EDA AND PREPROCESSING

[240]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

[ ]:
```
#The main objective of this project is to design and implement data preprocessing system .
#By performing effective data preprocessing the project aims to enhance the quality reliability and usefullness of the data for Machine learning
```

[ ]:
```
#The dataset employee.csv contains the details of employees
```

DATA EXPLORATION

[ ]:
```
#Explore the data, list down the unique values in each feature and find its length.
#Perform the statistical analysis and renaming of the columns.
```

[246]:
```
#Load the dataset
df=pd.read_csv("C:\\Users\\hp\\Downloads\\Employee.csv")
df
```

[246]:

	Company	Age	Salary	Place	Country	Gender
0	TCS	20.0	NaN	Chennai	India	0
1	Infosys	30.0	NaN	Mumbai	India	0
2	TCS	35.0	2300.0	Calcutta	India	0
3	Infosys	40.0	3000.0	Delhi	India	0
4	TCS	23.0	4000.0	Mumbai	India	0
...	...	...	...	...	...	...

	Company	Age	Salary	Place	Country	Gender
145	Infosys	44.0	4034.0	Delhi	India	1
146	TCS	33.0	5034.0	Mumbai	India	1
147	Infosys	22.0	8202.0	Cochin	India	0

148 rows × 6 columns

[9]:
```
df.head()
```

[9]:

	Company	Age	Salary	Place	Country	Gender
0	TCS	20.0	NaN	Chennai	India	0
1	Infosys	30.0	NaN	Mumbai	India	0
2	TCS	35.0	2300.0	Calcutta	India	0
3	Infosys	40.0	3000.0	Delhi	India	0
4	TCS	23.0	4000.0	Mumbai	India	0

[11]:
```
df.tail()
```

[11]:

	Company	Age	Salary	Place	Country	Gender
143	TCS	33.0	9024.0	Calcutta	India	1
144	Infosys	22.0	8787.0	Calcutta	India	1
145	Infosys	44.0	4034.0	Delhi	India	1
146	TCS	33.0	5034.0	Mumbai	India	1
147	Infosys	22.0	8202.0	Cochin	India	0

Home | Assignment EDA and Preprocessi

localhost:8888/notebooks/Assignment%20EDA%20and%20Preprocessing.ipynb?

jupyter Assignment EDA and Preprocessing Last Checkpoint: yesterday

Trusted

File Edit View Run Kernel Settings Help

Code JupyterLab Python 3 (ipykernel)

[15]:
```
df.shape
```

[15]:
```
(148, 6)
```

[17]:
```
df.dtypes
```

[17]:
```
Company     object
Age        float64
Salary     float64
Place       object
Country     object
Gender       int64
dtype: object
```

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

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 148 entries, 0 to 147
Data columns (total 6 columns):
 #   Column   Non-Null Count  Dtype  
---  ------   --------------  -----  
 0   Company  140 non-null    object 
 1   Age      130 non-null    float64
 2   Salary   124 non-null    float64
 3   Place    134 non-null    object 
 4   Country  148 non-null    object 
 5   Gender   148 non-null    int64  
dtypes: float64(2), int64(1), object(3)
memory usage: 7.1+ KB
```

[21]:
```
df.columns
```

[21]:
```
Index(['Company', 'Age', 'Salary', 'Place', 'Country', 'Gender'], dtype='object')
```

[25]:
```
df.nunique()
```

32°C Mostly sunny

Search

ENG IN

16:06 28-12-2024

Home | Assignment EDA and Preprocessi

localhost:8888/notebooks/Assignment%20EDA%20and%20Preprocessing.ipynb?

jupyter Assignment EDA and Preprocessing Last Checkpoint: yesterday

Trusted

File Edit View Run Kernel Settings Help

Code JupyterLab Python 3 (ipykernel)

[25]:
```
df.nunique()
```

[25]:
```
Company     6
Age        29
Salary     40
Place      11
Country     1
Gender      2
dtype: int64
```

[69]:
```
df['Company'].unique()
```

[69]:
```
array(['TCS', 'Infosys', 'CTS', nan, 'Tata Consultancy Services',
       'Congnizant', 'Infosys Pvt Lmt'], dtype=object)
```

[49]:
```
df['Age'].unique()
```

[49]:
```
array([20., 30., 35., 40., 23., nan, 34., 45., 18., 22., 32., 37., 50.,
       21., 46., 36., 26., 41., 24., 25., 43., 19., 38., 51., 31., 44.,
       33., 17.,  0., 54.])
```

[51]:
```
df['Salary'].unique()
```

[51]:
```
array([  nan, 2300., 3000., 4000., 5000., 6000., 7000., 8000., 9000.,
       1089., 1234., 3030., 3045., 3184., 4824., 5835., 7084., 8943.,
       8345., 9284., 9876., 2034., 7654., 2934., 4034., 5034., 8202.,
       9024., 4345., 6544., 6543., 3234., 4324., 5435., 5555., 8787.,
       3454., 5654., 5009., 5098., 3033.])
```

[57]:
```
df['Place'].unique()
```

[57]:
```
array(['Chennai', 'Mumbai', 'Calcutta', 'Delhi', 'Podicherry', 'Cochin',
       nan, 'Noida', 'Hyderabad', 'Bhopal', 'Nagpur', 'Pune'],
      dtype=object)
```

[55]:
```
df['Country'].unique()
```

32°C Mostly sunny

Search

ENG IN

16:06 28-12-2024

[51]:
```
array([  nan, 2300., 3000., 4000., 5000., 6000., 7000., 8000., 9000.,
       1089., 1234., 3030., 3045., 3184., 4824., 5835., 7084., 8943.,
       8345., 9284., 9876., 2034., 7654., 2934., 4034., 5034., 8202.,
       9024., 4345., 6544., 6543., 3234., 4324., 5435., 5555., 8787.,
       3454., 5654., 5009., 5098., 3033.])
```

[57]:
```
df['Place'].unique()
```

[57]:
```
array(['Chennai', 'Mumbai', 'Calcutta', 'Delhi', 'Podicherry', 'Cochin',
       nan, 'Noida', 'Hyderabad', 'Bhopal', 'Nagpur', 'Pune'],
      dtype=object)
```

[55]:
```
df['Country'].unique()
```

[55]:
```
array(['India'], dtype=object)
```

[59]:
```
df['Gender'].unique()
```

[59]:
```
array([0, 1], dtype=int64)
```

[63]:
```
len(df['Company'].unique())
```

[63]:
```
7
```

[73]:
```
len(df['Age'].unique())
```

[73]:
```
30
```

[75]:
```
len(df['Salary'].unique())
```

[75]:
```
41
```

[77]:
```
len(df['Place'].unique())
```

[77]:
```
12
```

Home | Assignment EDA and Preprocessi

localhost:8888/notebooks/Assignment%20EDA%20and%20Preprocessing.ipynb?

jupyter Assignment EDA and Preprocessing Last Checkpoint: yesterday

File Edit View Run Kernel Settings Help

Trusted

Code

JupyterLab Python 3 (ipykernel)

[81]:
```
df.describe()
```

[81]:

	Age	Salary	Gender
count	130.000000	124.000000	148.000000
mean	30.484615	5312.467742	0.222973
std	11.096640	2573.764683	0.417654
min	0.000000	1089.000000	0.000000
25%	22.000000	3030.000000	0.000000
50%	32.500000	5000.000000	0.000000
75%	37.750000	8000.000000	0.000000
max	54.000000	9876.000000	1.000000

[79]:
```
df.rename(columns={'Age':'Employee_Age'})
```

[79]:

	Company	Employee_Age	Salary	Place	Country	Gender
0	TCS	20.0	NaN	Chennai	India	0
1	Infosys	30.0	NaN	Mumbai	India	0
2	TCS	35.0	2300.0	Calcutta	India	0
3	Infosys	40.0	3000.0	Delhi	India	0
4	TCS	23.0	4000.0	Mumbai	India	0
...	...	...	...	...	...	...
143	TCS	33.0	9024.0	Calcutta	India	1

32°C Mostly sunny

Search

ENG IN

16:06 28-12-2024

DATA CLEANING

[ ]:
```
# Find the missing and inappropriate values, treat them appropriately.
# Remove all duplicate rows.
# Find the outliers.
# Replace the value 0 in age as NaN
# Treat the null values in all columns using any measures
```

[85]:
```
#Find missing Values
print("Missing Values:" , df.isnull().sum())
```

```
Missing Values: Company     8
Age        18
Salary     24
Place      14
Country     0
Gender      0
dtype: int64
```

[101]:
```
#Replace 0 in age with Nan
df['Age'].replace(0, np.nan)
```

[101]:
```
0      20.0
1      30.0
2      35.0
3      40.0
4      23.0
       ...
143    33.0
144    22.0
145    44.0
146    33.0
147    22.0
Name: Age, Length: 148, dtype: float64
```

[55]:
```
#Treat Null values
df.fillna(df.mode())
```

Home | Assignment EDA and Preprocessi

localhost:8888/notebooks/Assignment%20EDA%20and%20Preprocessing.ipynb?

jupyter Assignment EDA and Preprocessing Last Checkpoint: yesterday

File Edit View Run Kernel Settings Help

Trusted

Code

JupyterLab Python 3 (ipykernel)

[55]:
```
#Treat Null values
df.fillna(df.mode())
df
```

[55]:

	Company	Age	Salary	Place	Country	Gender
0	TCS	20.0	NaN	Chennai	India	0
1	Infosys	30.0	NaN	Mumbai	India	0
2	TCS	35.0	2300.0	Calcutta	India	0
3	Infosys	40.0	3000.0	Delhi	India	0
4	TCS	23.0	4000.0	Mumbai	India	0
...	...	...	...	...	...	...
143	TCS	33.0	9024.0	Calcutta	India	1
144	Infosys	22.0	8787.0	Calcutta	India	1
145	Infosys	44.0	4034.0	Delhi	India	1
146	TCS	33.0	5034.0	Mumbai	India	1
147	Infosys	22.0	8202.0	Cochin	India	0

148 rows × 6 columns

[31]:
```
#Find the Outliers using Boxplot
sns.boxplot(df['Salary'])
```

[31]:
```
<Axes: ylabel='Salary'>
```

32°C Mostly sunny

Search

ENG IN

16:06 28-12-2024

Home | Assignment EDA and Preprocessi

localhost:8888/notebooks/Assignment%20EDA%20and%20Preprocessing.ipynb?

jupyter Assignment EDA and Preprocessing Last Checkpoint: yesterday

File Edit View Run Kernel Settings Help

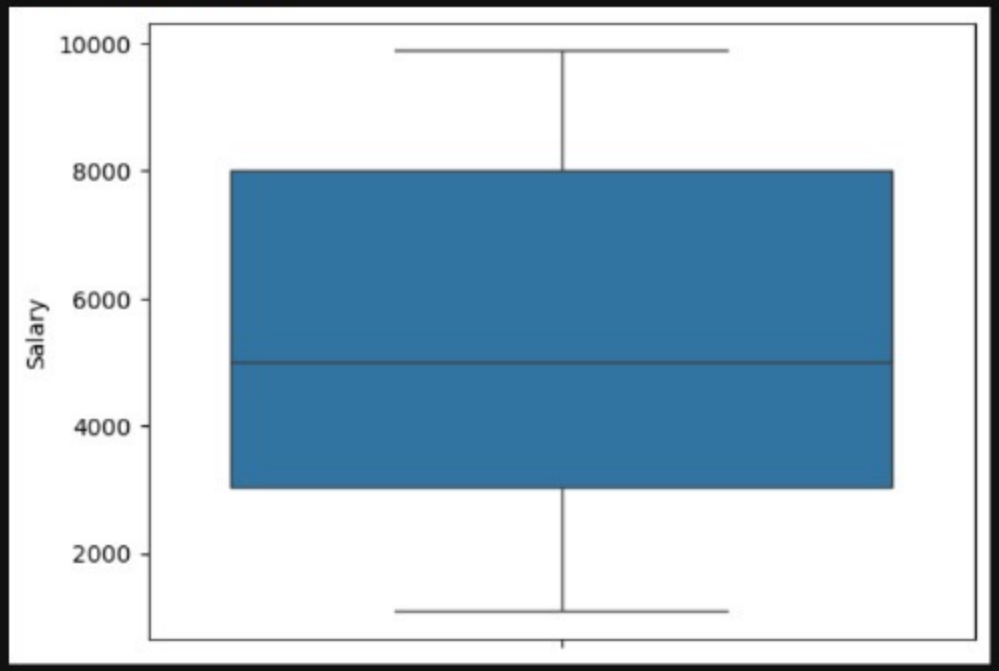
Trusted

Code

JupyterLab Python 3 (ipykernel)

[31]:
```
#Find the Outliers using Boxplot
sns.boxplot(df['Salary'])
```

[31]: <Axes: ylabel='Salary'>



[39]:
```
for i in df.columns:
    plt.figure(figsize=(12,5))

    # Histogram
    plt.subplot(1, 2, 1)
    sns.histplot(df[i], kde=True)
```

32°C Mostly sunny

Search

ENG IN

16:06 28-12-2024

Home | Assignment EDA and Preprocessi

localhost:8888/notebooks/Assignment%20EDA%20and%20Preprocessing.ipynb?

jupyter Assignment EDA and Preprocessing Last Checkpoint: yesterday

Trusted

File Edit View Run Kernel Settings Help

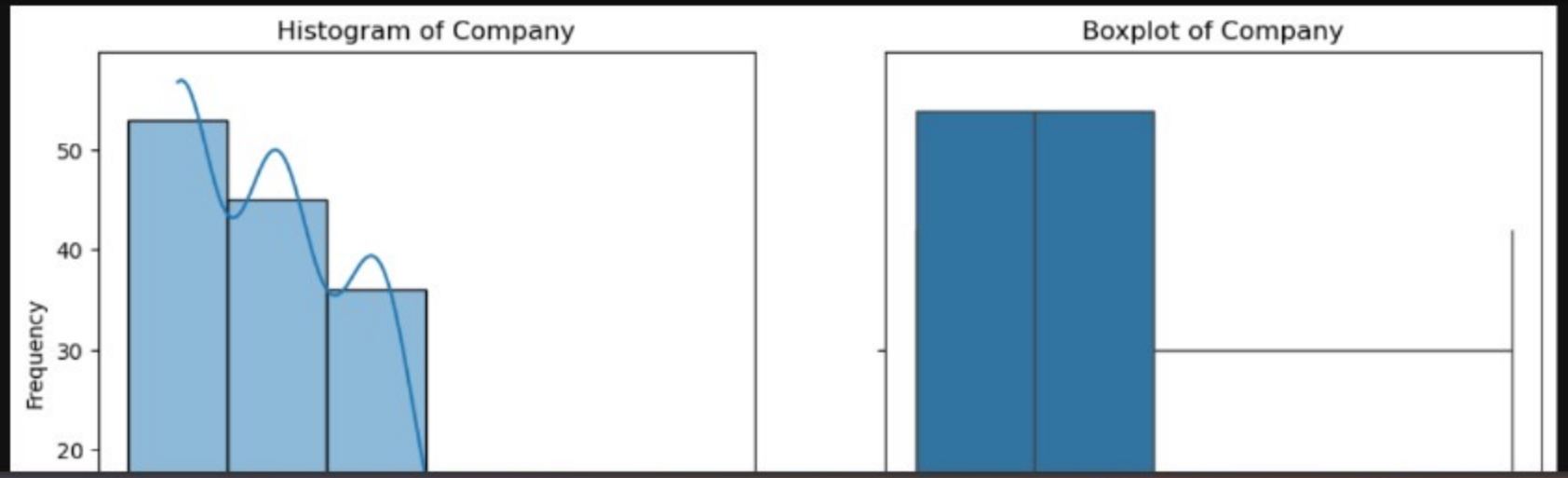
Code JupyterLab Python 3 (ipykernel)

[39]:
```
for i in df.columns:
    plt.figure(figsize=(12,5))

    # Histogram
    plt.subplot(1, 2, 1)
    sns.histplot(df[i], kde=True)
    plt.xlabel(i)
    plt.xticks(rotation=90)
    plt.ylabel('Frequency')
    plt.title(f'Histogram of {i}')

    # Boxplot
    plt.subplot(1, 2, 2)
    sns.boxplot(x=df[i])
    plt.title(f'Boxplot of {i}')
    plt.xticks(rotation=90)

    plt.show()
```



32°C Mostly sunny

Search

ENG IN

16:06 28-12-2024

Home | Assignment EDA and Preprocessi

localhost:8888/notebooks/Assignment%20EDA%20and%20Preprocessing.ipynb?

jupyter Assignment EDA and Preprocessing Last Checkpoint: yesterday

Trusted

File Edit View Run Kernel Settings Help

Code JupyterLab Python 3 (ipykernel)

[15]:
```
#Find Outliers
Upper_limit = df['Salary'].mean() + 3*df['Salary'].std()
lower_limit = df['Salary'].mean() - 3*df['Salary'].std()
print('Upper_limit:',Upper_limit)
print('lower_limit:',lower_limit)
```

```
Upper_limit: 13033.761789635413
lower_limit: -2408.8263057644463
```

[17]:
```
Outliers=df.loc[(df['Salary'] < lower_limit) | (df['Salary'] > Upper_limit)]
Outliers
```

[17]:

Company	Age	Salary	Place	Country	Gender

[19]:
```
#Remove Outliers
df_trimmed=df.loc[(df['Salary'] >= lower_limit) &  (df['Salary'] <= Upper_limit)]
df_trimmed
```

[19]:

	Company	Age	Salary	Place	Country	Gender
2	TCS	35.0	2300.0	Calcutta	India	0
3	Infosys	40.0	3000.0	Delhi	India	0
4	TCS	23.0	4000.0	Mumbai	India	0
5	Infosys	NaN	5000.0	Calcutta	India	0
6	TCS	NaN	6000.0	Chennai	India	1
...	...	...	...	...	...	...
143	TCS	33.0	9024.0	Calcutta	India	1
144	Infosys	22.0	8787.0	Calcutta	India	1
145	Infosys	44.0	4034.0	Delhi	India	1

32°C Mostly sunny

Search

ENG IN

16:06 28-12-2024

DATA ANALYSIS

[ ]:
```
# Filter the data with age >40 and salary<5000
# Plot the chart with age and salary
# Count the number of people from each place and represent it visually
```

[72]:
```
#Filter data with age>40 and salary<5000
filtered_data=df[(df['Age'] > 40) & (df['Salary'] < 5000)]
filtered_data
```

[72]:

	Company	Age	Salary	Place	Country	Gender
21	Infosys	50.0	3184.0	Delhi	India	0
32	Infosys	45.0	4034.0	Calcutta	India	0
39	Infosys	41.0	3000.0	Mumbai	India	0
50	Infosys	41.0	3000.0	Chennai	India	0
57	Infosys	51.0	3184.0	Hyderabad	India	0
68	Infosys	43.0	4034.0	Mumbai	India	0
75	Infosys	44.0	3000.0	Cochin	India	0
86	Infosys	41.0	3000.0	Delhi	India	0
93	Infosys	54.0	3184.0	Mumbai	India	0
104	Infosys	44.0	4034.0	Delhi	India	0
122	Infosys	44.0	3234.0	Mumbai	India	0
129	Infosys	50.0	3184.0	Calcutta	India	0
138	CTS	44.0	3033.0	Cochin	India	0
140	Infosys	44.0	4034.0	Hyderabad	India	0

Home | Assignment EDA and Preprocessi

localhost:8888/notebooks/Assignment%20EDA%20and%20Preprocessing.ipynb?

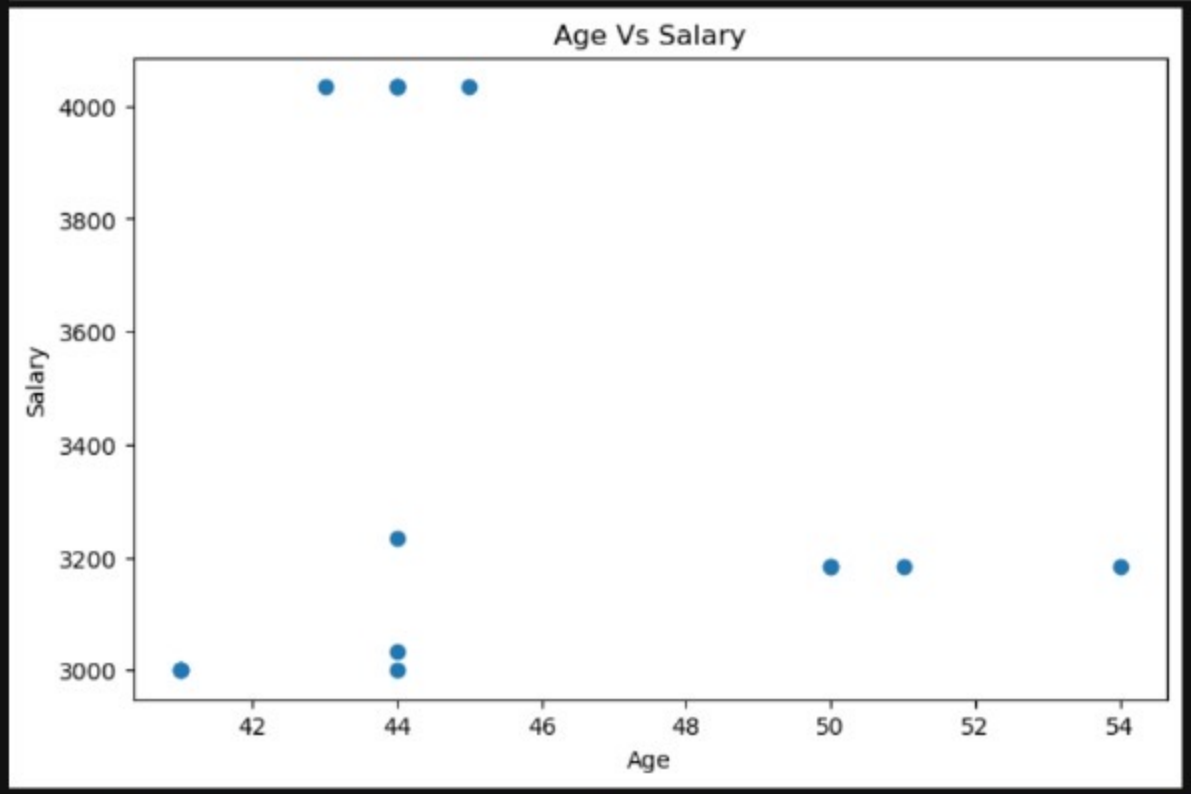
jupyter Assignment EDA and Preprocessing Last Checkpoint: yesterday

Trusted

File Edit View Run Kernel Settings Help

Code JupyterLab Python 3 (ipykernel)

[74]:
```
#Plot chart with age and salary
plt.figure(figsize=(8,5))
plt.scatter(filtered_data['Age'],filtered_data['Salary'])
plt.xlabel('Age')
plt.ylabel('Salary')
plt.title('Age Vs Salary')
plt.show()
```



32°C Mostly sunny

Search

ENG IN

16:06 28-12-2024

Home | Assignment EDA and Preprocessi

localhost:8888/notebooks/Assignment%20EDA%20and%20Preprocessing.ipynb?

jupyter Assignment EDA and Preprocessing Last Checkpoint: yesterday

Trusted

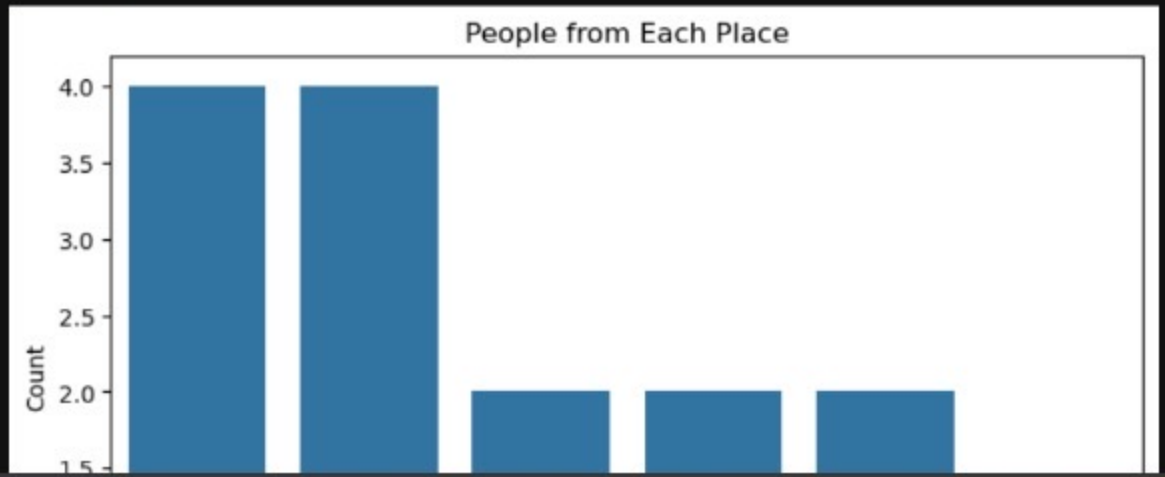
File Edit View Run Kernel Settings Help

Code JupyterLab Python 3 (ipykernel)

[78]:
```
#Count Number of People from each place
df1= filtered_data['Place'].value_counts()
df1
```

[78]:
```
Place
Delhi       4
Mumbai      4
Calcutta    2
Hyderabad   2
Cochin      2
Chennai     1
Name: count, dtype: int64
```

[80]:
```
#Represent the count visually using bar chart
plt.figure(figsize=(8,5))
sns.barplot(x=df1.index,y=df1.values)
plt.xlabel('Place')
plt.ylabel('Count')
plt.title('People from Each Place')
plt.show()
```



32°C Mostly sunny

Search

ENG IN

16:07 28-12-2024

Home | Assignment EDA and Preprocessi

localhost:8888/notebooks/Assignment%20EDA%20and%20Preprocessing.ipynb?

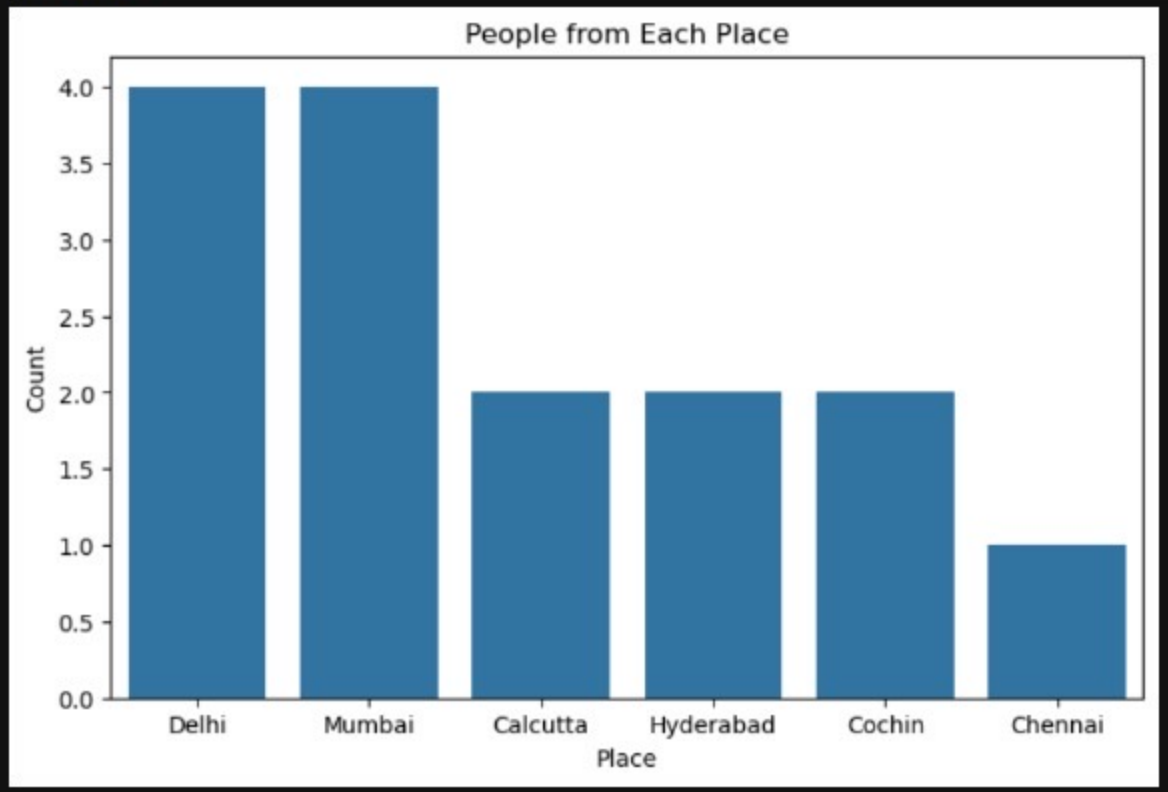
jupyter Assignment EDA and Preprocessing Last Checkpoint: yesterday

Trusted

File Edit View Run Kernel Settings Help

Code | JupyterLab | Python 3 (ipykernel)

[80]:
```
#Represent the count visually using bar chart
plt.figure(figsize=(8,5))
sns.barplot(x=df1.index,y=df1.values)
plt.xlabel('Place')
plt.ylabel('Count')
plt.title('People from Each Place')
plt.show()
```



32°C Mostly sunny

Search

ENG IN

16:07 28-12-2024

Home | Assignment EDA and Preprocessi

localhost:8888/notebooks/Assignment%20EDA%20and%20Preprocessing.ipynb?

jupyter Assignment EDA and Preprocessing Last Checkpoint: yesterday

Trusted

File Edit View Run Kernel Settings Help

Code JupyterLab Python 3 (ipykernel)

[146]:
```
import warnings
import sys
if not sys.warnoptions:
    warnings.simplefilter("ignore")
```

DATA ENCODING

[ ]:
```
#Convert categorical variables into numerical representations using techniques such as one-hot encoding,
#label encoding, making them suitable for analysis by machine learning algorithms.
```

[164]:
```
from sklearn.preprocessing import LabelEncoder
```

*[293]:
```
#Covert Categorical variables to Numeric
categorical_columns = ['Company','Place','Country']
label_encoder = LabelEncoder()
for column in categorical_columns:
    df[column] = label_encoder.fit_transform(df[column])
df
```

[293]:

	Company	Age	Salary	Place	Country	Gender
0	4	-0.948501	NaN	2	0	-0.535683
1	2	-0.043841	NaN	6	0	-0.535683
2	4	0.408489	-1.175200	1	0	-0.535683
3	2	0.860819	-0.902122	4	0	-0.535683
4	4	-0.677103	-0.512010	6	0	-0.535683
...	...	...	...	...	...	...
143	4	0.227557	1.447914	1	0	1.866775
144	2	-0.767569	1.355457	1	0	1.866775
145	2	1.222683	-0.498746	4	0	1.866775

32°C Mostly sunny

Search

ENG IN

16:07 28-12-2024

Home | Assignment EDA and Preprocessi

localhost:8888/notebooks/Assignment%20EDA%20and%20Preprocessing.ipynb?

jupyter Assignment EDA and Preprocessing Last Checkpoint: yesterday

Trusted

File Edit View Run Kernel Settings Help

Code JupyterLab Python 3 (ipykernel)

FEATURE SCALING

[ ]:
```
# After the process of encoding, perform the scaling of the features using
# standardscaler and minmaxscaler.
```

[242]:
```
import pandas as pd
from sklearn.preprocessing import StandardScaler
from sklearn.preprocessing import MinMaxScaler
```

*[268]:
```
#STANDARD SCALER
numeric_columns=df.select_dtypes(include=['int64','float']).columns
scaler = StandardScaler()
scaled_data= scaler.fit_transform(df[numeric_columns])
scaled_data
```

```
[ 0.3694207 , -0.88456066, -0.53568323],
[ 1.76547852, -0.83034108, -0.53568323],
[-0.85803509, -0.19055719, -0.53568323],
[        nan,  0.20384617, -0.53568323],
[        nan,  0.69109622,  1.86677489],
[-0.67710311,  1.41631466,  1.86677489],
[ 0.31802274,  1.18302761,  1.86677489],
[ 1.31314859,  1.54934289,  1.86677489],
[-0.67710311,  1.78028927,  1.86677489],
[ 0.40848872, -1.27897002, -0.53568323],
[ 1.40361457,  0.91346013, -0.53568323],
[-0.94850107, -0.92786911, -0.53568323],
[ 1.31314859, -0.49874577, -0.53568323],
[ 0.49895471, -0.10863364, -0.53568323],
[-0.40570515,  1.12724157, -0.53568323],
[ 0.40848872,  1.44791374,  1.86677489],
[ 0.13709076,         nan, -0.53568323],
[ 0.40848872,         nan, -0.53568323],
[ 0.31802274, -1.17520019, -0.53568323],
```

*[270]:
```
#MIN MAX SCALER
numeric_columns=df.select_dtypes(include=['int64','float']).columns
```

32°C Mostly sunny | Search | ENG IN | 16:07 28-12-2024

Home | Assignment EDA and Preprocessi

localhost:8888/notebooks/Assignment%20EDA%20and%20Preprocessing.ipynb?

jupyter Assignment EDA and Preprocessing Last Checkpoint: yesterday

Trusted

File Edit View Run Kernel Settings Help

Code JupyterLab Python 3 (ipykernel)

```
[ 1.76547852, -0.83034108, -0.53568323],
[-0.85803509, -0.19055719, -0.53568323],
[        nan,  0.20384617, -0.53568323],
[        nan,  0.69109622,  1.86677489],
[-0.67710311,  1.41631466,  1.86677489],
[ 0.31802274,  1.18302761,  1.86677489],
[ 1.31314859,  1.54934289,  1.86677489],
[-0.67710311,  1.78028927,  1.86677489],
[ 0.40848872, -1.27897002, -0.53568323],
[ 1.40361457,  0.91346013, -0.53568323],
[-0.94850107, -0.92786911, -0.53568323],
[ 1.31314859, -0.49874577, -0.53568323],
[ 0.49895471, -0.10863364, -0.53568323],
[-0.40570515,  1.12724157, -0.53568323],
[ 0.40848872,  1.44791374,  1.86677489],
[ 0.13709076,         nan, -0.53568323],
[ 0.40848872,         nan, -0.53568323],
[ 0.31802274, -1.17520019, -0.53568323],
```

•[270]:

```
#MIN MAX SCALER
numeric_columns=df.select_dtypes(include=['int64','float']).columns
Scaler = MinMaxScaler()
Scaled_data = Scaler.fit_transform(df[numeric_columns])
print(Scaled_data)
```

```
[0.35185185 0.01650165 0.        ]
[0.75925926 0.21748037 0.        ]
[0.44444444 0.21748037 0.        ]
[0.38888889 0.2208945  0.        ]
[0.64814815 0.44508934 0.        ]
[0.38888889 0.62080346 0.        ]
[0.59259259 0.74712644 0.        ]
[0.7037037  0.22260157 0.        ]
[0.94444444 0.23842039 0.        ]
[0.42592593 0.42505975 0.        ]
[       nan 0.54011608 0.        ]
[       nan 0.68225788 0.        ]
```

32°C Mostly sunny

Search

ENG IN

16:07 28-12-2024