SESSION 8: DATA CLEANING IN EXCEL-2 (HANDLING OUTLIERS)

SESSION OVERVIEW:

In this session, the students will be able to:

- Understand the concepts related to the variability of data
- Understand what are outliers and why we need to handle them
- Understand how to find and handle outliers in the dataset.
- Understand quartile function and their working.
- Understand briefly about feature engineering.

KEY TOPICS AND EXAMPLE:

1. Data variability/ dispersibility methods:

Before getting into data imputation methods, we need to understand the dispersibility or the variability of the data. Understanding variability of the data will help us understand the data and help us understand which data imputation method should be used. Thus, we have to understand the concepts of variance and standard variance.

a. Variance: (10 mins)

In statistics, variance is a measure of how spread out a dataset is. More specifically, it measures the average squared difference between each data point and the mean of the dataset. Variance is represented by the symbol σ^2 for a population and s^2 for a sample. Variance is commonly used in statistics to describe the variability or spread of a dataset. It is a useful tool for comparing the spread of two or more datasets, as well as for identifying outliers or extreme values in a dataset. The formula for variance depends on whether you are calculating the variance of a population or a sample.

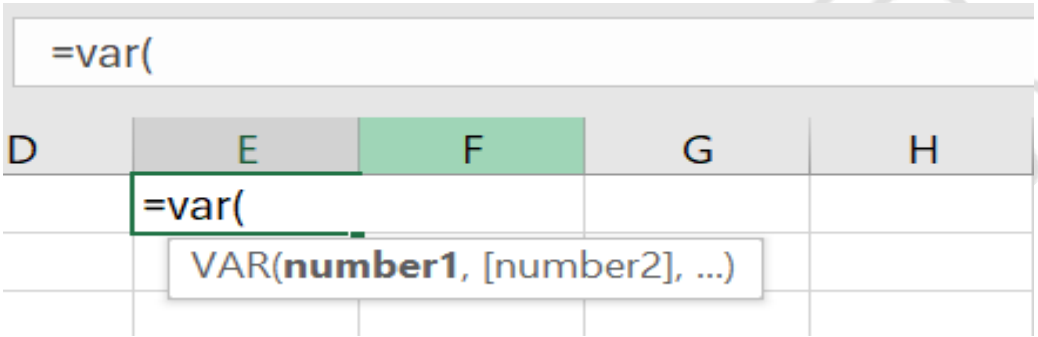


Figure: Represents the VARIANCE function in excel

This dataset will be used to explain this topic.

Inferences:

- If a dataset has high variance, it suggests that the data points are spread out over a wide range. Imputing missing values with mean or median imputation methods might not always be appropriate, as these measures might not capture the diversity of the data effectively. In such cases, more sophisticated imputation methods might be considered.

- For datasets with low variance, where data points are closely clustered around the mean, simpler imputation methods like mean or median imputation might suffice without significantly altering the dataset's overall characteristics.

b. Standard deviation: (10 mins)

Standard deviation is a statistical measure that is used to quantify the amount of variability or dispersion in a set of data. It is defined as the square root of the variance and is typically denoted by the symbol σ (sigma).

The standard deviation tells us how spread out the data is from the mean, or average value. A low standard deviation means that the data points tend to be close to the mean, while a high standard deviation means that the data points are spread out over a wider range.

To calculate the standard deviation, first find the mean of the data set. Then, for each data point, subtract the mean and square the result. Next, find the average of these squared differences, which is the variance. Finally, take the square root of the variance to get the standard deviation. The formula for the standard deviation is:

The higher the standard deviation the more variability or spread you have in your data. The larger your standard deviation, the more spread or variation in your data. Small standard deviations mean that most of your data is clustered around the mean.

This dataset will be used to explain this topic.

Example to help interpret:
Community A has ages: [20, 21, 22, 78, 80]
Community B has ages: [45, 46, 47, 53, 54]

Ask students to calculate VAR and STDEV for both these.

For Community A, the mean age might be around 44 years, but because two individuals are significantly older than the rest (78 and 80 years old), the variance will be high. The large gap between these ages and the mean indicates a high spread in the data, signifying high variance.

For Community B, the ages are more clustered around the mean (approximately 49 years), leading to a lower variance. The ages are relatively close to each other and to the mean, indicating the data points are more uniform, and thus, the variance is low.

Using these examples, both variance and standard deviation can be explained. Maybe 2 basic age datasets, one with middle aged population and other with young and old population.

Problem: (5 min)

Imagine you are a meteorologist studying daily temperature patterns in a city over a month. You have collected temperature data for each day, recording the high temperatures in degrees Celsius. Your goal is to analyze the variability in daily temperatures to understand the climate patterns and fluctuations.(**Dataset**)

Conclusions:

(Comment for instructor: The conclusion should be conveyed very clearly to the students how the mean, variance and standard deviation helps us to understand the spreadability or variability of the data.)

Mean (Average): The mean value of 16 suggests that, on average, the data points are centered around this value.

Standard Deviation: The standard deviation of around 3 indicates the average deviation of data points from the mean. In this case, a standard deviation of 3 suggests that the data points are relatively close to the mean, with most values falling within approximately 3 units of the mean.

Variance: The variance is a measure of the dispersion of data points around the mean. In this case, a variance of around 12 suggests that the data points are moderately dispersed from the mean. Since variance is the square of the standard deviation, a variance of 12 corresponds to a standard deviation of around 3.

How to impute missing values? (10 min)

1. If the Standard deviation is similar/ near to Average or bigger value, then we replace the missing value with Median
2. If the Standard deviation is less than the average value or has a small value that means values are clustered near to Average, then we replace the missing value with the Average

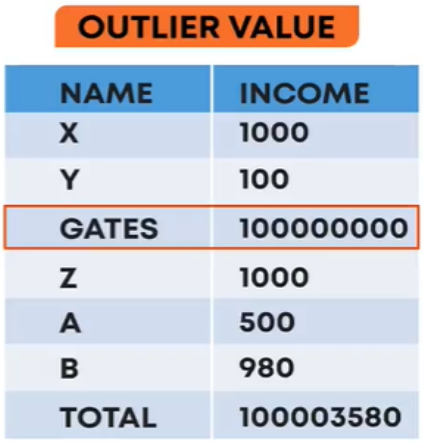
The ABC company **Dataset** is used to explain all the above functions and to understand how a business problem is solved.

2. Understand what are outliers and why we need to handle them:

a. What are outliers? (15 mins)

- An Outlier is an observation or data point significantly different from other observations or data points in a dataset.

- Various factors can lead to the occurrence of outliers, such as natural variations in the data, measurement errors, or, data entry errors.



b. Missing values vs outlier?

- Missing data and outliers are both issues that can affect the quality of a dataset and the accuracy of the analysis performed on it. However, they are different problems that require distinct approaches to handle them.

- Missing data refers to the absence of values in a dataset, which can occur for a variety of reasons such as data entry errors or incomplete surveys. Missing data can cause problems such as reduced sample size, biased results, and inaccurate analysis.
  To handle missing data, different techniques such as **imputation, exclusion** can be used to fill in the missing values.

- On the other hand, Outliers are extreme values that deviate significantly from the rest of the data in a dataset. Outliers can be caused by measurement errors, natural variation, or rare events and can skew the results of the analysis by affecting the mean, standard deviation, and other statistical measures.
  To handle outliers, techniques such as the **Quartile Function** can be used to identify and **remove** them from the dataset or **account for them in the analysis.**

c. Why should the outlier problem be handled before analysis?

- **Prevent misleading interpretation:** Outliers can significantly skew summary statistics such as the mean and standard deviation, leading to inaccurate representations of the central tendency and variability of the data. Removing or adjusting outliers can provide a more accurate summary of the dataset.

- **Data Quality**: Outliers may indicate errors in data collection or measurement, data entry mistakes, or rare but meaningful observations. Addressing outliers improves the overall quality and integrity of the dataset, enhancing the reliability of subsequent analyses and interpretations.

- **Discover Interesting Insights:** Not all outliers are errors; some may represent valuable or interesting variations within the dataset. Analyzing outliers can lead to insights about unusual cases or phenomena. For example, in medical data, outliers could indicate patients who responded unusually well or poorly to a treatment, meriting further investigation.

d. Impact of removing the outliers from the data.

The impact of removal of outliers from the dataset:

i. **Loss of Information:** Outliers may contain valuable information or insights about rare events, anomalies, or important observations in the data. Removing outliers without careful consideration can result in the loss of valuable information that could be relevant to the analysis objectives.

ii. **Biased Results:** If outliers are removed indiscriminately or without justification, it may introduce bias into the analysis results.

iii. **Reduced Generalizability:** Removing outliers may result in a dataset that is less representative of the underlying population, especially if the outliers are genuine observations rather than errors or anomalies.

3. Understand quartile function and its working: (10 mins)

a. What is quartile function?

The Quartile Function calculates the quartiles of a dataset, which divides the data into four equal groups. It splits the data range into four sections of equal size. This function can determine the minimum, maximum, first, second, and third quartiles.

b. Syntax for quartile in Excel:

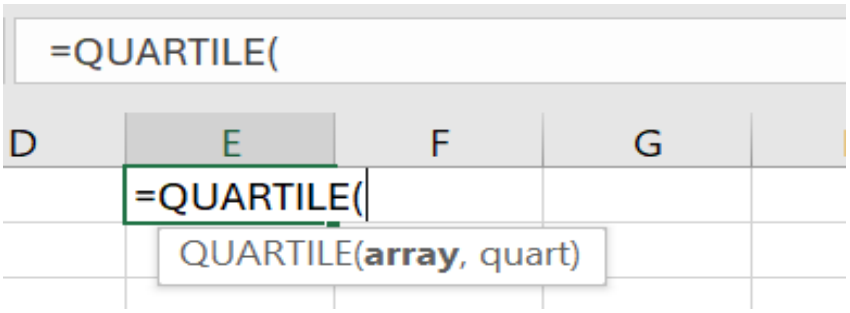


Figure: Represents the QUARTILE function in Excel

The QUARTILE function syntax has the following arguments:

- Array Required. The array or cell range of numeric values for which you want the quartile value.
- Quart Required. Indicates which value to return.

c. Different types of quartile functions:

The four quartiles that divide a data set into quartiles are:

1. The lowest 25% of numbers.
2. The next lowest is 25% of the numbers (up to the median).
3. The second highest 25% of numbers (above the median).
4. The highest 25% of numbers.

If quart equals	QUARTILE returns
0	Minimum value
1	First quartile (25th percentile)
2	Median value (50th percentile)
3	Third quartile (75th percentile)
4	Maximum value

Example:

1. This **Dataset** is used to demonstrate the example for QUARTILE function in Excel.

	A	B	C	D	E
1	SALES REVENUE		QUART VALUE	RESULT	EXPLANATION
2	45000		0	11000	The QUART "0" value will return minimum value from the sales revenue data.
3	23000		1	23500	For value 1, the function will return the first quartile or 25th percentile value from the dataset.
4	32000		2	38500	For value 2, the function will return the second quartile or 50th percentile value from the dataset.
5	11000		3	55250	For value 3, the function will return the second quartile or 75th percentile value from the dataset.
6	67000		4	89000	The QUART "4" value will return maximum value from the sales revenue data.
7	89000				
8	50000				
9	17000				
10	25000				
11	57000				

2. This **dataset** can be explained for better understanding.

Conclusion of the dataset:

This information means out of all the test results data 25% of the data falls between 58 and 71.75.
Another 25% of the data falls between 71.75 and 83.5 and so on.

*(**Additional information:***

- *If array is empty, QUARTILE returns the #NUM! error value.*
- *If quart is not an integer, it is truncated.*
- *If quart < 0 or if quart > 4, QUARTILE returns the #NUM! error value.*
- *MIN, MEDIAN, and MAX return the same value as QUARTILE when quart is equal to 0 (zero), 2, and 4, respectively)*

4. Understand how to find and handle outliers in Excel:(30 mins)

a. Steps to find outliers in dataset:

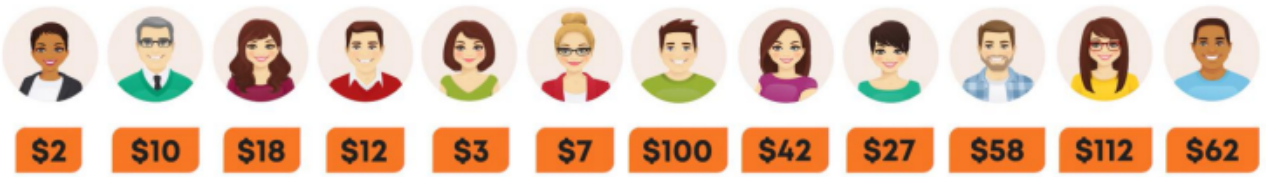
In order to identify outliers, it is necessary to follow these steps:

- Step 1: Calculate all the quartile values (Q1, Q2, Q3).
  For example, to find the quartiles, with a given dataset, {1, 2, 3, 4, 5, 6, 7, 8, 9, 10}. The median (Q2) is the middle value, which is 5. Q1 is the median of the lower half of the data, which is (2, 3, 4, 5), Q1 = 3. Q3 is the median of the upper half of the data, which is (6, 7, 8, 9), Q3 = 7.
- Step 2: Find Interquartile Range (IQR) for the required column.
  Formula: IQR = Q3 – Q1
- Step 3: In the new column apply this condition:
  =IF(OR(Cell_valueQ3+1.5IQR),1,0)
- Step 4: In the new column, apply the filter or use =COUNTIF()
- Step 5: Calculate the number of outliers present.

b. Examples:

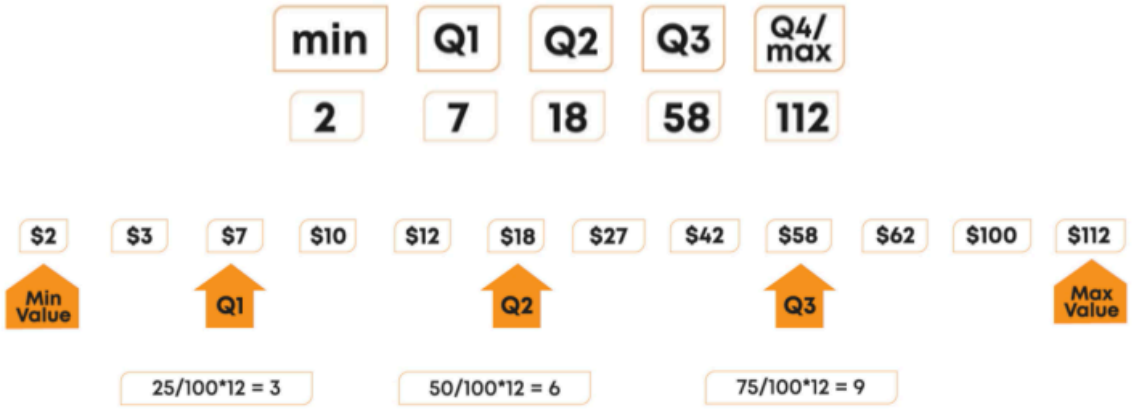
Case-1

i. Given Dataset: The dataset contains information on 12 customers and the corresponding amount they spent on purchasing fruits.



ii. First, sort the given data/column in increasing order.

iii. Calculate all the quartile values. Here Q1 = 7, Q2 = 18, Q3 = 58



iv. Find the Interquartile Range (IQR). Here IQR = 51

v. In the new column apply this condition: =IF(OR(Cell_valueQ3+1.5IQR),1,0)
This condition means that, Value < Q1 – 1.5 * IQR or Value > Q3 + 1.5 *IQR are Outliers.
In the given dataset, data points that fall below -69.5 or above 134.5 will be classified as outliers. You can conclude that there are no outliers present in this given dataset.

vi. If we have to calculate the number of outliers in the dataset, then we have to use the COUNTIF function in the dataset.

Case-2

This dataset can be used to explain the steps included to find outliers in the dataset.

Case-3

Example: ABC Food company (**Dataset**)

c. **Box Plot:**

You can use box plots to find outliers in a dataset. To insert a box plot in MS Excel:

I. In Excel, go to the "Insert" tab on the ribbon.
II. Click on "Insert" > "Waterfall or Stock" (found under the "Charts" group). Select the "Box and Whisker" chart option.
III. Excel will automatically generate a box plot based on the selected data range.

IV. You can further customize the appearance of the box plot by right-clicking on various elements (e.g., axes, data points) and accessing the formatting options.

Reading a box plot involves interpreting the various components of the plot to understand the distribution of the data.
Identify the median: The median is represented by a vertical line inside the box. It indicates the middle value of the data, separating the bottom 50% from the top 50%.

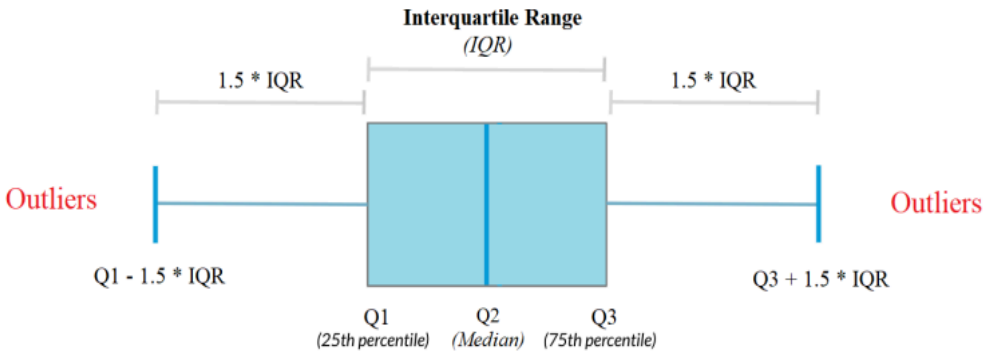


Figure: Identify Outlier in Data

d. <u>How to handle the outliers:</u>

The best way to address outliers is to **reach out to the stakeholders** to have suggestions. Also, we can perform two analyses simultaneously, one with the outliers and another without the outliers and whichever analysis seems more accurate to the stakeholders that can be used by them.