

```
# Name : Manoj Ramrao Pandit
#Roll no : 647 F(f3)
```

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

```
all_data=pd.read_csv("/content/sample_data/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())
```

| | Order ID | Product | Quantity Ordered | Price Each | Order Date | Purchase Address |
|----|----------|---------|------------------|------------|------------|------------------|
| 36 | NaN | NaN | NaN | NaN | NaN | NaN |
| 51 | NaN | NaN | NaN | NaN | NaN | NaN |

```
all_data.shape
```

(69, 6)

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

(67, 6)

```
#get rid of text order date column
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 DatePurchase 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]

#Make column correct type
all_data['Quantity Ordered']=pd.to_numeric(all_data['Quantity Ordered'])
all_data['Price Each'] = pd.to_numeric(all_data['Price Each'])
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['month 2']=pd.to_datetime(all_data['Order Date']).dt.month
all_data.head()
```

| | Order ID | Product | Quantity Ordered | Price Each | Order Date | Purchase Address | month 2 |
|---|----------|----------------------------|------------------|------------|------------------|--------------------------------------|---------|
| 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 |
| 2 | 176560.0 | Wired Headphones | 1.0 | 11.99 | 04-12-2019 14:38 | 669 Spruce St, Los Angeles, CA 90001 | 4 |

```
#add city column
def get_city(address):
    return address.split(",")[1].strip(" ")

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 2 | 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) |
| 2 | 176560.0 | Wired Headphones | 1.0 | 11.99 | 04-12-2019 14:38 | 669 Spruce St, Los Angeles, CA 90001 | 4 | Los Angeles (CA) |

```
#what was the best month for sales?how much was_earned that month?
all_data['Sales'] = all_data['Quantity Ordered'].astype('int')*all_data['Price Each'].astype('float')
all_data.groupby(['month 2']).sum()
```

```
<ipython-input-12-4ed187ff8fcf>:3: FutureWarning: The default value of numeric_only in DataFrameGroupBy
all_data.groupby(['month 2']).sum()
```

| | Order ID | Quantity Ordered | Price Each | Sales |
|---------|-----------|------------------|------------|---------|
| month 2 | | | | |
| 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 |

```
#2)WHICH CITY SOLD MOST PRODUCT?
Dummyscity=all_data.groupby(['City'])
print(Dummyscity)
#city_Max=all_data.groupby(['city']).sum()
#print(max(city_max))
```

```
<pandas.core.groupby.generic.DataFrameGroupBy object at 0x7fa92445b490>
```

```
#What products are most often sold together
df=all_data[all_data['Order ID'].duplicated(keep=False)]
df['Grouped']=df.groupby('Order ID')['Product'].transform(lambda x:','.join(x))
df2=df[['Order ID','Grouped']].drop_duplicates()
print(df['Grouped'])
```

```
1 Google Phone,Wired Headphones
2 Google Phone,Wired Headphones
Name: Grouped, dtype: object
<ipython-input-15-54c3911aa784>:3: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead
```

```
See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user\_guide/indexing.html#returning-a-view-versus
df['Grouped']=df.groupby('Order ID')['Product'].transform(lambda x:','.join(x))
```

```
from itertools import combinations
from collections import Counter

count=Counter()

for row in df2['Grouped']:
    row_list=row.split(',')
    count.update(Counter(combinations(row_list,2)))

for key,value in count.most_common(10):
    print(key,value)
```

```
('Google Phone', 'Wired Headphones') 1
```

```
product_group=all_data.groupby('Product')
quantity_ordered=product_group.sum()['Quantity Ordered']
```

```
<ipython-input-18-11142b314e0e>:2: FutureWarning: The default value of numeric_only in DataFrameGroupBy.sum is deprecated. In a fut
quantity_ordered=product_group.sum()['Quantity Ordered']
```

```
print(quantity_ordered)
```

```
Product
AA Batteries (4-pack)      64.0
AAA Batteries (4-pack)    109.0
Apple AirPods Headphones    3.0
Bose SoundSport Headphones  3.0
Google Phone                1.0
Lightning Charging Cable    4.0
USB-C Charging Cable        8.0
Wired Headphones            7.0
Name: Quantity Ordered, dtype: float64
```

```
prices=all_data.groupby('Product').mean()['Price Each']
```

```
<ipython-input-20-1f4f73bca841>:1: FutureWarning: The default value of numeric_only in DataFrameGroupBy.mean is deprecated. In a fu  
prices=all_data.groupby('Product').mean()['Price Each']
```

```
print(prices)
```

```
Product  
AA Batteries (4-pack)      3.84  
AAA Batteries (4-pack)     2.99  
Apple AirPods Headphones  150.00  
Bose SoundSport Headphones 99.99  
Google Phone              600.00  
Lightning Charging Cable   14.95  
USB-C Charging Cable       11.95  
Wired Headphones           11.99  
Name: Price Each, dtype: float64
```