DATA ANALYSIS ASSIGNMENT 2

August 24, 2020

[5]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

[4]:
```
#importing the data
data = pd.read_excel("C:/Users/Admin/Desktop/Data Analyst Assignment 2.xlsx")
data.head()
```

[4]:

	SKU	2019-08-01 00:00:00	2019-08-02 00:00:00	2019-08-03 00:00:00	\
0	2527	1551	1613	1613	
1	3042	2240	2330	2423	
2	3086	891	909	936	
3	3155	1628	1628	1563	
4	3166	369	354	343	

	2019-08-04 00:00:00	2019-08-05 00:00:00	2019-08-06 00:00:00	\
0	1532	1517	1441	
1	2350	2374	2398	
2	889	880	906	
3	1532	1563	1563	
4	340	330	337	

	2019-08-07 00:00:00	2019-08-08 00:00:00	2019-08-09 00:00:00	...	\
0	1383	1424	1381	...	
1	2278	2210	2232	...	
2	870	879	844	...	
3	1532	1593	1513	...	
4	350	343	340	...	

	2020-07-22 00:00:00	2020-07-23 00:00:00	2020-07-24 00:00:00	\
0	1929	1948	1929	
1	7096	7025	6955	
2	829	829	812	
3	1719	1805	1895	
4	335	328	325	

	2020-07-25 00:00:00	2020-07-26 00:00:00	2020-07-27 00:00:00	\

```
0                  1987                  2027                  1926
1                  6746                  6881                  6606
2                   788                   788                   772
3                  1895                  1990                  2030
4                   319                   316                   319

   2020-07-28 00:00:00  2020-07-29 00:00:00  2020-07-30 00:00:00  \
0                 1926                 1849                 1904
1                 6804                 6872                 6872
2                  811                  803                  763
3                 2050                 2132                 2111
4                  329                  336                  349

   2020-07-31 00:00:00
0                 1828
1                 7078
2                  763
3                 2027
4                  352

[5 rows x 367 columns]
```

[ ]:
```
#Melting the data wrt SKU to make it effective for data analysing
Sales= data.melt(id_vars=["SKU"], var_name="Date", value_name="Sales").
 ↪reset_index(drop=True)
```

[16]:
```
#Converting the date column to datatype datettime
Sales.Date = pd.to_datetime(Sales.Date)
Sales.set_index('Date', inplace=True)
```

[34]:
```
monthly_sales_by_SKU = Sales.groupby('SKU').resample('M').sum().reset_index()
monthly_sales_by_SKU
```

[34]:
```
           SKU       Date  Sales
0         2527 2019-08-31  41702
1         2527 2019-09-30  36880
2         2527 2019-10-31  30582
3         2527 2019-11-30  21916
4         2527 2019-12-31  23804
...        ...        ...    ...
2959  WIM51234 2020-03-31     62
2960  WIM51234 2020-04-30     60
2961  WIM51234 2020-05-31     62
2962  WIM51234 2020-06-30     60
2963  WIM51234 2020-07-31     62

[2964 rows x 3 columns]
```

[22]:
```
Quaterly_Sales_by_SKU = Sales.groupby('SKU').resample('Q').sum().reset_index()
Quaterly_Sales_by_SKU
```

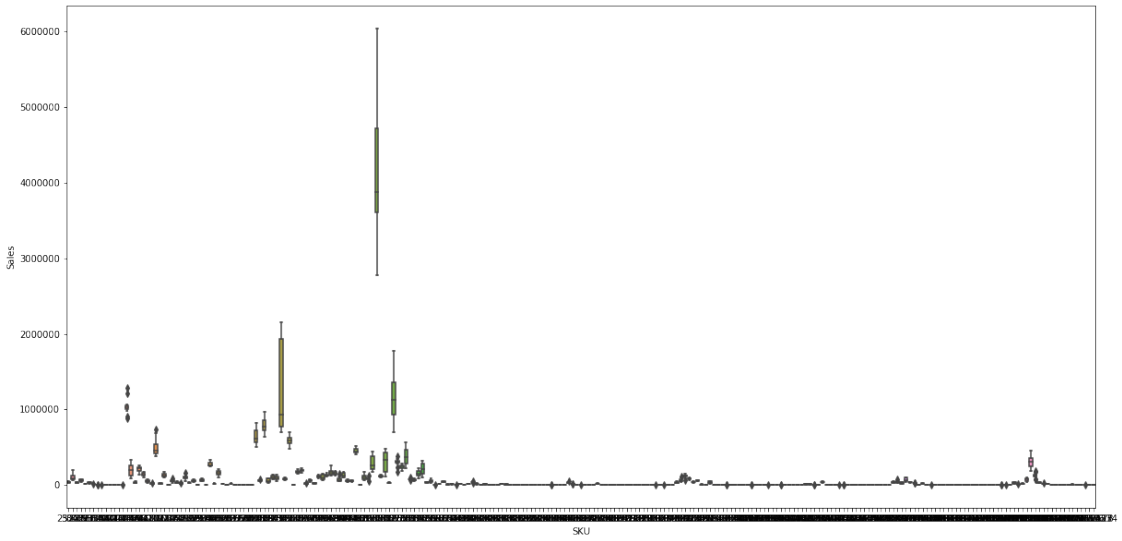
[22]:
```
            SKU       Date   Sales
0          2527 2019-09-30   78582
1          2527 2019-12-31   76302
2          2527 2020-03-31   75979
3          2527 2020-06-30  131890
4          2527 2020-09-30   55826
...         ...        ...     ...
1230  WIM51234 2019-09-30     122
1231  WIM51234 2019-12-31     184
1232  WIM51234 2020-03-31     182
1233  WIM51234 2020-06-30     182
1234  WIM51234 2020-09-30      62

[1235 rows x 3 columns]
```

[18]:
```
top_three_Monthly_sales_data = Sales.resample('M').sum()['Sales'].nlargest(3)
top_three_Monthly_sales_data
```

[18]:
```
Date
2020-03-31    20234570
2019-12-31    19536270
2020-01-31    19170099
Name: Sales, dtype: int64
```

[67]:
```
#Calculating the outliers with seaborn
a4_dims = (20, 10)
ax = plt.subplots(figsize=a4_dims)
sns.boxplot(x='SKU', y='Sales', data=monthly_sales_by_SKU)
```

[67]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x1ea92d88448>
```



[ ]: