Assignment 16

Kartik Thakur

In [61]:
```
# 1. Merge the two files in one dataframe
# 2. Clean the data
# 3. Change the float column into integer type
# 4. Get the month value from the order date
# 5. Which was the most productive month in terms of sales?
# 6.  Which city had the highest number of sales?
# 7. At what time people did purchase the product?
# 8. What is the average purchase by city?
```

In [62]:
```
import pandas as pd
```

1. Merge the two files in one dataframe

In [27]:
```
# File 1 -contains purchase data of december
sale_dec=pd.read_csv('Sales_December_2019.csv')
sale_dec.head()
```

Out[27]:

	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address
0	295665	Macbook Pro Laptop	1	1700	12/30/19 00:01	136 Church St, New York City, NY 10001
1	295666	LG Washing Machine	1	600	12/29/19 07:03	562 2nd St, New York City, NY 10001
2	295667	USB-C Charging Cable	1	11.95	12/12/19 18:21	277 Main St, New York City, NY 10001
3	295668	27in FHD Monitor	1	149.99	12/22/19 15:13	410 6th St, San Francisco, CA 94016
4	295669	USB-C Charging Cable	1	11.95	12/18/19 12:38	43 Hill St, Atlanta, GA 30301

In [28]:
```
sale_dec.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 25117 entries, 0 to 25116
Data columns (total 6 columns):
 #   Column            Non-Null Count  Dtype 
---  ------            --------------  ----- 
 0   Order ID          25037 non-null  object
 1   Product           25037 non-null  object
 2   Quantity Ordered  25037 non-null  object
 3   Price Each        25037 non-null  object
 4   Order Date        25037 non-null  object
 5   Purchase Address  25037 non-null  object
dtypes: object(6)
memory usage: 1.1+ MB
```

In [29]:
```
#  File 2 is imported
sale_june=pd.read_csv('Sales_June_2019.csv')
sale_june.head()
```

Out[29]:

	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address
0	209921	USB-C Charging Cable	1	11.95	06/23/19 19:34	950 Walnut St, Portland, ME 04101
1	209922	Macbook Pro Laptop	1	1700	06/30/19 10:05	80 4th St, San Francisco, CA 94016
2	209923	ThinkPad Laptop	1	999.99	06/24/19 20:18	402 Jackson St, Los Angeles, CA 90001
3	209924	27in FHD Monitor	1	149.99	06/05/19 10:21	560 10th St, Seattle, WA 98101
4	209925	Bose SoundSport Headphones	1	99.99	06/25/19 18:58	545 2nd St, San Francisco, CA 94016

In [30]:
```
# Merged File
sale=pd.concat([sale_june,sale_dec])
sale.head()
```

Out[30]:

	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address
0	209921	USB-C Charging Cable	1	11.95	06/23/19 19:34	950 Walnut St, Portland, ME 04101
1	209922	Macbook Pro Laptop	1	1700	06/30/19 10:05	80 4th St, San Francisco, CA 94016
2	209923	ThinkPad Laptop	1	999.99	06/24/19 20:18	402 Jackson St, Los Angeles, CA 90001
3	209924	27in FHD Monitor	1	149.99	06/05/19 10:21	560 10th St, Seattle, WA 98101
4	209925	Bose SoundSport Headphones	1	99.99	06/25/19 18:58	545 2nd St, San Francisco, CA 94016

2. Clean the data

In [31]:
```
sale.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 38739 entries, 0 to 25116
Data columns (total 6 columns):
 #   Column            Non-Null Count  Dtype 
---  ------            --------------  ----- 
 0   Order ID          38616 non-null  object
 1   Product           38616 non-null  object
 2   Quantity Ordered  38616 non-null  object
 3   Price Each        38616 non-null  object
 4   Order Date        38616 non-null  object
 5   Purchase Address  38616 non-null  object
dtypes: object(6)
memory usage: 2.1+ MB
```

In [32]:
```
# Checking for present null values
sale.isnull().sum()
```

Out[32]:
```
Order ID            123
Product             123
Quantity Ordered    123
Price Each          123
Order Date          123
Purchase Address    123
dtype: int64
```

In [33]:
```
# Making a copy of original file
sale2=sale.copy()
sale2.head()
```

Out[33]:

	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address
0	209921	USB-C Charging Cable	1	11.95	06/23/19 19:34	950 Walnut St, Portland, ME 04101
1	209922	Macbook Pro Laptop	1	1700	06/30/19 10:05	80 4th St, San Francisco, CA 94016
2	209923	ThinkPad Laptop	1	999.99	06/24/19 20:18	402 Jackson St, Los Angeles, CA 90001
3	209924	27in FHD Monitor	1	149.99	06/05/19 10:21	560 10th St, Seattle, WA 98101
4	209925	Bose SoundSport Headphones	1	99.99	06/25/19 18:58	545 2nd St, San Francisco, CA 94016

In [34]:
```
# Droping OR removing the null values
sale2.dropna(inplace=True)
```

In [35]:
```
sale2.isnull().sum()
```

Out[35]:
```
Order ID            0
Product             0
Quantity Ordered    0
Price Each          0
Order Date          0
Purchase Address    0
dtype: int64
```

In [36]:
```
#Hence File is cleaned of null Values
```

In [37]:
```
sale2[sale2['Order ID']=='Order ID']
```

Out[37]:

	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address
158	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address
990	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address
1679	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address
1684	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address
3126	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address
...	...	...	...	...	...	...
23198	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address
23337	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address
23748	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address
24192	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address
24222	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address

71 rows × 6 columns

In [38]:
```
sale2.drop(sale2[sale2['Order ID']=='Order ID'].index,inplace=True)
sale2.head()
```

Out[38]:

	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address
0	209921	USB-C Charging Cable	1	11.95	06/23/19 19:34	950 Walnut St, Portland, ME 04101
1	209922	Macbook Pro Laptop	1	1700	06/30/19 10:05	80 4th St, San Francisco, CA 94016
2	209923	ThinkPad Laptop	1	999.99	06/24/19 20:18	402 Jackson St, Los Angeles, CA 90001
3	209924	27in FHD Monitor	1	149.99	06/05/19 10:21	560 10th St, Seattle, WA 98101
4	209925	Bose SoundSport Headphones	1	99.99	06/25/19 18:58	545 2nd St, San Francisco, CA 94016

In [39]:
```
sale2.reset_index()
```

Out[39]:

	index	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address
0	0	209921	USB-C Charging Cable	1	11.95	06/23/19 19:34	950 Walnut St, Portland, ME 04101
1	1	209922	Macbook Pro Laptop	1	1700	06/30/19 10:05	80 4th St, San Francisco, CA 94016
2	2	209923	ThinkPad Laptop	1	999.99	06/24/19 20:18	402 Jackson St, Los Angeles, CA 90001
3	3	209924	27in FHD Monitor	1	149.99	06/05/19 10:21	560 10th St, Seattle, WA 98101
4	4	209925	Bose SoundSport Headphones	1	99.99	06/25/19 18:58	545 2nd St, San Francisco, CA 94016
...	...	...	...	...	...	...	...
38495	25112	319666	Lightning Charging Cable	1	14.95	12/11/19 20:58	14 Madison St, San Francisco, CA 94016
38496	25113	319667	AA Batteries (4-pack)	2	3.84	12/01/19 12:01	549 Willow St, Los Angeles, CA 90001
38497	25114	319668	Vareebadd Phone	1	400	12/09/19 06:43	273 Wilson St, Seattle, WA 98101
38498	25115	319669	Wired Headphones	1	11.99	12/03/19 10:39	778 River St, Dallas, TX 75001
38499	25116	319670	Bose SoundSport Headphones	1	99.99	12/21/19 21:45	747 Chestnut St, Los Angeles, CA 90001

38500 rows × 7 columns

In [40]:
```
# hence the file is cleaned of duplicate or unneeded values
```

3. Change the object type column into integer type or float type

In [41]:
```
sale2.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 38500 entries, 0 to 25116
Data columns (total 6 columns):
 #   Column            Non-Null Count  Dtype 
---  ------            --------------  ----- 
 0   Order ID          38500 non-null  object
 1   Product           38500 non-null  object
 2   Quantity Ordered  38500 non-null  object
 3   Price Each        38500 non-null  object
 4   Order Date        38500 non-null  object
 5   Purchase Address  38500 non-null  object
dtypes: object(6)
memory usage: 2.1+ MB
```

In [42]:
```
## Change the data type to int
convert=["Order ID","Quantity Ordered"]
for x in convert:
    sale2[x]=sale2[x].astype("int64")
sale2.head()
```

Out[42]:

	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address
0	209921	USB-C Charging Cable	1	11.95	06/23/19 19:34	950 Walnut St, Portland, ME 04101
1	209922	Macbook Pro Laptop	1	1700	06/30/19 10:05	80 4th St, San Francisco, CA 94016
2	209923	ThinkPad Laptop	1	999.99	06/24/19 20:18	402 Jackson St, Los Angeles, CA 90001
3	209924	27in FHD Monitor	1	149.99	06/05/19 10:21	560 10th St, Seattle, WA 98101
4	209925	Bose SoundSport Headphones	1	99.99	06/25/19 18:58	545 2nd St, San Francisco, CA 94016

In [43]:
```
sale2["Order ID"].dtype
```

Out[43]:
```
dtype('int64')
```

In [44]:
```
## Change the data type to float
convert=["Price Each"]
for col in convert:
    sale2[col]=sale2[col].astype("float64")
```

In [45]:
```
sale2['Price Each'].dtype
```

Out[45]:
```
dtype('float64')
```

In [46]:
```
sale2.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 38500 entries, 0 to 25116
Data columns (total 6 columns):
 #   Column            Non-Null Count  Dtype  
---  ------            --------------  -----  
 0   Order ID          38500 non-null  int64  
 1   Product           38500 non-null  object 
 2   Quantity Ordered  38500 non-null  int64  
 3   Price Each        38500 non-null  float64
 4   Order Date        38500 non-null  object 
 5   Purchase Address  38500 non-null  object 
dtypes: float64(1), int64(2), object(3)
memory usage: 2.1+ MB
```

4. Get the month value from the order date

In [47]:
```
l=[]
for x in sale2['Order Date']:
    d=x.split('/')
    l.append(d)
l
```

Out[47]:
```
[['06', '23', '19 19:34'],
 ['06', '30', '19 10:05'],
 ['06', '24', '19 20:18'],
 ['06', '05', '19 10:21'],
 ['06', '25', '19 18:58'],
 ['06', '28', '19 20:04'],
 ['06', '28', '19 00:07'],
 ['06', '16', '19 21:30'],
 ['06', '28', '19 10:56'],
 ['06', '02', '19 11:22'],
 ['06', '24', '19 13:55'],
 ['06', '12', '19 14:36'],
 ['06', '07', '19 15:39'],
 ['06', '13', '19 20:53'],
 ['06', '09', '19 11:13'],
 ['06', '15', '19 12:21'],
 ['06', '29', '19 18:01'],
 ['06', '15', '19 12:29'],
 ['06', '15', '19 12:29'],
```

In [48]:
```
m=[]
for i in l:
    m.append(i[0])
m
```

Out[48]:
```
['06',
 '06',
 '06',
 '06',
 '06',
 '06',
 '06',
 '06',
 '06',
 '06',
 '06',
 '06',
 '06',
 '06',
 '06',
 '06',
 '06',
 '06',
 '06',
 '06',
 '06',
```

In [49]:
```
sale2['Month']=m
sale2
```

Out[49]:

	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address	Month
0	209921	USB-C Charging Cable	1	11.95	06/23/19 19:34	950 Walnut St, Portland, ME 04101	06
1	209922	Macbook Pro Laptop	1	1700.00	06/30/19 10:05	80 4th St, San Francisco, CA 94016	06
2	209923	ThinkPad Laptop	1	999.99	06/24/19 20:18	402 Jackson St, Los Angeles, CA 90001	06
3	209924	27in FHD Monitor	1	149.99	06/05/19 10:21	560 10th St, Seattle, WA 98101	06
4	209925	Bose SoundSport Headphones	1	99.99	06/25/19 18:58	545 2nd St, San Francisco, CA 94016	06
...	...	...	...	...	...	...	...
25112	319666	Lightning Charging Cable	1	14.95	12/11/19 20:58	14 Madison St, San Francisco, CA 94016	12
25113	319667	AA Batteries (4-pack)	2	3.84	12/01/19 12:01	549 Willow St, Los Angeles, CA 90001	12
25114	319668	Vareebadd Phone	1	400.00	12/09/19 06:43	273 Wilson St, Seattle, WA 98101	12
25115	319669	Wired Headphones	1	11.99	12/03/19 10:39	778 River St, Dallas, TX 75001	12
25116	319670	Bose SoundSport Headphones	1	99.99	12/21/19 21:45	747 Chestnut St, Los Angeles, CA 90001	12

38500 rows × 7 columns

5. Which was the most productive month in terms of sales?

In [50]:
```
sale2['Sales']=sale2['Quantity Ordered']*sale2['Price Each']
sale2
```

Out[50]:

	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address	Month	Sales
0	209921	USB-C Charging Cable	1	11.95	06/23/19 19:34	950 Walnut St, Portland, ME 04101	06	11.95
1	209922	Macbook Pro Laptop	1	1700.00	06/30/19 10:05	80 4th St, San Francisco, CA 94016	06	1700.00
2	209923	ThinkPad Laptop	1	999.99	06/24/19 20:18	402 Jackson St, Los Angeles, CA 90001	06	999.99
3	209924	27in FHD Monitor	1	149.99	06/05/19 10:21	560 10th St, Seattle, WA 98101	06	149.99
4	209925	Bose SoundSport Headphones	1	99.99	06/25/19 18:58	545 2nd St, San Francisco, CA 94016	06	99.99
...	...	...	...	...	...	...	...	...
25112	319666	Lightning Charging Cable	1	14.95	12/11/19 20:58	14 Madison St, San Francisco, CA 94016	12	14.95
25113	319667	AA Batteries (4-pack)	2	3.84	12/01/19 12:01	549 Willow St, Los Angeles, CA 90001	12	7.68
25114	319668	Vareebadd Phone	1	400.00	12/09/19 06:43	273 Wilson St, Seattle, WA 98101	12	400.00
25115	319669	Wired Headphones	1	11.99	12/03/19 10:39	778 River St, Dallas, TX 75001	12	11.99
25116	319670	Bose SoundSport Headphones	1	99.99	12/21/19 21:45	747 Chestnut St, Los Angeles, CA 90001	12	99.99

38500 rows × 8 columns

In [51]:
```
sale2.groupby('Month')["Quantity Ordered"].sum().sort_values(ascending=False).
```

Out[51]:
```
Month
12    28056
Name: Quantity Ordered, dtype: int64
```

In [52]:
```
# Hence, we can say that in December , We Had the Higher numbers of Sales.
```

6. Which city had the highest number of sales?

In [53]:
```
y=[]
for x in sale2['Purchase Address']:
    m=x.split(',')
    y.append(m)
y
```

Out[53]:
```
[['950 Walnut St', ' Portland', ' ME 04101'],
 ['80 4th St', ' San Francisco', ' CA 94016'],
 ['402 Jackson St', ' Los Angeles', ' CA 90001'],
 ['560 10th St', ' Seattle', ' WA 98101'],
 ['545 2nd St', ' San Francisco', ' CA 94016'],
 ['386 Lake St', ' Seattle', ' WA 98101'],
 ['29 Lincoln St', ' Los Angeles', ' CA 90001'],
 ['350 9th St', ' New York City', ' NY 10001'],
 ['612 Meadow St', ' Portland', ' OR 97035'],
 ['625 Ridge St', ' Los Angeles', ' CA 90001'],
 ['761 14th St', ' New York City', ' NY 10001'],
 ['858 8th St', ' Boston', ' MA 02215'],
 ['932 Lakeview St', ' San Francisco', ' CA 94016'],
 ['387 Dogwood St', ' Boston', ' MA 02215'],
 ['530 Cedar St', ' Boston', ' MA 02215'],
 ['769 Dogwood St', ' Dallas', ' TX 75001'],
 ['675 Maple St', ' Dallas', ' TX 75001'],
 ['766 Meadow St', ' Seattle', ' WA 98101'],
 ['766 Meadow St', ' Seattle', ' WA 98101'],
```

In [54]:
```
z=[]
for i in y:
    z.append(i[-2])
z
```

Out[54]:
```
[' Portland',
 ' San Francisco',
 ' Los Angeles',
 ' Seattle',
 ' San Francisco',
 ' Seattle',
 ' Los Angeles',
 ' New York City',
 ' Portland',
 ' Los Angeles',
 ' New York City',
 ' Boston',
 ' San Francisco',
 ' Boston',
 ' Boston',
 ' Dallas',
 ' Dallas',
 ' Seattle',
 ' Seattle',
```

In [55]:
```
sale2['City']=z
```

In [56]:
```
sale2.groupby('City')["Quantity Ordered"].sum().sort_values(ascending=False).h
```

Out[56]:
```
City
 San Francisco    10462
Name: Quantity Ordered, dtype: int64
```

7. At what time people mostly purchase the product?

In [57]:
```
o=[]
for i in l:
    o.append(i[2][3:8])
o
```

Out[57]:
```
['19:34',
 '10:05',
 '20:18',
 '10:21',
 '18:58',
 '20:04',
 '00:07',
 '21:30',
 '10:56',
 '11:22',
 '13:55',
 '14:36',
 '15:39',
 '20:53',
 '11:13',
 '12:21',
 '18:01',
 '12:29',
 '12:29',
```

In [58]:
```
sale2['Time']=o
sale2
```

Out[58]:

	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address	Month	Sales	City	T
0	209921	USB-C Charging Cable	1	11.95	06/23/19 19:34	950 Walnut St, Portland, ME 04101	06	11.95	Portland	19
1	209922	Macbook Pro Laptop	1	1700.00	06/30/19 10:05	80 4th St, San Francisco, CA 94016	06	1700.00	San Francisco	10
2	209923	ThinkPad Laptop	1	999.99	06/24/19 20:18	402 Jackson St, Los Angeles, CA 90001	06	999.99	Los Angeles	20
3	209924	27in FHD Monitor	1	149.99	06/05/19 10:21	560 10th St, Seattle, WA 98101	06	149.99	Seattle	10
4	209925	Bose SoundSport Headphones	1	99.99	06/25/19 18:58	545 2nd St, San Francisco, CA 94016	06	99.99	San Francisco	18
...	...	...	...	...	...	...	...	...	...	
25112	319666	Lightning Charging Cable	1	14.95	12/11/19 20:58	14 Madison St, San Francisco, CA 94016	12	14.95	San Francisco	20
25113	319667	AA Batteries (4-pack)	2	3.84	12/01/19 12:01	549 Willow St, Los Angeles, CA 90001	12	7.68	Los Angeles	12
25114	319668	Vareebadd Phone	1	400.00	12/09/19 06:43	273 Wilson St, Seattle, WA 98101	12	400.00	Seattle	06
25115	319669	Wired Headphones	1	11.99	12/03/19 10:39	778 River St, Dallas, TX 75001	12	11.99	Dallas	10
25116	319670	Bose SoundSport Headphones	1	99.99	12/21/19 21:45	747 Chestnut St, Los Angeles, CA 90001	12	99.99	Los Angeles	21

38500 rows × 10 columns

In [59]:
```
sale2['Time'].value_counts().head(1)
```

Out[59]:
```
19:46    61
Name: Time, dtype: int64
```

8. What is the average purchase by city?

In [60]:
```
sale2.groupby('City')["Quantity Ordered"].mean().sort_values(ascending=False)
```

Out[60]:
```
City
 Dallas           1.136986
 Boston           1.135651
 Portland         1.135292
 Austin           1.125545
 New York City    1.124517
 San Francisco    1.123376
 Seattle          1.120792
 Los Angeles      1.119067
 Atlanta          1.115933
Name: Quantity Ordered, dtype: float64
```

In [ ]:

In [ ]: