Analiza prodaje u prodavnici elektronskih uređaja

Importovanje neophodnih biblioteka

In [79]:
```
import pandas as pd
```

Spajanje podataka iz izvora u jedan DataFrame

In [80]:
```
df1 = pd.read_csv('Sales_January_2019.csv')
df2 = pd.read_csv('Sales_February_2019.csv')
df3 = pd.read_csv('Sales_March_2019.csv')
df4 = pd.read_csv('Sales_April_2019.csv')
df5 = pd.read_csv('Sales_May_2019.csv')
df6 = pd.read_csv('Sales_June_2019.csv')
df7 = pd.read_csv('Sales_July_2019.csv')
df8 = pd.read_csv('Sales_August_2019.csv')
df9 = pd.read_csv('Sales_September_2019.csv')
df10 = pd.read_csv('Sales_October_2019.csv')
df11 = pd.read_csv('Sales_November_2019.csv')
df12 = pd.read_csv('Sales_December_2019.csv')

df = pd.concat([df1, df2, df3, df4, df5, df6, df7, df8, df9, df10, df11, df12])
```

Učitavanje novog Dataframe

In [81]:
```
df.head()
```

Out[81]:

	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address
0	141234	iPhone	1	700	01/22/19 21:25	944 Walnut St, Boston, MA 02215
1	141235	Lightning Charging Cable	1	14.95	01/28/19 14:15	185 Maple St, Portland, OR 97035
2	141236	Wired Headphones	2	11.99	01/17/19 13:33	538 Adams St, San Francisco, CA 94016
3	141237	27in FHD Monitor	1	149.99	01/05/19 20:33	738 10th St, Los Angeles, CA 90001
4	141238	Wired Headphones	1	11.99	01/25/19 11:59	387 10th St, Austin, TX 73301

Čišćenje podataka

Prvi korak je da obrišemo sta treba da se očisti. Prilikom obavljanja operacija javljaju se greške zbog Null vrednosti koje se pojavljuju u DataFrame. Neophodno ih je ukloniti.

Uklanjanje NaN vrednosti

In [87]:
```
df_null = df[df.isna().any(axis=1)]
display(df_null.head())

df = df.dropna(how='all')
df.head()
```

	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address
664	NaN	NaN	NaN	NaN	NaN	NaN
678	NaN	NaN	NaN	NaN	NaN	NaN
797	NaN	NaN	NaN	NaN	NaN	NaN
876	NaN	NaN	NaN	NaN	NaN	NaN
1299	NaN	NaN	NaN	NaN	NaN	NaN

Out[87]:

	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address
0	141234	iPhone	1	700	01/22/19 21:25	944 Walnut St, Boston, MA 02215
1	141235	Lightning Charging Cable	1	14.95	01/28/19 14:15	185 Maple St, Portland, OR 97035
2	141236	Wired Headphones	2	11.99	01/17/19 13:33	538 Adams St, San Francisco, CA 94016
3	141237	27in FHD Monitor	1	149.99	01/05/19 20:33	738 10th St, Los Angeles, CA 90001
4	141238	Wired Headphones	1	11.99	01/25/19 11:59	387 10th St, Austin, TX 73301

In [88]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 186305 entries, 0 to 25115
Data columns (total 6 columns):
 #   Column            Non-Null Count   Dtype 
---  ------            --------------   ----- 
 0   Order ID          186305 non-null  object
 1   Product           186305 non-null  object
 2   Quantity Ordered  186305 non-null  object
 3   Price Each        186305 non-null  object
 4   Order Date        186305 non-null  object
 5   Purchase Address  186305 non-null  object
dtypes: object(6)
memory usage: 9.9+ MB
```

Uklanjanje teksta ('Or') u 'Order date' koloni.

In [89]:
```
df = df[df['Order Date'].str[0:2]!='Or']
```

Dodavanje kolonama ispravni tip podataka

In [91]:
```
df['Quantity Ordered'] = pd.to_numeric(df['Quantity Ordered'])
df['Price Each'] = pd.to_numeric(df['Price Each'])
df['Order Date'] = pd.to_datetime(df['Order Date'],  errors='coerce')
```

Dodavanje novih kolona

Dodavanje kolone 'Month'

In [97]:
```
df['Month'] = df['Order Date'].dt.month
df.head()
```

Out[97]:

	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address	Month
0	141234	iPhone	1	700.00	2019-01-22 21:25:00	944 Walnut St, Boston, MA 02215	1
1	141235	Lightning Charging Cable	1	14.95	2019-01-28 14:15:00	185 Maple St, Portland, OR 97035	1
2	141236	Wired Headphones	2	11.99	2019-01-17 13:33:00	538 Adams St, San Francisco, CA 94016	1
3	141237	27in FHD Monitor	1	149.99	2019-01-05 20:33:00	738 10th St, Los Angeles, CA 90001	1
4	141238	Wired Headphones	1	11.99	2019-01-25 11:59:00	387 10th St, Austin, TX 73301	1

Dodavanje kolone 'City'

In [99]:
```
def get_city(address):
    return address.split(",")[1].strip(" ")

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

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

Out[99]:

	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address	Month	City
0	141234	iPhone	1	700.00	2019-01-22 21:25:00	944 Walnut St, Boston, MA 02215	1	Boston (MA)
1	141235	Lightning Charging Cable	1	14.95	2019-01-28 14:15:00	185 Maple St, Portland, OR 97035	1	Portland (OR)
2	141236	Wired Headphones	2	11.99	2019-01-17 13:33:00	538 Adams St, San Francisco, CA 94016	1	San Francisco (CA)
3	141237	27in FHD Monitor	1	149.99	2019-01-05 20:33:00	738 10th St, Los Angeles, CA 90001	1	Los Angeles (CA)
4	141238	Wired Headphones	1	11.99	2019-01-25 11:59:00	387 10th St, Austin, TX 73301	1	Austin (TX)

Istraživanje podataka i vizualizacija

Pitanje 1: Koji je najbolji mesec za prodaju?

In [100]:
```
df['Sales'] = df['Quantity Ordered'].astype('int') * df['Price Each'].astype('float')
```

In [101]:
```
df.groupby(['Month']).sum()
```

Out[101]:

Month	Quantity Ordered	Price Each	Sales
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

In [102]:
```
import matplotlib.pyplot as plt

months = range(1,13)
print(months)

plt.bar(months,df.groupby(['Month']).sum()['Sales'])
plt.xticks(months)
plt.ylabel('Sales in USD ($)')
plt.xlabel('Month number')
plt.show()
```

```
range(1, 13)
```

Najbolji mesec za prodaju proizvoda je decembar.

Pitanje 2: U kojem gradu je prodato najviše proizvoda?

In [103]:
```
df.groupby(['City']).sum()
```

Out[103]:

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	4.471893e+05	17144	4.497583e+05
Portland (OR)	11303	1.860558e+06	70621	1.870732e+06
San Francisco (CA)	50239	8.211461e+06	315520	8.262204e+06
Seattle (WA)	16553	2.733296e+06	104941	2.747755e+06

In [104]:
```
import matplotlib.pyplot as plt

keys = [city for city, df in df.groupby(['City'])]

plt.bar(keys,df.groupby(['City']).sum()['Sales'])
plt.ylabel('Sales in USD ($)')
plt.xlabel('Month number')
plt.xticks(keys, rotation='vertical', size=8)
plt.show()
```

San Francisco je grad u kojem je prodato najviše proizvoda

Pitanje 3: U koje vreme promovisati proizvode da bi se povećala prodaja?

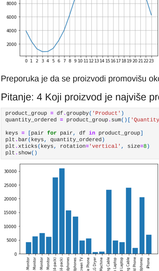
In [105]:
```
# dodavanje 'Hour' kolone
df['Hour'] = pd.to_datetime(df['Order Date']).dt.hour
df['Minute'] = pd.to_datetime(df['Order Date']).dt.minute
df['Count'] = 1
df.head()
```

Out[105]:

	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address	Month	City	Sales	Hour	Minute	Count
0	141234	iPhone	1	700.00	2019-01-22 21:25:00	944 Walnut St, Boston, MA 02215	1	Boston (MA)	700.00	21	25	1
1	141235	Lightning Charging Cable	1	14.95	2019-01-28 14:15:00	185 Maple St, Portland, OR 97035	1	Portland (OR)	14.95	14	15	1
2	141236	Wired Headphones	2	11.99	2019-01-17 13:33:00	538 Adams St, San Francisco, CA 94016	1	San Francisco (CA)	23.98	13	33	1
3	141237	27in FHD Monitor	1	149.99	2019-01-05 20:33:00	738 10th St, Los Angeles, CA 90001	1	Los Angeles (CA)	149.99	20	33	1
4	141238	Wired Headphones	1	11.99	2019-01-25 11:59:00	387 10th St, Austin, TX 73301	1	Austin (TX)	11.99	11	59	1

In [106]:
```
keys = [pair for pair, df in df.groupby(['Hour'])]

plt.plot(keys, df.groupby(['Hour']).count()['Count'])
plt.xticks(keys)
plt.grid()
plt.show()
```



Preporuka je da se proizvodi promovišu oko 11h ili 19h.

Pitanje: 4 Koji proizvod je najviše prodavan?

In [107]:
```
product_group = df.groupby('Product')
quantity_ordered = product_group.sum()['Quantity Ordered']

keys = [pair for pair, df in product_group]
plt.bar(keys, quantity_ordered)
plt.xticks(keys, rotation='vertical', size=8)
plt.show()
```

Proizvod AAA Batteries je najprodavaniji.

Zaključak

U projektu sam koristio Python biblioteke Pandas i Matplotlib za analizu i vizualizaciju podataka o prodaji elektronskih uređaja. Podaci sadrže stotine hiljada kupovina u prodavnici elektronskih uređaja razvrstanih po mesecima, vrsti proizvoda, ceni, adresi kupovine itd. Prikazani su odgovori na neka poslovna pitanja koja mogu dati jasnije uvide o daljem poslovanju i prodaji proizvoda.