Preksha - A - Patel
60004210126
C2-1
comps.

Machine Learning (23)

Exp - 1 : Data processing

Aim :- To perform data preprocessing in terms of handling, missing data, removing outliers, eliminating duplicate rows and modifying the datatype, etc.

Description of Experiment :-

Python is an easy-to-learn programming language, which makes it the most preferred choice for beginners in Data Science, Data Analytics, and Machine Learning. It also has a great community of online learners and excellent data-centric libraries. With so much data being generated, it becomes important that the data we use for Data Science applications like Machine Learning & predictive Modeling is clean. But what do we mean by clean data? And what makes data dirty in the first place? Dirty data simply means data that is erroneous. Duplicacy of records, incomplete or outdated data, and improper parsing can make data dirty. This data needs to be cleaned. Data cleaning (or data cleansing) refers to the process of "cleaning" this dirty data, by identifying errors in the data & then rectifying them. Data cleaning is an important step in and Machine Learning project, & we will cover some basic data cleaning techniques (in Python).

Cleaning Data in Python

We will now separate the numeric columns from the categorical columns

Missing values :- We will start by calculating the percentage of values missing in each column, & then storing this information in a DataFrame

P.T.O

Drop observations :- One way could be to drop these observations that contain any null value in them for any of the columns. This will work when the percentage of missing values in each column is very less.

Remove columns (features) :- Another way to tackle missing values in a dataset would be to drop those columns or features that have a significant percentage of values missing.

Impute missing values :- There is still missing data left in our dataset. We will now impute the missing values in each numerical column with the median value of that column.

Outliers :- An outlier is an unusual observation that lies away from the majority of the data. Outliers can affect the performance of a Machine Learning model significantly.

Duplicate records :- Data can sometimes contain duplicate values. It is important to remove duplicate records from your dataset before you proceed with any Machine Learning project. In our data, since the ~~data~~ ID column is a unique identifier, we will drop duplicate records by considering all but the ID column.

Fixing data type :- often in the dataset, values are not stored in the correct data type. This can create a problem in later stages, and we may not get the desired output or may get errors while execution.

SDS | Page No.
Date / /

Conclusion :-

In the "Data Preprocessing" experiment, we focused on enhancing the quality of our dataset for effective use in Data Science and Machine Learning. We addressed issues such as missing values, outliers, and duplicate records through techniques like imputation, observation removal, and data type correction. These steps are crucial to ensure clean, accurate data, laying a strong foundation for successful data analysis and predictive ~~analysis~~ modeling in Python.

NAME:-PREKSHA PATEL
SAPID:-60004210126
BRANCH:-COMPUTER ENGINEERING

MACHINE LEARNING

EXPERIMENT-01

CODE AND OUTPUT:-

```
import pandas as pd
```

Reading a csv file

```
[2] df=pd.read_csv("food_coded.csv")
```

```
[3] df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 129 entries, 0 to 128
Data columns (total 61 columns):
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   GPA                           127 non-null    object
 1   Gender                        129 non-null    int64
 2   breakfast                     129 non-null    int64
 3   calories_chicken              129 non-null    int64
 4   calories_day                  109 non-null    float64
 5   calories_scone                128 non-null    float64
 6   coffee                        129 non-null    int64
 7   comfort_food                  128 non-null    object
 8   comfort_food_reasons          128 non-null    object
 9   comfort_food_reasons_coded    110 non-null    float64
 10  cook                          126 non-null    float64
 11  comfort_food_reasons_coded.1  129 non-null    int64
 12  cuisine                       111 non-null    float64
 13  diet_current                  128 non-null    object
 14  diet_current_coded            129 non-null    int64
 15  drink                         127 non-null    float64
 16  eating_changes                126 non-null    object
 17  eating_changes_coded          129 non-null    int64
 18  eating_changes_coded1         129 non-null    int64
 19  eating_out                    129 non-null    int64
 20  employment                    120 non-null    float64
 21  ethnic_food                   129 non-null    int64
 22  exercise                      116 non-null    float64
 23  father_education              128 non-null    float64
 24  father_profession             126 non-null    object
 25  fav_cuisine                   127 non-null    object
 26  fav_cuisine_coded             129 non-null    int64
 27  fav_food                      127 non-null    float64
 28  food_childhood                128 non-null    object
 29  fries                         129 non-null    int64
 30  fruit_day                     129 non-null    int64
 31  grade_level                   129 non-null    int64
 32  greek_food                    129 non-null    int64
 33  healthy_feeling               129 non-null    int64
 34  healthy_meal                  128 non-null    object
```

Printing the dataset

```
[4] df
```

	GPA	Gender	breakfast	calories_chicken	calories_day	calories_scone	coffee	comfort_food	comfort_food_reasons	c
0	2.4	2	1	430	NaN	315.0	1	none	we dont have comfort	
1	3.654	1	1	610	3.0	420.0	2	chocolate, chips, ice cream	Stress, bored, anger	
2	3.3	1	1	720	4.0	420.0	2	frozen yogurt, pizza, fast food	stress, sadness	
3	3.2	1	1	430	3.0	420.0	2	Pizza, Mac and cheese, ice cream	Boredom	
4	3.5	1	1	720	2.0	420.0	2	Ice cream, chocolate, chips	Stress, boredom, cravings	
...	...	...	...	...	...	...	...	...	...	
124	3.9	1	1	430	NaN	315.0	2	Chocolates, pizza, and Ritz.	hormones, Premenstrual syndrome.	
125	2.4	2	1	430	NaN	315.0	1	none	we dont have comfort	
126	3.654	1	1	610	3.0	420.0	2	chocolate, chips, ice cream	Stress, bored, anger	
127	3.3	1	1	720	4.0	420.0	2	frozen yogurt, pizza, fast food	stress, sadness	
128	3.2	1	1	430	3.0	420.0	2	Pizza, Mac and cheese, ice cream	Boredom	

129 rows × 61 columns

Filling missing values

```
[17] df.fillna(method='ffill',inplace=True)
```

```
[6] df
```

	GPA	Gender	breakfast	calories_chicken	calories_day	calories_scone	coffee	comfort_food	comfort_food_reasons	c
0	2.4	2	1	430	NaN	315.0	1	none	we dont have comfort	
1	3.654	1	1	610	3.0	420.0	2	chocolate, chips, ice cream	Stress, bored, anger	
2	3.3	1	1	720	4.0	420.0	2	frozen yogurt, pizza, fast food	stress, sadness	
3	3.2	1	1	430	3.0	420.0	2	Pizza, Mac and cheese, ice cream	Boredom	
4	3.5	1	1	720	2.0	420.0	2	Ice cream, chocolate, chips	Stress, boredom, cravings	
...	...	...	...	...	...	...	...	...	...	
124	3.9	1	1	430	4.0	315.0	2	Chocolates, pizza, and Ritz.	hormones, Premenstrual syndrome.	
125	2.4	2	1	430	4.0	315.0	1	none	we dont have comfort	
126	3.654	1	1	610	3.0	420.0	2	chocolate, chips, ice cream	Stress, bored, anger	
127	3.3	1	1	720	4.0	420.0	2	frozen yogurt, pizza, fast food	stress, sadness	
128	3.2	1	1	430	3.0	420.0	2	Pizza, Mac and cheese, ice cream	Boredom	

129 rows × 61 columns

Dropping observations

```
[7]  df=df.dropna(subset = ['father_profession'])
```

```
[8]  df.info()
```

```
 1   Gender                        129 non-null    int64
 2   breakfast                     129 non-null    int64
 3   calories_chicken              129 non-null    int64
 4   calories_day                  128 non-null    float64
 5   calories_scone                129 non-null    float64
 6   coffee                        129 non-null    int64
 7   comfort_food                  129 non-null    object
 8   comfort_food_reasons          129 non-null    object
 9   comfort_food_reasons_coded    129 non-null    float64
 10  cook                          129 non-null    float64
 11  comfort_food_reasons_coded.1  129 non-null    int64
 12  cuisine                       128 non-null    float64
 13  diet_current                  129 non-null    object
 14  diet_current_coded            129 non-null    int64
 15  drink                         129 non-null    float64
 16  eating_changes                129 non-null    object
 17  eating_changes_coded          129 non-null    int64
 18  eating_changes_coded1         129 non-null    int64
 19  eating_out                    129 non-null    int64
 20  employment                    129 non-null    float64
 21  ethnic_food                   129 non-null    int64
 22  exercise                      129 non-null    float64
 23  father_education              129 non-null    float64
 24  father_profession             129 non-null    object
 25  fav_cuisine                   129 non-null    object
 26  fav_cuisine_coded             129 non-null    int64
 27  fav_food                      129 non-null    float64
 28  food_childhood                129 non-null    object
 29  fries                         129 non-null    int64
 30  fruit_day                     129 non-null    int64
 31  grade_level                   129 non-null    int64
 32  greek_food                    129 non-null    int64
 33  healthy_feeling               129 non-null    int64
 34  healthy_meal                  129 non-null    object
 35  ideal_diet                    129 non-null    object
 36  ideal_diet_coded              129 non-null    int64
 37  income                        129 non-null    float64
 38  indian_food                   129 non-null    int64
 39  italian_food                  129 non-null    int64
 40  life_rewarding                129 non-null    float64
 41  marital_status                129 non-null    float64
```

```
[9] #remove columns
    df.drop('father_profession',axis=1,inplace=True)
```

```
[10] df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 129 entries, 0 to 128
Data columns (total 60 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   GPA                             129 non-null    object
 1   Gender                          129 non-null    int64
 2   breakfast                       129 non-null    int64
 3   calories_chicken                129 non-null    int64
 4   calories_day                    128 non-null    float64
 5   calories_scone                  129 non-null    float64
 6   coffee                          129 non-null    int64
 7   comfort_food                    129 non-null    object
 8   comfort_food_reasons            129 non-null    object
 9   comfort_food_reasons_coded      129 non-null    float64
 10  cook                            129 non-null    float64
 11  comfort_food_reasons_coded.1    129 non-null    int64
 12  cuisine                         128 non-null    float64
 13  diet_current                    129 non-null    object
 14  diet_current_coded              129 non-null    int64
 15  drink                           129 non-null    float64
 16  eating_changes                  129 non-null    object
 17  eating_changes_coded            129 non-null    int64
 18  eating_changes_coded1           129 non-null    int64
 19  eating_out                      129 non-null    int64
 20  employment                      129 non-null    float64
 21  ethnic_food                     129 non-null    int64
 22  exercise                        129 non-null    float64
 23  father_education                129 non-null    float64
 24  fav_cuisine                     129 non-null    object
 25  fav_cuisine_coded               129 non-null    int64
 26  fav_food                        129 non-null    float64
 27  food_childhood                  129 non-null    object
 28  fries                           129 non-null    int64
 29  fruit_day                       129 non-null    int64
 30  grade_level                     129 non-null    int64
 31  greek_food                      129 non-null    int64
 32  healthy_feeling                 129 non-null    int64
 33  healthy_meal                    129 non-null    object
 34  ideal_diet                      129 non-null    object
 35  ideal_diet_coded                129 non-null    int64
 36  income                          129 non-null    float64
 37  indian_food                     129 non-null    int64
 38  italian_food                    129 non-null    int64
 39  life_rewarding                  129 non-null    float64
```

```
[11] df.drop('mother_profession',axis=1,inplace=True)
     df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 129 entries, 0 to 128
Data columns (total 59 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   GPA                             129 non-null    object
 1   Gender                          129 non-null    int64
 2   breakfast                       129 non-null    int64
 3   calories_chicken                129 non-null    int64
 4   calories_day                    128 non-null    float64
 5   calories_scone                  129 non-null    float64
 6   coffee                          129 non-null    int64
 7   comfort_food                    129 non-null    object
 8   comfort_food_reasons            129 non-null    object
 9   comfort_food_reasons_coded      129 non-null    float64
 10  cook                            129 non-null    float64
 11  comfort_food_reasons_coded.1    129 non-null    int64
 12  cuisine                         128 non-null    float64
 13  diet_current                    129 non-null    object
 14  diet_current_coded              129 non-null    int64
 15  drink                           129 non-null    float64
 16  eating_changes                  129 non-null    object
 17  eating_changes_coded            129 non-null    int64
 18  eating_changes_coded1           129 non-null    int64
 19  eating_out                      129 non-null    int64
 20  employment                      129 non-null    float64
 21  ethnic_food                     129 non-null    int64
 22  exercise                        129 non-null    float64
 23  father_education                129 non-null    float64
 24  fav_cuisine                     129 non-null    object
 25  fav_cuisine_coded               129 non-null    int64
 26  fav_food                        129 non-null    float64
 27  food_childhood                  129 non-null    object
 28  fries                           129 non-null    int64
 29  fruit_day                       129 non-null    int64
 30  grade_level                     129 non-null    int64
 31  greek_food                      129 non-null    int64
 32  healthy_feeling                 129 non-null    int64
 33  healthy_meal                    129 non-null    object
 34  ideal_diet                      129 non-null    object
 35  ideal_diet_coded                129 non-null    int64
 36  income                          129 non-null    float64
 37  indian_food                     129 non-null    int64
 38  italian_food                    129 non-null    int64
 39  life_rewarding                  129 non-null    float64
 40  marital_status                  129 non-null    float64
```

```
df.isnull().sum()
```

```
GPA                                0
Gender                             0
breakfast                          0
calories_chicken                   0
calories_day                       1
calories_scone                     0
coffee                             0
comfort_food                       0
comfort_food_reasons               0
comfort_food_reasons_coded         0
cook                               0
comfort_food_reasons_coded.1       0
cuisine                            1
diet_current                       0
diet_current_coded                 0
drink                              0
eating_changes                     0
eating_changes_coded               0
eating_changes_coded1              0
eating_out                         0
employment                         0
ethnic_food                        0
exercise                           0
father_education                   0
fav_cuisine                        0
fav_cuisine_coded                  0
fav_food                           0
food_childhood                     0
fries                              0
fruit_day                          0
grade_level                        0
greek_food                         0
healthy_feeling                    0
healthy_meal                       0
ideal_diet                         0
ideal_diet_coded                   0
income                             0
indian_food                        0
italian_food                       0
life_rewarding                     0
marital_status                     0
meals_dinner_friend                0
mother_education                   0
nutritional_check                  0
on_off_campus                      0
parents_cook                       0
pay_meal_out                       0
persian_food                       0
self_percention_weight             0
```

Filling NA values:-

```
[13] df=df.fillna(df.mean().iloc[0])
     df
```

```
<ipython-input-13-5c49c0aee89f>:1: FutureWarning: The default value of numeric_only in DataFrame.mean is deprecated. In a
  df=df.fillna(df.mean().iloc[0])
```

	GPA	Gender	breakfast	calories_chicken	calories_day	calories_scone	coffee	comfort_food	comfort_food_reasons	c
0	2.4	2	1	430	1.387597	315.0	1	none	we dont have comfort	
1	3.654	1	1	610	3.000000	420.0	2	chocolate, chips, ice cream	Stress, bored, anger	
2	3.3	1	1	720	4.000000	420.0	2	frozen yogurt, pizza, fast food	stress, sadness	
3	3.2	1	1	430	3.000000	420.0	2	Pizza, Mac and cheese, ice cream	Boredom	
4	3.5	1	1	720	2.000000	420.0	2	Ice cream, chocolate, chips	Stress, boredom, cravings	
...	...	...	...	...	...	...	...	...	...	
124	3.9	1	1	430	4.000000	315.0	2	Chocolates, pizza, and Ritz.	hormones, Premenstrual syndrome.	
125	2.4	2	1	430	4.000000	315.0	1	none	we dont have comfort	
126	3.654	1	1	610	3.000000	420.0	2	chocolate, chips, ice cream	Stress, bored, anger	
127	3.3	1	1	720	4.000000	420.0	2	frozen yogurt, pizza, fast food	stress, sadness	
128	3.2	1	1	430	3.000000	420.0	2	Pizza, Mac and cheese, ice cream	Boredom	

Dropping duplicates:-

```
[14] #duplicates
     df_no_duplicates = df.drop_duplicates()
```

```
[15] df
```

	GPA	Gender	breakfast	calories_chicken	calories_day	calories_scone	coffee	comfort_food	comfort_food_reasons	c
0	2.4	2	1	430	1.387597	315.0	1	none	we dont have comfort	
1	3.654	1	1	610	3.000000	420.0	2	chocolate, chips, ice cream	Stress, bored, anger	
2	3.3	1	1	720	4.000000	420.0	2	frozen yogurt, pizza, fast food	stress, sadness	
3	3.2	1	1	430	3.000000	420.0	2	Pizza, Mac and cheese, ice cream	Boredom	
4	3.5	1	1	720	2.000000	420.0	2	Ice cream, chocolate, chips	Stress, boredom, cravings	
...	...	...	...	...	...	...	...	...	...	
124	3.9	1	1	430	4.000000	315.0	2	Chocolates, pizza, and Ritz.	hormones, Premenstrual syndrome.	
125	2.4	2	1	430	4.000000	315.0	1	none	we dont have comfort	
126	3.654	1	1	610	3.000000	420.0	2	chocolate, chips, ice cream	Stress, bored, anger	
127	3.3	1	1	720	4.000000	420.0	2	frozen yogurt, pizza, fast food	stress, sadness	
128	3.2	1	1	430	3.000000	420.0	2	Pizza, Mac and cheese, ice cream	Boredom	

129 rows × 59 columns

Removing Outliers:-

```
def remove_outliers(column):
    Q1 = column.quantile(0.25)
    Q3 = column.quantile(0.75)
    IQR = Q3 - Q1
    lower_bound = Q1 - 1.5 * IQR
    upper_bound = Q3 + 1.5 * IQR
    return column[(column >= lower_bound) & (column <= upper_bound)]
columns_to_process = [ 'cook', 'coffee']
for column in columns_to_process:
    df[column] = remove_outliers(df[column])
```

```
[19] df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 129 entries, 0 to 128
Data columns (total 61 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   GPA                             129 non-null    object
 1   Gender                          129 non-null    int64
 2   breakfast                       129 non-null    int64
 3   calories_chicken                129 non-null    int64
 4   calories_day                    128 non-null    float64
 5   calories_scone                  129 non-null    float64
 6   coffee                          97 non-null     float64
 7   comfort_food                    129 non-null    object
 8   comfort_food_reasons            129 non-null    object
 9   comfort_food_reasons_coded      129 non-null    float64
 10  cook                            121 non-null    float64
 11  comfort_food_reasons_coded.1    129 non-null    int64
 12  cuisine                         128 non-null    float64
 13  diet_current                    129 non-null    object
 14  diet_current_coded              129 non-null    int64
 15  drink                           129 non-null    float64
 16  eating_changes                  129 non-null    object
 17  eating_changes_coded            129 non-null    int64
 18  eating_changes_coded1           129 non-null    int64
 19  eating_out                      129 non-null    int64
 20  employment                      129 non-null    float64
 21  ethnic_food                     129 non-null    int64
 22  exercise                        129 non-null    float64
 23  father_education                129 non-null    float64
 24  father_profession               129 non-null    object
 25  fav_cuisine                     129 non-null    object
 26  fav_cuisine_coded               129 non-null    int64
 27  fav_food                        129 non-null    float64
 28  food_childhood                  129 non-null    object
 29  fries                           129 non-null    int64
 30  fruit_day                       129 non-null    int64
 31  grade_level                     129 non-null    int64
 32  greek_food                      129 non-null    int64
 33  healthy_feeling                 129 non-null    int64
 34  healthy_meal                    129 non-null    object
 35  ideal_diet                      129 non-null    object
 36  ideal_diet_coded                129 non-null    int64
 37  income                          129 non-null    float64
 38  indian_food                     129 non-null    int64
 39  italian_food                    129 non-null    int64
 40  life_rewarding                  129 non-null    float64
 41  marital_status                  129 non-null    float64
 42  meals_dinner_friend             129 non-null    object
 43  mother_education                129 non-null    float64
```

```
df.describe()
```

	Gender	breakfast	calories_chicken	calories_day	calories_scone	coffee	comfort_food_reasons_coded	coo
count	129.000000	129.000000	129.000000	128.000000	129.000000	97.0	129.000000	121.00000
mean	1.387597	1.108527	576.395349	3.031250	501.124031	2.0	2.612403	2.62809
std	0.489101	0.312258	131.090642	0.651316	227.373110	0.0	1.901110	0.88630
min	1.000000	1.000000	265.000000	2.000000	315.000000	2.0	1.000000	1.00000
25%	1.000000	1.000000	430.000000	3.000000	420.000000	2.0	2.000000	2.00000
50%	1.000000	1.000000	610.000000	3.000000	420.000000	2.0	2.000000	3.00000
75%	2.000000	1.000000	720.000000	3.000000	420.000000	2.0	3.000000	3.00000
max	2.000000	2.000000	720.000000	4.000000	980.000000	2.0	9.000000	4.00000

8 rows × 47 columns