120	129900
9	7420	118000
10	11200	129500
11	11924	345000
12	12968	144000
13	10652	279500
14	10920	157000
15	6120	132000
16	11241	149000
17	10791	90000
18	13695	159000
19	7560	139000

Example:

Covariance

In this task, we will calculate the covariance between `LotArea` and `SalePrice` features

Instructions

- Create a new dataframe with columns `LotArea` and `SalePrice` with the first 20 observations and store it in variable `new`
- Calculate the mean for `LotArea` and `SalePrice`, make use of `mean()` method from pandas and store it in variable `mean_lotarea` and `mean_saleprice`
- Calculate the difference between each observation and mean for `LotArea` and `SalePrice`, store the difference series in two different variables named `diff_lotarea` and `diff_saleprice` respectively.
- Multiply `diff_lotarea` and `diff_saleprice` and calculate the sum, store the value in variable `summation`
- Calculate covariance by dividing `summation` by the number of observation pairs, store the same in variable `covariance` and print the same.

```

2  new = df[['LotArea', 'SalePrice']].iloc[:20,:].copy()
3
4  mean_lotarea = new['LotArea'].mean()
5  mean_saleprice = new['SalePrice'].mean()
6  diff_lotarea = (new['LotArea']-mean_lotarea)
7  diff_saleprice = (new['SalePrice']-mean_saleprice)
8  summation = (diff_lotarea*diff_saleprice).sum()
9  n = new.shape[0]
10
11 covariance = summation/n
12 print('Mean of LotArea : ', mean_lotarea)
13 print('Mean of SalePrice : ', mean_saleprice)
14 print('Covariance : ',covariance)
15 # Code ends here

```

RESULT

```

Mean of LotArea : 10415.1
Mean of SalePrice : 181270.0
Covariance : 41420373.0

```


What is half of the difference between the first and third quartiles?

Ans: The semi-interquartile range is a measure of spread or dispersion.

It is computed as one half the difference between the Q3 and the Q1. The formula for semi-interquartile range is therefore: $(Q3-Q1)/2$.

Skewness

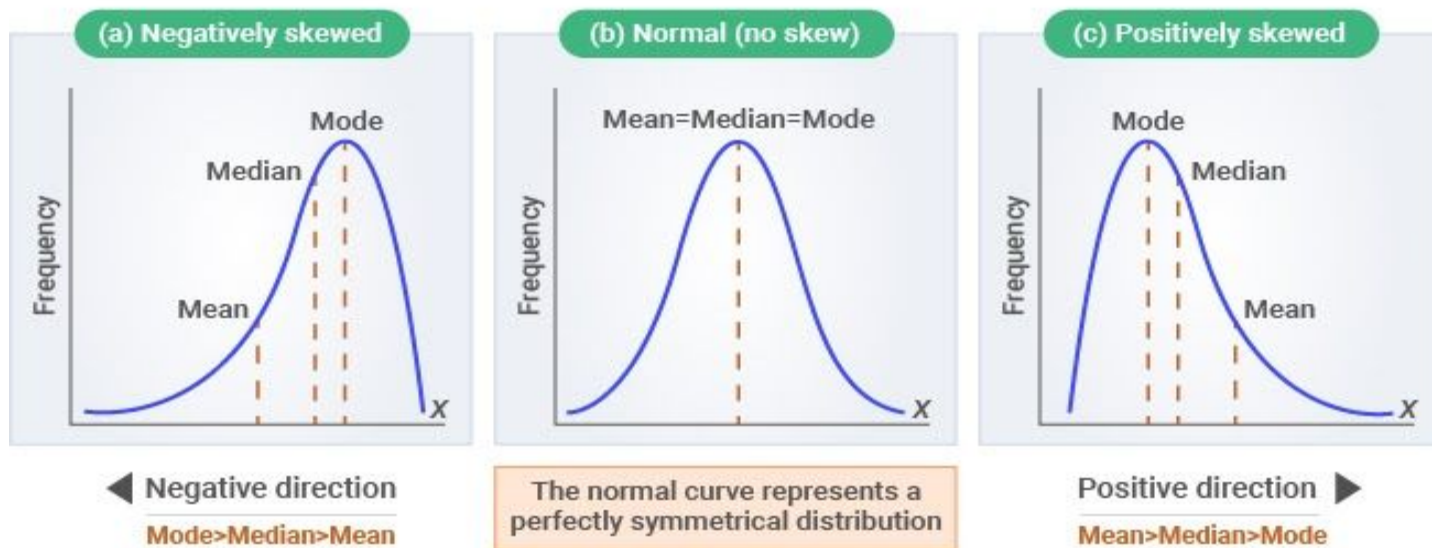
In the previous chapter, we observed that there are many outliers for the variable `SalePrice`. The measures of central tendency and dispersion fail to give us an intuition on how the data is distributed around the mean. Two data sets can have equal mean and standard deviation but very different frequency distributions.

A frequency distribution that is not symmetrical is called skewed. In a skewed data set extreme value (outliers) move to one side (tail) of the frequency distribution.

- If the extreme values move towards the right end of the distribution then it is called as a right-skewed or a positively skewed distribution.
- If the extreme values move towards the left end of the distribution then it is called a left-skewed or a negatively skewed distribution.

The relative position of mean, median and mode for different data distributions is shown in the following chart.

Position of Mean Median Mode



For a right skewed data,

$$\text{Mean} > \text{Median} > \text{Mode}$$

$$\text{Mean} > \text{Median} > \text{Mode}$$

For a left skewed data,

$$\text{Mean} < \text{Median} < \text{Mode}$$

$$\textit{Mean} < \textit{Median} < \textit{Mode}$$

For a perfect normal distribution,

$$\text{Mean} = \text{Median} = \text{Mode}$$

$$\textit{Mean} = \textit{Median} = \textit{Mode}$$

Real-life examples of skewed distributions include :

- Negative/Left Skewness - The distribution of retirement ages of individuals follows a left-skewed distribution since a majority of the individuals will tend to retire at the age of 58-60 and very few individuals will retire at an early age
- Positive/Right Skewness - The distribution of salaries of employees at an organization will be right-skewed since the majority of employees will have lesser salaries compared to a few employees at senior positions who will have higher salaries.

Kurtosis

The other measure that describes the shape of a distribution is called kurtosis.

Traditionally, kurtosis was defined as a measure of distribution's flatness or peakedness. However, Peter Westfall (2014) has been on a bit of a crusade to change this perception. According to Westfall, it's the tails that mostly account for kurtosis, not the central peak.

Dr. Donald Wheeler defines kurtosis as

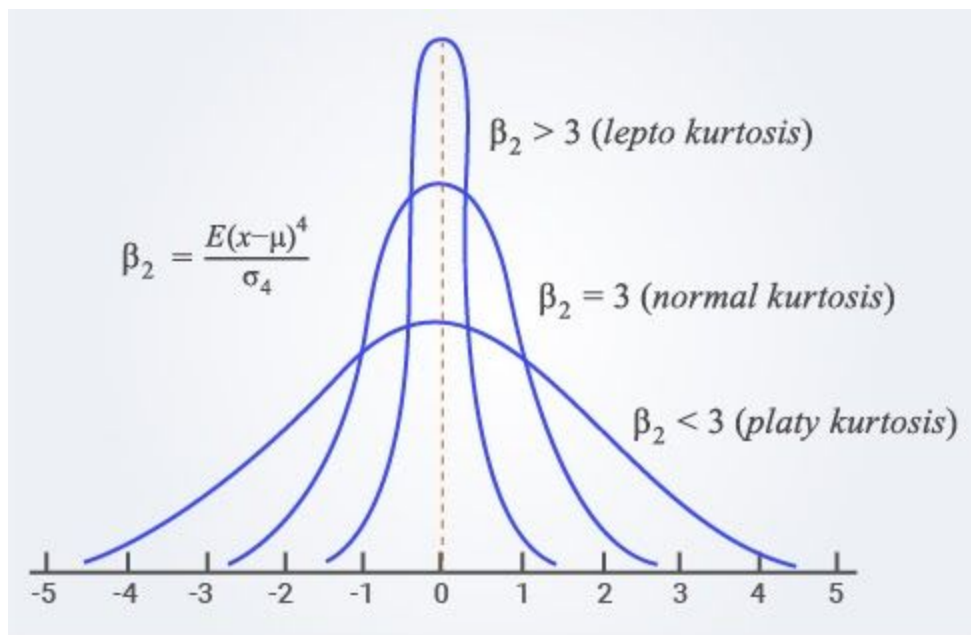
“The kurtosis parameter is a measure of the combined weight of the tails relative to the rest of the distribution.”

Although the weight of tails and peakedness are correlated, kurtosis cannot be used as a measure of peakedness.

The kurtosis of a normal distribution is 3, most of the statistical packages report excess kurtosis over the normal distribution.

- A normal distribution has kurtosis exactly 3 (excess kurtosis exactly 0). Any distribution with kurtosis ≈ 3 (excess ≈ 0) is called mesokurtic.
- A distribution with kurtosis < 3 (excess kurtosis < 0) is called platykurtic. Compared to a normal distribution, its tails are shorter and thinner, and often its central peak is lower and broader.
- A distribution with kurtosis > 3 (excess kurtosis > 0) is called leptokurtic. Compared to a normal distribution, its tails are longer and fatter, and often its central peak is higher and sharper.

Two or more distributions may have an identical mean, variation, and skewness but they may show different degrees of kurtosis. The following chart describes the different values of kurtosis and their respective names.



7.1.1 Skewness and Kurtosis

LESSON PRACTICE

Check whether GarageArea is skewed

Plot a histogram for GarageArea and identify whether the data is skewed.

Instructions

- Use the `hist()` method to plot a histogram for `GarageArea`, keep `bins=20` as parameter

Skills Covered:

Probability and Statistics

Reference Solution

```
1 plt.figure(figsize=(10, 6))
2 plt.hist(df.GarageArea, bins=20)
```

CODE DATA

```
1 import matplotlib.pyplot as plt
2 plt.hist(df['GarageArea'], bins=20)
3
```

We can see from the histogram that there are extreme values present on the right end of the histogram, hence we can say that the data is positively skewed

CONTINUE

> TRY IT SUBMIT

(array([43, 125, 135, 126, 208, 239, 184, 86, 68, 43, 49,

Defining Correlation

Most of the time while analyzing data, it is important to know the relationship between the variables.

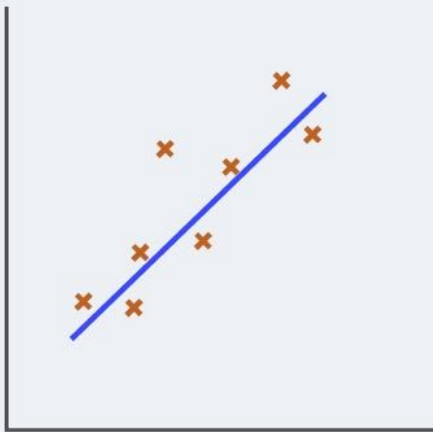
For example: in our data set if we can find a relationship between `SalePrice` and `LotArea`, we can make a rough estimate of the `SalePrice` based on `LotArea`.

Correlation is used to analyze the strength and direction of the relationship between two quantitative variables. We can answer the following questions based on the correlation

- Is there an association between two or more variables? If yes, what is the form and degree of this relationship?
- Is the relationship strong or significant enough to be useful to arrive at a desirable conclusion?

Correlation between two variables is said to be **positive** when their values change in the same direction and **negative** when their values change in opposite directions. The value of a correlation coefficient lies between -1 to 1, -1 being perfectly negatively correlated and 1 being perfectly positively correlated.

Positive Correlation



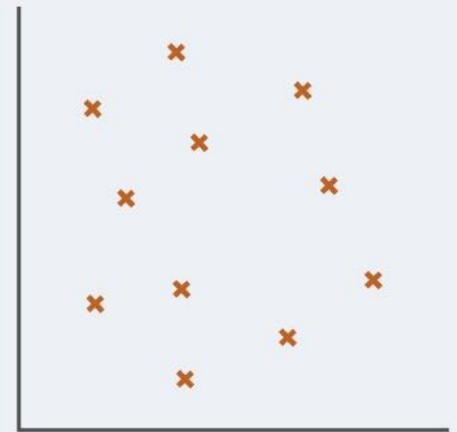
The points lie close to a straight line, which has a positive gradient.
This shows that as one variable **increases** the other **increases**.

Negative Correlation



The points lie close to a straight line, which has a negative gradient.
This shows that as one variable **increases** the other **decreases**.

No Correlation



There is no pattern to the points.
This shows that there is no connection between the two variables.

However, correlation does not imply causation. There may be, for example, an unknown factor that influences both variables similarly.

Causation indicates that one event is the result of the occurrence of the other event; i.e. there is a causal relationship between the two events. This is also referred to as cause and effect.

A statistically significant correlation has been reported, for example, between yellow cars and a lower incidence of accidents. That does not indicate that yellow cars are safer, but just that fewer yellow cars are involved in accidents. A third factor, such as the personality type of the purchaser of yellow cars, is more likely to be responsible than the color of the paint itself.

Karl Pearson's Correlation Coefficient

Karl Pearson's correlation coefficient measures the correlation between quantitative variables. Mathematically, Pearson's correlation coefficient can be represented as

$$r = \frac{\text{covariance}(x, y)}{\sigma_x \sigma_y}$$

where

$$\text{covariance}(x, y) = \frac{1}{n} \sum (x - \bar{x})(y - \bar{y})$$

$$\text{standard deviation of } x, \sigma_x = \sqrt{\frac{\sum (x - \bar{x})^2}{n}}$$

$$\text{standard deviation of } y, \sigma_y = \sqrt{\frac{\sum (y - \bar{y})^2}{n}}$$

Therefore

$$r = \frac{\frac{1}{n} \sum (x - \bar{x})(y - \bar{y})}{\sqrt{\frac{\sum (x - \bar{x})^2}{n}} \sqrt{\frac{\sum (y - \bar{y})^2}{n}}} = \frac{n \sum xy - \sum x \sum y}{\sqrt{n \sum x^2 - (\sum x)^2} \sqrt{n \sum y^2 - (\sum y)^2}}$$

Let's brave ourselves and calculate the Pearson's Correlation Co-efficient for our example of Heights and Weights

Height (in cm)	Weight (in kg)
173	80
176	86
167	65
178	82
180	115
167	62
156	36

177	83
172	77
175	81

Since we already know the values for

Feature	Value
Covariance(Height, Weight)	120.63
Standard Deviation for Weight	19.131
Standard Deviation for Height	6.75

The Pearson Correlation Coefficient between Height and Weight can be :

$$r = \frac{\text{covariance}(\text{Height}, \text{Weight})}{\sigma_{\text{Height}} \sigma_{\text{Weight}}} = \frac{120.63}{6.75 * 19.131} = 0.93$$



- Pearson's correlation coefficient is used only when two variables are linearly related
- The value of the coefficient is affected by the extreme values in the dataset
- Computation is complex and takes more computation time
- Pearson's correlation should be used only if data is normally distributed, extreme values in data affects the value of the coefficient

Spearman' Rank Correlation coefficient

Spearman's correlation coefficient can be used when the data is skewed on ordinal in nature and is robust when extreme values are present.


As the name suggests we need to rank the observations before calculating this correlation coefficient. The ranks are assigned by taking either the highest or the lowest value as rank one and so on for the values of both the variables. Further, we need to calculate the difference between the ranks of these observations.

Mathematically Spearman's Rank Correlation can be represented as

$$R = 1 - \frac{6 \sum d^2}{n(n^2 - 1)}$$

where **d** is the difference between the rank of the observations.

In the following table, we have taken 10 observations of LotArea and SalePrice from our dataset and calculated the Spearman's rank correlation

-  We assign rank 1 to the lowest value in LotArea and accordingly rank the remaining observations. We repeat the same process for SalePrice
- We calculate the difference between the ranks of each pair of observations in column d and then calculate its square
 - Now we can calculate the Spearman's rank correlation as

$$R = 1 - \frac{6 \times 62}{10(10^2 - 1)} = 0.63$$

	Rank LotArea	Rank SalePrice	LotArea	SalePrice	d	d^2
0	3.0	7.0	8450	208500	-4.0	16.0
1	5.0	5.0	9600	181500	0.0	0.0
2	8.0	8.0	11250	223500	0.0	0.0
3	4.0	3.0	9550	140000	1.0	1.0
4	10.0	9.0	14260	250000	1.0	1.0
5	9.0	4.0	14115	143000	5.0	25.0
6	6.0	10.0	10084	307000	-4.0	16.0
7	7.0	6.0	10382	200000	1.0	1.0
8	1.0	2.0	6120	129900	-1.0	1.0
9	2.0	1.0	7420	118000	1.0	1.0

1. You have given new dataframe `new_df` from an existing data frame with columns `LotArea` and `SalePrice`. (This will make the following tasks much simpler)

Pearson's Correlation Coefficient

2. Calculate the covariance between two variables by using `cov()` method from pandas, be careful to take an appropriate value from the dataframe returned by `cov()` method. Save the value into a variable named `covariance`
3. Calculate the standard deviation of `LotArea` and `SalePrice` using `std()` method from pandas and store the values in variables `std_LotArea` and `std_SalePrice` respectively
4. Calculate the Pearson's correlation coefficient, save the value in variable `pearson` and print the same.

Spearman Rank Correlation Coefficient

5. Calculate the rank for all the observations using `rank()` method from pandas. Note that the `axis` parameter must be set to 0.
6. Create a column `d^2` and store the squared value of the difference between the ranks calculated in the previous step
7. Calculate the sum of column `d^2` and store the same in `d_square`
8. Calculate the Spearman rank correlation, store the value in variable `spearman` and print the same.

```

1 newdf = df[['LotArea','SalePrice']].copy()
2 # Calculating Pearson correlation coefficient
3 covariance = newdf.cov().iloc[0,1]
4 print(covariance)
5 std_LotArea = newdf.LotArea.std()
6 print(std_LotArea)
7 std_SalePrice = newdf.SalePrice.std()
8 print(std_SalePrice)
9 pearson = covariance/(std_LotArea*std_SalePrice)
10 print("Pearson's Correlation Coefficient : ", pearson)
11
12 # Calculating Spearman rank correlation coefficient
13 ranks = newdf.rank(axis=0)
14 print(ranks)
15 ranks['d^2'] = (ranks['LotArea'] - ranks['SalePrice'])**2
16 d_square = ranks['d^2'].sum()
17 print(d_square)
18 n = ranks.shape[0]
19 spearman = 1 - ((6*d_square)/(n**3 - n))
20 print("Spearman's Rank Correlation Coefficient : ", spearman)

```

OUTPUT

RESULT

```

204159593.81452188
10214.70213252844
79023.8905997513
Pearson's Correlation Coefficient : 0.2529214590904538

```

Correlation

Correlation is a measure of the linear relationship between two variables. It is defined for a sample as the following and takes value between +1 and -1 inclusive:

$$r = \frac{s_{xy}}{\sqrt{s_{xx}}\sqrt{s_{yy}}}$$

s_{xy} , s_{xx} , s_{yy} are defined as:

$$s_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

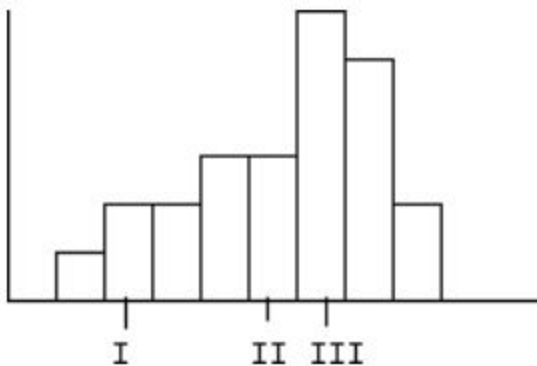
$$s_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$$

$$s_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2$$

It can also be understood as the cosine of the angle formed by the ordinary least square line determined in both variable dimensions. Explore this concept through

Questions:

1. For the following histogram, what is the proper ordering of the mean, median, and mode?



Note: The graph is NOT numerically precise -only the relative positions are important!!!

Ans: I = mean II = median III = mode

Explanation:

This is a left skewed distribution therefore mean would be less than median which make I as mean , II as median and III is the most frequent entry(bar is the highest) which results it in to be qualified as mode.

- 2.The primary outcome of the hypertension study is blood pressure. Is blood pressure a qualitative or quantitative variable ?

Ans: Quantitative

Explanation:

For the quantitative variables, the numbers are the actual numerical measure of the variable in question. In simple words, you can quantify blood pressure using numbers(A blood pressure of 140/90 mmHg or higher indicates high blood pressure).

- 3.A curve with a high kurtosis value has a higher peak.

Ans: True

Explanation:

Kurtosis is a measure of the “peakedness” of the probability distribution. An increased kurtosis (>3) can be visualized as a thin “bell” with a high peak whereas a decreased kurtosis corresponds to a broadening of the peak and “thickening” of the tails.

- 4.Suppose a database on companies contain the following four variables.

- Industry name
- Number of employees
- Number of job vacancies

You were asked to compute the mean value of these variables, where possible. What is the number of variables for which the mean can be computed meaningfully?

Ans: 2

Explanation:

Number of employees and Number of job vacancies are numeric variables so its mean can be calculated. Mean can't be calculated with industry names (It's not numeric)

5. A researcher wishes to calculate the average height of patients suffering from a particular disease. From patient records, the mean was computed as 156 cm, and standard deviation as 5 cm. Further investigation reveals that the scale was misaligned, and that all readings are 2 cm too large, e.g., a patient whose height is really 180 cm was measured as 182 cm. Furthermore, the researcher would like to work with statistics based on metres. The correct mean and standard deviation are:

Ans: 1.54m, .05m

Explanation:

Just as every measurement is reduced by 2 cm, the mean (average height) will also reduce by 2 cm. Adding a constant to each value in a data set does not change the distance between values so the standard deviation remains the same.

6. What is the variance for the following discrete data [2, 6, 8, 3, 7, 9, 1, 4] ?

Ans: 7.5

Explanation:

- Your first step is to find the Mean: $40/8 = 5$
- To calculate the Variance, take difference of each value with mean, square it, and then average the result: $60/8 = 7.5$

7. Considering the various types of tables and charts, which table, chart, diagram or plot would you use to depict categorical data for two variables in a visual format?

Ans: side-by-side bar chart

Explanation:

A histogram plot is generally used for frequency plotting of a single variable. Data concerning two categorical variables may be communicated (not visualised) using a two-way table, also known as a contingency table. A stem-and-leaf display or stem-and-leaf plot is a device for presenting quantitative data in a graphical format. Data concerning two categorical variables can be visualized using a segmented bar chart or a clustered bar chart. A clustered bar chart is also known as a side-by-side bar chart.

8. If arithmetic mean is multiplied to coefficient of variation then resulting value is classified as

Ans: standard deviation

Explanation:

Coefficient of Variation(CV)= standard derivation(std)/arithmetic mean(am).If we are asked to multiply CV with am, we get std.

9.The mean and standard deviation of 20 observations are found to be 10 and 2 respectively. On rechecking it was found that an observation 8 was incorrect. Calculate the correct mean if the wrong item is omitted.

Ans: 10.10

Explanation:

- Following is how we get the new mean:
 - Mean = summation of data points/total number of data points
 - Therefore we get, summation of data points as 200.
 - Subtracting the incorrect observation 8 from 200, we get 192 and using the formula of mean we get the new mean, that is $192/19$ as 10.1.

10. What are blood pressure values an example of ?

Ans: Continuous

Explanation:

Blood pressure can be measured to as many decimal places as the measuring instrument allows and therefore are continuous