GRIP:The Sparks Foundation

Data Science and Business Analytics Intern

Author: Sudeshna Saha

In [1]:

```
#import all the necessary libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

In [2]:

```
#load the data
df=pd.read_csv("SampleSuperstore.csv")
```

In [3]:

```
df.head()
```

Out[3]:

	Ship Mode	Segment	Country	City	State	Postal Code	Region	Category	Sub-Category	
0	Second Class	Consumer	United States	Henderson	Kentucky	42420	South	Furniture	Bookcases	26
1	Second Class	Consumer	United States	Henderson	Kentucky	42420	South	Furniture	Chairs	73
2	Second Class	Corporate	United States	Los Angeles	California	90036	West	Office Supplies	Labels	1
3	Standard Class	Consumer	United States	Fort Lauderdale	Florida	33311	South	Furniture	Tables	95
4	Standard Class	Consumer	United States	Fort Lauderdale	Florida	33311	South	Office Supplies	Storage	2

In [4]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9994 entries, 0 to 9993
Data columns (total 13 columns):
 #   Column        Non-Null Count  Dtype  
---  ------        --------------  -----  
 0   Ship Mode     9994 non-null   object 
 1   Segment       9994 non-null   object 
 2   Country       9994 non-null   object 
 3   City          9994 non-null   object 
 4   State         9994 non-null   object 
 5   Postal Code   9994 non-null   int64  
 6   Region        9994 non-null   object 
 7   Category      9994 non-null   object 
 8   Sub-Category  9994 non-null   object 
 9   Sales         9994 non-null   float64
 10  Quantity      9994 non-null   int64  
 11  Discount      9994 non-null   float64
 12  Profit        9994 non-null   float64
dtypes: float64(3), int64(2), object(8)
memory usage: 1015.1+ KB
```

In [5]:

```
df.isnull().sum()
```

Out[5]:

```
Ship Mode       0
Segment         0
Country         0
City            0
State           0
Postal Code     0
Region          0
Category        0
Sub-Category    0
Sales           0
Quantity        0
Discount        0
Profit          0
dtype: int64
```

In [6]:

```
df.shape
```

Out[6]:

```
(9994, 13)
```

In [7]:

```
df.describe()
```

Out[7]:

	Postal Code	Sales	Quantity	Discount	Profit
count	9994.000000	9994.000000	9994.000000	9994.000000	9994.000000
mean	55190.379428	229.858001	3.789574	0.156203	28.656896
std	32063.693350	623.245101	2.225110	0.206452	234.260108
min	1040.000000	0.444000	1.000000	0.000000	-6599.978000
25%	23223.000000	17.280000	2.000000	0.000000	1.728750
50%	56430.500000	54.490000	3.000000	0.200000	8.666500
75%	90008.000000	209.940000	5.000000	0.200000	29.364000
max	99301.000000	22638.480000	14.000000	0.800000	8399.976000

In [8]:

```
df.columns
```

Out[8]:

```
Index(['Ship Mode', 'Segment', 'Country', 'City', 'State', 'Postal Code',
       'Region', 'Category', 'Sub-Category', 'Sales', 'Quantity', 'Discoun
t',
       'Profit'],
      dtype='object')
```

In [19]:

```
df["Region"].value_counts()
```

Out[19]:

```
West       3203
East       2848
Central    2323
South      1620
Name: Region, dtype: int64
```

In [12]:

```
df["Country"].value_counts()
```

Out[12]:

```
United States    9994
Name: Country, dtype: int64
```

In [13]:

```
df["City"].value_counts()
```

Out[13]:

```
New York City     915
Los Angeles       747
Philadelphia      537
San Francisco     510
Seattle           428
                 ...
Waterloo            1
Manhattan           1
Chapel Hill         1
Orland Park         1
Jefferson City      1
Name: City, Length: 531, dtype: int64
```

In [18]:

```
df["State"].value_counts()
```

Out[18]:

```
California              2001
New York                1128
Texas                    985
Pennsylvania             587
Washington               506
Illinois                 492
Ohio                     469
Florida                  383
Michigan                 255
North Carolina           249
Virginia                 224
Arizona                  224
Georgia                  184
Tennessee                183
Colorado                 182
Indiana                  149
Kentucky                 139
Massachusetts            135
New Jersey               130
Oregon                   124
Wisconsin                110
Maryland                 105
Delaware                  96
Minnesota                 89
Connecticut               82
Oklahoma                  66
Missouri                  66
Alabama                   61
Arkansas                  60
Rhode Island              56
Mississippi               53
Utah                      53
Louisiana                 42
South Carolina            42
Nevada                    39
Nebraska                  38
New Mexico                37
Iowa                      30
New Hampshire             27
Kansas                    24
Idaho                     21
Montana                   15
South Dakota              12
Vermont                   11
District of Columbia      10
Maine                      8
North Dakota               7
West Virginia              4
Wyoming                    1
Name: State, dtype: int64
```

In [20]:

```
df["Category"].value_counts()
```

Out[20]:

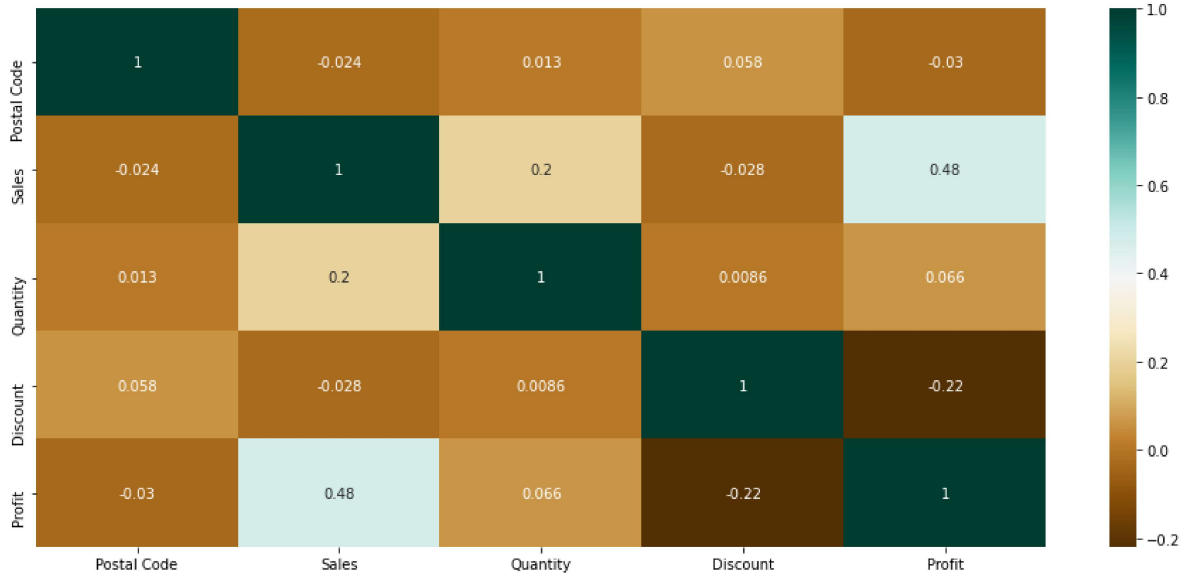
```
Office Supplies    6026
Furniture          2121
Technology         1847
Name: Category, dtype: int64
```

In [15]:

```
#check the correlation of the data
plt.figure(figsize=(16,7))
sns.heatmap(df.corr(),cmap="BrBG",annot=True)
```

Out[15]:

```
<AxesSubplot:>
```



In [21]:

```
#check duplicates
df.drop_duplicates(keep='first',inplace=True)
df.shape
```

Out[21]:

```
(9977, 13)
```

In [ ]:

```

```