Darshan UNIVERSITY
योग: कर्मसु कौशलम्

Data Mining

Lab - 4

137 | Vishal Baraiya | 23010101014

Step 1. Import the necessary libraries

In [25]:
```
import pandas as pd
import numpy as np
```

Step 2. Import the dataset from this address.

Step 3. Assign it to a variable called chipo.

In [4]:
```
chipo = pd.read_csv("https://raw.githubusercontent.com/justmarkham/DAT8/master/data
chipo
```

Out[4]:

	order_id	quantity	item_name	choice_description	item_price
0	1	1	Chips and Fresh Tomato Salsa	NaN	$2.39
1	1	1	Izze	[Clementine]	$3.39
2	1	1	Nantucket Nectar	[Apple]	$3.39
3	1	1	Chips and Tomatillo-Green Chili Salsa	NaN	$2.39
4	2	2	Chicken Bowl	[Tomatillo-Red Chili Salsa (Hot), [Black Beans...	$16.98
...	...	...	...	...	...
4617	1833	1	Steak Burrito	[Fresh Tomato Salsa, [Rice, Black Beans, Sour ...	$11.75
4618	1833	1	Steak Burrito	[Fresh Tomato Salsa, [Rice, Sour Cream, Cheese...	$11.75
4619	1834	1	Chicken Salad Bowl	[Fresh Tomato Salsa, [Fajita Vegetables, Pinto...	$11.25
4620	1834	1	Chicken Salad Bowl	[Fresh Tomato Salsa, [Fajita Vegetables, Lettu...	$8.75
4621	1834	1	Chicken Salad Bowl	[Fresh Tomato Salsa, [Fajita Vegetables, Pinto...	$8.75

4622 rows × 5 columns

Step 4. See the first 10 entries

In [5]:
```
chipo.head(10)
```

Out[5]:

	order_id	quantity	item_name	choice_description	item_price
0	1	1	Chips and Fresh Tomato Salsa	NaN	$2.39
1	1	1	Izze	[Clementine]	$3.39
2	1	1	Nantucket Nectar	[Apple]	$3.39
3	1	1	Chips and Tomatillo-Green Chili Salsa	NaN	$2.39
4	2	2	Chicken Bowl	[Tomatillo-Red Chili Salsa (Hot), [Black Beans...	$16.98
5	3	1	Chicken Bowl	[Fresh Tomato Salsa (Mild), [Rice, Cheese, Sou...	$10.98
6	3	1	Side of Chips	NaN	$1.69
7	4	1	Steak Burrito	[Tomatillo Red Chili Salsa, [Fajita Vegetables...	$11.75
8	4	1	Steak Soft Tacos	[Tomatillo Green Chili Salsa, [Pinto Beans, Ch...	$9.25
9	5	1	Steak Burrito	[Fresh Tomato Salsa, [Rice, Black Beans, Pinto...	$9.25

Step 5. What is the number of observations in the dataset?

In [9]:
```
# Solution 1
chipo.shape[0]
```

Out[9]: 4622

In [8]:
```
# Solution 2
chipo.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4622 entries, 0 to 4621
Data columns (total 5 columns):
 #   Column              Non-Null Count  Dtype 
---  ------              --------------  ----- 
 0   order_id            4622 non-null   int64 
 1   quantity            4622 non-null   int64 
 2   item_name           4622 non-null   object
 3   choice_description  3376 non-null   object
 4   item_price          4622 non-null   object
dtypes: int64(2), object(3)
memory usage: 180.7+ KB
```

Step 6. What is the number of columns in the dataset?

In [10]:
```
chipo.shape[1]
```

Out[10]: 5

Step 7. Print the name of all the columns.

In [11]:
```
chipo.columns
```

Out[11]:
```
Index(['order_id', 'quantity', 'item_name', 'choice_description',
       'item_price'],
      dtype='object')
```

Step 8. How is the dataset indexed?

In [12]:
```
chipo.index
```

Out[12]:
```
RangeIndex(start=0, stop=4622, step=1)
```

Step 9. Number of Unique Items ?

In [17]:
```
chipo["item_name"].nunique()
```

Out[17]: 50

In [18]:
```
chipo["item_name"].unique()
```

Out[18]:
```
array(['Chips and Fresh Tomato Salsa', 'Izze', 'Nantucket Nectar',
       'Chips and Tomatillo-Green Chili Salsa', 'Chicken Bowl',
       'Side of Chips', 'Steak Burrito', 'Steak Soft Tacos',
       'Chips and Guacamole', 'Chicken Crispy Tacos',
       'Chicken Soft Tacos', 'Chicken Burrito', 'Canned Soda',
       'Barbacoa Burrito', 'Carnitas Burrito', 'Carnitas Bowl',
       'Bottled Water', 'Chips and Tomatillo Green Chili Salsa',
       'Barbacoa Bowl', 'Chips', 'Chicken Salad Bowl', 'Steak Bowl',
       'Barbacoa Soft Tacos', 'Veggie Burrito', 'Veggie Bowl',
       'Steak Crispy Tacos', 'Chips and Tomatillo Red Chili Salsa',
       'Barbacoa Crispy Tacos', 'Veggie Salad Bowl',
       'Chips and Roasted Chili-Corn Salsa',
       'Chips and Roasted Chili Corn Salsa', 'Carnitas Soft Tacos',
       'Chicken Salad', 'Canned Soft Drink', 'Steak Salad Bowl',
       '6 Pack Soft Drink', 'Chips and Tomatillo-Red Chili Salsa', 'Bowl',
       'Burrito', 'Crispy Tacos', 'Carnitas Crispy Tacos', 'Steak Salad',
       'Chips and Mild Fresh Tomato Salsa', 'Veggie Soft Tacos',
       'Carnitas Salad Bowl', 'Barbacoa Salad Bowl', 'Salad',
       'Veggie Crispy Tacos', 'Veggie Salad', 'Carnitas Salad'],
      dtype=object)
```

Step 10. Which was the most-ordered item?

In [19]:
```
c = chipo.groupby('item_name')
c = c.sum()
c = c.sort_values(['quantity'],ascending=False)
c.head(1)
```

Out[19]:

item_name	order_id	quantity	choice_description	item_price
Chicken Bowl	713926	761	[Tomatillo-Red Chili Salsa (Hot), [Black Beans...	16.9810.98 11.258.75 8.4911.25 $8.75 ...

In [20]:
```
chipo.groupby('item_name')['quantity'].sum().sort_values(ascending=False).head(1)
```

Out[20]:
```
item_name
Chicken Bowl    761
Name: quantity, dtype: int64
```

Step 11. How many items were orderd in total?

In [21]:
```
chipo.quantity.sum()
```

Out[21]: 4972

Step 12. Turn the item price into a float

Step 12.a. Check the item price type

In [22]:
```
chipo.item_price.dtype
```

Out[22]: dtype('O')

Step 12.b. Create a lambda function and change the type of item price

In [23]:
```
dollarizer = lambda x: float(x[1:-1])
chipo.item_price = chipo.item_price.apply(dollarizer)
```

Step 12.c. Check the item price type

In [24]:
```
chipo.item_price.dtype
```

Out[24]: dtype('float64')

Step 14. How much was the revenue for the period in the dataset?

In [26]:
```
revenue = (chipo['quantity']* chipo['item_price']).sum()
print(f"Revenue Was: ${str(np.round(revenue,2))}")
```

```
Revenue Was: $39237.02
```

Step 15. How many orders were made ?

In [29]:
```
orders = chipo.order_id.value_counts().count()
orders
```

Out[29]:
```
1834
```

Step 17. How many different choice descriptions are there?

In [30]:
```
chipo["choice_description"].nunique()
```

Out[30]:
```
1043
```

Step 18. What items have been ordered more than 100 times?

In [32]:
```
items = chipo.groupby('item_name')['quantity'].sum()
items[items > 100]
```

Out[32]:
```
item_name
Bottled Water                  211
Canned Soda                    126
Canned Soft Drink              351
Chicken Bowl                   761
Chicken Burrito                591
Chicken Salad Bowl             123
Chicken Soft Tacos             120
Chips                          230
Chips and Fresh Tomato Salsa   130
Chips and Guacamole            506
Side of Chips                  110
Steak Bowl                     221
Steak Burrito                  386
Name: quantity, dtype: int64
```

Step 19. What is the average revenue amount per order?

In [33]:
```
# Solution 1
chipo['revenue'] = chipo['quantity'] * chipo['item_price']
order_grouped = chipo.groupby(by=['order_id']).sum()
order_grouped['revenue'].mean()
```

Out[33]:
```
21.39423118865867
```

In [34]:
```
# Solution 2
chipo.groupby(by=['order_id']).sum()['revenue'].mean()
```

Out[34]:
```
21.39423118865867
```

In [ ]:

In [ ]:

In [ ]:

In [ ]: