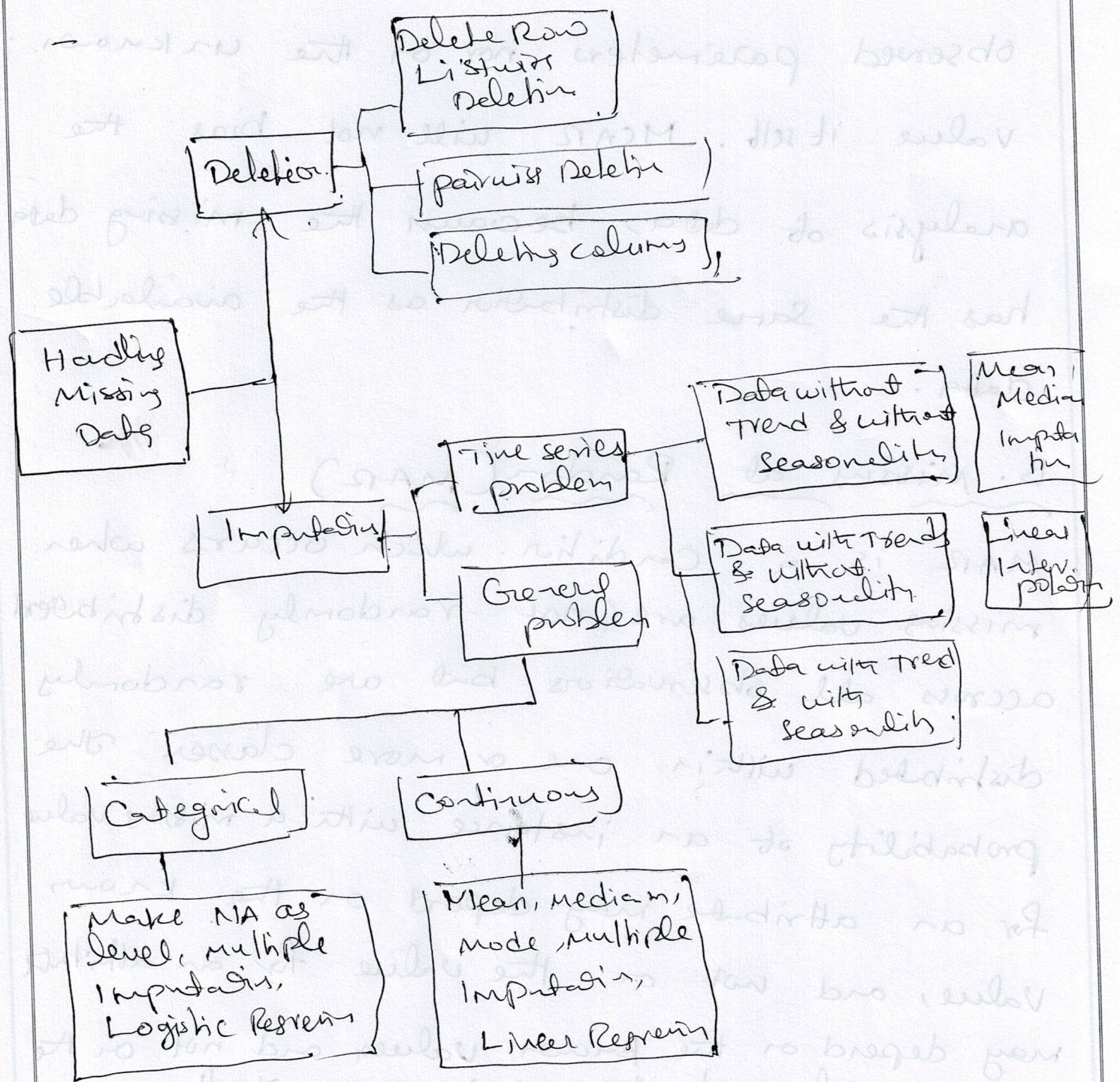
Subject: EDA
Faculty: D. Jyothi
Topic: Traditional methods of dealing with missing data

Class Notes

Unit No: II
Lecture No: 1
Link to Session
Planner (SP): S.No.... of SP
Book Reference:
Date Conducted:
Page No: 1

Traditional methods of dealing with data missing.



A. Missing completely at Random (MCAR)

MCAR is the strongest assumption for missing values of a dataset. The missing of a value neither depends on the observed parameters nor on the unknown value itself. MCAR will not bias the analysis of data, because the missing data has the same distribution as the available data.

B. Missing at Random (MAR)

MAR is a condition which occurs when missing values are not randomly distributed across all observations but are randomly distributed within one or more classes. The probability of an instance with a missing value for an attribute may depend on the known value, and not on the value for an attribute may depend on the known values, and not on the

Subject: EDA
Faculty: D. Jyothi
Topic: traditional methods of dealing with missing data

Class Notes

Unit No: III
Lecture No: 2
Link to Session
Planner (SP): S.No.... of SP
Book Reference:
Date Conducted:
Page No: 3

Not missing at Random (NMAR): not missing at Random is the most challenging form, occurs when missing values are not randomly distributed across observations. It is also called as non-ignorable missingness. The probability of an instance with a missing value for an attribute might depend on the value of the attribute.

List wise Deletion:

If a case has missing data for any of the variables, then simply delete that case from the analysis. Usually it is default in the statistical packages.

Advantages:

It can be used with any kind of statistical analysis and no special computational methods are required

Disadvantages: It can exclude a large fraction

Pairwise Deletion:

This method considers every feature independently. For each feature, all recorded values in each observation are considered and missing data are ignored using all available data between pair of variables to calculate covariance. Only delete "pairs" with specific missing data.

Advantage:

- simple
- All available data is used
- Smaller loss of cases than in the Listwise deletion.

Disadvantage:

- Covariance have different sample sizes different parts of the model have different degrees of freedom.

Subject: EDA
Faculty: D. Jyothi
Topic: Traditional methods of dealing with missing data

Class Notes

Unit No:
Lecture No: 3
Link to Session
Planner (SP): S.No.... of SP
Book Reference:
Date Conducted:
Page No: 5

C. Imputation Method.

The missing values are filled with estimated values, Each missing value is substituted for a reasonable guess, and then analysis is carried out as if there were no missing values. To deal with missing data patterns of missingness and assumptions are used to determine which methods can be used. using standard methods,

Advantage:

Mean imputation is one of the most frequently used techiques to treat missing data. The variables mean value of all known values of that attributes is substituted where the instance with missing attribute belongs

Regression Imputation:

All the missing values of a dataset replaced by the predicted value of that variable from a regression analysis based only on the complete cases.

In this model even if the mean parameters are correctly estimated, the variance parameters are underestimated because this method assumes no residual variance around the regression line predictor.

Disadvantages:

- The missing data are replaced by a linear regression function instead of replacing all missing data with a statistics.

Subject: EDA
Faculty: D. Jyothi
Topic: Methods of dealing with Missing data.

Class Notes

Unit No:
Lecture No: 4
Link to Session
Planner (SP): S.No.... of SP
Book Reference:
Date Conducted:
Page No: 14

Hot deck Imputation.

Missing data are replaced with values from input data vector that is nearest in terms of the attributes that are known in both patterns. By substituting differed observed values for each missing value it attempts to protect the distribution. The similar method of hot deck imputation is cold deck imputation method which takes other data source than current dataset.

Advantage:

- Full sample size is preserved

Disadvantage:

Fifth: artificially increases statistical power by assuming that similar units are actually identical.

Subject: EDA
Faculty: D. Jyothi
Topic: Missing data handling, improving the accuracy of analysis

Class Notes

Unit No: IV
Lecture No: 5
Link to Session
Planner (SP): S.No.... of SP
Book Reference:
Date Conducted:
Page No: 8

Missing data handling, Improving the accuracy of analysis Reading Datasets.

Missing data into the "sepal-length" column and then handle it to improve accuracy

Step1: Import Required Libraries and Load the Dataset

```
import pandas as pd
from sklearn.datasets import load_iris
import numpy as np
# Load dataset
    iris = load_iris()
iris_df = pd.DataFrame(data= np.c[iris["data"],
  iris["target"]], columns = iris["feature_names"] +
    ['target'])
```

```
# replace missing data with NaN
np.random.seed(0)
missing_indices = np.random.choice(iris_df.index,
    size=10, replace=False)
iris_df.loc[missing_indices, 'sepal length (cm)']
    = np.nan.
```

Step 2: Handle Missing Data

Imputing missing data with mean value of the "sepal length (cm)" column

```
mean_sepal_length = iris_df['sepal length (cm)'].
    mean()
iris_df['sepal length (cm)'].fillna(mean_sepal_
    length, inplace=True)
```

Subject: EDA
Faculty: D. Jyothi
Topic: Missing data handling, improving the accuracy of analysis

Class Notes

Unit No: III
Lecture No: 6
Link to Session Planner (SP): S.No.... of SP
Book Reference:
Date Conducted:
Page No: 16

Step 3: Analyze and Compare Result.

Analyze the modified dataset and results. Let's calculate the average sepal length before and after handling missing data.

```
# Calculate the average sepal length before handling missing data.
average_sepal_length_before = iris_df['sepal length (cm)'].mean()
print(f"Average Sepal Length Before Handling Missing data: {average_sepal_length_before:.2f}")
# Calculate the average sepal length after handling missing data
average_sepal_length_after = iris_df['sepal length (cm)'].mean()
```

print f (f"Average Sepal Length After Handling.

Missing Data: {average_sepal_length_after:.2f}")

Missing data in the "sepal length(CM)" column and imputing it with the mean value, we have improved the accuracy of our analysis of sepal lengths in the iris dataset.

Subject: EDA
Faculty: D. Jyothi
Topic: working with different file types (.csv)

Class Notes

Unit No: III
Lecture No: 7
Link to Session Planner (SP): S.No.... of SP
Book Reference:
Date Conducted:
Page No: 12

Working with different file types .txt .csv etc practical issues in Multiple Imputation.

Step:1 Import Required Libraries and Load the Dataset from CSV

- Assuming you have a CSV file named "iris.csv" with your dataset

```
# Load the dataset from a CSV file
file_path = 'iris.csv'  # Replace with the actual file path.
iris_df = pd.read_csv(file_path)

# checked the loaded dataset
print(iris_df.head())
```

Handling Missing data (Step 2)

imputing with mean value.

```
mean_sepal_length = iris_df['sepal_length'].
                    mean()

iris_df['sepal_Length'].fillna(mean_sepal_length,
                    inplace = True)
```

check if missing values are filled

```
print(iris_df.head())
```

Step3: Analyze and Compare Results

Calculate the Average Sepal Length before and after handling missing data:

calculate the Average Sepal Length before handling missing data

```
average_sepal_length_before = iris_df['sepal_
    length'].mean()

print(f'Average Sepal Length Before Handling missing
data: {average_sepal_length_before:.2f}')
```

Subject: EDA
Faculty: D. Jyothi
Topic: working with .txt file

Class Notes

Unit No: IV
Lecture No: 8
Link to Session
Planner (SP): S.No.... of SP
Book Reference:
Date Conducted:
Page No: 14

#Calculate average Sepal Length after handling missing data

```
average_sepal_length_after = iris_df['sepal_length'].mean()

print(f"Average Sepal Length After handling missing data : {average_sepal_length_after:.2f}")
```

Example to Load .txt file

Step1: Import Required Libraries and Load the Dataset from .txt.

"iris.txt" with the dataset and the data is space separated.

```
import pandas as pd

# load dataset from a .txt file

file_path = 'iris.txt'

iris_df = pd.read_csv(file_path, sep=' ')

# check loaded dataset

print(iris_df.head())
```

Step2: Handle missing Data

```
# impute missing values in the "sepal_length"
  column with the mean value

mean_sepal_length = iris_df['sepal_length'].
                    mean()

iris_df['sepal_length'].fillna(mean_sepal_
    length, inplace=true)

# check if missing values are filled
    print(iris_df.head())
```

Subject: EDA
Faculty: D. Jyothi
Topic: working with different bike types.

Class Notes

Unit No: III
Lecture No: 9.
Link to Session Planner (SP): S.No.... of SP
Book Reference:
Date Conducted:
Page No: 96

Step3: Analyze and Compare Results.

#Calculate Average sepal length before handling missing data

```
average_sepal_length_before = iris_df['sepal length'].mean()
print(f'Average Sepal Length Before handling Missing Data: {average_sepal_length_before:.2f}')
```

#calculating the average sepal length after handling missing data

```
average_sepal_length_after = iris_df['sepal length'].mean()
print(f'Average Sepal Length After Handling missing data: {average_sepal_length_after:.2f}')
```

Practical issues related to multiple imputation

Model selection:

Choosing the appropriate imputation model can be challenging - For Example, imputing missing math scores might require considering the relationship between math scores and other variables like science scores, age and gender.

Number of Imputations:

Decides on the number of imputations to impute. Lets say we decide to perform 5 imputations. The computational burden will increase because we're essentially creating 5 complete datasets with imputed values for analysis.

Subject: EDA
Faculty: D-Jyothi
Topic: Issues in multiple Imputation.

Class Notes

Unit No: III
Lecture No: 10
Link to Session
Planner (SP): S.No.... of SP
Book Reference:
Date Conducted:
Page No: 18

Imputation of Categorical variables

Gender is a categorical variable. Deciding how to impute missing values in a categorical variables requires a different approach. It might involve logistic regression or mode imputation, and the choice depends on the research question.

pooling of Result:

After imputation, you will need to poll the results correctly. For example, if you want to calculate the average math score across imputations, you can't simply average the imputed values because it might underestimate uncertainty.

Data privacy:

If you are working with sensitive data like age and gender, you need to ensure that the imputation process doesn't compromise data privacy. you might need to add noise or use privacy-preserving techniques.

Reporting and Interpretation:

You will need to be clearly report how you handled missing data, including details of imputation models, number of imputations, and how you pooled results. Interpretation should account for the uncertainty of imputation.

Subject: EDA
Faculty: D. Jyothi
Topic: Data Modelling

Class Notes

Unit No: III
Lecture No: 11
Link to Session
Planner (SP): S.No.... of SP
Book Reference:
Date Conducted:
Page No: 20

1. Univariate Data Modelling

- Histograms:

Create histograms to visualize the distribution of individual variables. This help in understanding the central tendency, spread, and shape of data

- Kernal Density plots.

Construct kernal density plots to estimate the probability density function of a continuous variable. This can reveal underlying patterns in the data distribution.

- Box plots: Generate box plots to identify potential outliers, visualize the spread of data, and detect skewness

2. Bivariate Data Modelling

- Scatter plot:

Construct scatter plots to explore the relationship between two continuous variables this help identify patterns, correlations, and potential outlier.

Pairplots: Use pairplots or scatter matrix plots to visualize the pairwise relationships between multiple variables simultaneously. This can reveal complex interactions.

Correlation Heatmaps:

Create correlation heatmaps to visualize the strength and direction of linear relationship between variables. This is particularly useful for identifying correlation features.

Subject:
Faculty:
Topic:

Class Notes

Unit No:
Lecture No:
Link to Session
Planner (SP): S.No.... of SP
Book Reference:
Date Conducted:
Page No: 22

Multivariate Data Modeling:

cluster analysis: Apply clustering techiques like k-means or hierarchical clustering to group similar data points together. This can reveal natural groupings within the data

principal Component Analysis (PCA): use PCA to reduce the dimensionality of data while preserving as much variance is possible.

Factor Analysis:

Employ factor analysis to identify underlying latent factors that explain patterns of correlations between observed variables

Time series modelling:

- Time series Decomposition
- Auto correlation & partial Auto correlation

Text data Modelling:

- word clouds
- topic clouds.

Geospatial Data modelling

- Geospatial visualization.

Interactive data modelling:

- Interactive visualization tools:

  Es: tableau, powerBI, python libraries like plotly and Bokeh to create dynamic visualization

Subject:
Faculty:
Topic: schema design.

Class Notes

Unit No: IV
Lecture No: 12
Link to Session
Planner (SP): S.No.... of SP
Book Reference:
Date Conducted:
Page No: 24

Data loading & storage:

Decide on the appropriate format for storing your data during the EDA phase. Common formats include CSV, Excel, SQL Databases, parquet, HDF5

Column Naming & Data types:

- Give meaningful names to columns to facilitate understanding clear descriptive column names make it easier to interpret the data. Assign appropriate data types to each column to ensure data consistency and enable correct data operations.

Data cleaning & transformation:

Address missing data by deciding on an appropriate strategy

- perform data transformations when necessary such as converting datatypes, normalizing values, creating derived variables.

Data Exploration variables:

Create metadata or annotations for columns to store additional information. such as units of measures, data sources, description.

- Identify key variables that are central to your analysis and focus your exploration on them

Data Relationships & joins

- Dataset includes multiple tables or datasets, define relationships between them to facilitate joining & merging of analysis

Subject:
Faculty:
Topic:

Class Notes

Unit No:
Lecture No:
Link to Session
Planner (SP): S.No.... of SP
Book Reference:
Date Conducted:
Page No: 26

Data Indexing:

Creating indexes on columns that are frequently used for filtering or sorting data. This can significantly speed up query performance during exploration.

→ Hierarchical Data:

Determine how to represent hierarchical structures for effective analysis

→ Version Control & Documentation

- Keep track of changes to your data schema and analysis code using version control system.
- Maintain detailed documentation that describes the schema, data sources, and any data preprocessing steps taken during EDA.