Imputation

Instructions:

Please share your answers filled inline in the word document. Submit code files wherever applicable.

Please ensure you update all the details:

Name: ULLI VENKATA SAI KUMAR

Batch Id: 04072024HYD10AM

Topic: Data Pre-Processing

Problem Statement:

Majority of the datasets have missing values, that might be because the data collected were not at regular intervals or the breakdown of instruments and so on. It is nearly impossible to build the proper model or in other words, get accurate results. The common techniques are either removing those records completely or substitute those missing values with the logical ones, there are various techniques to treat these types of problems.

1) Prepare the dataset using various techniques to solve the problem, explore all the techniques available and use them to see which gives the best result.

Hint: Go through this link: https://360digitmg.com/mindmap-data-science

```
import pandas as pd
import numpy as np
from sklearn.impute import SimpleImputer
from sklearn.impute import KNNImputer
df =pd.read_csv(r"claimants.csv")
# 1. Checking for missing values
missing_values = df.isnull().sum()
# 2. Removing rows with missing data
df_dropped = df.dropna()
# 3. Mean Imputation for missing values
mean_imputer = SimpleImputer(strategy='mean')
df_mean_imputed = pd.DataFrame(mean_imputer.fit_transform(df), columns=df.columns)
```

```
# 4. Mode Imputation
mode_imputer = SimpleImputer(strategy='most_frequent')
df_mode_imputed = pd.DataFrame(mode_imputer.fit_transform(df), columns=df.columns)
# 5. Forward Fill/Backward Fill
df_ffill = df.fillna(method='ffill')
df_bfill = df.fillna(method='bfill')
# 6. K-Nearest Neighbors (KNN) Imputation
knn_imputer = KNNImputer(n_neighbors=5)
df_knn_imputed = pd.DataFrame(knn_imputer.fit_transform(df), columns=df.columns)
# 7. Interpolation
df_interpolated = df.interpolate()
# Output the missing values count and imputed datasets
missing_values, df_dropped, df_mean_imputed, df_mode_imputed, df_ffill, df_bfill,
df_knn_imputed, df_interpolated
```

Name	Type	Size	Value
df_ffill	DataFrame	(1340, 7)	Column names: CASENUM, ATTORNEY, CLMSEX, CLMINSUR, SEATBELT, CLMAGE, L ...
df_interpolated	DataFrame	(1340, 7)	Column names: CASENUM, ATTORNEY, CLMSEX, CLMINSUR, SEATBELT, CLMAGE, L ...
df_knn_imputed	DataFrame	(1340, 7)	Column names: CASENUM, ATTORNEY, CLMSEX, CLMINSUR, SEATBELT, CLMAGE, L ...
df_mean_imputed	DataFrame	(1340, 7)	Column names: CASENUM, ATTORNEY, CLMSEX, CLMINSUR, SEATBELT, CLMAGE, L ...
df_mode_imputed	DataFrame	(1340, 7)	Column names: CASENUM, ATTORNEY, CLMSEX, CLMINSUR, SEATBELT, CLMAGE, L ...
knn_imputer	impute._knn.KNNImputer	1	KNNImputer object of sklearn.impute._knn module
mean_imputer	impute._base.SimpleImputer	1	SimpleImputer object of sklearn.impute._base module
missing_values	Series	(7,)	Series object of pandas.core.series module
mode_imputer	impute._base.SimpleImputer	1	SimpleImputer object of sklearn.impute._base module

Help | Variable Explorer | Plots | Files

Console 7/A

```
[1340 rows x 7 columns],
      CASENUM  ATTORNEY  CLMSEX  CLMINSUR  SEATBELT  CLMAGE    LOSS
0           5         0     0.0       1.0       0.0    50.0  34.940
1           3         1     1.0       0.0       0.0    18.0   0.891
2          66         1     0.0       1.0       0.0     5.0   0.330
3          70         0     0.0       1.0       1.0    31.0   0.037
4          96         1     0.0       1.0       0.0    30.0   0.038
...       ...       ...     ...       ...       ...     ...     ...
1335    34100         1     0.0       1.0       0.0    31.0   0.576
1336    34110         0     1.0       1.0       0.0    46.0   3.705
1337    34113         1     1.0       1.0       0.0    39.0   0.099
1338    34145         0     1.0       0.0       0.0     8.0   3.177
1339    34153         1     1.0       1.0       0.0    30.0   0.688

[1340 rows x 7 columns])
```