Ventas

July 2, 2020

[1]:
```
import pandas as pd
import os
import matplotlib.pyplot as plt
```

1 Merge all data

[2]:
```
df = pd.read_csv('datos/Sales_April_2019.csv')
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 18383 entries, 0 to 18382
Data columns (total 6 columns):
 #   Column            Non-Null Count  Dtype 
---  ------            --------------  ----- 
 0   Order ID          18324 non-null  object
 1   Product           18324 non-null  object
 2   Quantity Ordered  18324 non-null  object
 3   Price Each        18324 non-null  object
 4   Order Date        18324 non-null  object
 5   Purchase Address  18324 non-null  object
dtypes: object(6)
memory usage: 861.8+ KB
```

[3]:
```
files = [file for file in os.listdir('datos')]
for file in files:
    print(file)
```

```
Sales_February_2019.csv
Sales_March_2019.csv
Sales_January_2019.csv
Sales_September_2019.csv
Sales_April_2019.csv
Sales_December_2019.csv
Sales_June_2019.csv
Sales_August_2019.csv
Sales_November_2019.csv
Sales_October_2019.csv
```

```
Sales_May_2019.csv
Sales_July_2019.csv
```

[4]:
```
df2 = pd.DataFrame()
for file in files:
    df1 = pd.read_csv('datos/'+file)
    df2 = pd.concat([df2,df1])
df2.head()
```

[4]:

	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address
0	150502	iPhone	1	700	02/18/19 01:35	866 Spruce St, Portland, ME 04101
1	150503	AA Batteries (4-pack)	1	3.84	02/13/19 07:24	18 13th St, San Francisco, CA 94016
2	150504	27in 4K Gaming Monitor	1	389.99	02/18/19 09:46	52 6th St, New York City, NY 10001
3	150505	Lightning Charging Cable	1	14.95	02/02/19 16:47	129 Cherry St, Atlanta, GA 30301
4	150506	AA Batteries (4-pack)	2	3.84	02/28/19 20:32	548 Lincoln St, Seattle, WA 98101

[5]:
```
df2.to_csv('All.csv',index=False)
```

2 Cual es el mejor mes en ventas y cuanto se gano?

3 Limpiar los NaN, drop rows of NaN, drop Or

[6]:
```
nan_df=df2[df2.isna().any(axis=1)]
nan_df.head()
```

[6]:

	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address
75	NaN	NaN	NaN	NaN	NaN	NaN
169	NaN	NaN	NaN	NaN	NaN	NaN
172	NaN	NaN	NaN	NaN	NaN	NaN
1488	NaN	NaN	NaN	NaN	NaN	NaN
1517	NaN	NaN	NaN	NaN	NaN	NaN

[7]:
```
df2 = df2.dropna(how='all')
```

[8]:
```
df2 = df2[df2['Order Date'].str[0:2] !='Or']
```

4 Agragar una columna para cada mes

[9]:
```
df2['Month']= df2['Order Date'].str[0:2]
df2['Month']= df2['Month'].astype('int32')
df2.head()
```

[9]:
```
  Order ID                   Product Quantity Ordered Price Each  \
0   150502                    iPhone                1        700
1   150503      AA Batteries (4-pack)                1       3.84
2   150504     27in 4K Gaming Monitor                1     389.99
3   150505   Lightning Charging Cable                1      14.95
4   150506      AA Batteries (4-pack)                2       3.84

       Order Date                      Purchase Address  Month
0  02/18/19 01:35     866 Spruce St, Portland, ME 04101      2
1  02/13/19 07:24  18 13th St, San Francisco, CA 94016      2
2  02/18/19 09:46   52 6th St, New York City, NY 10001      2
3  02/02/19 16:47      129 Cherry St, Atlanta, GA 30301      2
4  02/28/19 20:32     548 Lincoln St, Seattle, WA 98101      2
```

5 Agragar una columna de ventas

[10]:
```
df2['Quantity Ordered'] = pd.to_numeric(df2['Quantity Ordered'])
df2['Price Each'] = pd.to_numeric(df2['Price Each'])
```

[11]:
```
df2['Sales']= df2['Quantity Ordered']*df2['Price Each']
df2.head()
```

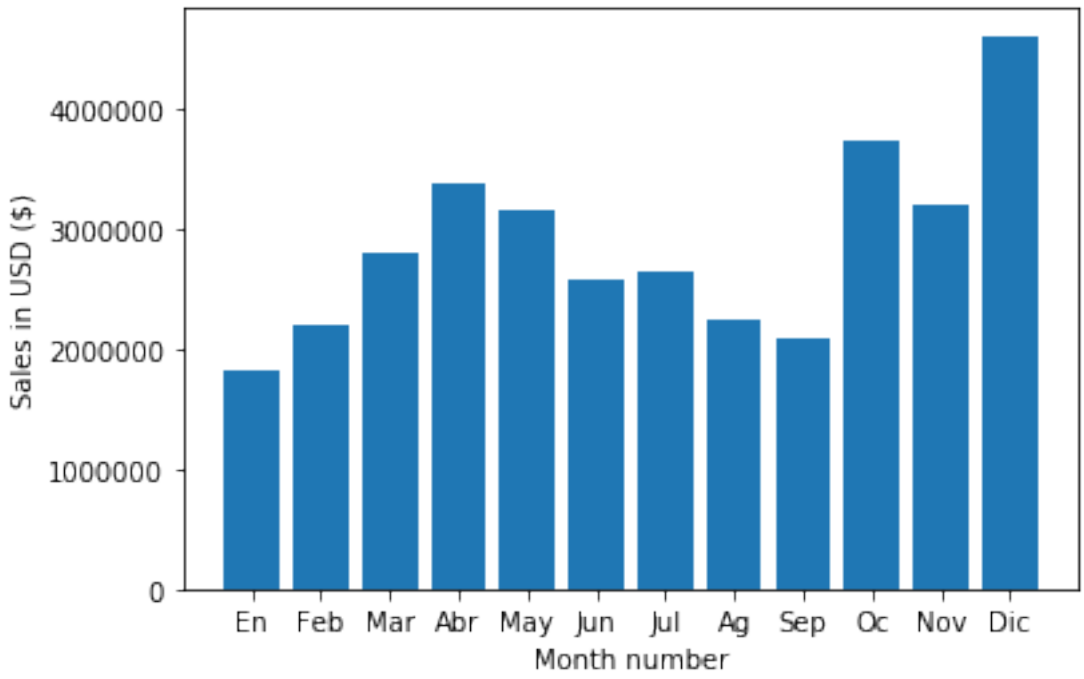
[11]:
```
  Order ID                   Product  Quantity Ordered  Price Each  \
0   150502                    iPhone                 1      700.00
1   150503      AA Batteries (4-pack)                 1        3.84
2   150504     27in 4K Gaming Monitor                 1      389.99
3   150505   Lightning Charging Cable                 1       14.95
4   150506      AA Batteries (4-pack)                 2        3.84

       Order Date                      Purchase Address  Month   Sales
0  02/18/19 01:35     866 Spruce St, Portland, ME 04101      2  700.00
1  02/13/19 07:24  18 13th St, San Francisco, CA 94016      2    3.84
2  02/18/19 09:46   52 6th St, New York City, NY 10001      2  389.99
3  02/02/19 16:47      129 Cherry St, Atlanta, GA 30301      2   14.95
4  02/28/19 20:32     548 Lincoln St, Seattle, WA 98101      2    7.68
```

[12]:
```
df2.groupby('Month').sum()
```

[12]:
```
       Quantity Ordered    Price Each         Sales
Month
```

1	10903	1.811768e+06	1.822257e+06
2	13449	2.188885e+06	2.202022e+06
3	17005	2.791208e+06	2.807100e+06
4	20558	3.367671e+06	3.390670e+06
5	18667	3.135125e+06	3.152607e+06
6	15253	2.562026e+06	2.577802e+06
7	16072	2.632540e+06	2.647776e+06
8	13448	2.230345e+06	2.244468e+06
9	13109	2.084992e+06	2.097560e+06
10	22703	3.715555e+06	3.736727e+06
11	19798	3.180601e+06	3.199603e+06
12	28114	4.588415e+06	4.613443e+06

[13]:
```
import numpy as np
results = df2.groupby('Month').sum()
months = range(1,13)
plt.bar(months, results['Sales'])
plt.xticks(np.arange(1,13), ('En', 'Feb', 'Mar', 'Abr', 'May', 'Jun', 'Jul',␣
 ↪'Ag', 'Sep', 'Oc', 'Nov', 'Dic'))
plt.xlabel('Month number')
plt.ylabel('Sales in USD ($)')
plt.show()
```



5.0.1 En que ciudad hubo más ventas

5.0.2 Aderir un columna de ciudad, usando .apply

[14]:
```
df2['City']=df2['Purchase Address'].apply(lambda x: x.split(',')[1] + ' (' + x.
↪split(',')[2].split(' ')[1] + ')')
df2.head()
```

[14]:

	Order ID	Product	Quantity Ordered	Price Each	\
0	150502	iPhone	1	700.00	
1	150503	AA Batteries (4-pack)	1	3.84	
2	150504	27in 4K Gaming Monitor	1	389.99	
3	150505	Lightning Charging Cable	1	14.95	
4	150506	AA Batteries (4-pack)	2	3.84	

	Order Date	Purchase Address	Month	Sales	\
0	02/18/19 01:35	866 Spruce St, Portland, ME 04101	2	700.00	
1	02/13/19 07:24	18 13th St, San Francisco, CA 94016	2	3.84	
2	02/18/19 09:46	52 6th St, New York City, NY 10001	2	389.99	
3	02/02/19 16:47	129 Cherry St, Atlanta, GA 30301	2	14.95	
4	02/28/19 20:32	548 Lincoln St, Seattle, WA 98101	2	7.68	

	City
0	Portland (ME)
1	San Francisco (CA)
2	New York City (NY)
3	Atlanta (GA)
4	Seattle (WA)

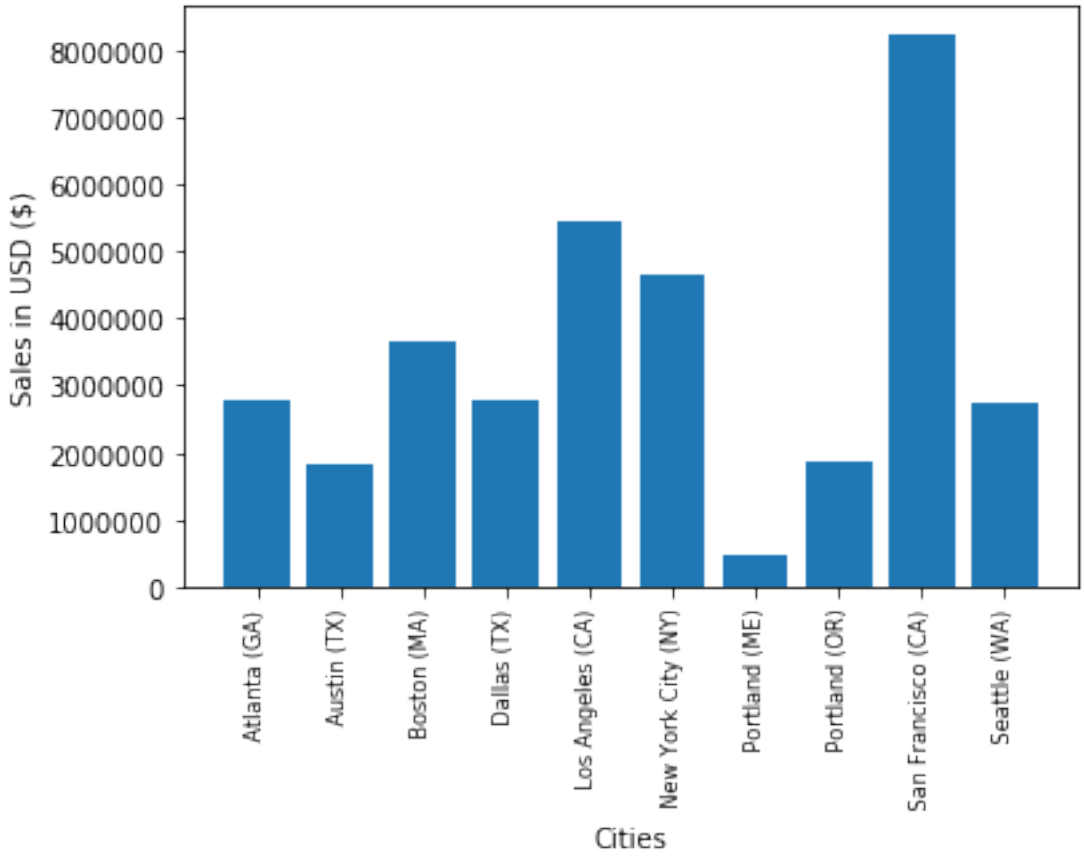
[15]:
```
df2.groupby('City').sum()
```

[15]:

City	Quantity Ordered	Price Each	Month	Sales
Atlanta (GA)	16602	2.779908e+06	104794	2.795499e+06
Austin (TX)	11153	1.809874e+06	69829	1.819582e+06
Boston (MA)	22528	3.637410e+06	141112	3.661642e+06
Dallas (TX)	16730	2.752628e+06	104620	2.767975e+06
Los Angeles (CA)	33289	5.421435e+06	208325	5.452571e+06
New York City (NY)	27932	4.635371e+06	175741	4.664317e+06
Portland (ME)	2750	4.471892e+05	17144	4.497583e+05
Portland (OR)	11303	1.860558e+06	70621	1.870732e+06
San Francisco (CA)	50239	8.211462e+06	315520	8.262204e+06
Seattle (WA)	16553	2.733296e+06	104941	2.747755e+06

[16]:
```
results2 = df2.groupby('City').sum()
cities = [city for city, df in df2.groupby('City')]
plt.bar(cities, results2['Sales'])
plt.xticks(cities, rotation = 'vertical', size=8)
```

```
plt.xlabel('Cities')
plt.ylabel('Sales in USD ($)')
plt.show()
```



6 A qué hora debemos mostrar publicidad para maximizar la probabilidad de que los clientes compren productos?

[17]:
```
df2['Order Date'] = pd.to_datetime(df2['Order Date']) #df2.drop(columns=['Order␣
↪date'])
df2.head()
```

[17]:

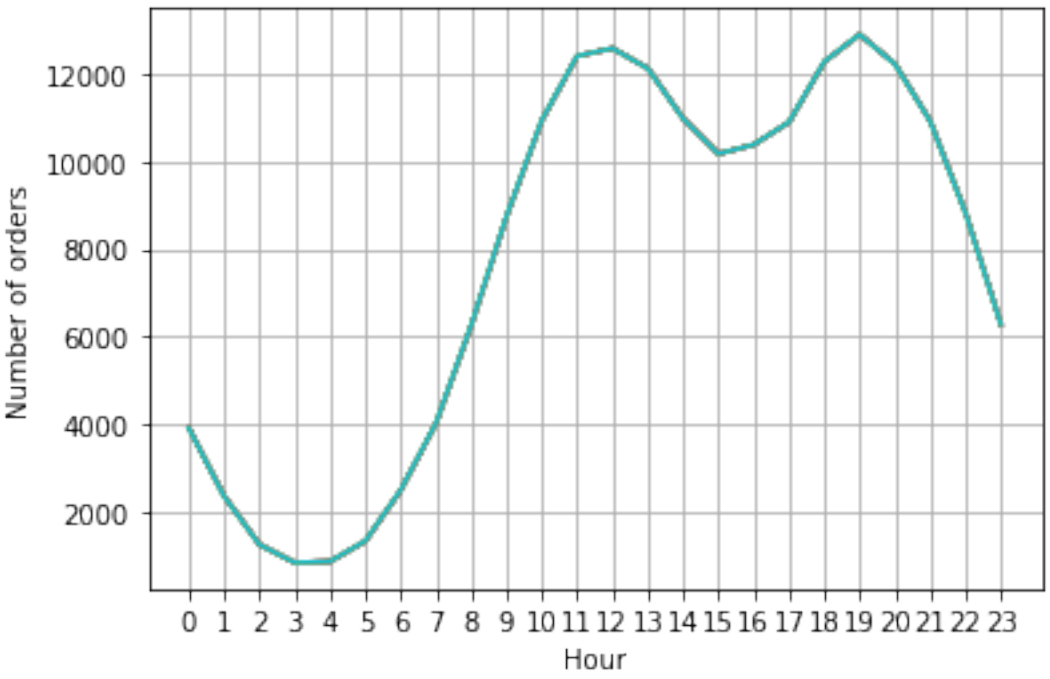
	Order ID	Product	Quantity Ordered	Price Each	\
0	150502	iPhone	1	700.00	
1	150503	AA Batteries (4-pack)	1	3.84	
2	150504	27in 4K Gaming Monitor	1	389.99	
3	150505	Lightning Charging Cable	1	14.95	
4	150506	AA Batteries (4-pack)	2	3.84	

```
            Order Date                       Purchase Address  Month   Sales  \
0  2019-02-18 01:35:00    866 Spruce St, Portland, ME 04101      2  700.00
1  2019-02-13 07:24:00  18 13th St, San Francisco, CA 94016      2    3.84
2  2019-02-18 09:46:00   52 6th St, New York City, NY 10001      2  389.99
3  2019-02-02 16:47:00     129 Cherry St, Atlanta, GA 30301      2   14.95
4  2019-02-28 20:32:00    548 Lincoln St, Seattle, WA 98101      2    7.68

                  City
0        Portland (ME)
1   San Francisco (CA)
2   New York City (NY)
3         Atlanta (GA)
4         Seattle (WA)
```

[18]:
```
df2['Hour'] = df2['Order Date'].dt.hour
df2['Minute'] = df2['Order Date'].dt.minute
df2.head()
```

[18]:
```
  Order ID                   Product  Quantity Ordered  Price Each  \
0   150502                    iPhone                 1      700.00
1   150503      AA Batteries (4-pack)                 1        3.84
2   150504    27in 4K Gaming Monitor                  1      389.99
3   150505  Lightning Charging Cable                  1       14.95
4   150506      AA Batteries (4-pack)                 2        3.84

            Order Date                       Purchase Address  Month   Sales  \
0  2019-02-18 01:35:00    866 Spruce St, Portland, ME 04101      2  700.00
1  2019-02-13 07:24:00  18 13th St, San Francisco, CA 94016      2    3.84
2  2019-02-18 09:46:00   52 6th St, New York City, NY 10001      2  389.99
3  2019-02-02 16:47:00     129 Cherry St, Atlanta, GA 30301      2   14.95
4  2019-02-28 20:32:00    548 Lincoln St, Seattle, WA 98101      2    7.68

                  City  Hour  Minute
0        Portland (ME)     1      35
1   San Francisco (CA)     7      24
2   New York City (NY)     9      46
3         Atlanta (GA)    16      47
4         Seattle (WA)    20      32
```

[20]:
```
hours = [hour for hour, df in df2.groupby('Hour')]
plt.plot(hours, df2.groupby(['Hour']).count())
plt.xticks(hours)
plt.xlabel('Hour')
plt.ylabel('Number of orders')
plt.grid()
plt.show()
```


Number of orders
12000
10000
8000
6000
4000
2000
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
Hour