Data Preprocessing

2

```
%python
mount_point = "/mnt/azure_blob"
if not any(mount.mountPoint == mount_point for mount in dbutils.fs.mounts()):
    dbutils.fs.mount(
        source="wasbs://crimeprediction@crimepred.blob.core.windows.net",
        mount_point=mount_point,
        extra_configs={"fs.azure.account.key.crimepred.blob.core.windows.net": "ouDRpWreVVZXFlnpJpMqjAj50KXEODgCXmRy
        +9x013et0XzN0cHxnKx03MtKcxIDLfcoXCZRFibC+AStWSYeqg=="}
    )


df = pd.read_excel('/dbfs/mnt/azure_blob/United_States_Offense_Type_by_Agency_2023.xlsx')
df1 = pd.read_excel('/dbfs/mnt/azure_blob/United_States_Offense_Type_by_Agency_2022.xlsx')
df2 = df = pd.read_excel('/dbfs/mnt/azure_blob/United_States_Offense_Type_by_Agency_2021.xlsx')
df3= df = pd.read_excel('/dbfs/mnt/azure_blob/United_States_Offense_Type_by_Agency_2020.xlsx')


display(df)
```

Table

	State	Agency Type	Agency Name	Population1	Total Offenses	Crimes Against Persons
1	ARIZONA	Cities	Apache Junction	43385	2169	4
2	null	null	Casa Grande	59822	3841	11
3	null	null	Coolidge	13277	1316	2
4	null	null	Eagar	4944	91	
5	null	null	Gilbert	259629	11263	21
6	null	null	Huachuca City	1723	30	
7	null	null	Lake Havasu City	56243	2656	7
8	null	null	Mesa	527361	27198	66
9	null	null	Oro Valley	46634	1671	1
10	null	null	Parker	3224	214	
11	null	null	Payson	15869	1367	2
12	null	null	San Luis	35574	947	1
13	null	null	Somerton	16811	418	1
14	null	null	Surprise	144620	4806	11

7,399 rows

3

```
df['State'] = df['State'].fillna(method='ffill')
df = df.drop(columns=['Agency Type'])
df1['State'] = df1['State'].fillna(method='ffill')
df1 = df1.drop(columns=['Agency Type'])
df2['State'] = df2['State'].fillna(method='ffill')
df2 = df2.drop(columns=['Agency Type'])
df3['State'] = df3['State'].fillna(method='ffill')
df3 = df3.drop(columns=['Agency Type'])
```

4

```
df = df.fillna(0)
df1 = df1.fillna(0)
df2 = df2.fillna(0)
df3 = df3.fillna(0)
```

5

6

```
numerical_features = df1.select_dtypes(include=['number']).columns.tolist()
```

7

```
print("Numerical Features:")
print(numerical_features)
```

```
Numerical Features:
['Population1', 'Total\nOffenses', 'Crimes\nAgainst\nPersons', 'Crimes\nAgainst\nProperty', 'Crimes\nAgainst\nSociety', 'Assault
\nOffenses', 'Aggravated\nAssault', 'Simple\nAssault', 'Intimidation', 'Homicide\nOffenses', 'Murder and\nNonnegligent\nManslaugh
ter', 'Negligent\nMan-\nslaughter', 'Justifiable\nHomicide', 'Human\nTrafficking\nOffenses', 'Commercial\nSex Acts', 'Involuntary
\nServitude', 'Kidnapping/\nAbduction', 'Sex\nOffenses', 'Rape', 'Sodomy', 'Sexual\nAssault\nWith an\nObject', 'Fondling', 'Inces
t', 'Statutory\nRape', 'Arson', 'Bribery', 'Burglary/\nBreaking &\nEntering', 'Counter-\nfeiting/\nForgery', 'Destruction/\nDamag
e/\nVandalism\nof Property', 'Embezzle-\nment', 'Extortion/\nBlackmail', 'Fraud\nOffenses', 'False\nPretenses/\nSwindle/\nConfide
```

```
nce\nGame', 'Credit\nCard/\nAutomated\nTeller\nMachine\nFraud', 'Imper-\nsonation', 'Welfare\nFraud', 'Wire\nFraud', 'Identity \n
Theft', 'Hacking/\nComputer \nInvasion', 'Larceny/\nTheft\nOffenses', 'Pocket-\npicking', 'Purse-\nsnatching', 'Shop-\nlifting',
'Theft\nFrom\nBuilding', 'Theft\nFrom\nCoin Op-\nerated\nMachine\nor Device', 'Theft\nFrom\nMotor\nVehicle', 'Theft of \nMotor \n
Vehicle\nParts or\nAcces-\nsories', 'All\nOther\nLarceny', 'Motor\nVehicle\nTheft', 'Robbery', 'Stolen\nProperty\nOffenses', 'Ani
mal \nCruelty', 'Drug/\nNarcotic\nOffenses', 'Drug/\nNarcotic\nViolations', 'Drug\nEquipment\nViolations', 'Gambling\nOffenses',
'Betting/\nWagering', 'Operating/\nPromoting/\nAssisting\nGambling', 'Gambling\nEquipment\nViolations', 'Sports\nTampering', 'Por
-\nnography/\nObscene\nMaterial', 'Pros-\ntitution\nOffenses', 'Pros-\ntitution', 'Assisting or\nPromoting\nProstitution', 'Purch
asing\nProstitution', 'Weapon\nLaw\nViolations']
```

8

```
categorical_features = df1.select_dtypes(include=['object']).columns.tolist()
```

9

```
print("\nCategorical Features:")
print(categorical_features)
```

```
Categorical Features:
['State', 'Agency Name']
```

10

```
df['Year']= 2023
df1['Year']= 2022
df2['Year']= 2021
df3['Year']= 2020
```

11

```
(69, 69, 69, 69)
```

12

```
df['State'] = df['State'].fillna(method='ffill')
```

13

```
display(df[['State']].head(20))

# Fill down the 'State' column with forward fill to propagate non-null values
df['State'] = df['State'].fillna(method='ffill')

# Verify if the filling was successful
display(df[['State']].head(20))

# Additional check: Ensure there are no more NaN values in 'State'
print("Number of remaining NaN values in 'State':", df['State'].isna().sum())
```

Table

	State
1	ARIZONA
2	ARIZONA
3	ARIZONA
4	ARIZONA
5	ARIZONA
6	ARIZONA
7	ARIZONA
8	ARIZONA
9	ARIZONA
10	ARIZONA
11	ARIZONA
12	ARIZONA
13	ARIZONA
14	ARIZONA
15	ARIZONA

20 rows

Table

	State
1	ARIZONA
2	ARIZONA
3	ARIZONA
4	ARIZONA
5	ARIZONA
6	ARIZONA
7	ARIZONA
8	ARIZONA
9	ARIZONA

10	ARIZONA
11	ARIZONA
12	ARIZONA
13	ARIZONA
14	ARIZONA
15	ARIZONA

20 rows

```
Number of remaining NaN values in 'State': 0
```

14

```
display(df3)
```

Table

	State	Agency Name	1.2 Population1	1.2 Total Offenses	1.2 Crimes Against Persons	1.2 Crimes Against Pr
1	ARIZONA	Apache Junction	43385	2169	449	
2	ARIZONA	Casa Grande	59822	3841	1108	
3	ARIZONA	Coolidge	13277	1316	202	
4	ARIZONA	Eagar	4944	91	25	
5	ARIZONA	Gilbert	259629	11263	2176	
6	ARIZONA	Huachuca City	1723	30	2	
7	ARIZONA	Lake Havasu City	56243	2656	794	
8	ARIZONA	Mesa	527361	27198	6604	
9	ARIZONA	Oro Valley	46634	1671	137	
10	ARIZONA	Parker	3224	214	19	
11	ARIZONA	Payson	15869	1367	276	
12	ARIZONA	San Luis	35574	947	109	
13	ARIZONA	Somerton	16811	418	128	
14	ARIZONA	Surprise	144620	4806	1199	

7,399 rows

15

```
/databricks/spark/python/pyspark/sql/pandas/conversion.py:510: UserWarning: createDataFrame attempted Arrow optimization becau
se 'spark.sql.execution.arrow.pyspark.enabled' is set to true; however, failed by the reason below:
  Expected bytes, got a 'int' object
Attempting non-optimization as 'spark.sql.execution.arrow.pyspark.fallback.enabled' is set to true.
  warn(msg)
```

Table

	State	Agency Name	Population1	Total Offenses	Crimes Against Persons	Crimes Against Pr
1	ARIZONA	Apache Junction	43385	2169	449	
2	ARIZONA	Casa Grande	59822	3841	1108	
3	ARIZONA	Coolidge	13277	1316	202	
4	ARIZONA	Eagar	4944	91	25	
5	ARIZONA	Gilbert	259629	11263	2176	
6	ARIZONA	Huachuca City	1723	30	2	
7	ARIZONA	Lake Havasu City	56243	2656	794	
8	ARIZONA	Mesa	527361	27198	6604	
9	ARIZONA	Oro Valley	46634	1671	137	
10	ARIZONA	Parker	3224	214	19	
11	ARIZONA	Payson	15869	1367	276	
12	ARIZONA	San Luis	35574	947	109	
13	ARIZONA	Somerton	16811	418	128	
14	ARIZONA	Surprise	144620	4806	1199	

6,430+ rows | Truncated data due to byte limit

16

```
Updated file saved to: /dbfs/mnt/azure_blob/Final_Dataset.xlsx
```