❖ Day-8:-Task:-Data Preprocessing

❖ Data Preprocessing:

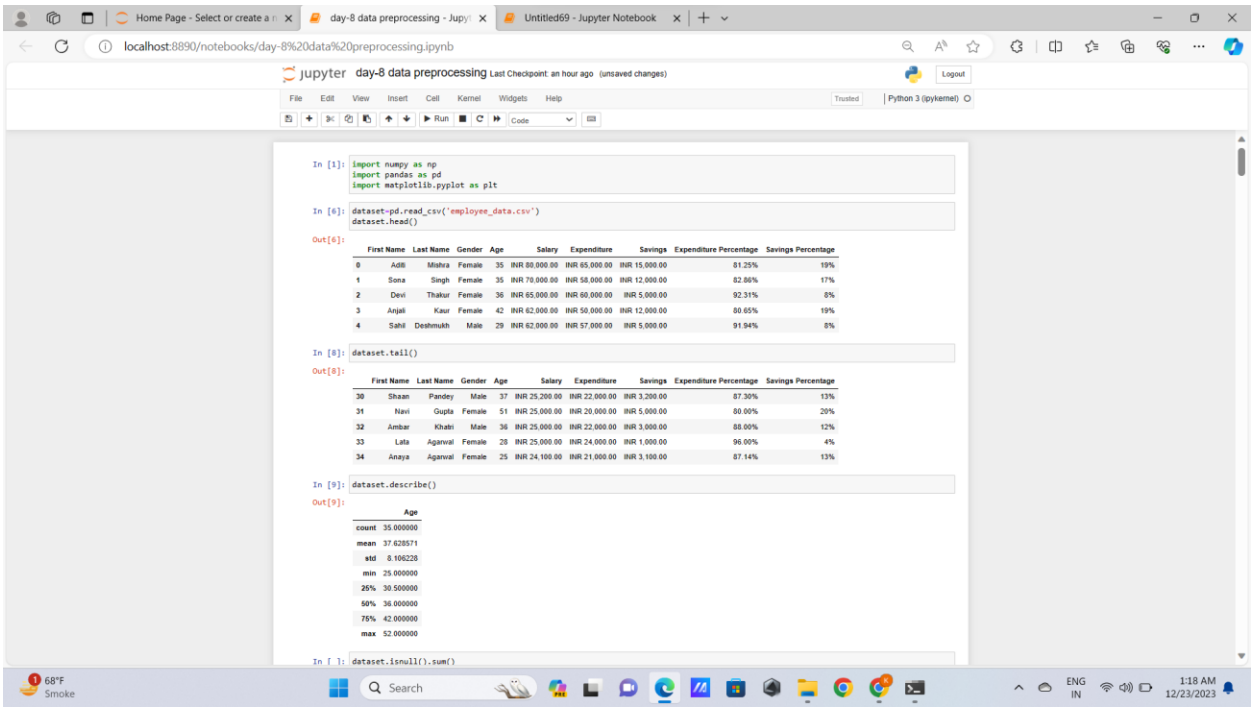
Data preprocessing is a crucial step in the data analysis and machine learning pipeline. It involves cleaning and transforming raw data into a format that is suitable for analysis or model training.

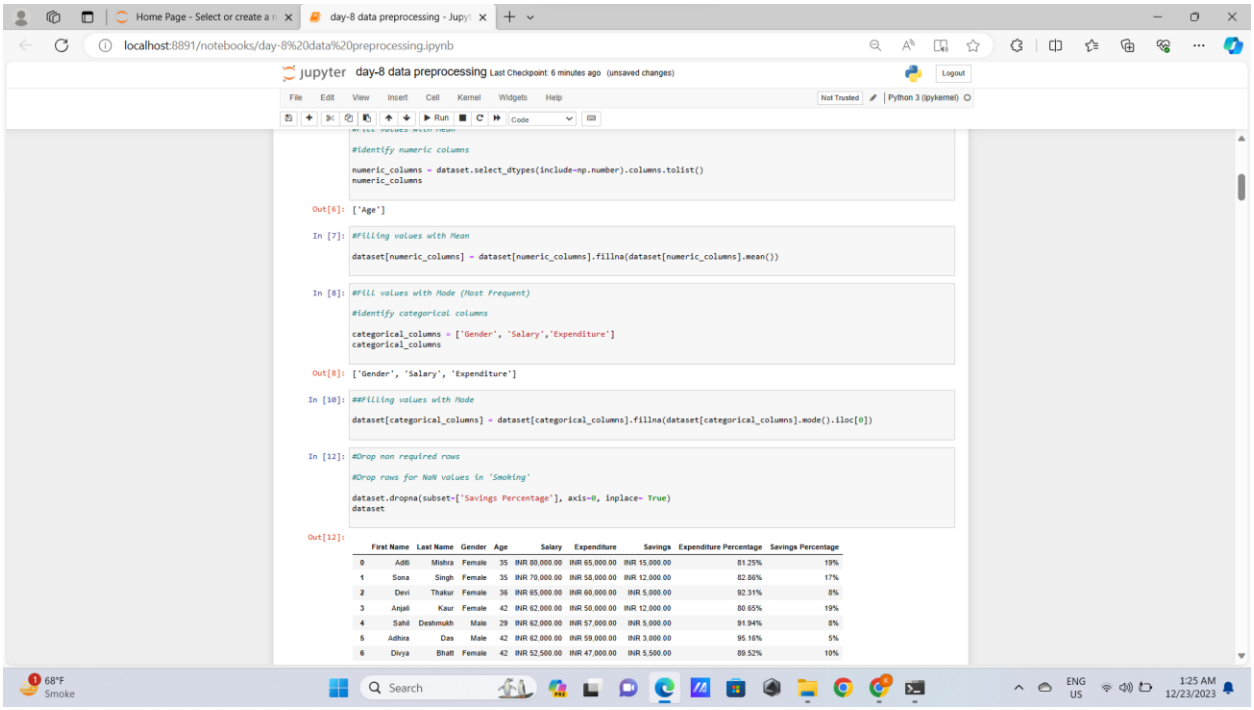
❖ Data processing working involves:

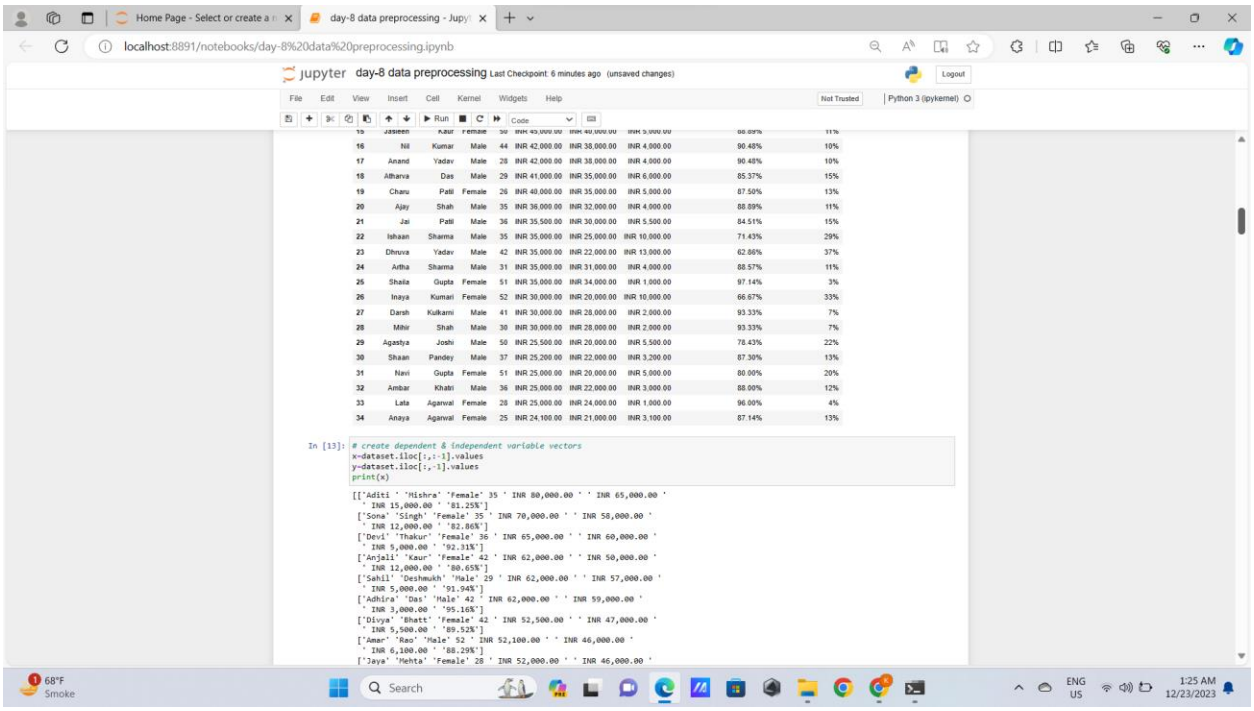
I. Handling Missing Data: Identify and deal with missing values.
II. Handling Duplicate Data: Detect and remove duplicate records.
III. Handling outliers: Address outliers through methods like z-score or IQR.
IV. Feature Scaling: standardize or normalize numerical features.
V. Categorical Variable Encoding: convert categorical variable into numerical format.
VI. Handling Text data: Transform text data for analysis or modeling.
VII. Feature Engineering: create or transform features for improved performance.
VIII. Data splitting: divide the dataset into training and testing sets.
IX. Data Transformation: Apply necessary transformations for modeling.
X. Handling imbalanced data: Address class imbalance if needed.
XI. Data Visualization: visualize data for insights.

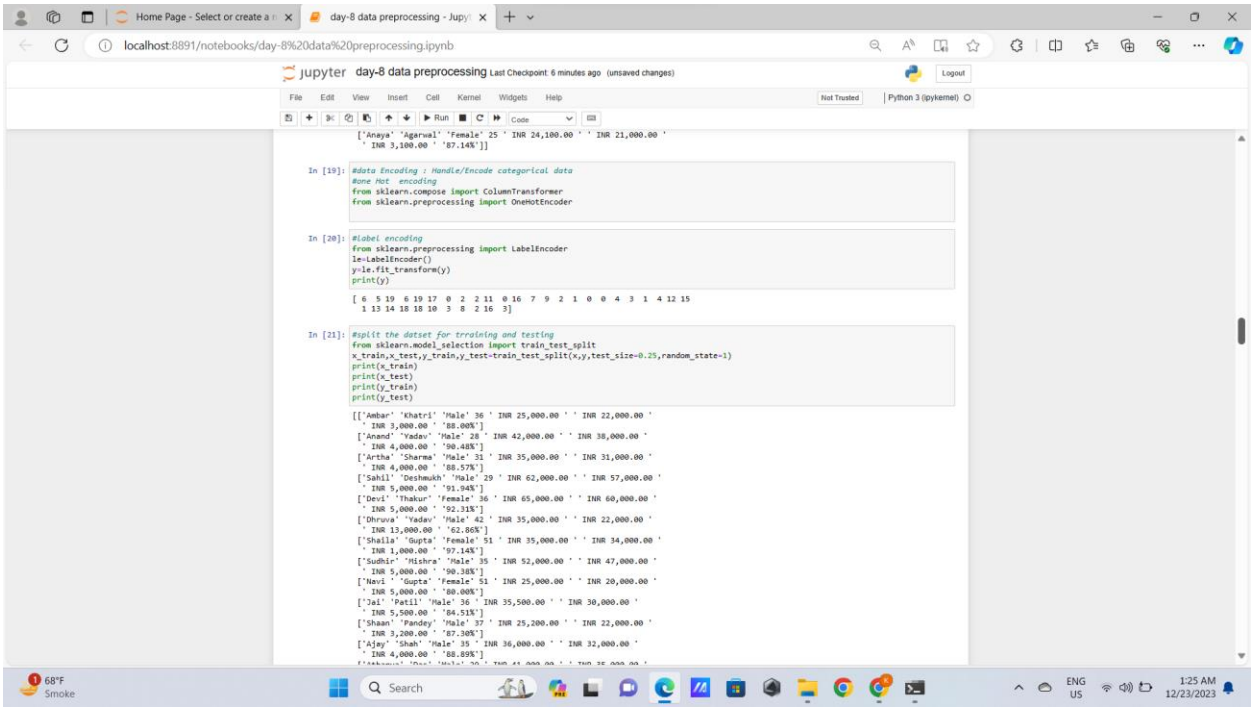
❖Description:

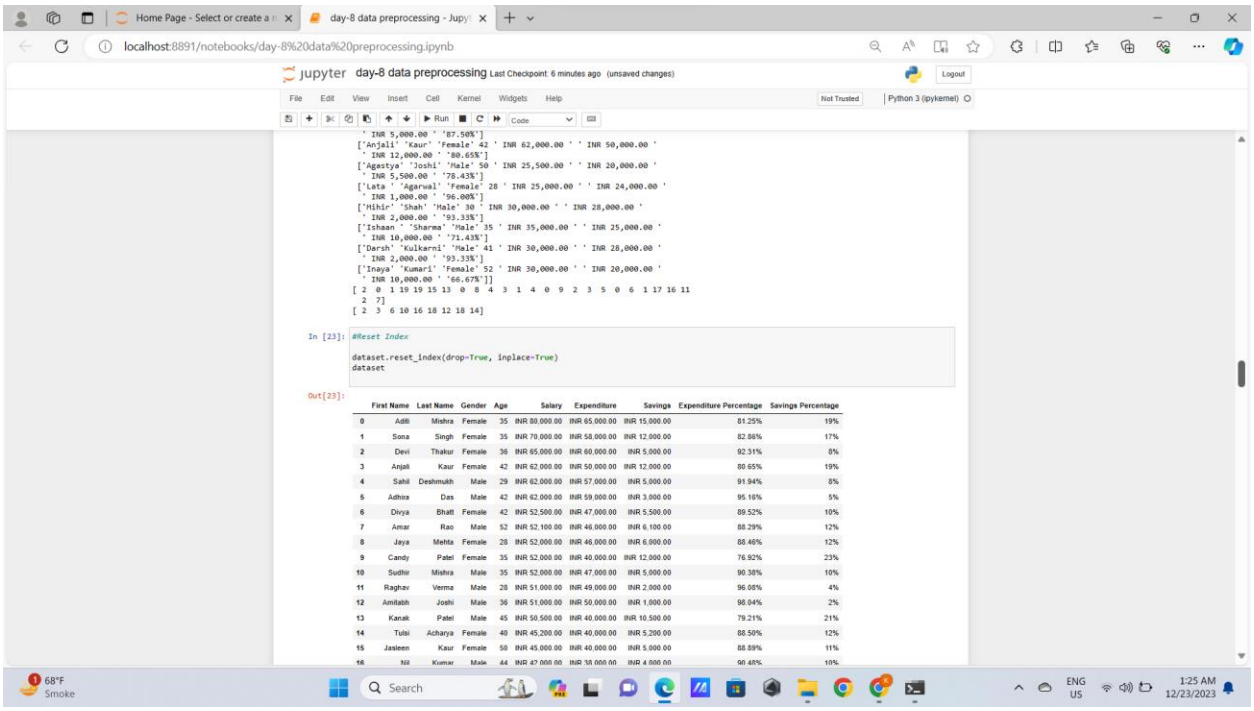
I'm thrilled to apply this newfound knowledge to real-world scenarios, and the prospect of further honing these skills through the Skill Boost Internship Program adds an extra layer of excitement. Here's to the journey ahead and the exciting challenges awaiting in the Skill Boost Internship Program(www.Batweb.com).

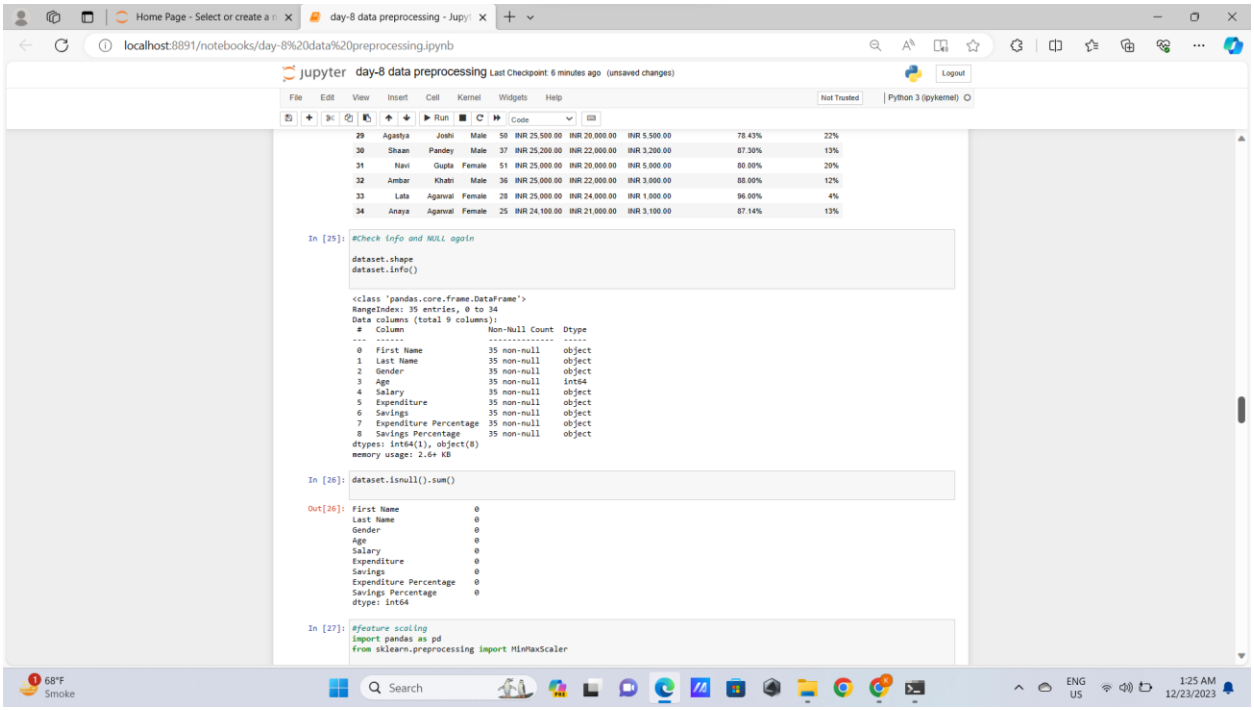

Home Page - Select or create a n
day-8 data preprocessing - Jupyt
Untitled69 - Jupyter Notebook
localhost:8890/notebooks/day-8%20data%20preprocessing.ipynb
jupyter day-8 data preprocessing Last Checkpoint: an hour ago (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code
In [1]:
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
In [6]:
dataset=pd.read_csv('employee_data.csv')
dataset.head()
Out[6]:
First Name Last Name Gender Age Salary Expenditure Savings Expenditure Percentage Savings Percentage
0 Aditi Mishra Female 35 INR 80,000.00 INR 65,000.00 INR 15,000.00 81.25% 19%
1 Sona Singh Female 35 INR 70,000.00 INR 58,000.00 INR 12,000.00 82.86% 17%
2 Devi Thakur Female 36 INR 65,000.00 INR 60,000.00 INR 5,000.00 92.31% 8%
3 Anjali Kaur Female 42 INR 62,000.00 INR 50,000.00 INR 12,000.00 80.65% 19%
4 Sahil Deshmukh Male 29 INR 62,000.00 INR 57,000.00 INR 5,000.00 91.94% 8%
In [8]:
dataset.tail()
Out[8]:
First Name Last Name Gender Age Salary Expenditure Savings Expenditure Percentage Savings Percentage
30 Shaan Pandey Male 37 INR 25,200.00 INR 22,000.00 INR 3,200.00 87.30% 13%
31 Navi Gupta Female 51 INR 25,000.00 INR 20,000.00 INR 5,000.00 80.00% 20%
32 Ambar Khatri Male 36 INR 25,000.00 INR 22,000.00 INR 3,000.00 88.00% 12%
33 Lata Agarwal Female 28 INR 25,000.00 INR 24,000.00 INR 1,000.00 96.00% 4%
34 Anaya Agarwal Female 25 INR 24,100.00 INR 21,000.00 INR 3,100.00 87.14% 13%
In [9]:
dataset.describe()
Out[9]:
Age
count 35.000000
mean 37.628571
std 8.106228
min 25.000000
25% 30.500000
50% 36.000000
75% 42.000000
max 52.000000
In [ ]:
dataset.isnull().sum()
68°F
Smoke
Search
ENG
IN
1:18 AM
12/23/2023


Home Page - Select or create a n
day-8 data preprocessing - Jupyt
localhost:8891/notebooks/day-8%20data%20preprocessing.ipynb
jupyter day-8 data preprocessing Last Checkpoint: 6 minutes ago (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Code
#identify numeric columns
numeric_columns = dataset.select_dtypes(include=np.number).columns.tolist()
numeric_columns
Out[6]:
['Age']
In [7]:
#Filling values with Mean
dataset[numeric_columns] = dataset[numeric_columns].fillna(dataset[numeric_columns].mean())
In [8]:
#Fill values with Mode (Most Frequent)
#identify categorical columns
categorical_columns = ['Gender', 'Salary','Expenditure']
categorical_columns
Out[8]:
['Gender', 'Salary', 'Expenditure']
In [10]:
##Filling values with Mode
dataset[categorical_columns] = dataset[categorical_columns].fillna(dataset[categorical_columns].mode().iloc[0])
In [12]:
#Drop non required rows
#Drop rows for NaN values in 'Smoking'
dataset.dropna(subset=['Savings Percentage'], axis=0, inplace= True)
dataset
Out[12]:
First Name Last Name Gender Age Salary Expenditure Savings Expenditure Percentage Savings Percentage
0 Aditi Mishra Female 35 INR 80,000.00 INR 65,000.00 INR 15,000.00 81.25% 19%
1 Sona Singh Female 35 INR 70,000.00 INR 58,000.00 INR 12,000.00 82.86% 17%
2 Devi Thakur Female 36 INR 65,000.00 INR 60,000.00 INR 5,000.00 92.31% 8%
3 Anjali Kaur Female 42 INR 62,000.00 INR 50,000.00 INR 12,000.00 80.65% 19%
4 Sahil Deshmukh Male 29 INR 62,000.00 INR 57,000.00 INR 5,000.00 91.94% 8%
5 Adhira Das Male 42 INR 62,000.00 INR 59,000.00 INR 3,000.00 95.16% 5%
6 Divya Bhatt Female 42 INR 52,500.00 INR 47,000.00 INR 5,500.00 89.52% 10%
68°F
Smoke
Search
ENG
US
1:25 AM
12/23/2023


Home Page - Select or create a
day-8 data preprocessing - Jupy
localhost:8891/notebooks/day-8%20data%20preprocessing.ipynb
jupyter day-8 data preprocessing Last Checkpoint: 6 minutes ago (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Code
16 Nil Kumar Male 44 INR 42,000.00 INR 38,000.00 INR 4,000.00 90.48% 10%
17 Anand Yadav Male 28 INR 42,000.00 INR 38,000.00 INR 4,000.00 90.48% 10%
18 Atharva Das Male 29 INR 41,000.00 INR 35,000.00 INR 6,000.00 85.37% 15%
19 Charu Patil Female 26 INR 40,000.00 INR 35,000.00 INR 5,000.00 87.50% 13%
20 Ajay Shah Male 35 INR 36,000.00 INR 32,000.00 INR 4,000.00 88.89% 11%
21 Jai Patil Male 36 INR 35,500.00 INR 30,000.00 INR 5,500.00 84.51% 15%
22 Ishaan Sharma Male 35 INR 35,000.00 INR 25,000.00 INR 10,000.00 71.43% 29%
23 Dhruva Yadav Male 42 INR 35,000.00 INR 22,000.00 INR 13,000.00 62.86% 37%
24 Artha Sharma Male 31 INR 35,000.00 INR 31,000.00 INR 4,000.00 88.57% 11%
25 Shaila Gupta Female 51 INR 35,000.00 INR 34,000.00 INR 1,000.00 97.14% 3%
26 Inaya Kumari Female 52 INR 30,000.00 INR 20,000.00 INR 10,000.00 66.67% 33%
27 Darsh Kulkarni Male 41 INR 30,000.00 INR 28,000.00 INR 2,000.00 93.33% 7%
28 Mihir Shah Male 30 INR 30,000.00 INR 28,000.00 INR 2,000.00 93.33% 7%
29 Agastya Joshi Male 50 INR 25,500.00 INR 20,000.00 INR 5,500.00 78.43% 22%
30 Shaan Pandey Male 37 INR 25,200.00 INR 22,000.00 INR 3,200.00 87.30% 13%
31 Navi Gupta Female 51 INR 25,000.00 INR 20,000.00 INR 5,000.00 80.00% 20%
32 Ambar Khatri Male 36 INR 25,000.00 INR 22,000.00 INR 3,000.00 88.00% 12%
33 Lata Agarwal Female 28 INR 25,000.00 INR 24,000.00 INR 1,000.00 96.00% 4%
34 Anaya Agarwal Female 25 INR 24,100.00 INR 21,000.00 INR 3,100.00 87.14% 13%
In [13]:
# create dependent & independent variable vectors
x=dataset.iloc[:,:-1].values
y=dataset.iloc[:,-1].values
print(x)
[['Aditi ' 'Hishra' 'Female' 35 ' INR 80,000.00 ' ' INR 65,000.00 '
' INR 15,000.00 ' '81.25%']
['Sona' 'Singh' 'Female' 35 ' INR 70,000.00 ' ' INR 58,000.00 '
' INR 12,000.00 ' '82.86%']
['Devi' 'Thakur' 'Female' 36 ' INR 65,000.00 ' ' INR 60,000.00 '
' INR 5,000.00 ' '92.31%']
['Anjali' 'Kaur' 'Female' 42 ' INR 62,000.00 ' ' INR 50,000.00 '
' INR 12,000.00 ' '80.65%']
['Sahil' 'Deshmukh' 'Male' 29 ' INR 62,000.00 ' ' INR 57,000.00 '
' INR 5,000.00 ' '91.94%']
['Adhira' 'Das' 'Male' 42 ' INR 62,000.00 ' ' INR 59,000.00 '
' INR 3,000.00 ' '95.16%']
['Divya' 'Bhatt' 'Female' 42 ' INR 52,500.00 ' ' INR 47,000.00 '
' INR 5,500.00 ' '89.52%']
['Amar' 'Rao' 'Male' 52 ' INR 52,100.00 ' ' INR 46,000.00 '
' INR 6,100.00 ' '88.29%']
['Jaya' 'Mehta' 'Female' 28 ' INR 52,000.00 ' ' INR 46,000.00 '
68°F Smoke
Search
ENG US
1:25 AM 12/23/2023


Home Page - Select or create a
day-8 data preprocessing - Jupy
localhost:8891/notebooks/day-8%20data%20preprocessing.ipynb
jupyter day-8 data preprocessing Last Checkpoint: 6 minutes ago (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Code
['Anaya' 'Agarwal' 'Female' 25 ' INR 24,100.00 ' ' INR 21,000.00 '
' INR 3,100.00 ' '87.14%']]
In [19]:
#data Encoding : Handle/Encode categorical data
#one Hot encoding
from sklearn.compose import ColumnTransformer
from sklearn.preprocessing import OneHotEncoder
In [20]:
#label encoding
from sklearn.preprocessing import LabelEncoder
le=LabelEncoder()
y=le.fit_transform(y)
print(y)
[6 5 19 6 19 17 0 2 2 11 0 16 7 9 2 1 0 0 4 3 1 4 12 15
1 13 14 18 18 10 3 8 2 16 3]
In [21]:
#split the dataset for trraining and testing
from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.25,random_state=1)
print(x_train)
print(x_test)
print(y_train)
print(y_test)
[['Ambar' 'Khatri' 'Male' 36 ' INR 25,000.00 ' ' INR 22,000.00 '
' INR 3,000.00 ' '88.00%']
['Anand' 'Yadav' 'Male' 28 ' INR 42,000.00 ' ' INR 38,000.00 '
' INR 4,000.00 ' '90.48%']
['Artha' 'Sharma' 'Male' 31 ' INR 35,000.00 ' ' INR 31,000.00 '
' INR 4,000.00 ' '88.57%']
['Sahil' 'Deshmukh' 'Male' 29 ' INR 62,000.00 ' ' INR 57,000.00 '
' INR 5,000.00 ' '91.94%']
['Devi' 'Thakur' 'Female' 36 ' INR 65,000.00 ' ' INR 60,000.00 '
' INR 5,000.00 ' '92.31%']
['Dhruva' 'Yadav' 'Male' 42 ' INR 35,000.00 ' ' INR 22,000.00 '
' INR 13,000.00 ' '62.86%']
['Shaila' 'Gupta' 'Female' 51 ' INR 35,000.00 ' ' INR 34,000.00 '
' INR 1,000.00 ' '97.14%']
['Sudhir' 'Hishra' 'Male' 35 ' INR 52,000.00 ' ' INR 47,000.00 '
' INR 5,000.00 ' '90.38%']
['Navi' 'Gupta' 'Female' 51 ' INR 25,000.00 ' ' INR 20,000.00 '
' INR 5,000.00 ' '80.00%']
['Jai' 'Patil' 'Male' 36 ' INR 35,500.00 ' ' INR 30,000.00 '
' INR 5,500.00 ' '84.51%']
['Shaan' 'Pandey' 'Male' 37 ' INR 25,200.00 ' ' INR 22,000.00 '
' INR 3,200.00 ' '87.30%']
['Ajay' 'Shah' 'Male' 35 ' INR 36,000.00 ' ' INR 32,000.00 '
' INR 4,000.00 ' '88.89%']
68°F Smoke
Search
ENG US
1:25 AM 12/23/2023


Home Page - Select or create a n
day-8 data preprocessing - Jupyt
localhost:8891/notebooks/day-8%20data%20preprocessing.ipynb
jupyter day-8 data preprocessing Last Checkpoint: 6 minutes ago (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Code
' INR 5,000.00 ' '87.50%']
['Anjali' 'Kaur' 'Female' 42 ' INR 62,000.00 ' ' INR 50,000.00 '
' INR 12,000.00 ' '80.65%']
['Agastya' 'Joshi' 'Male' 50 ' INR 25,500.00 ' ' INR 20,000.00 '
' INR 5,500.00 ' '78.43%']
['Lata ' 'Agarwal' 'Female' 28 ' INR 25,000.00 ' ' INR 24,000.00 '
' INR 1,000.00 ' '96.00%']
['Mihir' 'Shah' 'Male' 30 ' INR 30,000.00 ' ' INR 28,000.00 '
' INR 2,000.00 ' '93.33%']
['Ishaan ' 'Sharma' 'Male' 35 ' INR 35,000.00 ' ' INR 25,000.00 '
' INR 10,000.00 ' '71.43%']
['Darsh' 'Kulkarni' 'Male' 41 ' INR 30,000.00 ' ' INR 28,000.00 '
' INR 2,000.00 ' '93.33%']
['Inaya' 'Kumari' 'Female' 52 ' INR 30,000.00 ' ' INR 20,000.00 '
' INR 10,000.00 ' '66.67%']]
[2 0 1 19 19 15 13 0 8 4 3 1 4 0 9 2 3 5 0 6 1 17 16 11
2 7]
[2 3 6 10 16 18 12 18 14]
In [23]:
#Reset Index
dataset.reset_index(drop=True, inplace=True)
dataset
Out[23]:
First Name Last Name Gender Age Salary Expenditure Savings Expenditure Percentage Savings Percentage
0 Aditi Mishra Female 35 INR 80,000.00 INR 65,000.00 INR 15,000.00 81.25% 19%
1 Sona Singh Female 35 INR 70,000.00 INR 58,000.00 INR 12,000.00 82.86% 17%
2 Devi Thakur Female 36 INR 65,000.00 INR 60,000.00 INR 5,000.00 92.31% 8%
3 Anjali Kaur Female 42 INR 62,000.00 INR 50,000.00 INR 12,000.00 80.65% 19%
4 Sahil Deshmukh Male 29 INR 62,000.00 INR 57,000.00 INR 5,000.00 91.94% 8%
5 Adhira Das Male 42 INR 62,000.00 INR 59,000.00 INR 3,000.00 95.16% 5%
6 Divya Bhatt Female 42 INR 52,500.00 INR 47,000.00 INR 5,500.00 89.52% 10%
7 Amar Rao Male 52 INR 52,100.00 INR 46,000.00 INR 6,100.00 88.29% 12%
8 Jaya Mehta Female 28 INR 52,000.00 INR 46,000.00 INR 6,000.00 88.46% 12%
9 Candy Patel Female 35 INR 52,000.00 INR 40,000.00 INR 12,000.00 76.92% 23%
10 Sudhir Mishra Male 35 INR 52,000.00 INR 47,000.00 INR 5,000.00 90.38% 10%
11 Raghav Verma Male 28 INR 51,000.00 INR 49,000.00 INR 2,000.00 96.08% 4%
12 Amitabh Joshi Male 36 INR 51,000.00 INR 50,000.00 INR 1,000.00 98.04% 2%
13 Kanak Patel Male 45 INR 50,500.00 INR 40,000.00 INR 10,500.00 79.21% 21%
14 Tulsi Acharya Female 40 INR 45,200.00 INR 40,000.00 INR 5,200.00 88.50% 12%
15 Jasleen Kaur Female 50 INR 45,000.00 INR 40,000.00 INR 5,000.00 88.89% 11%
16 Nil Kumar Male 44 INR 47,000.00 INR 38,000.00 INR 4,000.00 90.48% 10%
68°F Smoke
Search
ENG US
1:25 AM 12/23/2023


Home Page - Select or create a n
day-8 data preprocessing - Jupyt
localhost:8891/notebooks/day-8%20data%20preprocessing.ipynb
jupyter day-8 data preprocessing Last Checkpoint: 6 minutes ago (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Code
29 Agastya Joshi Male 50 INR 25,500.00 INR 20,000.00 INR 5,500.00 78.43% 22%
30 Shaan Pandey Male 37 INR 25,200.00 INR 22,000.00 INR 3,200.00 87.30% 13%
31 Navi Gupta Female 51 INR 25,000.00 INR 20,000.00 INR 5,000.00 80.00% 20%
32 Ambar Khatri Male 36 INR 25,000.00 INR 22,000.00 INR 3,000.00 88.00% 12%
33 Lata Agarwal Female 28 INR 25,000.00 INR 24,000.00 INR 1,000.00 96.00% 4%
34 Anaya Agarwal Female 25 INR 24,100.00 INR 21,000.00 INR 3,100.00 87.14% 13%
In [25]:
#Check info and NULL again
dataset.shape
dataset.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 35 entries, 0 to 34
Data columns (total 9 columns):
Column Non-Null Count Dtype
--- ------ -------------- -----
0 First Name 35 non-null object
1 Last Name 35 non-null object
2 Gender 35 non-null object
3 Age 35 non-null int64
4 Salary 35 non-null object
5 Expenditure 35 non-null object
6 Savings 35 non-null object
7 Expenditure Percentage 35 non-null object
8 Savings Percentage 35 non-null object
dtypes: int64(1), object(8)
memory usage: 2.6+ KB
In [26]:
dataset.isnull().sum()
Out[26]:
First Name 0
Last Name 0
Gender 0
Age 0
Salary 0
Expenditure 0
Savings 0
Expenditure Percentage 0
Savings Percentage 0
dtype: int64
In [27]:
#feature scaling
import pandas as pd
from sklearn.preprocessing import MinMaxScaler
68°F Smoke
Search
ENG US
1:25 AM 12/23/2023

Home Page - Select or create a n | day-8 data preprocessing - Jupyt

localhost:8891/notebooks/day-8%20data%20preprocessing.ipynb

jupyter **day-8 data preprocessing** Last Checkpoint: 7 minutes ago (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

Run Code

```
import pandas as pd
from sklearn.preprocessing import MinMaxScaler

# Sample data
data = {
    'First Name': ['Aditi', 'Sona', 'Devi', 'Anjali', 'Sahil'],
    'Last Name': ['Mishra', 'Singh', 'Thakur', 'Kaur', 'Deshmukh'],
    'Gender': ['Female', 'Female', 'Female', 'Female', 'Male'],
    'Age': [35, 35, 36, 42, 29],
    'Salary': [80000, 70000, 65000, 62000, 62000],
    'Expenditure': [65000, 58000, 60000, 50000, 57000],
    'Savings': [15000, 12000, 5000, 12000, 5000],
    'Expenditure Percentage': [81.25, 82.86, 92.31, 80.65, 91.94],
    'Savings Percentage': [19, 17, 8, 19, 8]
}

df = pd.DataFrame(data)

# Selecting numerical features for scaling
numerical_features = ['Age', 'Salary', 'Expenditure', 'Savings', 'Expenditure Percentage', 'Savings Percentage']

# Initialize the MinMaxScaler
scaler = MinMaxScaler()

# Fit and transform the scaler on the selected numerical features
df[numerical_features] = scaler.fit_transform(df[numerical_features])

# Display the scaled data
print(df)
```

```
  First Name Last Name  Gender       Age    Salary  Expenditure  Savings  \
0      Aditi    Mishra  Female  0.461538  1.000000     1.000000      1.0
1       Sona     Singh  Female  0.461538  0.444444     0.533333      0.7
2       Devi    Thakur  Female  0.538462  0.166667     0.666667      0.0
3     Anjali      Kaur  Female  1.000000  0.000000     0.000000      0.7
4      Sahil  Deshmukh    Male  0.000000  0.000000     0.466667      0.0

   Expenditure Percentage  Savings Percentage
0                0.051458            1.000000
1                0.189537            0.818182
2                1.000000            0.000000
3                0.000000            1.000000
4                0.968268            0.000000
```

In [28]:
```
plt.hist(dataset['Age'],bins=15)
plt.show()
```

68°F Smoke | Search | ENG US | 1:25 AM 12/23/2023

Home Page - Select or create a n | day-8 data preprocessing - Jupyt

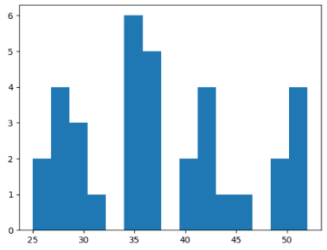
localhost:8891/notebooks/day-8%20data%20preprocessing.ipynb

jupyter **day-8 data preprocessing** Last Checkpoint: 7 minutes ago (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

Run Code

In [28]:
```
plt.hist(dataset['Age'],bins=15)
plt.show()
```



In [29]:
```
#quantile
lowerlimit=dataset['Age'].quantile(0.05)
lowerlimit
dataset[dataset['Age']<lowerLimit]
```

Out[29]:

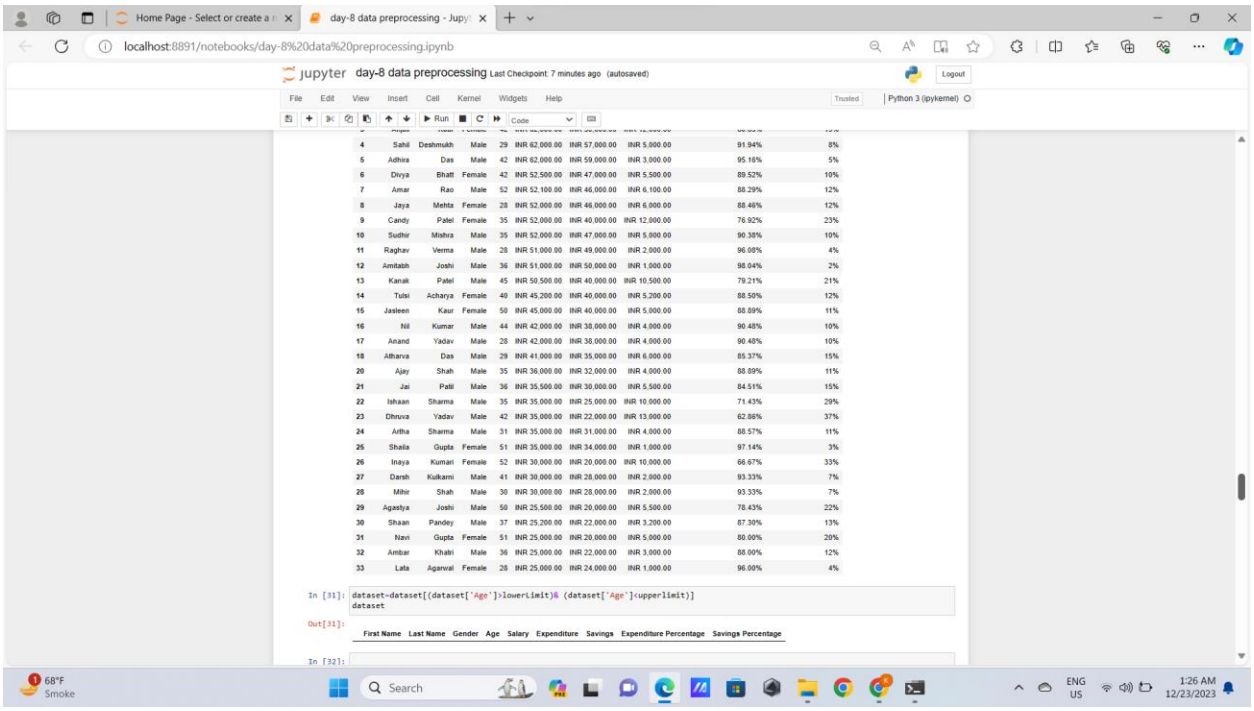
	First Name	Last Name	Gender	Age	Salary	Expenditure	Savings	Expenditure Percentage	Savings Percentage
19	Charu	Patil	Female	26	INR 40,000.00	INR 35,000.00	INR 5,000.00	87.50%	13%
34	Anaya	Agarwal	Female	25	INR 24,100.00	INR 21,000.00	INR 3,100.00	87.14%	13%

In [30]:
```
upperlimit=dataset['Age'].quantile(0.05)
upperlimit
dataset[dataset['Age']>upperlimit]
```

Out[30]:

	First Name	Last Name	Gender	Age	Salary	Expenditure	Savings	Expenditure Percentage	Savings Percentage
0	Aditi	Mishra	Female	35	INR 80,000.00	INR 65,000.00	INR 15,000.00	81.25%	19%
1	Sona	Singh	Female	35	INR 70,000.00	INR 58,000.00	INR 12,000.00	82.86%	17%
2	Devi	Thakur	Female	36	INR 65,000.00	INR 60,000.00	INR 5,000.00	92.31%	8%

68°F Smoke | Search | ENG US | 1:26 AM 12/23/2023



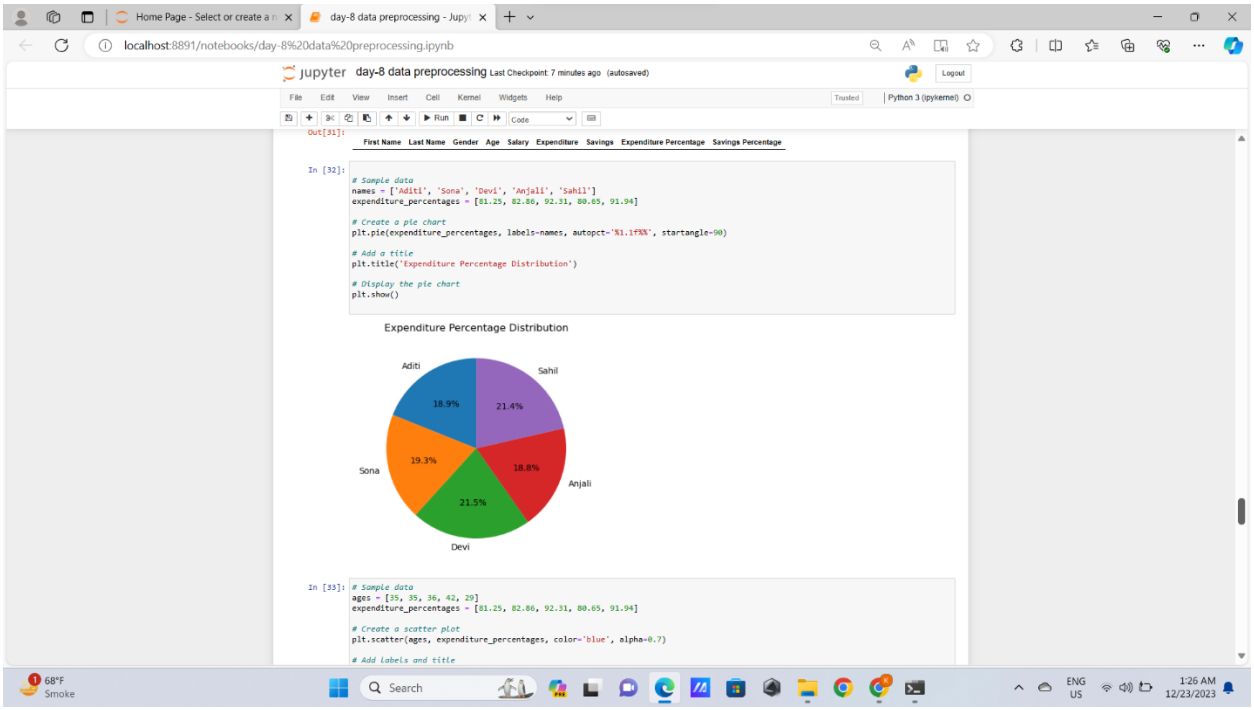
	First Name	Last Name	Gender	Age	Salary	Expenditure	Savings	Expenditure Percentage	Savings Percentage
4	Sahil	Deshmukh	Male	29	INR 62,000.00	INR 57,000.00	INR 5,000.00	91.94%	8%
5	Adhira	Das	Male	42	INR 62,000.00	INR 59,000.00	INR 3,000.00	95.16%	5%
6	Divya	Bhatt	Female	42	INR 52,500.00	INR 47,000.00	INR 5,500.00	89.52%	10%
7	Amar	Rao	Male	52	INR 52,100.00	INR 46,000.00	INR 6,100.00	88.29%	12%
8	Jaya	Mehta	Female	28	INR 52,000.00	INR 46,000.00	INR 6,000.00	88.46%	12%
9	Candy	Patel	Female	35	INR 52,000.00	INR 40,000.00	INR 12,000.00	76.92%	23%
10	Sudhir	Mishra	Male	35	INR 52,000.00	INR 47,000.00	INR 5,000.00	90.38%	10%
11	Raghav	Verma	Male	28	INR 51,000.00	INR 49,000.00	INR 2,000.00	96.08%	4%
12	Amitabh	Joshi	Male	36	INR 51,000.00	INR 50,000.00	INR 1,000.00	98.04%	2%
13	Kanak	Patel	Male	45	INR 50,500.00	INR 40,000.00	INR 10,500.00	79.21%	21%
14	Tulsi	Acharya	Female	40	INR 45,200.00	INR 40,000.00	INR 5,200.00	88.50%	12%
15	Jasleen	Kaur	Female	50	INR 45,000.00	INR 40,000.00	INR 5,000.00	88.89%	11%
16	Nil	Kumar	Male	44	INR 42,000.00	INR 38,000.00	INR 4,000.00	90.48%	10%
17	Anand	Yadav	Male	28	INR 42,000.00	INR 38,000.00	INR 4,000.00	90.48%	10%
18	Atharva	Das	Male	29	INR 41,000.00	INR 35,000.00	INR 6,000.00	85.37%	15%
20	Ajay	Shah	Male	35	INR 36,000.00	INR 32,000.00	INR 4,000.00	88.89%	11%
21	Jai	Patil	Male	36	INR 35,500.00	INR 30,000.00	INR 5,500.00	84.51%	15%
22	Ishaan	Sharma	Male	35	INR 35,000.00	INR 25,000.00	INR 10,000.00	71.43%	29%
23	Dhruva	Yadav	Male	42	INR 35,000.00	INR 22,000.00	INR 13,000.00	62.86%	37%
24	Artha	Sharma	Male	31	INR 35,000.00	INR 31,000.00	INR 4,000.00	88.57%	11%
25	Shaila	Gupta	Female	51	INR 35,000.00	INR 34,000.00	INR 1,000.00	97.14%	3%
26	Inaya	Kumari	Female	52	INR 30,000.00	INR 20,000.00	INR 10,000.00	66.67%	33%
27	Darsh	Kulkarni	Male	41	INR 30,000.00	INR 28,000.00	INR 2,000.00	93.33%	7%
28	Mihir	Shah	Male	30	INR 30,000.00	INR 28,000.00	INR 2,000.00	93.33%	7%
29	Agastya	Joshi	Male	50	INR 25,500.00	INR 20,000.00	INR 5,500.00	78.43%	22%
30	Shaan	Pandey	Male	37	INR 25,200.00	INR 22,000.00	INR 3,200.00	87.30%	13%
31	Navi	Gupta	Female	51	INR 25,000.00	INR 20,000.00	INR 5,000.00	80.00%	20%
32	Ambar	Khatri	Male	36	INR 25,000.00	INR 22,000.00	INR 3,000.00	88.00%	12%
33	Lata	Agarwal	Female	28	INR 25,000.00	INR 24,000.00	INR 1,000.00	96.00%	4%

```
In [31]: dataset=dataset[(dataset['Age']>lowerLimit)& (dataset['Age']<upperlimit)]
         dataset
```

Out[31]:

First Name	Last Name	Gender	Age	Salary	Expenditure	Savings	Expenditure Percentage	Savings Percentage

In [32]:



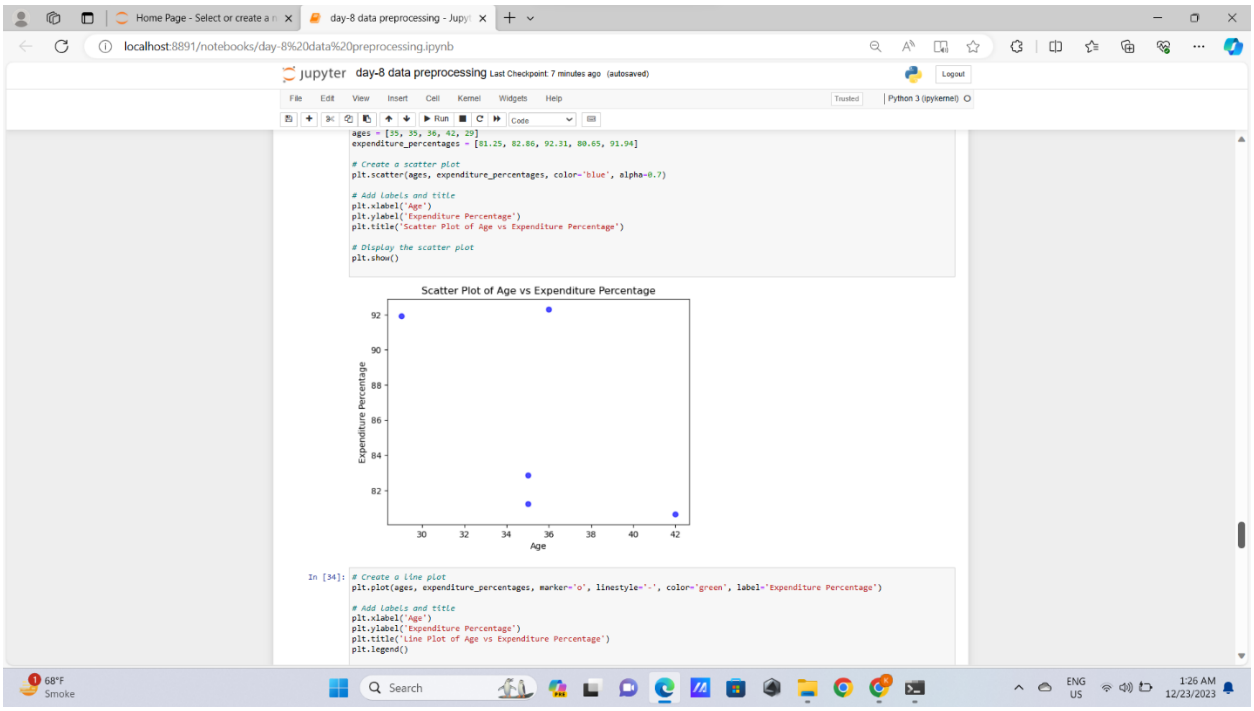
Out[31]:

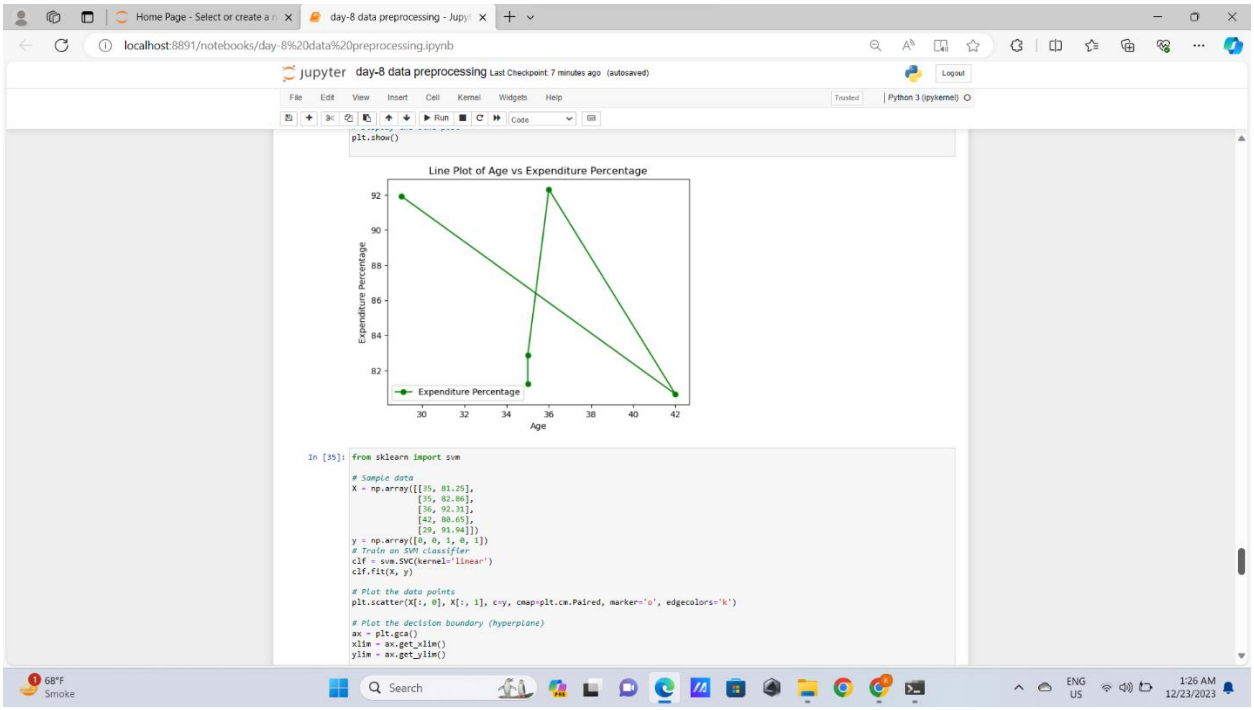
First Name	Last Name	Gender	Age	Salary	Expenditure	Savings	Expenditure Percentage	Savings Percentage

```
In [32]: # Sample data
names = ['Aditi', 'Sona', 'Devi', 'Anjali', 'Sahil']
expenditure_percentages = [81.25, 82.86, 92.31, 80.65, 91.94]

# Create a pie chart
plt.pie(expenditure_percentages, labels=names, autopct='%1.1f%%', startangle=90)

# Add a title
plt.title('Expenditure Percentage Distribution')

# Display the pie chart
plt.show()
```



```
In [33]: # Sample data
ages = [35, 35, 36, 42, 29]
expenditure_percentages = [81.25, 82.86, 92.31, 80.65, 91.94]

# Create a scatter plot
plt.scatter(ages, expenditure_percentages, color='blue', alpha=0.7)

# Add labels and title
```


Home Page - Select or create a n
day-8 data preprocessing - Jupyt
localhost:8891/notebooks/day-8%20data%20preprocessing.ipynb
jupyter day-8 data preprocessing Last Checkpoint: 7 minutes ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code
ages = [35, 35, 36, 42, 29]
expenditure_percentages = [81.25, 82.86, 92.31, 80.65, 91.94]
# Create a scatter plot
plt.scatter(ages, expenditure_percentages, color='blue', alpha=0.7)
# Add labels and title
plt.xlabel('Age')
plt.ylabel('Expenditure Percentage')
plt.title('Scatter Plot of Age vs Expenditure Percentage')
# Display the scatter plot
plt.show()
Scatter Plot of Age vs Expenditure Percentage
Expenditure Percentage
92
90
88
86
84
82
30
32
34
36
38
40
42
Age
In [34]:
# Create a line plot
plt.plot(ages, expenditure_percentages, marker='o', linestyle='-', color='green', label='Expenditure Percentage')
# Add labels and title
plt.xlabel('Age')
plt.ylabel('Expenditure Percentage')
plt.title('Line Plot of Age vs Expenditure Percentage')
plt.legend()
68°F
Smoke
Search
ENG
US
1:26 AM
12/23/2023


Home Page - Select or create a n
day-8 data preprocessing - Jupyt
localhost:8891/notebooks/day-8%20data%20preprocessing.ipynb
jupyter day-8 data preprocessing Last Checkpoint: 7 minutes ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code
plt.show()
Line Plot of Age vs Expenditure Percentage
Expenditure Percentage
92
90
88
86
84
82
30
32
34
36
38
40
42
Age
Expenditure Percentage
In [35]:
from sklearn import svm
# Sample data
X = np.array([[35, 81.25],
[35, 82.86],
[36, 92.31],
[42, 80.65],
[29, 91.94]])
y = np.array([0, 0, 1, 0, 1])
# Train an SVM classifier
clf = svm.SVC(kernel='linear')
clf.fit(X, y)
# Plot the data points
plt.scatter(X[:, 0], X[:, 1], c=y, cmap=plt.cm.Paired, marker='o', edgecolors='k')
# Plot the decision boundary (hyperplane)
ax = plt.gca()
xlim = ax.get_xlim()
ylim = ax.get_ylim()
68°F
Smoke
Search
ENG
US
1:26 AM
12/23/2023

Home Page - Select or create a n | day-8 data preprocessing - Jupyt

localhost:8891/notebooks/day-8%20data%20preprocessing.ipynb

jupyter **day-8 data preprocessing** Last Checkpoint: 8 minutes ago (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

Run Code

In [35]:
```
from sklearn import svm

# Sample data
X = np.array([[35, 81.25],
              [35, 82.86],
              [36, 92.31],
              [42, 80.65],
              [29, 91.94]])
y = np.array([0, 0, 1, 0, 1])
# Train an SVM classifier
clf = svm.SVC(kernel='linear')
clf.fit(X, y)

# Plot the data points
plt.scatter(X[:, 0], X[:, 1], c=y, cmap=plt.cm.Paired, marker='o', edgecolors='k')

# Plot the decision boundary (hyperplane)
ax = plt.gca()
xlim = ax.get_xlim()
ylim = ax.get_ylim()

# Create grid to evaluate model
xx, yy = np.meshgrid(np.linspace(xlim[0], xlim[1], 50),
                     np.linspace(ylim[0], ylim[1], 50))

# Decision function to get the hyperplane
Z = clf.decision_function(np.c_[xx.ravel(), yy.ravel()])
Z = Z.reshape(xx.shape)

# Plot decision boundary and margins
ax.contour(xx, yy, Z, colors='k', levels=[-1, 0, 1], alpha=0.5, linestyles=['--', '-', '--'])

# Highlight support vectors
ax.scatter(clf.support_vectors_[:, 0], clf.support_vectors_[:, 1], s=100,
           linewidth=1, facecolors='none', edgecolors='k')

# Add labels and title
plt.xlabel('Age')
plt.ylabel('Expenditure Percentage')
plt.title('Hyperplane Plot for SVM Classifier')

# Display the plot
plt.show()
```

68°F Smoke | Search | ENG US | 1:26 AM 12/23/2023

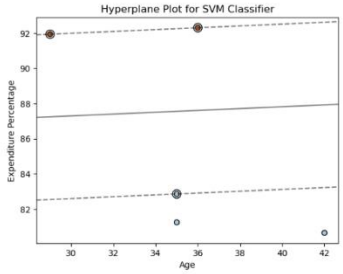
Home Page - Select or create a n | day-8 data preprocessing - Jupyt

localhost:8891/notebooks/day-8%20data%20preprocessing.ipynb

jupyter **day-8 data preprocessing** Last Checkpoint: 8 minutes ago (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

Run Code



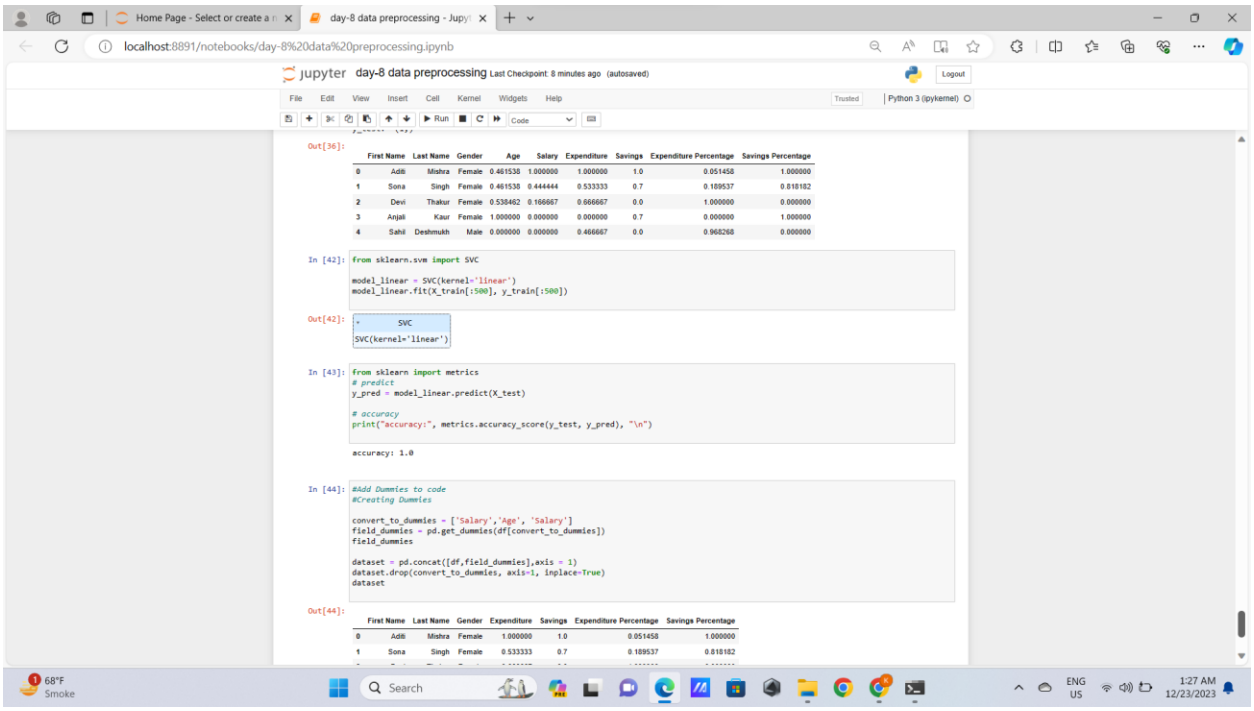
In [36]:
```
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=1)

print("X_train: ", X_train.shape)
print("X_test: ", X_test.shape)
print("y_train: ", y_train.shape)
print("y_test: ", y_test.shape)
df
```

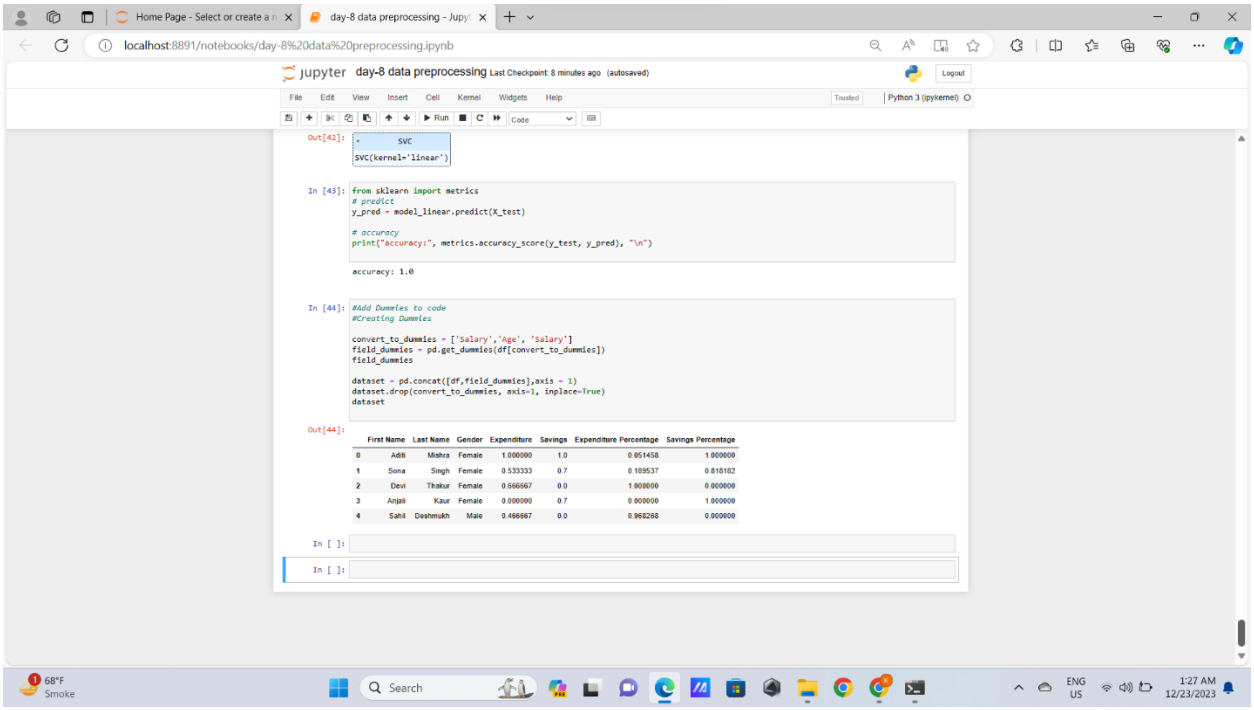
```
X_train:  (4, 2)
X_test:  (1, 2)
y_train:  (4,)
y_test:  (1,)
```

Out[36]:

	First Name	Last Name	Gender	Age	Salary	Expenditure	Savings	Expenditure Percentage	Savings Percentage
0	Aditi	Mishra	Female	0.461538	1.000000	1.000000	1.0	0.051455	1.000000
1	Sona	Singh	Female	0.461538	0.444444	0.533333	0.7	0.189537	0.818182
2	Devi	Thakur	Female	0.538462	0.166667	0.866667	0.0	1.000000	0.000000

68°F Smoke | Search | ENG US | 1:26 AM 12/23/2023


Home Page - Select or create a n
day-8 data preprocessing - Jupyt
localhost:8891/notebooks/day-8%20data%20preprocessing.ipynb
jupyter day-8 data preprocessing Last Checkpoint: 8 minutes ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code
Out[36]:
First Name Last Name Gender Age Salary Expenditure Savings Expenditure Percentage Savings Percentage
0 Aditi Mishra Female 0.461538 1.000000 1.000000 1.0 0.051458 1.000000
1 Sona Singh Female 0.461538 0.444444 0.533333 0.7 0.189537 0.818182
2 Devi Thakur Female 0.538462 0.166667 0.666667 0.0 1.000000 0.000000
3 Anjali Kaur Female 1.000000 0.000000 0.000000 0.7 0.000000 1.000000
4 Sahil Deshmukh Male 0.000000 0.000000 0.466667 0.0 0.968268 0.000000
In [42]:
from sklearn.svm import SVC
model_linear = SVC(kernel='linear')
model_linear.fit(X_train[:500], y_train[:500])
Out[42]:
SVC
SVC(kernel='linear')
In [43]:
from sklearn import metrics
# predict
y_pred = model_linear.predict(X_test)
# accuracy
print("accuracy:", metrics.accuracy_score(y_test, y_pred), "\n")
accuracy: 1.0
In [44]:
#Add Dummies to code
#Creating Dummies
convert_to_dummies = ['Salary','Age', 'Salary']
field_dummies = pd.get_dummies(df[convert_to_dummies])
field_dummies
dataset = pd.concat([df,field_dummies],axis = 1)
dataset.drop(convert_to_dummies, axis=1, inplace=True)
dataset
Out[44]:
First Name Last Name Gender Expenditure Savings Expenditure Percentage Savings Percentage
0 Aditi Mishra Female 1.000000 1.0 0.051458 1.000000
1 Sona Singh Female 0.533333 0.7 0.189537 0.818182
68°F
Smoke
Search
ENG
US
1:27 AM
12/23/2023


Home Page - Select or create a n
day-8 data preprocessing - Jupyt
localhost:8891/notebooks/day-8%20data%20preprocessing.ipynb
jupyter day-8 data preprocessing Last Checkpoint: 8 minutes ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code
Out[42]:
SVC
SVC(kernel='linear')
In [43]:
from sklearn import metrics
# predict
y_pred = model_linear.predict(X_test)
# accuracy
print("accuracy:", metrics.accuracy_score(y_test, y_pred), "\n")
accuracy: 1.0
In [44]:
#Add Dummies to code
#Creating Dummies
convert_to_dummies = ['Salary','Age', 'Salary']
field_dummies = pd.get_dummies(df[convert_to_dummies])
field_dummies
dataset = pd.concat([df,field_dummies],axis = 1)
dataset.drop(convert_to_dummies, axis=1, inplace=True)
dataset
Out[44]:
First Name Last Name Gender Expenditure Savings Expenditure Percentage Savings Percentage
0 Aditi Mishra Female 1.000000 1.0 0.051458 1.000000
1 Sona Singh Female 0.533333 0.7 0.189537 0.818182
2 Devi Thakur Female 0.666667 0.0 1.000000 0.000000
3 Anjali Kaur Female 0.000000 0.7 0.000000 1.000000
4 Sahil Deshmukh Male 0.466667 0.0 0.968268 0.000000
In [ ]:
In [ ]:
68°F
Smoke
Search
ENG
US
1:27 AM
12/23/2023