FINDING MISSING VALUES IN DATA

Introduction: -

Analyse raw data and provide insights from the data for business intelligence.

Process (Python & Ms Excel): -

1. Cleaning, Transforming the data and finding missing values: - Data from Data Assignment.xlsx file is being transferred into python by reading the data in the python data frame. The xlsx sheet consists of 3 worksheets i.e. (Associate ABC, Associate KLM, Associate XYZ) have been first separated into 3 individual csv files.
2. Python files: - (Using Jupyter Notebook for making python scripts (. ipynb file type).
    1) Time Interpolation 1.ipynb (Associate ABC is being read in the script).
    2) Time Interpolation 2.ipynb (Associate KLM is being read in the script).
    3) Time Interpolation 3.ipynb (Associate XYZ is being read in the script).
    4) Final_Table.ipynb (3 output csv files generated using python are concatenated).
3. In these above mentioned python files in Time Interpolation 1.ipynb file the necessary libraries have been imported like (pandas, numpy , matplotlib, datetime, seaborn and sklearn (Simple Imputer)).
    1) Reading the data from the csv file (xlsx converted to csv file).
    2) Converting date values datatype from object to datetime and minutes in the (time spent on LG (mins)) column to hour and minutes.
    3) Then using interpolate method based on time a machine learning method to find the values which are missing. Making Date column as index column for interpolation.
    4) Then checking the variance with the original data to identify the variance occurring with the new generated values. Therefore, by using the interpolation method based on time we get very less to no variance comparing with original dataset. And this is being check using line graph using matplotlib library.
    5) For Further analysis checking the values showing gaps from the original data in a line chart (leads by date), and the continuous line chart (leads by date) with generated values filling the gaps in the data.
    6) Then the data has been loaded into new csv file named associate_ABC_interpolate_time.csv.
4. Similarly for Time Interpolation 2.ipynb we do the same as above and we obtain the results in associate_KLM_interpolate_time.csv file and for Time Interpolation 3.ipynb we do the same as above and we obtain the results in associate_XYZ_interpolate_time.csv file.
5. In Final_Table.ipynb file all the outputs generated using above 3 python scripts have been read in this file (3 csv files output generated) and transformation like adding a separate column known as associate names and other datetime transformations. Here all the 3 data frames have concatenated for getting a single and distinguishable result.
6. And other transformations which are needed are done in Ms excel and Power Bi like making of column known as hours in Ms excel derived from column (Time_spend_on_LG_in_mins) when divided by 60 (for getting hours).

7. Using Jupyter Notebook for python scripts and the python files are in **(.ipynb)** form according to jupyter notebook.

Power BI: -

It is the data visualization tool used to obtain results consisting of KPI, trends and insights from the data. File name PowerBI(Dashboard).

1. All the 4 output csv files generated using python are being loaded into the Power Bi. Four pages (report view) have been made into the power bi file consisting of data visualization (Main Report, Associate ABC, Associate KLM, Associate XYZ).
2. The data visualized using charts like staked bar chart, line chart, pie chart, cards, and pivot table. Some of the charts which are also further extendable using drill down options (arrows shown on the top & bottom parts of the chart when chart is selected) for more information.

Results: -

1. Total No. of Leads Generated by Each Associate: -
    - Associate ABC: - 523 leads.
    - Associate KLM: - 671 leads.
    - Associate XYZ: - 716 leads.
2. Total No. of Leaves taken by each associate: -
    - Associate ABC: - 16 leaves.
    - Associate KLM: - 23 leaves.
    - Associate XYZ: - 16 leaves.
3. Average No. of leads generated by each associate: -
    - Associate ABC: - 6 leads.
    - Associate KLM: - 9 leads.
    - Associate XYZ: - 11 leads.
4. The associate that has been the most consistent in leads generation is Associate XYZ by generating 716 leads with an average of 11 leads generation and taking minimum leaves that is 16 leaves.
5. Removing missing values from the dataset for analysis, yes I use it when the loss of data occurs due to server issue, internet problem or person entering the data manually forgets to enter some values then it is very useful from removing missing values from the dataset for data analysis.
    - Rationale for my answer would be like removing mean, median imputation for replacing missing values with it.
    - Another example would be if three columns having data and one column is date column and other two columns are having numeric values, if in only one column from these two numeric data columns some data is missing, so we can find the missing data using correlation method between the columns.
6. My recommendations for the business development team would be please see the trends and the kpi's generated to improve the area of business process lacking behind in a certain way.
7. All the filenames are being mentioned in this document, processes done using Python, Ms Excel and dashboard/report made using Power bi.