Exploratory Analysis on Sales Data.. by AL

In [1]:
```
# import data analysis packages for visualization
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
```

In [2]:
```
# import file and create data frame
df = pd.read_csv("C:\\Users\\Zen\\Documents\\PracticeData\\FilteredSalesData.csv")
```

What are the data types and how many records are there?

In [3]:
```
# Shown info for data types of columns
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 199 entries, 0 to 198
Data columns (total 16 columns):
 #   Column          Non-Null Count  Dtype 
---  ------          --------------  ----- 
 0   Order_Date      199 non-null    object
 1   OrderId         199 non-null    object
 2   Item_ID         199 non-null    object
 3   Item            199 non-null    object
 4   Qty_Sold        199 non-null    int64 
 5   Actual Price    199 non-null    int64 
 6   Discount Price  199 non-null    int64 
 7   Discount        199 non-null    object
 8   Date_Shipd      199 non-null    object
 9   Time_Shipd      199 non-null    object
 10  Loc_ID          199 non-null    object
 11  City            199 non-null    object
 12  Country         199 non-null    object
 13  Region          171 non-null    object
 14  Sales_ID        199 non-null    object
 15  SalesPerson     199 non-null    object
dtypes: int64(3), object(13)
memory usage: 25.0+ KB
```

What entries are missing from the region column?

In [4]:
```
# limited data output
df.loc[df['Region'].isna()].head(5)
```

Out[4]:

	Order_Date	OrderId	Item_ID	Item	Qty_Sold	Actual Price	Discount Price	Discount	Date_Sh
1	12/15/2013	I-1164	I-04	Toaster	43	50	40	20.00%	1/26/2
3	1/15/2014	I-1019	I-03	Ceiling fan	53	150	150	0.00%	3/11/2
4	1/16/2014	I-1133	I-11	Vacuum Cleaner	94	250	175	30.00%	2/8/2
10	3/14/2014	I-1199	I-11	Vacuum Cleaner	88	250	193	22.80%	5/1/2
21	6/9/2014	I-1088	I-03	Ceiling fan	90	150	150	0.00%	7/15/2

In [5]:
```
# Convert date and time columns for analysis, re run info cell to verify checks
df['Date_Shipd']= pd.to_datetime(df['Date_Shipd'])
df['Order_Date']= pd.to_datetime(df['Order_Date'])
df['Time_Shipd']= pd.to_timedelta(df['Time_Shipd'])
```

What does the data frame look like?

In [6]:
```
# show top 5 rows of data frame to see what data looks like
df.head()
```

Out[6]:

	Order_Date	OrderId	Item_ID	Item	Qty_Sold	Actual Price	Discount Price	Discount	Date_Shi
0	2013-11-11	I-1092	I-07	Washing Machine	34	800	712	11.00%	2014-01-
1	2013-12-15	I-1164	I-04	Toaster	43	50	40	20.00%	2014-01-
2	2013-12-26	I-1184	I-04	Toaster	87	50	44	12.00%	2014-01-
3	2014-01-15	I-1019	I-03	Ceiling fan	53	150	150	0.00%	2014-03-
4	2014-01-16	I-1133	I-11	Vacuum Cleaner	94	250	175	30.00%	2014-02-

Who is the Best Salesperson?

In [7]:
```
# Create list of salesperon and how many items they sold and total value
df[['SalesPerson','Qty_Sold','Actual Price']].groupby(by=['SalesPerson'],sort=True)
```

Out[7]:

SalesPerson	Qty_Sold	Actual Price
Nicholas Holloway	36	800
Ronald Butler	917	6160
Anthony Connolly	943	9140
Abdul Heywood	1050	6260
Alen Dinan	1066	6900
Robin Hall	1130	6470
Chandrakant Atkins	1131	7370
Gary Shaw	1153	3670
Philip Dewar	1229	10140
Robert Arnold	1470	6590

What country are the sales people from?

In [8]:
```
# temporaraly set index to get grouping of sales names
df[['Country','SalesPerson']].set_index('Country').groupby(by=['Country','SalesPers
```

Out[8]:

Country	SalesPerson
Argentina	**Abdul Heywood**
	Alen Dinan
	Philip Dewar
Australia	**Alen Dinan**
	Anthony Connolly
...	**...**
USA	**Ronald Butler**
Vietnam	**Abdul Heywood**
	Alen Dinan
	Chandrakant Atkins
	Robert Arnold

141 rows × 0 columns

What item sold the most units?

In [9]:
```
# List of item qty that sold
df[['Item','Qty_Sold']].groupby(by=['Item'],sort=True).sum().sort_values(by='Qty_So
```

Out[9]:

Item	Qty_Sold
Dishwasher	478
Washing Machine	690
Coffee grinder	713
Ceiling fan	726
Air conditioner	727
Blender	795
Toaster	798
Refrigerator	817
Vacuum Cleaner	829
Oven	975
Microwave	1099
Iron	1478

What item brought in the higest gross profits?

In [10]:
```
# List of item qty that sold
df[['Item','Actual Price']].groupby(by=['Item'],sort=True).sum().sort_values(by='Ac
```

Out[10]:

Item	Actual Price
Refrigerator	17000
Washing Machine	12000
Air conditioner	9800
Oven	9500
Dishwasher	4000
Vacuum Cleaner	3750
Ceiling fan	2250
Microwave	1600
Coffee grinder	1260
Iron	840
Blender	750
Toaster	750

What country are we shipping the most units to?

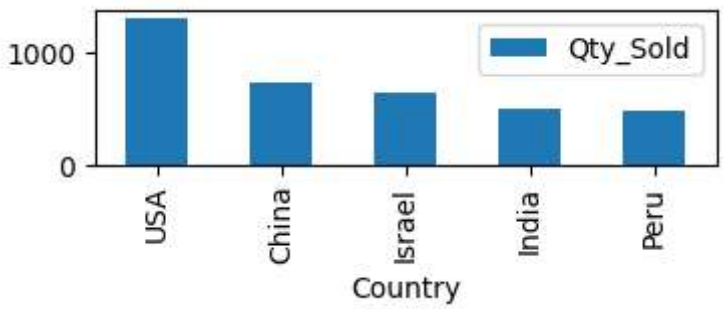
In [11]:
```
df[['Country','Qty_Sold']].groupby(by='Country').sum().sort_values(by='Qty_Sold',as
```

Out[11]:

Country	Qty_Sold
USA	1321
China	739
Israel	643
India	505
Peru	490

In [12]:
```
df[['Country','Qty_Sold']].groupby(by='Country').sum().sort_values(by='Qty_Sold',as
```

Out[12]:
```
<Axes: xlabel='Country'>
```



What country is placing the most orders?

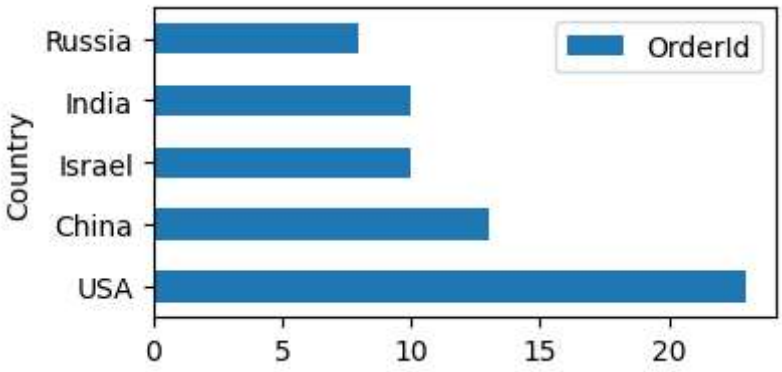
In [13]:
```
df[['Country','OrderId']].groupby(by=['Country']).count().sort_values(by='OrderId',
```

Out[13]:

Country	OrderId
USA	23
China	13
Israel	10
India	10
Russia	8

In [14]:
```
df[['Country','OrderId']].groupby(by=['Country']).count().sort_values(by='OrderId',
```

Out[14]:
```
<Axes: ylabel='Country'>
```



What was the amount of discounts given year over year?

In [15]:
```
# Adding a total discount column to df and seeing yearly values
df['total_discount']= df['Actual Price'] - df['Discount Price']
df[['Order_Date','Qty_Sold','Actual Price','Discount Price','total_discount']].resa
```

Out[15]:

Order_Date	Qty_Sold	Actual Price	Discount Price	total_discount
2013-12-31	164	900	796	104
2014-12-31	2127	15180	12979	2201
2015-12-31	2088	9480	7639	1841
2016-12-31	1590	10450	9548	902
2017-12-31	2171	16370	14022	2348
2018-12-31	1985	11120	9918	1202

What are the country discount totals over the years?

In [16]:
```
df[['Country','Order_Date','total_discount']].set_index(['Country']).groupby(by='Co
```

Out[16]:

Country	Order_Date	total_discount
Argentina	**2013-12-31**	88
	2014-12-31	40
	2015-12-31	0
	2016-12-31	2
	2017-12-31	430
...	**...**	...
Vietnam	**2014-12-31**	8
	2015-12-31	2
	2016-12-31	0
	2017-12-31	2
	2018-12-31	4

137 rows × 1 columns

What are the stats on the total discounts given?

In [17]:
```
# df discount stats
df['total_discount'].describe()
```

Out[17]:
```
count    199.00000
mean      43.20603
std       90.93378
min        0.00000
25%        3.00000
50%        8.00000
75%       27.00000
max      500.00000
Name: total_discount, dtype: float64
```

What orders were above the mean discount price and who gave the most discounts?

In [18]:
```
df.loc[(df['total_discount']> 43)].groupby(by='SalesPerson').sum(numeric_only=True)
```

Out[18]:

SalesPerson	Qty_Sold	Actual Price	Discount Price	total_discount
Philip Dewar	364	7100	4782	2318
Alen Dinan	276	4500	3328	1172
Chandrakant Atkins	295	5250	4107	1143
Robert Arnold	297	3550	2838	712
Anthony Connolly	167	4600	3913	687
Abdul Heywood	190	2600	2060	540
Gary Shaw	99	800	512	288
Nicholas Holloway	36	800	584	216
Ronald Butler	68	1000	910	90
Robin Hall	89	500	455	45