Data Mining Lab-1 (Data Preprocessing)

DataSet:

Used Cricket DataSet

Link: /content/cricket.odb

Source Code:

```python
import pandas as pd
import numpy as np

df=pd.read_csv('/content/odb.csv')
df1['SR'].replace('Unknown', np.nan, inplace=True)
df1['SR'].fillna(df['SR'].mean().round(2), inplace=True)
df1['BF'].fillna(df1['BF'].median(),inplace=True)
df1['50'].fillna(df1['50'].median(),inplace=True)
df1['0'].fillna(df1['0'].mean(),inplace=True)
df1['Runs'].fillna(df1['Runs'].median(),inplace=True)
import matplotlib.pyplot as plt
xaxis=df1['Mat']
yaxis=df1['SR']
plt.bar(xaxis,yaxis)
plt.show()
```

Screenshots

```
[ ] df.isnull().sum()
```

	0
Unnamed: 0	0
Player	0
Span	0
Mat	0
Inns	0
NO	2
Runs	9
HS	0
Ave	0
BF	31
SR	22
100	0
50	11
0	11
4s	11
6s	11

dtype: int64

```
df1['BF'].fillna(df1['BF'].median(),inplace=True)
df1['50'].fillna(df1['50'].median(),inplace=True)
df1['0'].fillna(df1['0'].mean(),inplace=True)
df1['Runs'].fillna(df1['Runs'].median(),inplace=True)
```

/tmp/ipython-input-58-2754236557.py:1: FutureWarning: A value is trying to be set on a copy of a DataFrame or Series through chained assignment using an inplace method.
The behavior will change in pandas 3.0. This inplace method will never work because the intermediate object on which we are setting values always behaves as a copy.

For example, when doing 'df[col].method(value, inplace=True)', try using 'df.method({col: value}, inplace=True)' or df[col] = df[col].method(value) instead, to perform t

    df1['BF'].fillna(df1['BF'].median(),inplace=True)
/tmp/ipython-input-58-2754236557.py:2: FutureWarning: A value is trying to be set on a copy of a DataFrame or Series through chained assignment using an inplace method.
The behavior will change in pandas 3.0. This inplace method will never work because the intermediate object on which we are setting values always behaves as a copy.

For example, when doing 'df[col].method(value, inplace=True)', try using 'df.method({col: value}, inplace=True)' or df[col] = df[col].method(value) instead, to perform t

    df1['50'].fillna(df1['50'].median(),inplace=True)
/tmp/ipython-input-58-2754236557.py:3: FutureWarning: A value is trying to be set on a copy of a DataFrame or Series through chained assignment using an inplace method.
The behavior will change in pandas 3.0. This inplace method will never work because the intermediate object on which we are setting values always behaves as a copy.

For example, when doing 'df[col].method(value, inplace=True)', try using 'df.method({col: value}, inplace=True)' or df[col] = df[col].method(value) instead, to perform t

    df1['0'].fillna(df1['0'].mean(),inplace=True)
/tmp/ipython-input-58-2754236557.py:4: FutureWarning: A value is trying to be set on a copy of a DataFrame or Series through chained assignment using an inplace method.
The behavior will change in pandas 3.0. This inplace method will never work because the intermediate object on which we are setting values always behaves as a copy.

For example, when doing 'df[col].method(value, inplace=True)', try using 'df.method({col: value}, inplace=True)' or df[col] = df[col].method(value) instead, to perform t
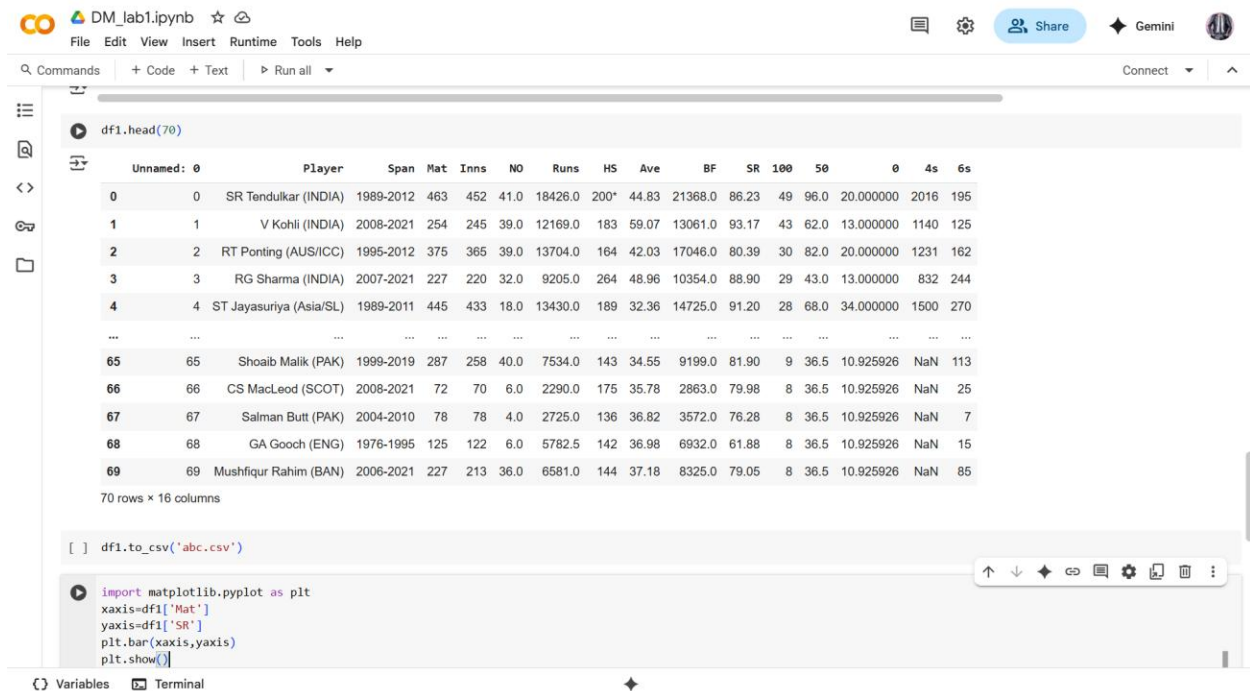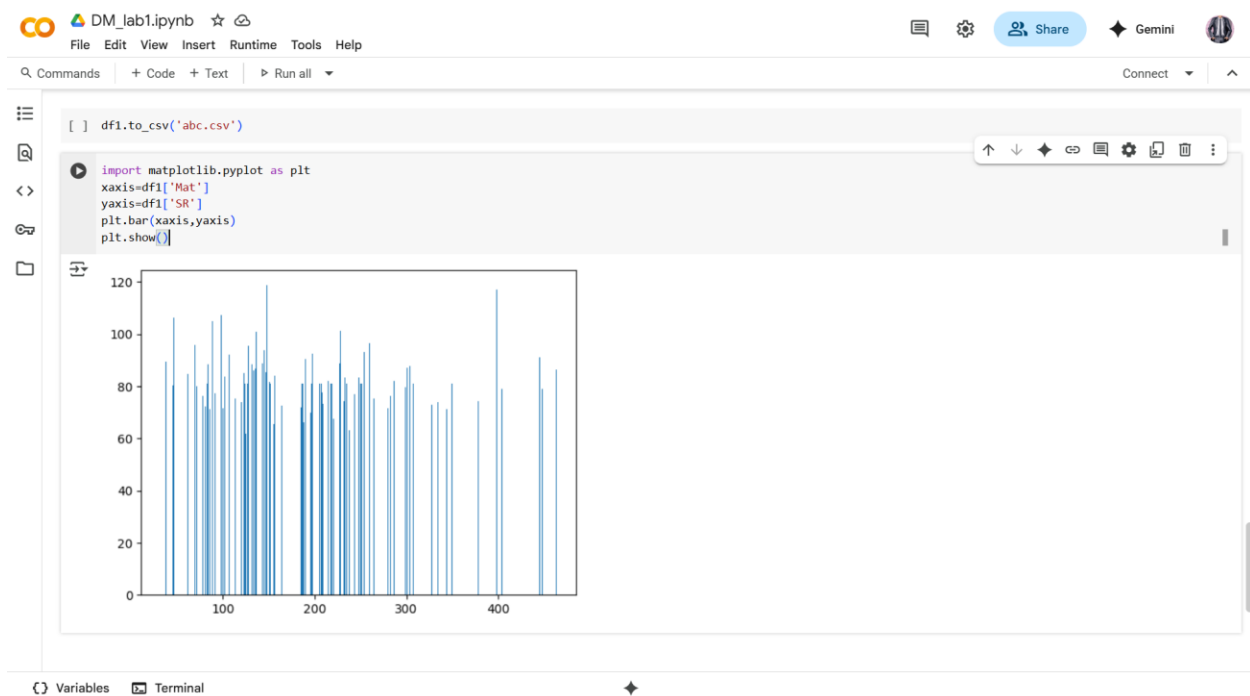
    df1['Runs'].fillna(df1['Runs'].median(),inplace=True)

{} Variables   🖵 Terminal   ✦

```
df1.head(70)
```

	Unnamed: 0	Player	Span	Mat	Inns	NO	Runs	HS	Ave	BF	SR	100	50	0	4s	6s
0	0	SR Tendulkar (INDIA)	1989-2012	463	452	41.0	18426.0	200*	44.83	21368.0	86.23	49	96.0	20.000000	2016	195
1	1	V Kohli (INDIA)	2008-2021	254	245	39.0	12169.0	183	59.07	13061.0	93.17	43	62.0	13.000000	1140	125
2	2	RT Ponting (AUS/ICC)	1995-2012	375	365	39.0	13704.0	164	42.03	17046.0	80.39	30	82.0	20.000000	1231	162
3	3	RG Sharma (INDIA)	2007-2021	227	220	32.0	9205.0	264	48.96	10354.0	88.90	29	43.0	13.000000	832	244
4	4	ST Jayasuriya (Asia/SL)	1989-2011	445	433	18.0	13430.0	189	32.36	14725.0	91.20	28	68.0	34.000000	1500	270
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
65	65	Shoaib Malik (PAK)	1999-2019	287	258	40.0	7534.0	143	34.55	9199.0	81.90	9	36.5	10.925926	NaN	113
66	66	CS MacLeod (SCOT)	2008-2021	72	70	6.0	2290.0	175	35.78	2863.0	79.98	8	36.5	10.925926	NaN	25
67	67	Salman Butt (PAK)	2004-2010	78	78	4.0	2725.0	136	36.82	3572.0	76.28	8	36.5	10.925926	NaN	7
68	68	GA Gooch (ENG)	1976-1995	125	122	6.0	5782.5	142	36.98	6932.0	61.88	8	36.5	10.925926	NaN	15
69	69	Mushfiqur Rahim (BAN)	2006-2021	227	213	36.0	6581.0	144	37.18	8325.0	79.05	8	36.5	10.925926	NaN	85

70 rows × 16 columns

```
df1.to_csv('abc.csv')
```

```
import matplotlib.pyplot as plt
xaxis=df1['Mat']
yaxis=df1['SR']
plt.bar(xaxis,yaxis)
plt.show()
```

{} Variables   ⌨ Terminal

Ober

```
df1.to_csv('abc.csv')
```

```
import matplotlib.pyplot as plt
xaxis=df1['Mat']
yaxis=df1['SR']
plt.bar(xaxis,yaxis)
plt.show()
```


{} Variables   ⌨ Terminal

Observations:

1: we can say that sachin tedulkar is highest run getter

2:many cricketers have played more than 400 matches