



## Data Mining

### Lab - 4

Jeet Bhalodi (23031701006)

#### Step 1. Import the necessary libraries

```
In [1]: import pandas as pd  
import numpy as nump
```

#### Step 2. Import the dataset from this [address](https://raw.githubusercontent.com/justmarkham/DAT8/master/data/chipo).

```
In [7]: chipo = pd.read_csv("https://raw.githubusercontent.com/justmarkham/DAT8/master/data/chipo")
```

Out[7]:

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

4622 rows × 5 columns

### Step 3. Assign it to a variable called chipo.

```
In [9]: chipo = pd.read_csv("https://raw.githubusercontent.com/justmarkham/DAT8/master/data/chipo")
```

Out[9]:

	order_id	quantity	item_name	choice_description	item_price
<b>0</b>	1	1	Chips and Fresh Tomato Salsa	NaN	\$2.39
<b>1</b>	1	1	Izze	[Clementine]	\$3.39
<b>2</b>	1	1	Nantucket Nectar	[Apple]	\$3.39
<b>3</b>	1	1	Chips and Tomatillo-Green Chili Salsa	NaN	\$2.39
<b>4</b>	2	2	Chicken Bowl	[Tomatillo-Red Chili Salsa (Hot), [Black Beans...	\$16.98
...	...	...	...	...	...
<b>4617</b>	1833	1	Steak Burrito	[Fresh Tomato Salsa, [Rice, Black Beans, Sour ...	\$11.75
<b>4618</b>	1833	1	Steak Burrito	[Fresh Tomato Salsa, [Rice, Sour Cream, Cheese...	\$11.75
<b>4619</b>	1834	1	Chicken Salad Bowl	[Fresh Tomato Salsa, [Fajita Vegetables, Pinto...	\$11.25
<b>4620</b>	1834	1	Chicken Salad Bowl	[Fresh Tomato Salsa, [Fajita Vegetables, Lettu...	\$8.75
<b>4621</b>	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 [11]: `chipo.head(10)`

Out[11]:

	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 [13]: `# Solution 1`  
`chipo.shape[0]`

Out[13]: 4622

In [15]: `# 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 [17]: `chipo.shape[1]`

Out[17]: 5

## Step 7. Print the name of all the columns.

In [19]: `chipo.columns`

Out[19]: Index(['order\_id', 'quantity', 'item\_name', 'choice\_description',  
'item\_price'],  
dtype='object')

## Step 8. How is the dataset indexed?

In [21]: `chipo.index`

Out[21]: RangeIndex(start=0, stop=4622, step=1)

## Step 9. Number of Unique Items ?

In [23]: `chipo["item_name"].nunique()`

Out[23]: 50

## Step 10. Which was the most-ordered item?

In [38]: `orderd=chipo.groupby("item_name").sum().sort_values(['quantity'])`  
`orderd[['order_id','quantity']].tail(1)`  
  
`# orderd=chipo.groupby("item_name").sum().sort_values(['quantity'],ascending=False)`  
`# orderd[['order_id','quantity']].head(1)`

Out[38]:

	order_id	quantity
item_name		
Chicken Bowl	713926	761

## Step 11. How many items were orderd in total?

In [25]: `chipo["quantity"].sum()`

Out[25]: 4972

## Step 12. Turn the item price into a float

### Step 12.a. Check the item price type

In [94]: `print(chipo['item_price'].dtypes)`

float64

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

```
In [56]: m = lambda x: float(x[1:])
chipo['item_price'] = chipo['item_price'].apply(m)
chipo
```

```
Out[56]:
```

	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 12.c. Check the item price type

```
In [58]: chipo['item_price'].dtypes
```

```
Out[58]: dtype('float64')
```

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

```
In [76]: chipo['revenue'] = chipo['quantity'] * chipo['item_price']
revenue = chipo['revenue'].sum()
print("Revenue was: $",revenue)
```

Revenue was: \$ 39237.02

## Step 15. How many orders were made ?

```
In [44]: chipo['order_id'].nunique()
```

```
Out[44]: 1834
```

## Step 17. How many different choice descriptions are there?

```
In [46]: chipo['choice_description'].nunique()
```

```
Out[46]: 1043
```

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

```
In [92]: ic = chipo.groupby('item_name')['quantity'].sum()
pi = ic[ic > 100]
pi
```

```
Out[92]: 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 [82]: # Solution 1
revenue_per_order = chipo.groupby('order_id')['revenue'].sum()
average_revenue = revenue_per_order.mean()
average_revenue
```

```
Out[82]: 21.39423118865867
```

```
In [ ]:
```