NEXTHIKES IT SOLUTIONS

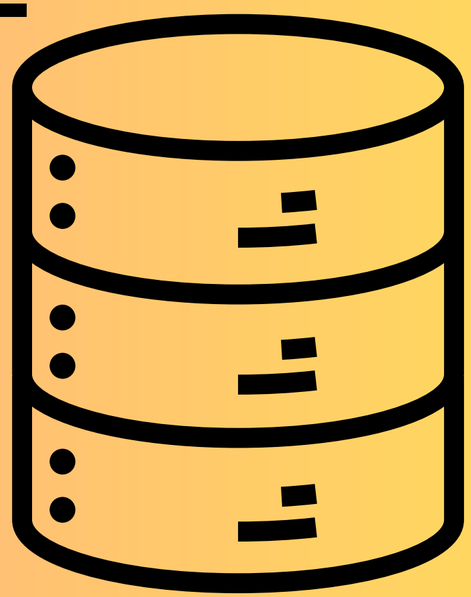
BY SHIVANI SABHARWAL

PROJECT-2

PRE-PROCESSING OF DATA -SETS

TECHNICAL SKILLS-

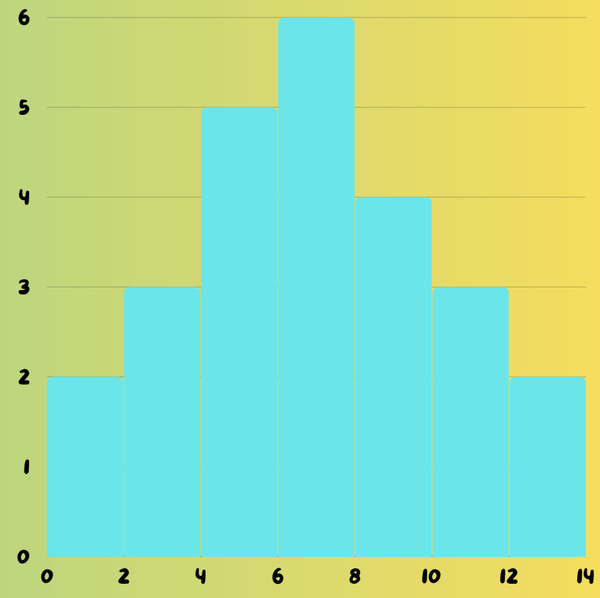
1.DATA ACQUISITION
2.DATA WRANGLING



PRE-PROCESSING OF DATA -SETS

USING PYTHON LIBRARIES

1. USING NUMPY
2. USING PANDAS
3. USING MATPLOT LIB
4. USING SEABORNE



STEPS -

1.DOWNLOAD DATASET1 AND PRE-PROCESS IT.

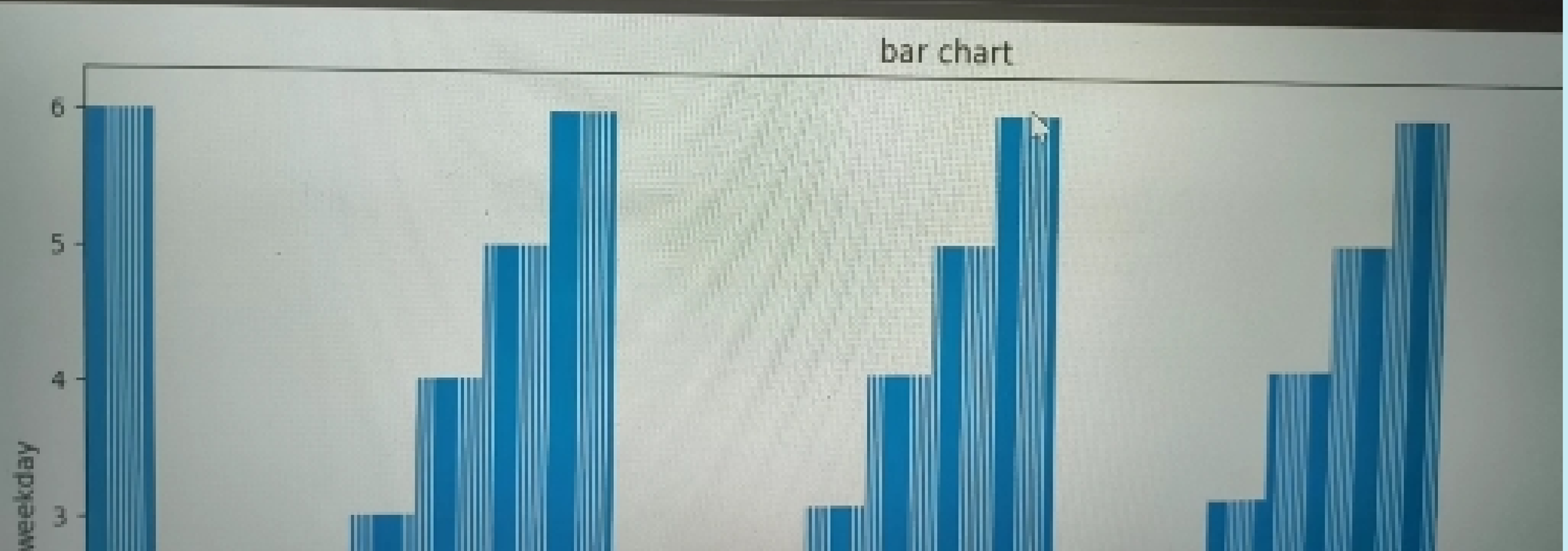
2.FOR PRE-PROCESSING -
A.HANDLE MISSING VALUES IF ANY.
B.CHECK IF THERE ARE DUPLICATE VALUES.
C.DATA TYPE GIVEN IS CORRECT OR NOT.

3.CREATE VISUALS EDA FOR GIVEN DATA SET.

4.DOWNLOAD DATASET2 AND PRE-PROCESS IT IN SAME WAY.

DATASETI

```
plt.figure(figsize=(12,6))
sns.barplot(x='instant',y='weekday',data=ds)
plt.title('bar chart')
plt.xlabel('instant')
plt.ylabel('weekday')
plt.show()
```

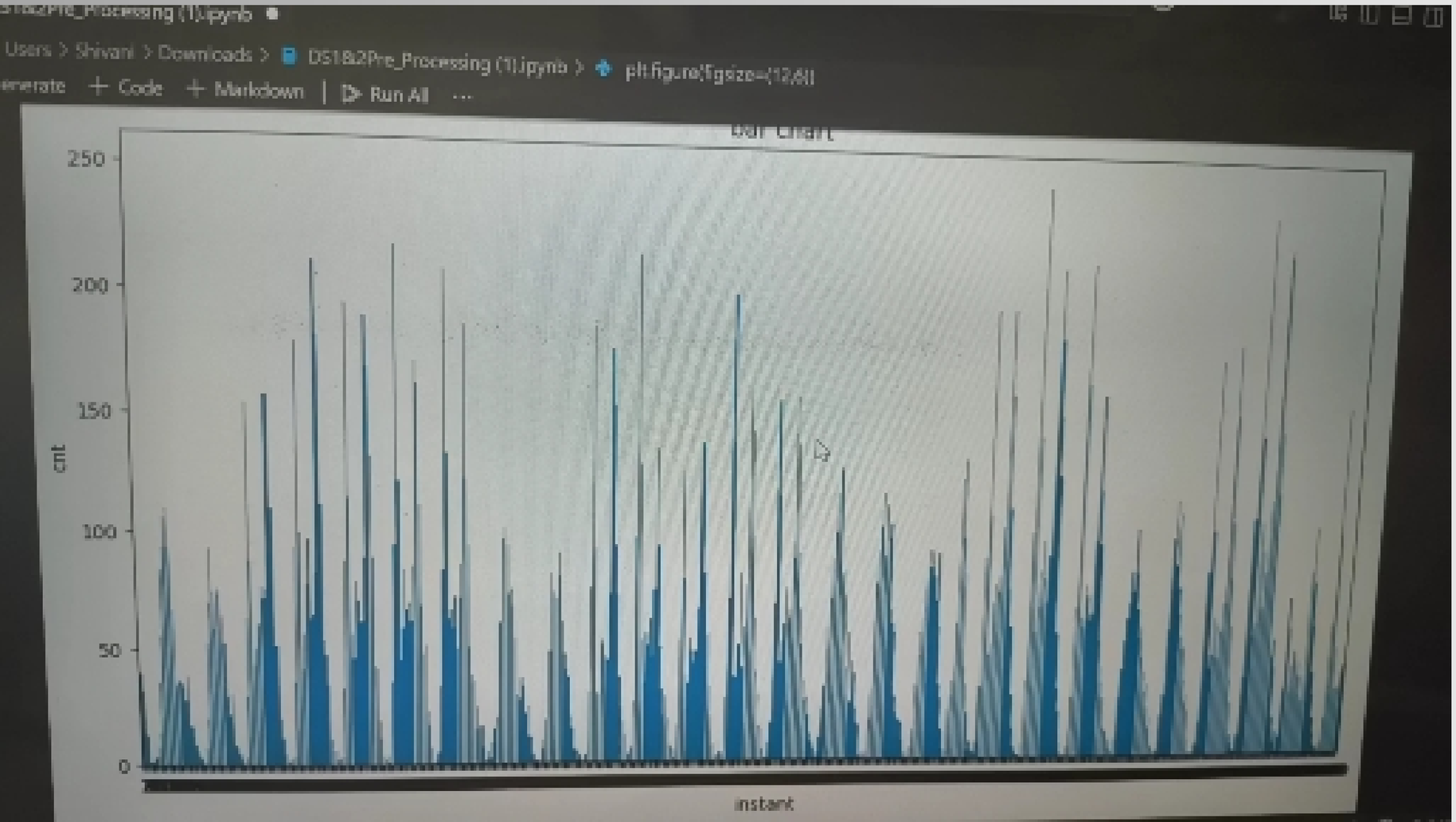


C: > Users > Shivani > Downloads > 05182Pw_Processing (1).ipynb > plt.figure(figsize=(12,6))

Generate + Code + Markdown | Run All ...

Ln 2, Col 19 (26 selected)

DATASET2


Users > Shivani > Downloads > DS1&2Pre_Processing (1).ipynb > plt.figure(figsize=(12,6))
+ Code + Markdown | Run All ...
Bar Chart
250
200
150
100
50
0
cnt
instant

MERGE BOTH THE DATA SETSWITH THE COMMON COLUMN I.E INSTANT

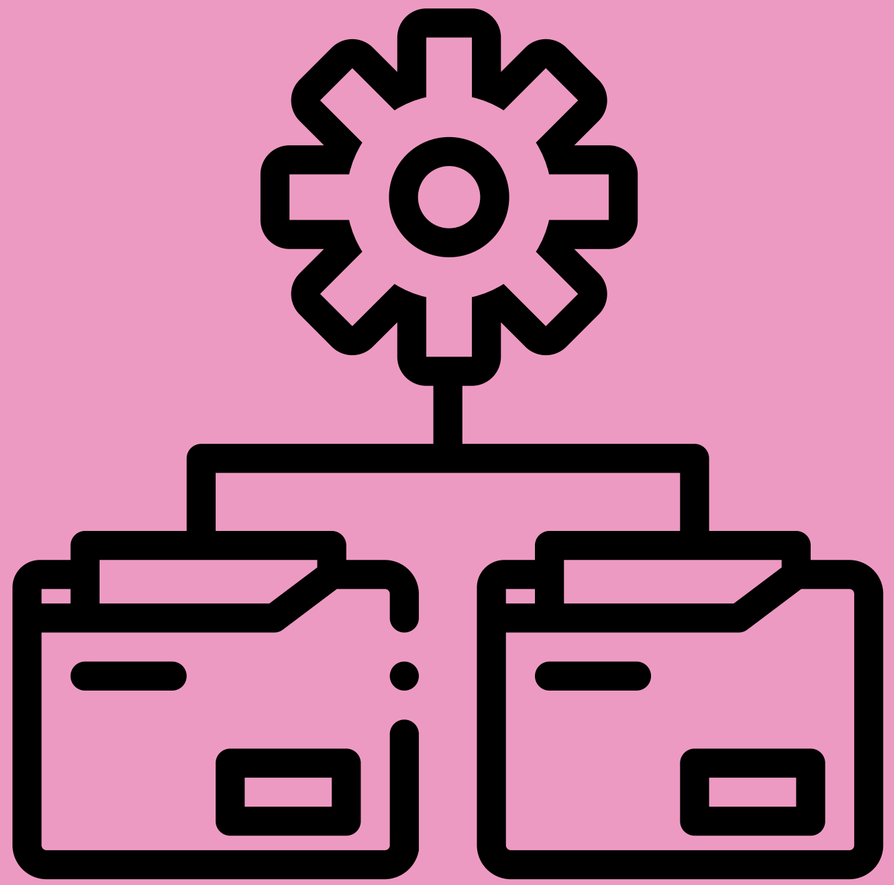
```
master_data = pd.merge(ds,df,on='instant',how = 'inner')
print(master_data)
```

```
      instant      dteday  season  yr  mnth  hr  holiday  weekday  weathersit \
0           1  01-01-2011       1   0     1   0    False        6           1
1           2  01-01-2011       1   0     1   1    False        6           1
2           3  01-01-2011       1   0     1   2    False        6           1
3           4  01-01-2011       1   0     1   3    False        6           1
4           5  01-01-2011       1   0     1   4    False        6           1
..        ...         ...     ...  ..   ...  ..      ...      ...         ...
605       606  28-01-2011       1   0     1  11    False        5           3
606       607  28-01-2011       1   0     1  12    False        5           3
607       608  28-01-2011       1   0     1  13    False        5           3
608       609  28-01-2011       1   0     1  14    False        5           3
609       610  28-01-2011       1   0     1  15    False        5           2

     temp  Unnamed: 0   atemp   hum  windspeed  casual  registered  cnt
0    0.24           0  0.2879  0.81     0.0000       3          13   16
1    0.22           1  0.2727  0.80     0.0000       8          32   40
2    0.22           2  0.2727  0.80     0.0000       5          27   32
3    0.24           3  0.2879  0.75     0.0000       3          10   13
4    0.24           4  0.2879  0.75     0.0000       0           1    1
..    ...         ...     ...   ...        ...     ...         ...  ...
605  0.18         605  0.2121  0.93     0.1045       0          30   30
606  0.18         606  0.2121  0.93     0.1045       1          28   29
607  0.18         607  0.2121  0.93     0.1045       0          31   31
```

LOADING DATASET INTO CSV FILE

LOADING DATA INTO CSV FORMAT

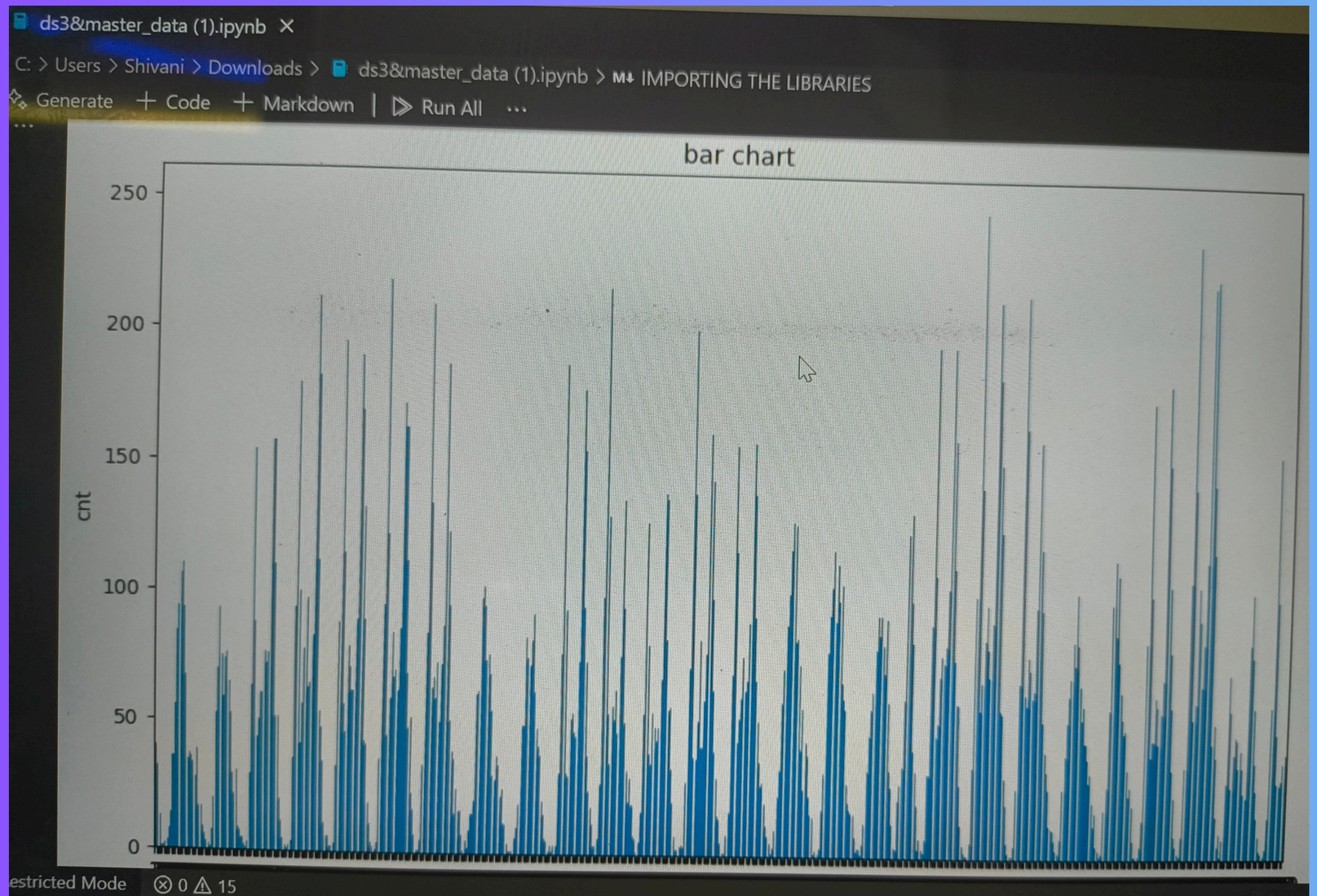
```
ds2 = pd.DataFrame(ds2)
ds2.to_csv('master_data.csv')
```



1.OPEN A NEW NOTEBOOK
2.DOWNLOAD DATASET3
3.PRE-PEOCESS DATASET3
4.CREATE VISUALS FOR DATASET3.

VISUALIZATON.OF DATASET3

VISUALIZATION OF MASTER DATASET



MERGING OF DATASET3 AND MASTER DATASET

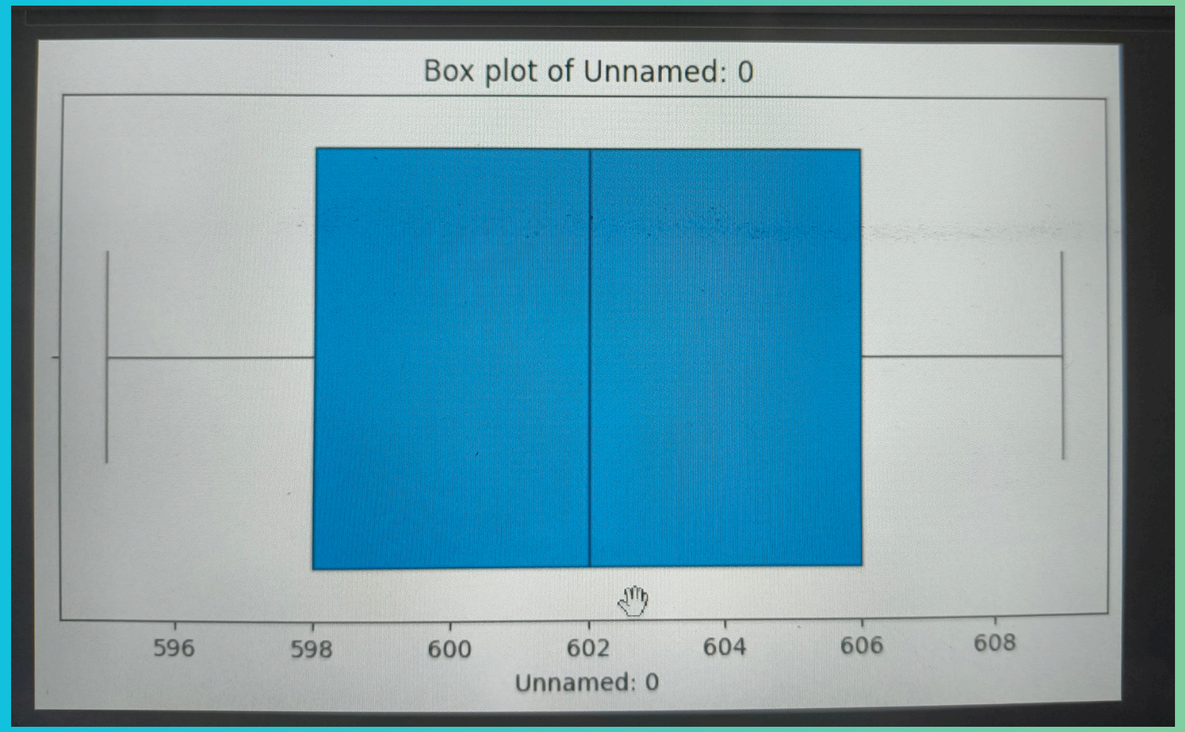
```
Final_data = pd.merge(ds,df,on='dteday',how = 'inner')
print(Final_data)
```

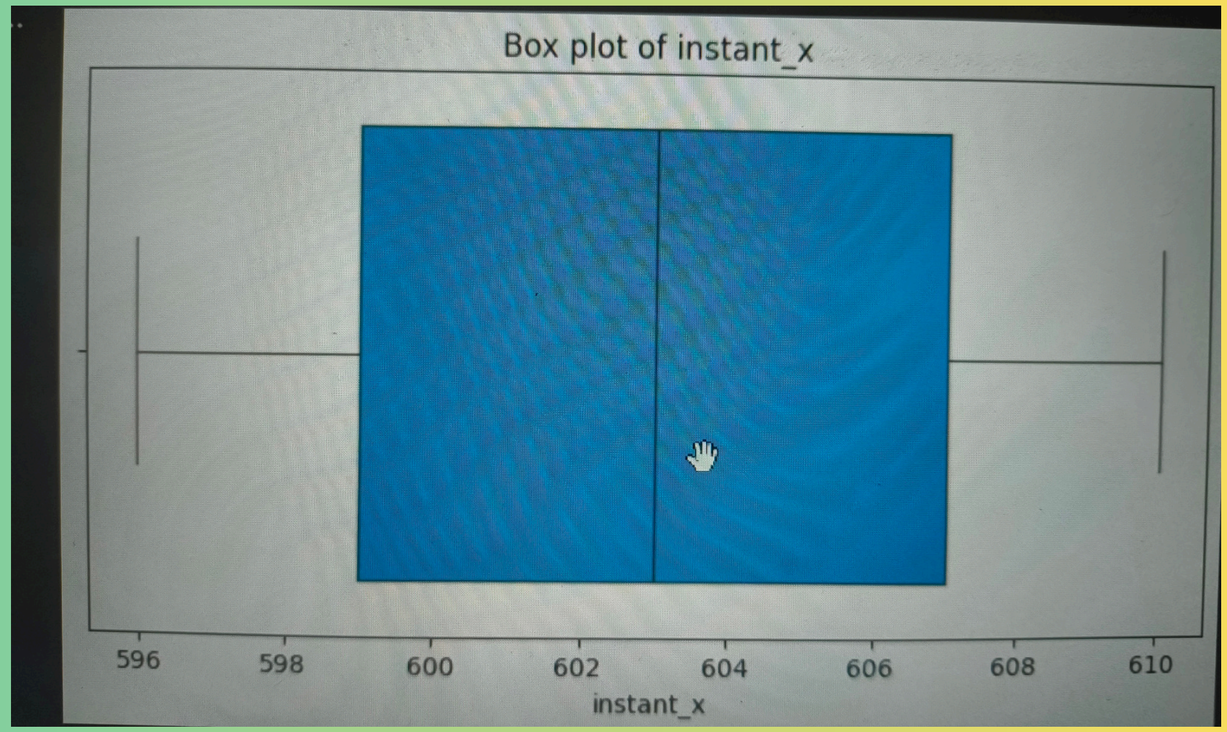
```
      Unnamed: 0  instant_x      dteday  season_x  yr_x  mnth_x  hr_x  \
0            595        596  28-01-2011         1     0       1     0
1            595        596  28-01-2011         1     0       1     0
2            595        596  28-01-2011         1     0       1     0
3            595        596  28-01-2011         1     0       1     0
4            595        596  28-01-2011         1     0       1     0
..           ...        ...         ...       ...   ...     ...   ...
115          609        610  28-01-2011         1     0       1    15
116          609        610  28-01-2011         1     0       1    15
117          609        610  28-01-2011         1     0       1    15
118          609        610  28-01-2011         1     0       1    15
119          609        610  28-01-2011         1     0       1    15

     holiday_x  weekday_x  weathersit_x  ...  holiday_y  weekday_y  \
0        False          5             2  ...      False          5
1        False          5             2  ...      False          5
2        False          5             2  ...      False          5
3        False          5             2  ...      False          5
4        False          5             2  ...      False          5
..         ...        ...           ...  ...        ...        ...
```

LOAD DATA INTO CSV FILES

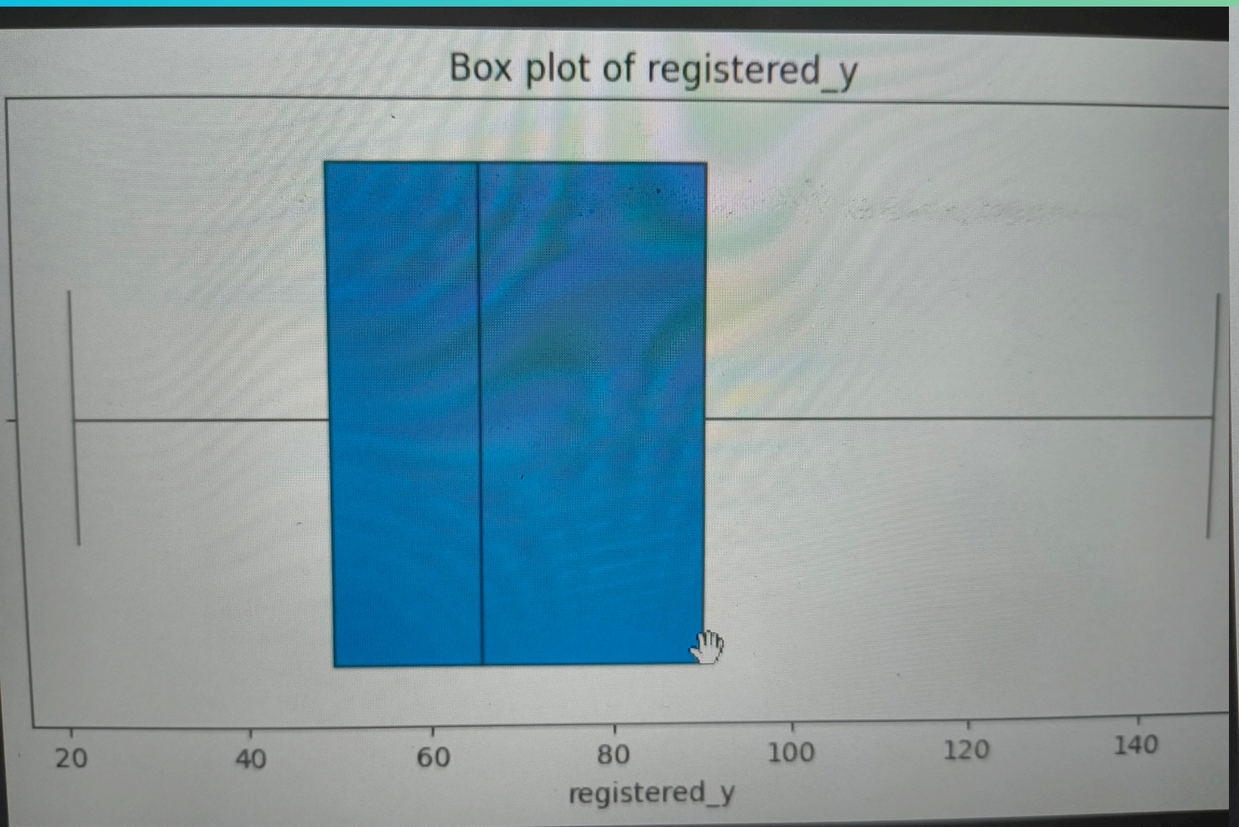
```
Final_data.to_csv('Final_data.csv')
```

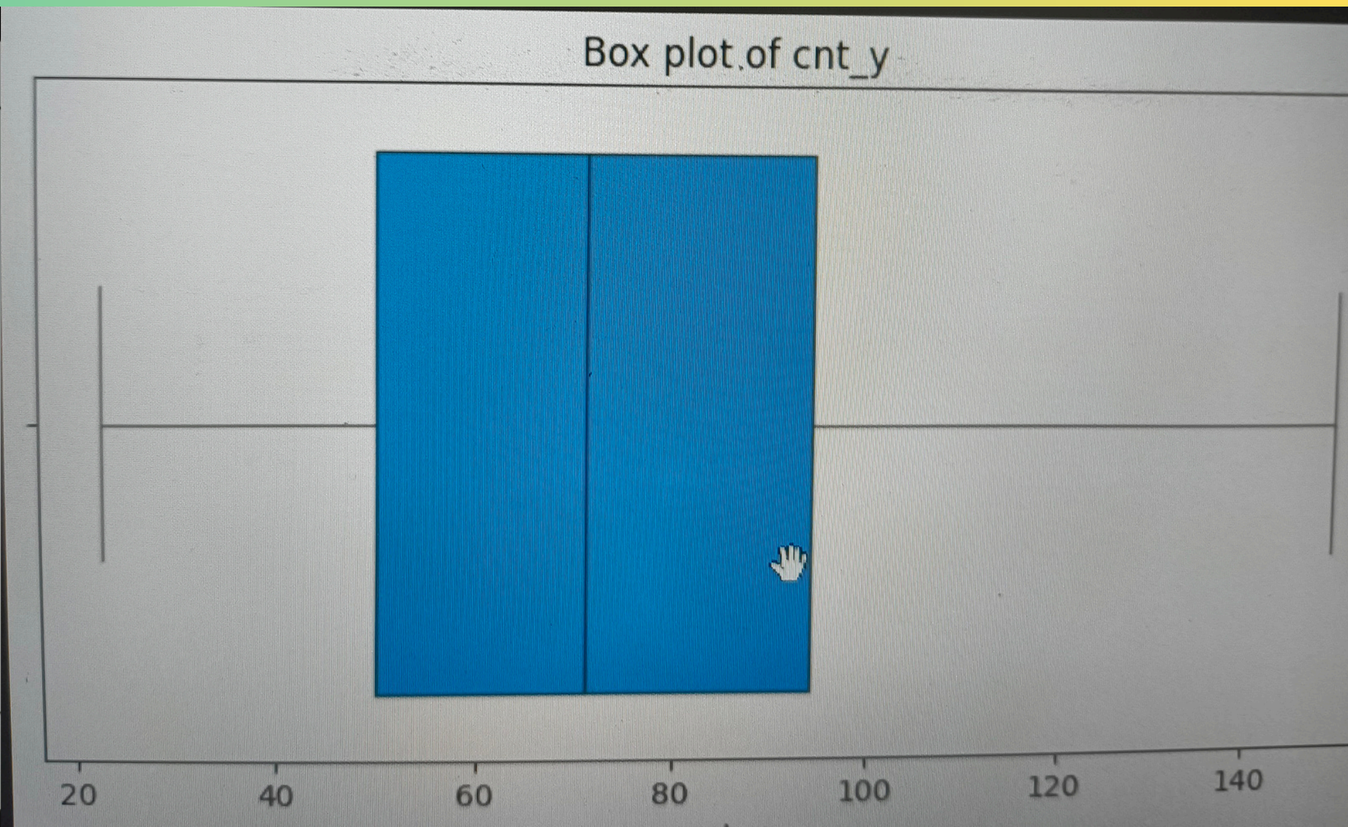
CHECKING IF THERE ARE ANY OUTLIERS.

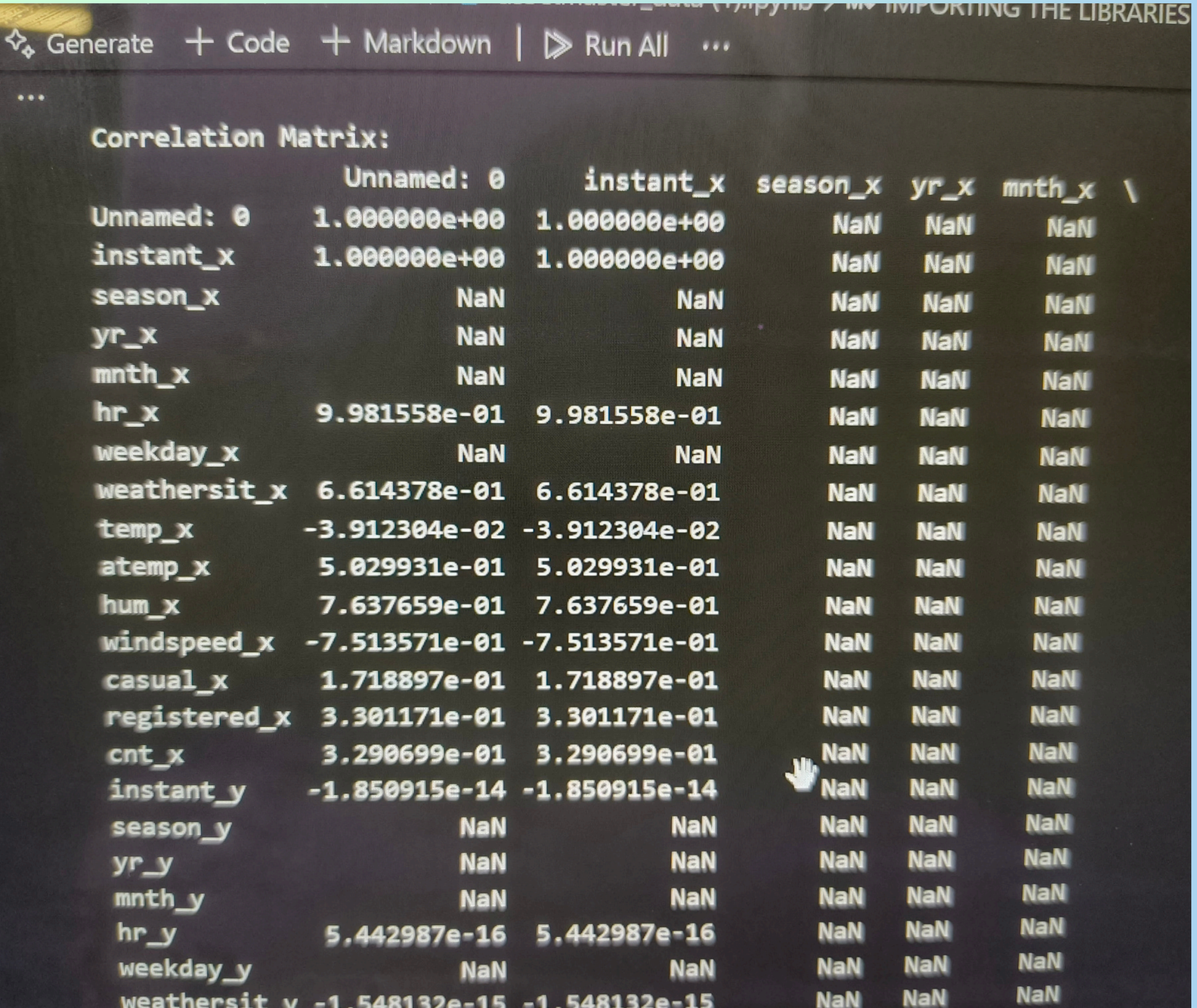




CHECKING IF THERE ARE ANY OUTLIERS.







Correlation Matrix:

	Unnamed: 0	instant_x	season_x	yr_x	mnth_x	\
Unnamed: 0	1.000000e+00	1.000000e+00	NaN	NaN	NaN	
instant_x	1.000000e+00	1.000000e+00	NaN	NaN	NaN	
season_x	NaN	NaN	NaN	NaN	NaN	
yr_x	NaN	NaN	NaN	NaN	NaN	
mnth_x	NaN	NaN	NaN	NaN	NaN	
hr_x	9.981558e-01	9.981558e-01	NaN	NaN	NaN	
weekday_x	NaN	NaN	NaN	NaN	NaN	
weathersit_x	6.614378e-01	6.614378e-01	NaN	NaN	NaN	
temp_x	-3.912304e-02	-3.912304e-02	NaN	NaN	NaN	
atemp_x	5.029931e-01	5.029931e-01	NaN	NaN	NaN	
hum_x	7.637659e-01	7.637659e-01	NaN	NaN	NaN	
windspeed_x	-7.513571e-01	-7.513571e-01	NaN	NaN	NaN	
casual_x	1.718897e-01	1.718897e-01	NaN	NaN	NaN	
registered_x	3.301171e-01	3.301171e-01	NaN	NaN	NaN	
cnt_x	3.290699e-01	3.290699e-01	NaN	NaN	NaN	
instant_y	-1.850915e-14	-1.850915e-14	NaN	NaN	NaN	
season_y	NaN	NaN	NaN	NaN	NaN	
yr_y	NaN	NaN	NaN	NaN	NaN	
mnth_y	NaN	NaN	NaN	NaN	NaN	
hr_y	5.442987e-16	5.442987e-16	NaN	NaN	NaN	
weekday_y	NaN	NaN	NaN	NaN	NaN	
weathersit_y	-1.548132e-15	-1.548132e-15	NaN	NaN	NaN	

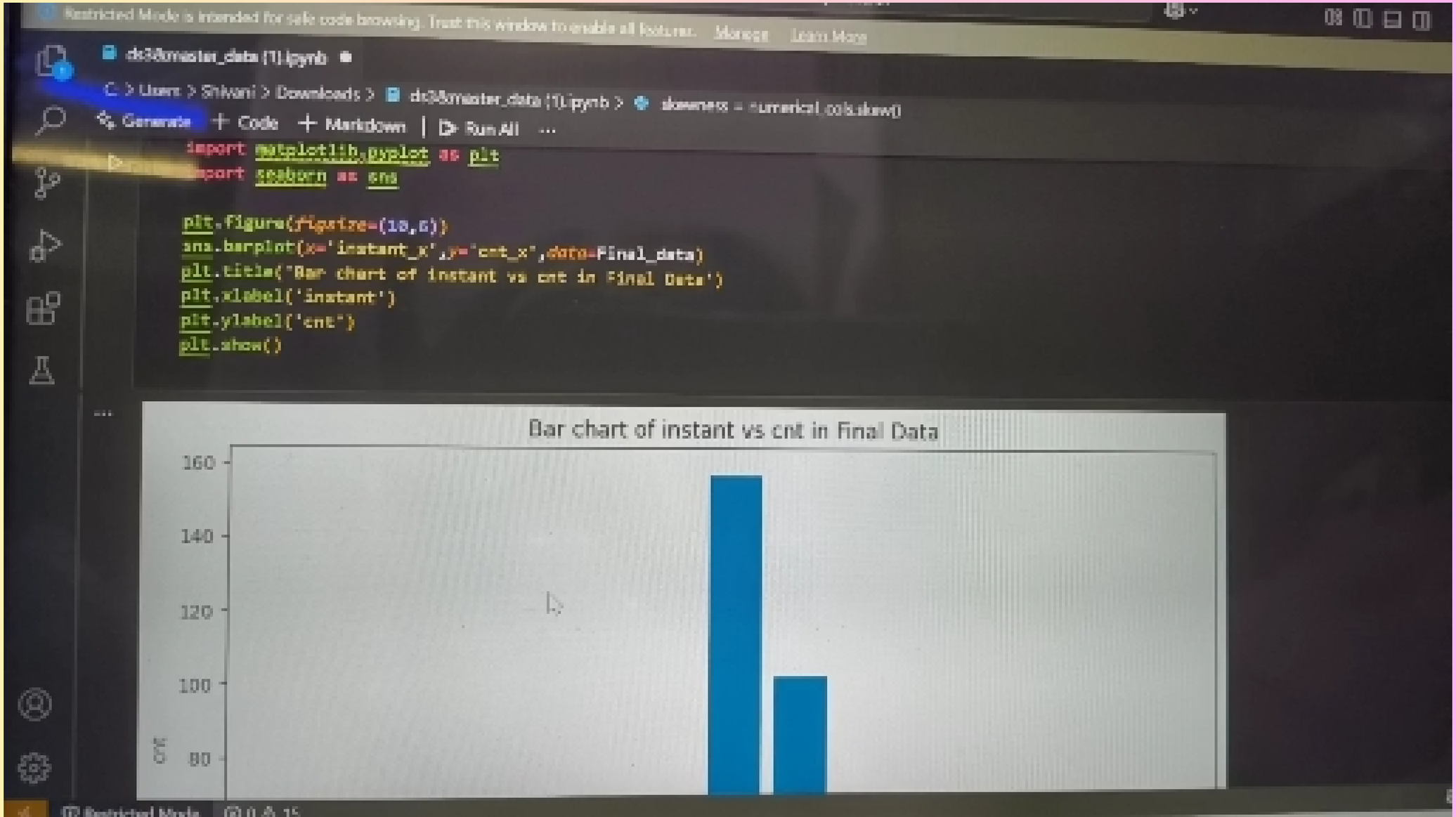
CORRELATIONS MATRIX

Skewness:

Unnamed: 0	0.000000
instant_x	0.000000
season_x	0.000000
yr_x	0.000000
mnth_x	0.000000
hr_x	-0.123522
weekday_x	0.000000
weathersit_x	0.413434
temp_x	0.152838
atemp_x	1.678722
hum_x	0.125971
windspeed_x	-1.179283
casual_x	2.165740
registered_x	1.715044
cnt_x	1.681389
instant_y	0.000000
season_y	0.000000
yr_y	0.000000
mnth_y	0.000000
hr_y	0.000000
weekday_y	0.000000
weathersit_y	0.522958
temp_y	-1.169369

SKEWNESS

VISUALS ON FINAL DATASET



REFERENCES -

1.PROJECT DOCUMENT
2.LINK GIVEN BY NEXTHIKES IT SOLUTIONS - HTTPS://PANDAS.PYDATA.ORG/DOCS/REFERENCE/API/PANDAS.DATAFRAME.TO_CSV.HTML

THANK YOU