data_sales-cleaning

August 28, 2020

[1]:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import os as os
import seaborn as sns
%matplotlib inline
```

merging data sales for 12 month

[2]:

```
files=[file for file in os.listdir('C:/Users/ENTER/Desktop/my work now/all data␣
 ↪before merge')]
all_data=pd.DataFrame()
for file in files:
    df =pd.read_csv('C:/Users/ENTER/Desktop/my work now/all data before merge/
 ↪'+file)
    all_data=pd.concat([all_data,df])
all_data.to_csv("C:/Users/ENTER/Desktop/my work now/all data before merge/
 ↪all_data.csv" , index=False)
```

cleaning data

first-removing nan value

[3]:

```
all_data = all_data.dropna()
```

more cleaning

[4]:

```
all_data = all_data.drop(all_data[all_data['Quantity Ordered']== 'Quantity␣
 ↪Ordered'].index)
all_data = all_data.drop(all_data[all_data['Price Each']== 'Price Each'].index)
all_data = all_data.drop(all_data[all_data['Order ID']== 'Order ID'].index)
```

modifing columns types

[5]:

```
all_data['Quantity Ordered']=pd.to_numeric(all_data['Quantity Ordered'])
all_data['Order ID']=pd.to_numeric(all_data['Order ID'])
all_data['Price Each']=pd.to_numeric(all_data['Price Each'])
```

[6]:

```
all_data.dtypes
```

[6]:

```
Order ID              int64
Product              object
Quantity Ordered      int64
Price Each          float64
Order Date           object
Purchase Address     object
dtype: object
```

extracting new columns

creating month column ,total_price for each order

[7]:

```
all_data['Purchase city']=all_data['Purchase Address'].str.split(',').str[1]
all_data.head()
```

[7]:

```
   Order ID                     Product  Quantity Ordered  Price Each  \
0    176558        USB-C Charging Cable                 2       11.95
2    176559  Bose SoundSport Headphones                 1       99.99
3    176560                Google Phone                 1      600.00
4    176560            Wired Headphones                 1       11.99
5    176561            Wired Headphones                 1       11.99

       Order Date                      Purchase Address Purchase city
0  04/19/19 08:46          917 1st St, Dallas, TX 75001        Dallas
2  04/07/19 22:30     682 Chestnut St, Boston, MA 02215        Boston
3  04/12/19 14:38  669 Spruce St, Los Angeles, CA 90001   Los Angeles
4  04/12/19 14:38  669 Spruce St, Los Angeles, CA 90001   Los Angeles
5  04/30/19 09:27     333 8th St, Los Angeles, CA 90001   Los Angeles
```

[8]:

```
all_data['Month']=all_data['Order Date'].str[0:2]
all_data['Month']=pd.to_numeric(all_data['Month'])
all_data['Total Price']=all_data['Quantity Ordered']* all_data['Price Each']

all_data.head()
```

[8]:

```
   Order ID                     Product  Quantity Ordered  Price Each  \
0    176558        USB-C Charging Cable                 2       11.95
2    176559  Bose SoundSport Headphones                 1       99.99
3    176560                Google Phone                 1      600.00
4    176560            Wired Headphones                 1       11.99
5    176561            Wired Headphones                 1       11.99

       Order Date                      Purchase Address Purchase city  Month  \
0  04/19/19 08:46          917 1st St, Dallas, TX 75001        Dallas      4
2  04/07/19 22:30     682 Chestnut St, Boston, MA 02215        Boston      4
3  04/12/19 14:38  669 Spruce St, Los Angeles, CA 90001   Los Angeles      4
```

```
4  04/12/19 14:38  669 Spruce St, Los Angeles, CA 90001   Los Angeles      4
5  04/30/19 09:27     333 8th St, Los Angeles, CA 90001   Los Angeles      4

   Total Price
0        23.90
2        99.99
3       600.00
4        11.99
5        11.99
```

[9]:

```
all_data.to_csv("C:/Users/ENTER/Desktop/my work now/all data before merge/
 ↪all_data_cleaning.csv" , index=False)
```