Q3

1)

Three Types of Missingness: MCAR, MAR, and MNAR

In data analysis, understanding the mechanism of missing data is crucial for correctly handling and inferring conclusions. Missing data can generally be categorized into three types based on its relationship with other variables: **MCAR**, **MAR**, and **MNAR**. Here are their definitions and implications:

1. MCAR (Missing Completely at Random)

- **Definition:** The data is missing completely at random, with no relationship to any observed or unobserved variables. That is, the missing data does not depend on any observed values or itself.
- **Implication:** MCAR missing data does not introduce any bias in the analysis because the missing values are independent of other variables in the dataset.

2. MAR (Missing at Random)

- **Definition:** The data is missing at random if the probability of a value being missing depends on observed data but not on the missing values themselves.
- **Implication:** While MAR data is not completely random, it still does not bias the analysis, as long as the missingness is accounted for by observed data.

3. MNAR (Missing Not at Random)

- **Definition:** The data is missing not at random if the probability of a value being missing depends on the unobserved value itself. In other words, the missingness is related to the value that is missing.
- **Implication:** MNAR introduces bias in the analysis, as the missingness is dependent on the data that is missing, making it more difficult to handle.

MCAR Test Logic

To determine if the missing data follows the **MCAR** mechanism, we use statistical tests like the **t-test** for numeric variables and **Chi-Square test** for categorical variables. Here's how the tests work:

1. **Numeric Variables (t-test):**
    - The t-test compares the distribution of the observed values between the

"missing" and "non-missing" groups.

- If the p-value from the t-test is greater than 0.05, we fail to reject the null hypothesis, meaning the missing data is likely MCAR.
- This suggests that the missing data does not depend on the values of the variable itself or any other variables, making it random.

2. **Categorical Variables (Chi-Square test):**
    - For categorical variables, the Chi-Square test is used to check if the distribution of missingness is independent of other categorical variables.
    - A p-value greater than 0.05 in the Chi-Square test suggests that the missing data does not depend on the observed categories, meaning the missing data is MCAR.

Conclusion

Based on the results of the tests performed for the variables:

- **BloodPressure** (numeric) – The t-test showed no significant difference between the "missing" and "non-missing" groups, indicating it is likely MCAR.
- **BloodSugar** (numeric) – Similarly, the t-test for this variable suggested that missing data is likely MCAR.
- **MedicalHistory** (categorical) – The Chi-Square test indicated that the missingness is likely independent of the observed categories, thus also supporting the hypothesis that the missing data is MCAR.

Final Conclusion: All three variables—**BloodPressure**, **BloodSugar**, and **MedicalHistory**—appear to have missing data that is **MCAR (Missing Completely at Random)**. This means that the missingness of these variables does not depend on either observed or missing values, and they can be treated without introducing significant bias in the analysis.

2)

Using three different methods to fill missing BloodSugar values:

1. Mean and Median Filling:

By calculating the mean and median of the BloodSugar field and using the .fillna() method to fill in the missing values.

2. KNN Filling:

Using the KNNImputer class for filling, with n_neighbors=5 indicating the selection of the 5 most similar neighbors for filling.

Data Comparison:

Using matplotlib to draw comparison charts, showing the distribution of the BloodSugar field in the original data and the filled data.

Different filling methods are represented by different colors (original data, mean filling, median filling, and KNN filling).

Chart and CSV Output:

The chart displays the changes in the distribution of the BloodSugar field after using different filling methods (mean, column median, KNN).

It can be intuitively seen which method better preserves the distribution characteristics of the original data and which method may introduce bias.

The filled csv file obtained: medical_dataset_filled.csv, with the last three columns representing the BloodSugar filling values from the three different methods.