Double-click (or enter) to edit

```
# from google.colab import drive
# drive.mount('/content/drive')
```

```
import numpy as np
import pandas as pd
```

```
all_data=pd.read_csv("/content/drive/MyDrive/Colab Notebooks/1686715083343_all_data.csv")
```

```
all_data.head()
```

	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address
0	176559.0	Bose SoundSport Headphones	1.0	99.99	04-07-2019 22:30	682 Chestnut St, Boston, MA 02215
1	176560.0	Google Phone	1.0	600.00	04-12-2019 14:38	669 Spruce St, Los Angeles, CA 90001
2	176560.0	Wired Headphones	1.0	11.99	04-12-2019 14:38	669 Spruce St, Los Angeles, CA 90001

```
all_data.shape
```

```
(69, 6)
```

```
# Find NAN
nan_df = all_data[all_data.isna().any(axis=1)]
display(nan_df.head())

all_data.shape

all_data = all_data.dropna(how='all')
all_data.head()

all_data.shape
```

Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address

```
(67, 6)
```

```
all_data = all_data[all_data['Order Date'].str[0:2]!='Or']
print(all_data)
```

```
    Order ID                     Product  Quantity Ordered  Price Each  \
0   176559.0  Bose SoundSport Headphones               1.0       99.99   
1   176560.0                Google Phone               1.0      600.00   
2   176560.0            Wired Headphones               1.0       11.99   
3   176561.0            Wired Headphones               1.0       11.99   
4   176562.0        USB-C Charging Cable               1.0       11.95   
..       ...                         ...               ...         ...   
64  259329.0    Lightning Charging Cable               1.0       14.95   
65  259330.0      AA Batteries (4-pack)                2.0        3.84   
66  259331.0    Apple Airpods Headphones               1.0      150.00   
67  259332.0    Apple Airpods Headphones               1.0      150.00   
68  259333.0  Bose SoundSport Headphones               1.0       99.99   

          Order Date                      Purchase Address  
0   04-07-2019 22:30       682 Chestnut St, Boston, MA 02215  
1   04-12-2019 14:38    669 Spruce St, Los Angeles, CA 90001  
2   04-12-2019 14:38    669 Spruce St, Los Angeles, CA 90001  
3      05/30/19 9:27       333 8th St, Los Angeles, CA 90001  
4     04/29/19 13:03  381 Wilson St, San Francisco, CA 94016  
..               ...                                     ...  
64  09-05-2019 19:00       480 Lincoln St, Atlanta, GA 30301  
65    09/25/19 22:01    763 Washington St, Seattle, WA 98101  
66     09/29/19 7:00     770 4th St, New York City, NY 10001  
67    09/16/19 19:21          782 Lake St, Atlanta, GA 30301  
68    09/19/19 18:03   347 Ridge St, San Francisco, CA 94016  
```

```
[67 rows x 6 columns]
```

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

```
all_data['Month'] = all_data['Order Date'].str[0:2]
all_data['Month'] = all_data['Month'].astype('int32')
all_data.head()
```

	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address	Month
0	176559.0	Bose SoundSport Headphones	1.0	99.99	04-07-2019 22:30	682 Chestnut St, Boston, MA 02215	4
1	176560.0	Google Phone	1.0	600.00	04-12-2019 14:38	669 Spruce St, Los Angeles, CA 90001	4
					04-12-	669 Spruce St	

```
def get_city(address):
  return address.split(",")[1].split(" ")
```

```
def get_state(address):
    return address.split(",")[2].split(" ")[1]

all_data['City'] = all_data['Purchase Address'].apply(lambda x: f"{get_city(x)}  ({get_state(x)})")
all_data.head()
```

	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address	Month	City
0	176559.0	Bose SoundSport Headphones	1.0	99.99	04-07-2019 22:30	682 Chestnut St, Boston, MA 02215	4	['', 'Boston'] (MA)
1	176560.0	Google Phone	1.0	600.00	04-12-2019 14:38	669 Spruce St, Los Angeles, CA 90001	4	['', 'Los', 'Angeles'] (CA)

```
all_data['Sales'] = all_data['Quantity Ordered'].astype('int') * all_data['Price Each'].astype('float')
```

```
all_data.groupby(['Month']).sum()
```

```
<ipython-input-28-dce0a735c05d>:1: FutureWarning: The default value of numeric_onl
  all_data.groupby(['Month']).sum()
```

	Order ID	Quantity Ordered	Price Each	Sales
Month				
4	7335546.0	123.0	885.80	1210.76
5	353124.0	2.0	111.98	111.98
6	184076.0	1.0	14.95	14.95
8	726962.0	9.0	23.92	50.83
9	2378802.0	17.0	591.44	616.62
10	550924.0	11.0	10.67	39.69
11	740314.0	19.0	13.66	65.31
12	550635.0	17.0	8.97	50.83

```
all_data['sales'] = all data['Quantity ordered'].astype(int) = all_data['Price Each'].astype('float')
```

```
    File "<ipython-input-2-4858d656c123>", line 1
      all_data['sales'] = all data['Quantity ordered'].astype(int) =
  all_data['Price Each'].astype('float')
                                      ^
```

```
all_data.groupby(['Month']).sum()
```

```
Dummycity = all_data.groupby(['city'])
print(Dummycity)
#city_max=all_data.groupby(['city]).sum()
#print(max(city_max))
```

```
df = all_data[all_data['Order ID'].duplicated(keep=False)]
df['Gouped'] = df.groupby('Order ID')['Product'].transform(lambda x:','.join(x))
df2 = df[['Order ID', 'Grouped']].drop_duplicaters()
print(df['Grouped'])
```