SESSION 7: DATA CLEANING IN EXCEL-1 (HANDLING MISSING VALUES)

SESSION OVERVIEW:

By the end of this session, students will be able to:

- Understand the importance of checking data types and perform the data type check function in MS Excel.
- Understand the concept of missing values and different methods to handle missing values.
- Perform all the functions related to handling of the missing values in MS Excel.

KEY TOPICS AND EXAMPLES:

1. **Checking data types: (15 min)**
    a. **Importance of checking data type:**

- **Data Accuracy**: Ensuring that each column has the correct data type helps maintain the accuracy of the data. For example, if a column intended for numeric values contains text, it can lead to errors in calculations or analyses.

  Example: If you're trying to sum a column of numbers but one of the numbers is entered as text (e.g., '5' instead of 5), Excel will ignore the text entry, leading to an incorrect total.

- **Analysis and Reporting**: Different data types require different treatment and analysis methods. For example, numeric data can be used in calculations and aggregations, while text data may require string manipulation or categorization. By identifying the data types upfront, analysts can better plan their analyses and reporting processes.

  Example: If you're creating a line chart to show trends over time but your dates are formatted as text, Excel may not plot the points in the correct order, or it might not recognize the text as dates at all, leading to a misleading or blank chart.

- **Compatibility with Functions and Formulas:** Excel functions and formulas often have specific requirements regarding data types. For instance, certain functions may only accept numeric values or dates. Ensuring that the data types match the requirements of the functions used in analyses or calculations is crucial for obtaining accurate results.

  Example: If dates are formatted as text, Excel will sort them alphabetically rather than chronologically, mixing up the order of events or data points crucial for time-series analysis.

(Most of the time Microsoft Excel does the automatic selection of the data types used in each column in the dataset, but it is always important to check the data type before-hand to have a clear understanding of the data and not get involved in some critical errors related to it.)

b. Function and syntax:

The dataset used for this topic.

For checking data type of any column in MS Excel we use the TYPE function. The TYPE function returns a numeric value, representing the datatype of the data. The numbers and the corresponding data types are as follows:

NUMBERS	DATA TYPES
1	Numeric/ Integer/ Date/ Time/ Empty cell
2	Text/ string
4	Boolean/ Logical
16	Error
64	Array

Steps to change the data type of the column:

- Select the column or range of cells containing the data you want to change the data type for.
- Right-click on the selected column or cells.
- In the context menu that appears, choose "Format Cells." (Shortcut to open the dialogue box CTRL+1)
- From the Category list on the left side, choose the desired data type.
- Once you've selected the desired data type, you may see additional options or formats depending on the chosen category. Adjust these options as needed.
- Click "OK" to apply the changes and close the Format Cells dialog box.

Sometimes, excel doesn't change the data type based on the steps mentioned here. Its crucial to use alternative methods in such cases:

Alternate method-1:

- First, select the entire column that contains the data you want to change
- Navigate to the "Data" tab on the Ribbon.
- Look for the "Data Tools" group.
- Click on "Text to Columns". This will open the Text to Columns wizard.
- In the first dialog box of the wizard, you're asked to choose the file type - * "Delimited" * "Fixed-Width".
- You can choose either as this feature is not relevant when changing data type. Click "Next".

- In the final step of the wizard, you can specify the data format for your column. You can select:
  - "Text" to treat your data as text, even if it looks like numbers or dates.
  - "Date" and specify the date format (DMY, MDY, etc.).

Add Alternate steps - 2 (Using formula):

- For converting text to numbers, you can use a formula like =VALUE(A1) where A1 contains the text you want to convert.
- For dates, if Excel hasn't recognized a date format, you can use the DATEVALUE function or custom formulas to parse and convert text to date formats**.**

2. Handling missing values:(30 min)

We may not always be evident that handling a missing value would have a direct impact on the analysis. The impact can be a bit subtle, but in that subtle impact if it changes the total analysis, then it becomes really important for you to handle missing values. So, it is always advisable to handle the missing values so that we are sure that any of the problems that could occur if we're not handling it correctly would lead the analysis in a different direction.
In this section, we are gonna talk about:

a. Why is it required to handle the missing values?

- **Many machine learning algorithms require complete datasets. So, if there are any missing values in any of the columns, that particular function or model in Excel or any of the machine learning models may not work.**

  Example:
  Imagine you're the manager of a grocery store, and you want to understand the preferences of your customers to improve your inventory management and marketing strategies. You collect data on the items purchased by each customer during their visits to your store.
  In this scenario we build predictive models to forecast future customer behavior or segment customers into different groups based on their shopping preferences. You may use machine learning algorithms to predict which products a customer is likely to purchase next or to identify high-value customer segments for targeted marketing campaigns.

- **Handling the missing values will impact the mathematical or analytical results for the business problem.**

  Problems which can be encountered:
  - **Accuracy**: Imagine you're calculating the average income of a group of people, but for some, the income is not recorded. If you just ignore

these missing values, the average income calculated could be higher or lower than the actual average for the entire group. By properly handling missing values, you can ensure a more accurate calculation.

- **Completeness:** Consider you're analyzing survey data where respondents were asked to rate a service from 1 to 5. If some responses are missing, you don't have a complete picture of customer satisfaction. Filling in these missing values (appropriately) or deciding how to account for them ensures that your analysis considers all aspects of the dataset.
- **Decision Making**: Let's say a hospital is analyzing patient data to predict health outcomes. If key information (like blood pressure readings) is missing for some patients, predictions made about their health could be incorrect. This could lead to inappropriate decisions about their care. Handling these missing values ensures that decisions are based on as complete a dataset as possible.
- **Bias Reduction:** If you simply remove all records with missing values, you might end up with a biased dataset. For example, if more young people tend to leave the age field blank in a survey, excluding these records could skew your analysis towards older respondents. By addressing missing values thoughtfully, you can reduce such biases.

Example:

Let's assume we have 100 customers and all of the customer's income is known to us. Also let's assume that out 100 customers, probably 5 missing values are available in the income column. What could be the reason?

- The first reason could be the team collecting that dataset were not able to get that particular data point from those customers.
- The other reason could be, the income of those 5 customers is 0. The team collecting the dataset, rather than putting 0 they have kept it blank.

So, these types of problems are going to impact the analysis greatly and further it will lead to wrong delivery of the analysis to the stakeholders.

b. What are the different types of missing values that can be encountered during analysis?

Some of the different types of missing values could be:

- Blank and null values
- Special characters
- Random numbers
- Large values (outliers, which we are going to study in the next session)

c. How to recognize if there is any missing values or not:

- Blank and null values:

The **dataset** can be used to explain different missing values recognizing techniques.

- **Using the filter operation**: We apply the filter operation to the dataset and in the dialogue box if there is a blank option available in the list, then we can conclude that there is presence of null or blank values.
- **Using COUNTBLANK:**
  The COUNTBLANK function in Excel is used to count the number of empty or blank cells within a specified range.
  The syntax is =COUNTBLANK(Range), where range represents the range of cells where the formula needs to be applied.

- **Using COUNTA function**:
  The COUNTA function in Excel is an aggregate function used to count the number of cells in a range that are not empty. It counts cells that contain any type of data, including text, numbers, logical values, error values, and empty text ("").
  So, once we can calculate the total number of cells in the dataset and secondly we calculate the total number of cells that are not empty (COUNTA) and then we take the difference between both the values. If there is a difference between both then, we conclude that the dataset has missing values and if the difference is zero then we can conclude that the dataset has no missing values.

d. Methods to solve missing value problem:

- **Leave the rows with missing values:**

  - This is one of the methods that we can use while handling the missing values. In this method, we keep the rows with the missing values.

  SCENARIO-1: If the dataset is related to some algorithmic models, Then we cannot leave the missing values as it is because the models will not run appropriately.

  SCENARIO-2: If the dataset is small and consists of 10-15% of the data as missing values, we cannot keep them as it is. It will impact the analysis largely.

  SCENARIO-3: If the dataset is huge enough and does not include any kind of algorithmic models, then we can leave the missing value rows as it is.

- **Delete the rows with missing values:**
  - If the rows with missing values are deleted then other corresponding columns with the data points will also get deleted which might impact the analysis.

- If the dataset is small, let's say 1000 rows of data and we have 200 rows in total which encountered missing values then we cannot delete those rows as it will impact the analysis massively.
- If the dataset is large enough, let's say millions of rows of dataset and 10-15 rows are encountered with the missing values, then we can delete those rows as the calculation and aggregation will not be impacted greatly.

- **Data imputation methods:**
  Let's say the stakeholder doesn't want us to delete the rows and also doesn't want us to leave those rows as it is. In this scenario, we introduce the data imputation method. In this method, we impute some values in those missing rows. Also, we perform some statistical calculations which lead us to some values related to the values in that particular column and replace those missing values rows with that number or values or text.
  - **Replace with 0:**
    - Let's suppose, we have student marks as a dataset where the marks of the students are mentioned and there are some rows which have missing values. In this situation, we can conclude that the instructor might have ignored the students who have received 0 marks in the test and thus, we can impute 0 in those rows.
    - If the dataset is small and we are doing some calculation or aggregation function in the dataset, then the calculations might not be accurate, Thus, it is always advisable to talk to the stakeholder and select the method of imputation accordingly.
  - **Replace with the most frequently occurring values:**
    Let's suppose we have the "marital status" column in a dataset and suppose we have missing values in that column, so here we cannot replace it with 0 or run some calculations to get to the solution. Here, the best way is to calculate the maximum number of marital status in that column and replace it with the missing rows as well. Let's suppose after calculation we recognised that the maximum repeated output in the column is "married" and thus we replace the missing rows with "married".
  - **Statistical calculations:**
    This is a very accurate and reliable method in the industry which makes the dataset more relevant to accept by the stakeholders. This method includes the following functions
    - **Mean Imputation:** Replace missing values with the mean (average) of the observed values in the dataset. This method assumes that the missing values are similar to the observed values in the dataset.
    - **Median Imputation**: Replace missing values with the median value of the observed values in the dataset.

- **Mode Imputation**: Replace missing values with the most frequently occurring value (mode) in the dataset.

3. Statistical functions used in solving missing value problems:

a. Average/mean: (10 min)

In statistics, the mean, also known as the arithmetic mean or average, is a measure of the central tendency of a set of numerical values. It is calculated by adding up all the values in the data set and then dividing the sum by the total number of values. The formula for calculating the mean is:

$$m = \frac{\text{Sum of the terms}}{\text{Number of terms}}$$

Here, m = mean. The mean is a useful tool for summarizing a set of data and understanding its general properties. It can be affected by extreme values, so it is important to consider other measures of central tendency, such as the median and mode, as well as the variability of the data when analyzing a dataset.
In this course, you will use the AVERAGE function in MS Excel to compute the mean.

Example:
Let's begin with a situation . Let’s say you're analyzing the dataset of houses for sale, you're focusing on the blank values in the square footage column of the houses. You notice that the square footage values are normally distributed and there aren't significant outliers — most houses fall within a similar range of size.

In this scenario, imputing these missing values with the mean square footage would be reasonable. This is because the mean would accurately reflect the "average" house size in your dataset, given the normal distribution and lack of extreme outliers.

Syntax in Excel:

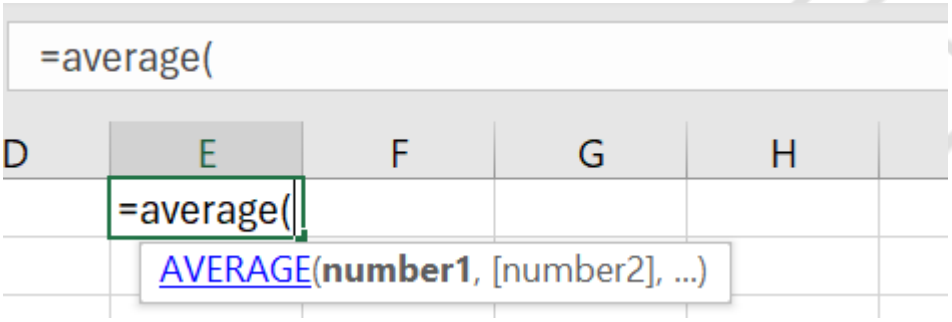

Figure: Represents the AVERAGE function in excel

This dataset will be used to explain this topic.

b. Median: (10 min)

The median is a statistical measure that represents the middle value of a dataset when it is arranged in order of magnitude. It is the value that divides the data set into two equal halves, such that half of the values are above the the median and half are below it.
To find the median, the data set is first arranged in ascending or descending order.
1. If the data set contains an odd number of values, then the median is the middle value. For example, the median for a sorted list of 15 observations is the 8th value.
2. If the data set contains an even number of values, then the median is the average of the two middle values. For example, the median for the sorted list of 16 observations is the average of the 8th and 9th values.

Unlike the mean, the median is not influenced by extreme values or outliers in the data set, making it a useful measure of central tendency in skewed or asymmetric distributions. A skewed or asymmetric distribution is a type of data distribution where the values are not evenly spread out around the average or middle of the data. In this type of distribution, the data tends to be concentrated on one side of the center, and the other side has fewer values that are more spread out.

Example:
Imagine you're analyzing a dataset of houses for sale, which includes information on the number of bedrooms. However, some listings are missing this information. Your data shows a significant range in the number of bedrooms, from 1 to 10, with a few outliers (e.g., mansions with 10+ bedrooms) that make the data right-skewed.
If you decide to fill these missing values with the mean number of bedrooms, the average might be artificially inflated by the outliers. Instead, by choosing the median, you use a value that more accurately reflects the central tendency of the majority of the houses in your dataset, thus providing a more accurate imputation for the missing values.

In this course, you will use the MEDIAN function in MS Excel to compute the mean.
Syntax:

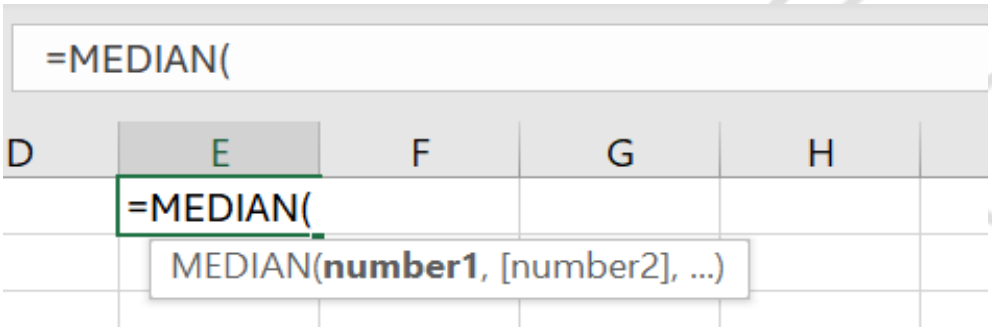

Figure: Represents the MEDIAN function in excel

This dataset will be used to explain this topic.

c. Mode: (5 min)

In statistics, the mode is a measure of central tendency that represents the most frequent value in a dataset. More specifically, the mode is the value that occurs with the highest frequency in a set of observations or data points. It is one of the three main measures of central tendency, along with the mean and median.
The mode is particularly useful when dealing with categorical or discrete data, such as the number of times a certain event occurs, or the most common color of cars in a parking lot. It is also useful when dealing with continuous data that can be grouped into categories or bins.
Unlike the mean and median, the mode does not take into account the actual values of the data, only their frequency of occurrence. This makes it less sensitive to outliers or extreme values that may affect the mean or median. However, it may not be a representative measure of central tendency if there are multiple modes in the dataset or if the frequency of the modes is close to each other. You can use the MODE function to compute the mode.

Example:
Suppose you're examining the feature that describes the heating type used in each house (e.g., gas, electric, solar). This is a categorical variable, and you notice that a significant number of houses use "gas" heating, making it the most common (mode) heating type in your dataset.

Using mode imputation to fill these missing values with "gas" would be appropriate. This approach assumes that the missing houses are likely to have the same heating type as the majority of houses in your dataset, which is a reasonable assumption given the categorical nature of this data

Syntax in MS Excel:

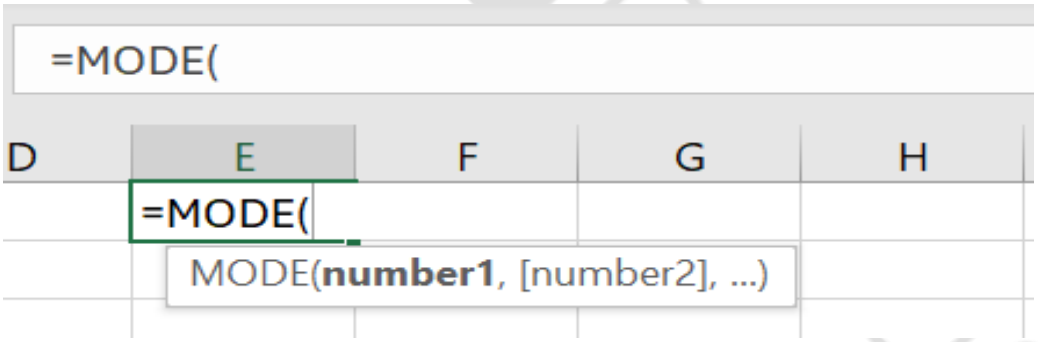

Figure: Represents the MODE function in excel

This dataset will be used to explain this topic.

d. Moving averages: (15 min)

1. What is a moving average?

A moving average, also called a moving mean or a rolling mean, is a calculation that relies on a series of averages from data subsets within an entire data set. It's a term statisticians, technical analysts and financial analysts use to describe changes to averages as new data becomes available. It explains how a data series changes over a set period. The moving average also updates to include recent data along with data points from predetermined intervals.

2. **Why calculate moving averages in excel?**

    - **To constantly update average prices:**
      Excel is a useful program for calculating and recording moving averages because it offers a comprehensive look at different data points over time. By showing the moving averages over set periods, like every 15 days, 100 days or 200 days, you can notice uptrends or downtrends that can inform important perceptions and decisions. Excel also makes it easier to find the best timeline for your data uses.

    - **To mitigate data errors associated with short-term fluctuations:**
      Another benefit of calculating a moving average is that short-term fluctuations won't affect overall numbers as much. Instead, technical analysts, financial analysts and data experts can rely on more consistent data and trends to inform their takeaways. This can mitigate risks associated with outlying data points.

    - **Imputation for seasonal disturbances:**
      Suppose the dataset tracks features that have seasonal variations, such as the average number of days houses are on the market before being sold, and some seasons are missing this information. A moving average could help impute these missing values by averaging the data from similar seasons in previous and following years, assuming that the market behavior follows a consistent pattern across similar seasons.

    Example of when moving averages should be used in excel:
    Imagine the dataset includes monthly sale prices of houses in a particular area over several years, but some months are missing data due to recording errors or no sales being recorded. A moving average could be used to impute these missing values by averaging the sale prices of the houses from the surrounding months. For example, if the sale price in June is missing, you could take the average of the prices in May and July (a simple moving average) or use more months to calculate the average (an extended moving average) to fill in the missing value for June.

3. **Steps for moving averages in Excel: (Dataset)**

Organize your data: Ensure your data is arranged in sequential order, typically with the values you want to calculate the moving average for in a single column.

Determine the period for the moving average: Decide on the number of periods (e.g., days, months, quarters) over which you want to calculate the moving average. This will determine the size of the moving average window.

Calculate the moving average:
For a simple moving average, enter the following formula in a separate column where you want the moving average to appear: =AVERAGE(range)
Replace "range" with the actual range of cells containing the data points you want to include in the moving average.

For example, if your data is in column A starting from A2, and you want to calculate a 3-period moving average, you would enter the formula =AVERAGE(A2:A4) in cell B4 (assuming your first moving average value appears in cell B4).
Drag the formula down to calculate the moving average for subsequent periods.
Adjust the formula for subsequent periods: As you drag the formula down to calculate the moving average for subsequent periods, ensure that the cell references adjust accordingly to maintain the correct moving average window.

Problem:

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: We have observed the mean, median and mode are almost equal in value.When the mean, median, and mode are almost the same, it means that there is no significant bias or distortion in the data. Analysts can have confidence in using any of these measures to represent the typical value or central tendency of the dataset.

Example: ABC company dataset