ASSIGNMENT 1

Name: Aakanksha Bhondve

Rollno: 322004

Grno: 21810939

Comp B1

Aim –

Perform the following operations using Python on suitable data sets. Read a data from different formats(like csv,xls),indexing and selecting data, sort data, describe attributes of data, checking data types of each column, counting unique values of data, format of each column, converting variable data type (e.g. from long to short, vise versa), identifying missing values and fill in the missing values.

Theory –

Numpy: Numpy is Python library that provides mathematical function to handle large dimension array. It provides various method/function for Array, Metrics, and linear algebra.

Pandas: Pandas is one of the most popular Python library for data manipulation and analysis. Pandas provide useful functions to manipulate large amount of structured data. Pandas provide easiest method to perform analysis. It provide large data structures and manipulating numerical tables and time series data. Pandas is a perfect tool for data wrangling. Pandas is designed for quick and easy data manipulation, aggregation, and visualization. There two data structures in Pandas –

Series – It Handle and store data in one-dimensional data.

DataFrame – It Handle and store Two dimensional data.

Matplotlib: Matplolib is another useful Python library for Data Visualization. Descriptive analysis and visualizing data is very important for any organization. Matplotlib provides various method to Visualize data in more effective way. Matplotlib allows to quickly make line graphs,

pie charts, histograms, and other professional grade figures. Using Matplotlib, one can customize every aspect of a figure. Matplotlib has interactive features like zooming and planning and saving the Graph in graphics format.

Code and Output –

In [1]:
```
import pandas as pd
import numpy as np
```

Reading file

In [2]:
```
file = pd.read_csv('telecom.csv')
file.head()
```

Out[2]:

	State	Account length	Area code	International plan	Voice mail plan	Number vmail messages	Total day minutes	Total day calls	Total day charge	Total eve minutes	Total eve calls	Total eve charge	Total night minutes	Total night calls	Total night charge	Total intl minutes	Total intl calls	Total intl charge	Cus s
0	KS	128	415	No	Yes	25	265.1	110	45.07	197.4	99	16.78	244.7	91	11.01	10.0	3	2.70	
1	OH	107	415	No	Yes	26	161.6	123	27.47	195.5	103	16.62	254.4	103	11.45	13.7	3	3.70	
2	NJ	137	415	No	No	0	243.4	114	41.38	121.2	110	10.30	162.6	104	7.32	12.2	5	3.29	
3	OH	84	408	Yes	No	0	299.4	71	50.90	61.9	88	5.26	196.9	89	8.86	6.6	7	1.78	
4	OK	75	415	Yes	No	0	166.7	113	28.34	148.3	122	12.61	186.9	121	8.41	10.1	3	2.73	

Indexing and Selecting Data

In [5]:
```
file[['State', 'Area code']]
```

Out[5]:

	State	Area code
0	KS	415
1	OH	415
2	NJ	415
3	OH	408
4	OK	415
...	...	...
3328	AZ	415
3329	WV	415
3330	RI	510
3331	CT	510
3332	TN	415

3333 rows × 2 columns

In [3]:
```
file.tail(5)
```

Out[3]:

	State	Account length	Area code	International plan	Voice mail plan	Number vmail messages	Total day minutes	Total day calls	Total day charge	Total eve minutes	Total eve calls	Total eve charge	Total night minutes	Total night calls	Total night charge	Total intl minutes	Total intl calls	Total intl charge
3328	AZ	192	415	No	Yes	36	156.2	77	26.55	215.5	126	18.32	279.1	83	12.56	9.9	6	2.67
3329	WV	68	415	No	No	0	231.1	57	39.29	153.4	55	13.04	191.3	123	8.61	9.6	4	2.59
3330	RI	28	510	No	No	0	180.8	109	30.74	288.8	58	24.55	191.9	91	8.64	14.1	6	3.81
3331	CT	184	510	Yes	No	0	213.8	105	36.35	159.6	84	13.57	139.2	137	6.26	5.0	10	1.35
3332	TN	74	415	No	Yes	25	234.4	113	39.85	265.9	82	22.60	241.4	77	10.86	13.7	4	3.70

In [4]:
```
file.iloc[1:5,0:2]
```

Out[4]:

	State	Account length
1	OH	107
2	NJ	137
3	OH	84
4	OK	75

Sorting Values

In [6]:
```
file['Total day minutes'].sort_values(ascending = False)
```

Out[6]:
```
365     350.8
985     346.8
2594    345.3
156     337.4
605     335.5
        ...
1986      7.9
2753      7.8
2736      2.6
1397      0.0
1345      0.0
Name: Total day minutes, Length: 3333, dtype: float64
```

In [8]:
```
file.sort_values(['Total day calls','Total day charge']).head()
```

Out[8]:

	State	Account length	Area code	International plan	Voice mail plan	Number vmail messages	Total day minutes	Total day calls	Total day charge	Total eve minutes	Total eve calls	Total eve charge	Total night minutes	Total night calls	Total night charge	Total intl minutes	Total intl calls	Total intl charge
1345	SD	98	415	No	No	0	0.0	0	0.00	159.6	130	13.57	167.1	88	7.52	6.8	1	1.84
1397	VT	101	510	No	No	0	0.0	0	0.00	192.1	119	16.33	168.8	95	7.60	7.2	4	1.94
1144	NH	155	408	No	No	0	216.7	30	36.84	144.3	125	12.27	135.3	106	6.09	10.8	1	2.92
1989	MT	124	415	No	Yes	30	144.5	35	24.57	262.3	101	22.30	226.5	82	10.19	12.0	7	3.24
692	NE	82	408	No	No	0	185.8	36	31.59	276.5	134	23.50	192.1	104	8.64	5.7	7	1.54

In [9]:
```
file.sort_values(['Total day calls','Total eve calls'], ascending=[True,False]).head()
```

Out[9]:

	State	Account length	Area code	International plan	Voice mail plan	Number vmail messages	Total day minutes	Total day calls	Total day charge	Total eve minutes	Total eve calls	Total eve charge	Total night minutes	Total night calls	Total night charge	Total intl minutes	Total intl calls	Total intl charge
1345	SD	98	415	No	No	0	0.0	0	0.00	159.6	130	13.57	167.1	88	7.52	6.8	1	1.84
1397	VT	101	510	No	No	0	0.0	0	0.00	192.1	119	16.33	168.8	95	7.60	7.2	4	1.94
1144	NH	155	408	No	No	0	216.7	30	36.84	144.3	125	12.27	135.3	106	6.09	10.8	1	2.92
1989	MT	124	415	No	Yes	30	144.5	35	24.57	262.3	101	22.30	226.5	82	10.19	12.0	7	3.24
692	NE	82	408	No	No	0	185.8	36	31.59	276.5	134	23.50	192.1	104	8.64	5.7	7	1.54

Describing the attributes of data

In [10]:
```
file.describe()
```

Out[10]:

	Account length	Area code	Number vmail messages	Total day minutes	Total day calls	Total day charge	Total eve minutes	Total eve calls	Total eve charge	Total night minutes	Total night calls	Tota
count	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000	3333.(
mean	101.064806	437.182418	8.099010	179.775098	100.435644	30.562307	200.980348	100.114311	17.083540	200.872037	100.107711	9.(
std	39.822106	42.371290	13.688365	54.467389	20.069084	9.259435	50.713844	19.922625	4.310668	50.573847	19.568609	2.2
min	1.000000	408.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	23.200000	33.000000	1.(
25%	74.000000	408.000000	0.000000	143.700000	87.000000	24.430000	166.600000	87.000000	14.160000	167.000000	87.000000	7.5
50%	101.000000	415.000000	0.000000	179.400000	101.000000	30.500000	201.400000	100.000000	17.120000	201.200000	100.000000	9.(
75%	127.000000	510.000000	20.000000	216.400000	114.000000	36.790000	235.300000	114.000000	20.000000	235.300000	113.000000	10.5
max	243.000000	510.000000	51.000000	350.800000	165.000000	59.640000	363.700000	170.000000	30.910000	395.000000	175.000000	17.1

In [13]:
```
file.shape
```

Out[13]:
```
(3333, 20)
```

Checking datatypes of each column

In [14]:
```
file.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3333 entries, 0 to 3332
Data columns (total 20 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   State                   3333 non-null   object
 1   Account length          3333 non-null   int64
 2   Area code               3333 non-null   int64
 3   International plan      3333 non-null   object
 4   Voice mail plan         3333 non-null   object
 5   Number vmail messages   3333 non-null   int64
 6   Total day minutes       3333 non-null   float64
 7   Total day calls         3333 non-null   int64
 8   Total day charge        3333 non-null   float64
 9   Total eve minutes       3333 non-null   float64
 10  Total eve calls         3333 non-null   int64
 11  Total eve charge        3333 non-null   float64
 12  Total night minutes     3333 non-null   float64
 13  Total night calls       3333 non-null   int64
 14  Total night charge      3333 non-null   float64
 15  Total intl minutes      3333 non-null   float64
 16  Total intl calls        3333 non-null   int64
 17  Total intl charge       3333 non-null   float64
 18  Customer service calls  3333 non-null   int64
 19  Churn                   3333 non-null   bool
dtypes: bool(1), float64(8), int64(8), object(3)
memory usage: 498.1+ KB
```

In [16]:
```
file.columns
```

Out[16]:
```
Index(['State', 'Account length', 'Area code', 'International plan',
       'Voice mail plan', 'Number vmail messages', 'Total day minutes',
       'Total day calls', 'Total day charge', 'Total eve minutes',
       'Total eve calls', 'Total eve charge', 'Total night minutes',
       'Total night calls', 'Total night charge', 'Total intl minutes',
       'Total intl calls', 'Total intl charge', 'Customer service calls',
       'Churn'],
      dtype='object')
```

In [17]:
```
file.dtypes
```

Out[17]:
```
State                    object
Account length            int64
Area code                 int64
International plan       object
Voice mail plan          object
Number vmail messages     int64
Total day minutes       float64
Total day calls           int64
Total day charge        float64
Total eve minutes       float64
Total eve calls           int64
Total eve charge        float64
Total night minutes     float64
Total night calls         int64
Total night charge      float64
Total intl minutes      float64
Total intl calls          int64
Total intl charge       float64
Customer service calls    int64
Churn                      bool
dtype: object
```

Checking and counting the unique values

In [18]:
```
file['State'].unique()
```

Out[18]:
```
array(['KS', 'OH', 'NJ', 'OK', 'AL', 'MA', 'MO', 'LA', 'WV', 'IN', 'RI',
       'IA', 'MT', 'NY', 'ID', 'VT', 'VA', 'TX', 'FL', 'CO', 'AZ', 'SC',
       'NE', 'WY', 'HI', 'IL', 'NH', 'GA', 'AK', 'MD', 'AR', 'WI', 'OR',
       'MI', 'DE', 'UT', 'CA', 'MN', 'SD', 'NC', 'WA', 'NM', 'NV', 'DC',
       'KY', 'ME', 'MS', 'TN', 'PA', 'CT', 'ND'], dtype=object)
```

In [19]:
```
file['State'].nunique()
```

Out[19]:
```
51
```

In [20]:
```
file.nunique()
```

Out[20]:
```
State                     51
Account length           212
Area code                  3
International plan         2
Voice mail plan            2
Number vmail messages     46
Total day minutes       1667
Total day calls          119
Total day charge        1667
Total eve minutes       1611
Total eve calls          123
Total eve charge        1440
Total night minutes     1591
Total night calls        120
Total night charge       933
Total intl minutes       162
Total intl calls          21
Total intl charge        162
Customer service calls    10
Churn                      2
dtype: int64
```

Converting variable datatypes

In [20]:
```
1 file.convert_dtypes().dtypes
```

Out[20]:
```
State                    string
Account length            Int64
Area code                 Int64
International plan       string
Voice mail plan          string
Number vmail messages     Int64
Total day minutes       float64
Total day calls           Int64
Total day charge        float64
Total eve minutes       float64
Total eve calls           Int64
Total eve charge        float64
Total night minutes     float64
Total night calls         Int64
Total night charge      float64
Total intl minutes      float64
Total intl calls          Int64
Total intl charge       float64
Customer service calls    Int64
Churn                   boolean
dtype: object
```

In [21]:
```
1 file['Churn'] = file['Churn'].astype(int)
```

Checking null values and filling them

In [27]:
```
1 orders = pd.read_excel('NewOrders.xlsx')
2 orders.head()
```

Out[27]:

	Row ID	Order ID	Order Date	Order Priority	Order Quantity	Sales	Discount	Ship Mode	Profit	Unit Price	...	Customer Name	Province	Region	Customer Segment	Product Category	Produc Cat
0	1	3	2010-10-13	Low	6	261.5400	0.04	Regular Air	-213.250	38.94	...	Muhammed MacIntyre	Nunavut	Nunavut	Small Business	Office Supplies	Stor Organi
1	2	6	2012-02-20	Not Specified	2	6.9300	0.01	Regular Air	-4.640	2.08	...	Ruben Dartt	Alberta	West	Corporate	Office Supplies	Sci Rule Trir
2	3	32	2011-07-15	High	26	2808.0800	0.07	Regular Air	1054.820	107.53	...	Liz Pelletier	Alberta	West	Corporate	Furniture	Furni
3	4	32	2011-07-15	High	24	1761.4000	0.09	Delivery Truck	-1748.560	70.89	...	Liz Pelletier	Alberta	West	Corporate	Furniture	
4	5	32	2011-07-15	High	23	160.2335	0.04	Regular Air	-85.129	7.99	...	Liz Pelletier	Alberta	West	Corporate	Technology	Telep Communi

5 rows × 21 columns

In [35]:
```
orders['Product Base Margin'].fillna( method ='ffill', inplace = True)
orders.head()
```

Out[35]:

	Row ID	Order ID	Order Date	Order Priority	Order Quantity	Sales	Discount	Ship Mode	Profit	Unit Price	...	Customer Name	Province	Region	Customer Segment	Product Category	Produc Cat
0	1	3	2010-10-13	Low	6	261.5400	0.04	Regular Air	-213.250	38.94	...	Muhammed MacIntyre	Nunavut	Nunavut	Small Business	Office Supplies	Stor Organi
1	2	6	2012-02-20	Not Specified	2	6.9300	0.01	Regular Air	-4.640	2.08	...	Ruben Dartt	Alberta	West	Corporate	Office Supplies	Sci Rule Trir
2	3	32	2011-07-15	High	26	2808.0800	0.07	Regular Air	1054.820	107.53	...	Liz Pelletier	Alberta	West	Corporate	Furniture	Furni
3	4	32	2011-07-15	High	24	1761.4000	0.09	Delivery Truck	-1748.560	70.89	...	Liz Pelletier	Alberta	West	Corporate	Furniture	
4	5	32	2011-07-15	High	23	160.2335	0.04	Regular Air	-85.129	7.99	...	Liz Pelletier	Alberta	West	Corporate	Technology	Telep Communi

5 rows × 21 columns

In [37]:
```
orders['Product Base Margin'].isnull()
```

Out[37]:
```
0       False
1       False
2       False
3       False
4       False
        ...
8394    False
8395    False
8396    False
8397    False
8398    False
Name: Product Base Margin, Length: 8399, dtype: bool
```