APSSDC

Andhra Pradesh State Skill Development Corporation

Skill AP
APSSDC

(https://apssdc.in)

Data Analysis with Pandas

Day Objectives

- Importing of Data from Multiple files
- Combining/merging of DataFrames (JOIN)
- Grouping
- Statistics
- Sorting data
- Data Visualization
- Data Cleaning / Data Preprocessing

In [1]:

```
import pandas as pd
```

In [2]:

```
df = pd.read_csv("dhs_daily_report.csv", index_col = 0)
df.head()
```

Out[2]:

	adult_families_in_shelter	adults_in_families_with_children_in_shelter	children_in_families_with_
0	1796	14607	
1	1803	14622	
2	1802	14611	
3	1801	14650	
4	1804	14694	

In [4]:

```
df.shape
```

Out[4]:

```
(1000, 14)
```

In [5]:

```
df.columns
```

Out[5]:

```
Index(['adult_families_in_shelter',
       'adults_in_families_with_children_in_shelter',
       'children_in_families_with_children_in_shelter', 'date_of_census',
       'families_with_children_in_shelter',
       'individuals_in_adult_families_in_shelter',
       'single_adult_men_in_shelter', 'single_adult_women_in_shelter',
       'total_adults_in_shelter', 'total_children_in_shelter',
       'total_individuals_in_families_with_children_in_shelter_',
       'total_individuals_in_shelter', 'total_single_adults_in_shelter',
       'individuals_in_shelter'],
      dtype='object')
```

In [6]:

```
df['adult_families_in_shelter'].max()
```

Out[6]:

```
2356
```

In [7]:

```
df['adult_families_in_shelter'].argmax()
```

Out[7]:

```
983
```

In [8]:

```
df['date_of_census'].iloc[983]
```

Out[8]:

```
'2016-06-06T00:00:00.000'
```

In [9]:

```
df['adult_families_in_shelter'].min()
```

Out[9]:

```
1796
```

In [10]:

```python
df['adult_families_in_shelter'].argmin()
```

Out[10]:

```
0
```

In [16]:

```python
(df['adult_families_in_shelter'] == df['adult_families_in_shelter'].min()).sum()
```

Out[16]:

```
1
```

In [17]:

```python
df.describe()
```

Out[17]:

	adult_families_in_shelter	adults_in_families_with_children_in_shelter	children_in_families_v
count	1000.000000	1000.000000	
mean	2074.955000	16487.932000	
std	148.020238	848.363772	
min	1796.000000	14607.000000	
25%	1906.000000	15831.500000	
50%	2129.000000	16836.000000	
75%	2172.250000	17118.250000	
max	2356.000000	17733.000000	

In [18]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1000 entries, 0 to 999
Data columns (total 14 columns):
 #   Column                                         Non-Null Cou
nt  Dtype
---  ------                                         ------------
--  -----
 0   adult_families_in_shelter                      1000 non-nul
l   int64
 1   adults_in_families_with_children_in_shelter    1000 non-nul
l   int64
 2   children_in_families_with_children_in_shelter  1000 non-nul
l   int64
 3   date_of_census                                 1000 non-nul
l   object
 4   families_with_children_in_shelter              1000 non-nul
l   int64
 5   individuals_in_adult_families_in_shelter       999 non-null
float64
```

In [19]:

```
df.head()
```

Out[19]:

ilies_with_children_in_shelter	children_in_families_with_children_in_shelter	date_of_census	familie
14607	21314	2013-08-21T00:00:00.000	
14622	21324	2013-08-22T00:00:00.000	
14611	21291	2013-08-23T00:00:00.000	
14650	21343	2013-08-24T00:00:00.000	
14694	21400	2013-08-25T00:00:00.000	

In [4]:

```
df['date_of_census'] = df['date_of_census'].astype('datetime64')
```

In [5]:

```python
df.sort_values('adult_families_in_shelter')
```

Out[5]:

	adult_families_in_shelter	adults_in_families_with_children_in_shelter	children_in_families_with_children_in_
0	1796	14607	
3	1801	14650	
2	1802	14611	
1	1803	14622	
7	1803	14647	
...	...	...	
997	2352	17202	
986	2353	17186	
984	2353	17125	

In [6]:

```python
help(df.sort_values)
```

```
Help on method sort_values in module pandas.core.frame:

sort_values(by, axis=0, ascending=True, inplace=False, kind='quicksort', n
a_position='last', ignore_index=False) method of pandas.core.frame.DataFra
me instance
    Sort by the values along either axis.

    Parameters
    ----------
            by : str or list of str
                Name or list of names to sort by.

                - if `axis` is 0 or `'index'` then `by` may contain index
                  levels and/or column labels.
                - if `axis` is 1 or `'columns'` then `by` may contain colu
mn
                  levels and/or index labels.

                .. versionchanged:: 0.23.0
```

In [8]:

```python
df2 = df.sort_values(['adult_families_in_shelter', 'children_in_families_with_children_
```

In [9]:

```
df2.head()
```

Out[9]:

```
   adult_families_in_shelter  adults_in_families_with_children_in_shelter  children_in_families_with_
0                       1796                                        14607
3                       1801                                        14650
2                       1802                                        14611
7                       1803                                        14647
1                       1803                                        14622
```

In [11]:

```
df.sort_index(ascending = False)
```

Out[11]:

```
     adult_families_in_shelter  adults_in_families_with_children_in_shelter  children_in_families_with_children_in_
999                       2346                                        17166
998                       2347                                        17173
997                       2352                                        17202
996                       2347                                        17219
995                       2341                                        17223
...                        ...                                          ...
  4                       1804                                        14694
  3                       1801                                        14650
  2                       1802                                        14611
```

In [ ]:

```

```

In [17]:

```
d1 = pd.read_excel("Day01_25Nov2020.xls", skiprows = 6)
d2 = pd.read_excel("Day02_26Nov2020.xls", skiprows = 6)
```

In [22]:

```
d1.shape, d2.shape
```

Out[22]:

```
((75, 7), (70, 7))
```

In [23]:

```
cdf = pd.concat([d1, d2])
```

In [24]:

```
cdf.shape
```

Out[24]:

```
(145, 7)
```

In [25]:

```
cdf = pd.concat([d1, d2], axis = 'columns')
```

In [26]:

```
cdf.shape
```

Out[26]:

```
(75, 14)
```

In [28]:

```
cdf.head()
```

Out[28]:

	Name	Email Address	Join Time	Leave Time	Time in Session (minutes)	Unnamed: 5	Unnamed: 6	Name
0	17X41A1202-Sai Harini-SRKIT	harini.akkineni@outlook.com	9:36 AM	12:14 PM	158	NaN	NaN	17X41A1202-Sai Harini-SRKIT
1	17X41A1203-Tejaswini Alapati-SRKIT	alapatitejaswini999@gmail.com	9:29 AM	12:13 PM	164	NaN	NaN	17X41A1203-Tejaswini Alapati-SRKIT
2	17X41A1204-Karthik-X4	NaN	9:19 AM	12:14 PM	174	NaN	NaN	17X41A1204-Karthik-X4
3	17X41A1207-Bollu Bhavana-X4	NaN	9:30 AM	11:52 AM	142	NaN	NaN	17X41A1207-Bollu Bhavana-X4

In [29]:

```
cdf.tail()
```

Out[29]:

	Name	Email Address	Join Time	Leave Time	Time in Session (minutes)	Unnamed: 5	Unnamed: 6	Name
70	Vikram Parmar	NaN	9:29 AM	9:30 AM	0	NaN	NaN	NaN
71	Vikram Parmar	NaN	9:53 AM	9:54 AM	1	NaN	NaN	NaN
72	Yashwanth	rajyashwanth27@gmail.com	9:31 AM	9:32 AM	0	NaN	NaN	NaN
73	navya	NaN	9:26 AM	9:29 AM	3	NaN	NaN	NaN
74	navya	NaN	9:29 AM	12:18 PM	168	NaN	NaN	NaN

In [31]:

```
d1.merge(d2, on = 'Name')
```

Out[31]:

	Name	Email Address_x	Join Time_x	Leave Time_x	Time in Session (minutes)_x	Unnamed: 5_x	Unnamed: 6_x	
0	17X41A1202-Sai Harini-SRKIT	harini.akkineni@outlook.com	9:36 AM	12:14 PM	158	NaN	NaN	harin
1	17X41A1203-Tejaswini Alapati-SRKIT	alapatitejaswini999@gmail.com	9:29 AM	12:13 PM	164	NaN	NaN	alapatite
2	17X41A1204-Karthik-X4	NaN	9:19 AM	12:14 PM	174	NaN	NaN	
3	17X41A1207-Bollu Bhavana-X4	NaN	9:30 AM	11:52 AM	142	NaN	NaN	

In [34]:

```
d1.merge(d2, left_on = 'Email Address', right_on = 'Name')
```

Out[34]:

Name_x	Email Address_x	Join Time_x	Leave Time_x	Time in Session (minutes)_x	Unnamed: 5_x	Unnamed: 6_x	Name_y	Emai Address_y

In [35]:

```
group = d1.groupby('Name')
```

In [37]:

```
group
```

Out[37]:

```
<pandas.core.groupby.generic.DataFrameGroupBy object at 0x0000023572A699D0>
```

In [38]:

```
group.sum()
```

17X41A1234-Divyasri-SRKIT	173	0.0	0.0
17X41A1235 R.V.Lahari SRKIT	163	0.0	0.0
17X41A1236-sk anjum Kousar-SRKIT	158	0.0	0.0
17X41A1239-Sheereen-SRKIT	165	0.0	0.0
17X41A1243-syed yasmeen-SRKIT	165	0.0	0.0
17X41A1247-V Ramya Sri - X4	166	0.0	0.0
17x41a1224-Gopichand(Srkit)-x4	52	0.0	0.0
17x41a1237 - Aslam- X4	39	0.0	0.0
17x41a1237-sk.aslam-x4	133	0.0	0.0
Abinaya	0	0.0	0.0
Amritha Mishra	14	0.0	0.0
Anil Kumar Teegala [APSSDC]	361	0.0	0.0
Bhuvaneswari Bonthu 1208	158	0.0	0.0

In [39]:

```python
group.count()
```

Out[39]:

	Email Address	Join Time	Leave Time	Time in Session (minutes)	Unnamed: 5	Unnamed: 6
Name						
17X41A1202-Sai Harini-SRKIT	1	1	1	1	0	0
17X41A1203-Tejaswini Alapati-SRKIT	1	1	1	1	0	0
17X41A1204-Karthik-X4	0	1	1	1	0	0
17X41A1207-Bollu Bhavana-X4	0	2	2	2	0	0
17X41A1209-HARSHA-X4	0	2	2	2	0	0
17X41A1210-Gopi Chand-SRKIT	1	1	1	1	0	0

In [40]:

```python
group = d1.groupby(['Name', 'Email Address'])
```

In [43]:

```python
df5 = group.sum()
df5.head()
```

Out[43]:

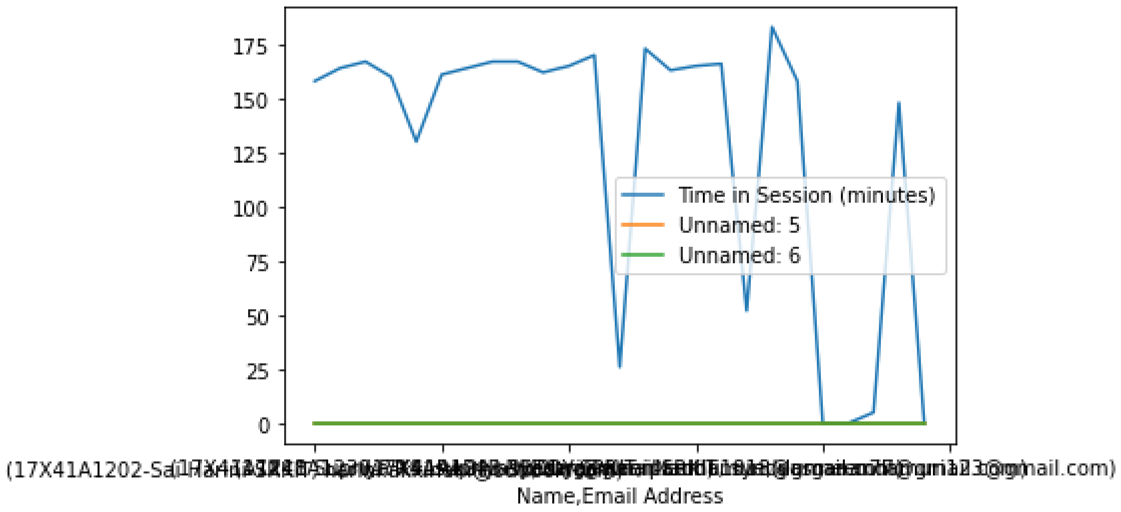
		Time in Session (minutes)	Unnamed: 5	Unnamed: 6
Name	**Email Address**			
17X41A1202-Sai Harini-SRKIT	**harini.akkineni@outlook.com**	158	0.0	0.0
17X41A1203-Tejaswini Alapati-SRKIT	**alapatitejaswini999@gmail.com**	164	0.0	0.0
17X41A1210-Gopi Chand-SRKIT	**gopichandcherukuri121@gmail.com**	167	0.0	0.0
17X41A1212-yashwanth-X4	**rajyashwanth27@gmail.com**	160	0.0	0.0
17X41A1214-Gorantla Ram Gopal-X4	**ramgopalgorantla43@gmail.com**	130	0.0	0.0

In [44]:

```
df5.plot()
```

Out[44]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x235745ec220>
```



- Finding and Removing Duplicates
- Identifying and Elemenating Outliers
- Identifying and working on missing values

Identifying and Elemenating Outliers

the data which is far away from the min/max value data

- Open Price
- Avg Price
- Total Price
- Closing Price

In [64]:

```
import numpy as np

stock = np.array([5.5, 5.2, 5.3, 5.6, 2.2, 5.5, 5.2, 5.3,10])
stock1 = np.array([5.5, 5.2, 5.3, 5.6, 5.525, 5.5, 5.2, 5.3])
stock.mean()
```

Out[64]:

```
5.533333333333333
```

In [58]:

```
np.median(stock), np.median(stock1)
```

Out[58]:

```
(5.3, 5.4)
```

In [57]:

```
stock.mean(), stock1.mean()
```

Out[57]:

```
(4.975, 5.390625)
```

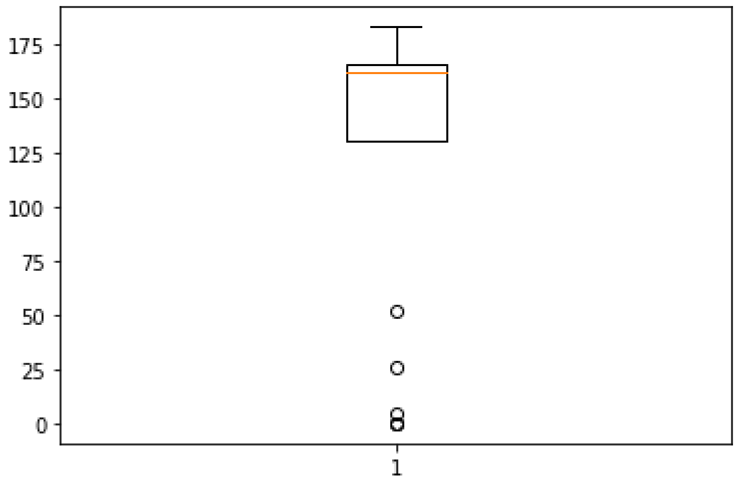
In [60]:

```
import matplotlib.pyplot as plt

plt.boxplot(df5['Time in Session (minutes)'])
```

Out[60]:

```
{'whiskers': [<matplotlib.lines.Line2D at 0x235778e12e0>,
  <matplotlib.lines.Line2D at 0x235778e1640>],
 'caps': [<matplotlib.lines.Line2D at 0x235778e19a0>,
  <matplotlib.lines.Line2D at 0x235778e1d00>],
 'boxes': [<matplotlib.lines.Line2D at 0x235778d2f40>],
 'medians': [<matplotlib.lines.Line2D at 0x235778ea0d0>],
 'fliers': [<matplotlib.lines.Line2D at 0x235778ea3d0>],
 'means': []}
```



In [62]:

```
df5['Time in Session (minutes)'].mean(), df5['Time in Session (minutes)'].median()
```

Out[62]:

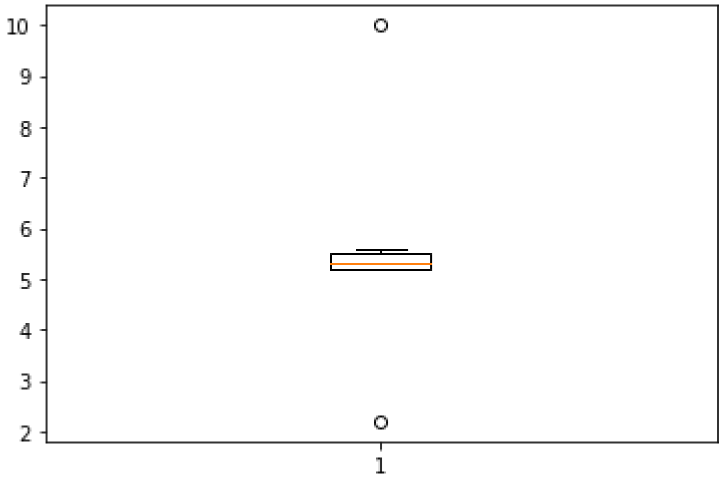
```
(126.96, 162.0)
```

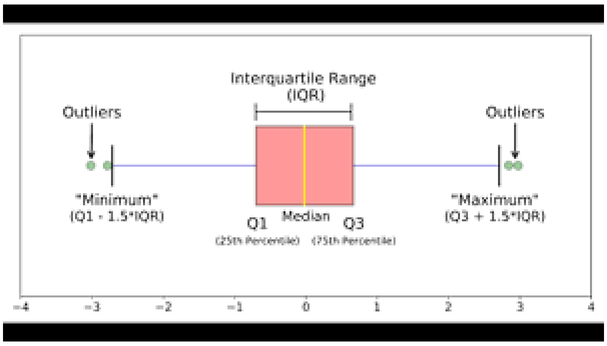
In [65]:

```
plt.boxplot(stock)
```

Out[65]:

```
{'whiskers': [<matplotlib.lines.Line2D at 0x23577990970>,
  <matplotlib.lines.Line2D at 0x23577990cd0>],
 'caps': [<matplotlib.lines.Line2D at 0x2357799a0a0>,
  <matplotlib.lines.Line2D at 0x2357799a400>],
 'boxes': [<matplotlib.lines.Line2D at 0x23577990610>],
 'medians': [<matplotlib.lines.Line2D at 0x2357799a760>],
 'fliers': [<matplotlib.lines.Line2D at 0x2357799aa60>],
 'means': []}
```





In [73]:

```
(df5['Time in Session (minutes)'].quantile(0.25)) - (1.5 * (df5['Time in Session (minut
```

Out[73]:

```
184.0
```

```
- Formulae: data < min and data > max

data[~ data < min or data > max]
```

```
- IQR - InterQuantileRange - Q3-Q1 -> 0.25 - 0.75
- Q1 = 0.25
- Q3 = 0.75
- min = Q1 - 1.5 * IQR = Q1 - 1.5 * (Q3 - Q1)
- max = Q3 + 1.5 * IQR = Q1 - 1.5 * (Q3 - Q1)
```

In [74]:

```
Q1 = df5['Time in Session (minutes)'].quantile(0.25)
Q3 = df5['Time in Session (minutes)'].quantile(0.75)

IQR = Q3 - Q1

mn = Q1 - 1.5 * IQR
mx = Q3 + 1.5 * IQR
```

In [77]:

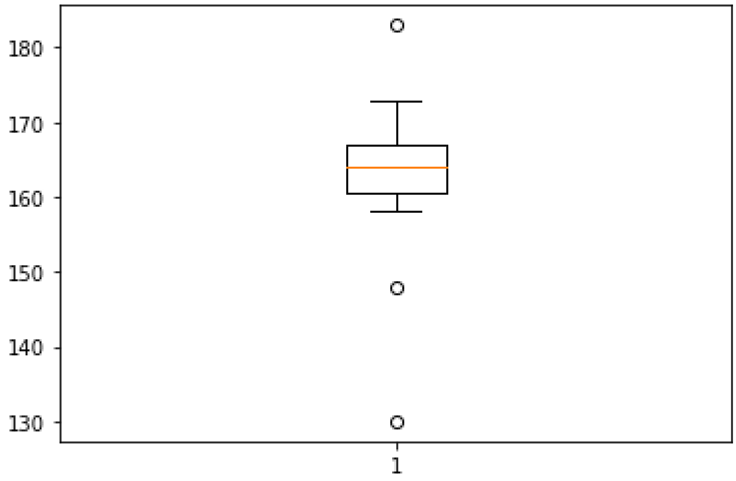
```
x = df5['Time in Session (minutes)'][~ ((df5['Time in Session (minutes)'] > mx) | (df5
```

In [78]:

```
plt.boxplot(x)
```

Out[78]:

```
{'whiskers': [<matplotlib.lines.Line2D at 0x235781854f0>,
  <matplotlib.lines.Line2D at 0x235781859d0>],
 'caps': [<matplotlib.lines.Line2D at 0x23578185cd0>,
  <matplotlib.lines.Line2D at 0x235781042b0>],
 'boxes': [<matplotlib.lines.Line2D at 0x23578185dc0>],
 'medians': [<matplotlib.lines.Line2D at 0x23578104610>],
 'fliers': [<matplotlib.lines.Line2D at 0x23578104910>],
 'means': []}
```



In [79]:

```
(df5['Time in Session (minutes)'] > mx)
```

Out[79]:

```
Name                                   Email Address
17X41A1202-Sai Harini-SRKIT            harini.akkineni@outlook.com        Fa
lse
17X41A1203-Tejaswini Alapati-SRKIT     alapatitejaswini999@gmail.com      Fa
lse
17X41A1210-Gopi Chand-SRKIT            gopichandcherukuri121@gmail.com    Fa
lse
17X41A1212-yashwanth-X4                rajyashwanth27@gmail.com           Fa
lse
17X41A1214-Gorantla Ram Gopal-X4       ramgopalgorantla43@gmail.com       Fa
lse
17X41A1218 Supriya X4                  supriya9kesari@gmail.com           Fa
lse
17X41A1219-Navya Sri - SRKIT           navyakota2000@gmail.com            Fa
lse
17X41A1225-M N Sandhya-SRKIT           sandhyameegada4@gmail.com          Fa
lse
17X41A1226-Mounika Munagala-SRKIT      mounika.munagala143@gmail.com      Fa
lse
17X41A1227-Durga Rani Nandamuri-SRKIT  durganandamuri123@gmail.com        Fa
lse
17X41A1230-P.P.SARADHI REDDY-SRKIT     pardhu1818@gmail.com               Fa
lse
17X41A1232  Maheswari Srkit            mahipottetti1234@gmail.com         Fa
lse
17X41A1233                             potumahesh1234@gmail.com           Fa
lse
17X41A1234-Divyasri-SRKIT              divyasripushadapu@gmail.com        Fa
lse
17X41A1235 R.V.Lahari SRKIT            lahariaug@gmail.com                Fa
lse
17X41A1243-syed yasmeen-SRKIT          syedyasmeen77@gmail.com            Fa
lse
17X41A1247-V Ramya Sri - X4            ramyavagicharla@gmail.com          Fa
lse
17x41a1224-Gopichand(Srkit)-x4         m.gopichand7777@gmail.com          Fa
lse
Anil Kumar Teegala [APSSDC]            aps.sdc.ml@gmail.com               Fa
lse
Bhuvaneswari Bonthu 1208               bhuvanabonthu123@gmail.com         Fa
lse
Durga Rani Nandamuri                   durganandamuri123@gmail.com        Fa
lse
Sairam                                 sairam.ummadisetti@gmail.com       Fa
lse
Surisetti Jayasai                      jaisainaidu123@gmail.com           Fa
lse
Vedasri                                vedasri.avirneni@gmail.com         Fa
lse
Yashwanth                              rajyashwanth27@gmail.com           Fa
lse
Name: Time in Session (minutes), dtype: bool
```

In [86]:

```
df5['Time in Session (minutes)'][~(df5['Time in Session (minutes)'] < mn)]
```

Out[86]:

```
Name                                   Email Address
17X41A1202-Sai Harini-SRKIT            harini.akkineni@outlook.com        15
8
17X41A1203-Tejaswini Alapati-SRKIT     alapatitejaswini999@gmail.com      16
4
17X41A1210-Gopi Chand-SRKIT            gopichandcherukuri121@gmail.com    16
7
17X41A1212-yashwanth-X4                rajyashwanth27@gmail.com           16
0
17X41A1214-Gorantla Ram Gopal-X4       ramgopalgorantla43@gmail.com       13
0
17X41A1218 Supriya X4                  supriya9kesari@gmail.com           16
1
17X41A1219-Navya Sri - SRKIT           navyakota2000@gmail.com            16
4
17X41A1225-M N Sandhya-SRKIT           sandhyameegada4@gmail.com          16
7
17X41A1226-Mounika Munagala-SRKIT      mounika.munagala143@gmail.com      16
7
17X41A1227-Durga Rani Nandamuri-SRKIT  durganandamuri123@gmail.com        16
2
17X41A1230-P.P.SARADHI REDDY-SRKIT     pardhu1818@gmail.com               16
5
17X41A1232  Maheswari Srkit            mahipottetti1234@gmail.com         17
0
17X41A1234-Divyasri-SRKIT              divyasripushadapu@gmail.com        17
3
17X41A1235 R.V.Lahari SRKIT            lahariaug@gmail.com                16
3
17X41A1243-syed yasmeen-SRKIT          syedyasmeen77@gmail.com            16
5
17X41A1247-V Ramya Sri - X4            ramyavagicharla@gmail.com          16
6
Anil Kumar Teegala [APSSDC]            aps.sdc.ml@gmail.com               18
3
Bhuvaneswari Bonthu 1208               bhuvanabonthu123@gmail.com         15
8
Vedasri                                vedasri.avirneni@gmail.com         14
8
Name: Time in Session (minutes), dtype: int64
```