Ex2 - Getting and Knowing your Data

This time we are going to pull data directly from the internet. Special thanks to: https://github.com/justmarkham for sharing the dataset and materials.

Step 1. Import the necessary libraries

```
import pandas as pd
```

Step 2. Import the dataset from this address.

Step 3. Assign it to a variable called chipo.

```
url = 'https://raw.githubusercontent.com/justmarkham/DAT8/master/data/chipotle.tsv'
chipo = pd.read_csv(url, sep='\t')
```

Step 4. See the first 10 entries

```
print(chipo.head(10))
```

```
   order_id  quantity                              item_name  \
0         1         1           Chips and Fresh Tomato Salsa
1         1         1                                   Izze
2         1         1                       Nantucket Nectar
3         1         1  Chips and Tomatillo-Green Chili Salsa
4         2         2                           Chicken Bowl
5         3         1                           Chicken Bowl
6         3         1                          Side of Chips
7         4         1                          Steak Burrito
8         4         1                       Steak Soft Tacos
9         5         1                          Steak Burrito

                                  choice_description item_price
0                                                NaN     $2.39
1                                       [Clementine]     $3.39
2                                            [Apple]     $3.39
3                                                NaN     $2.39
4  [Tomatillo-Red Chili Salsa (Hot), [Black Beans...    $16.98
5  [Fresh Tomato Salsa (Mild), [Rice, Cheese, Sou...    $10.98
6                                                NaN     $1.69
7  [Tomatillo Red Chili Salsa, [Fajita Vegetables...    $11.75
8  [Tomatillo Green Chili Salsa, [Pinto Beans, Ch...     $9.25
9  [Fresh Tomato Salsa, [Rice, Black Beans, Pinto...     $9.25
```

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

```
# Solution 1

print(len(chipo))
```

```
4622
```

```
# Solution 2
print(chipo.shape[0])
```

```
4622
```

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

```
print(chipo.shape[1])
```

```
5
```

Step 7. Print the name of all the columns.

```
print(chipo.columns.tolist())
```

```
['order_id', 'quantity', 'item_name', 'choice_description', 'item_price']
```

Step 8. How is the dataset indexed?

```
print(chipo.index)
```

```
RangeIndex(start=0, stop=4622, step=1)
```

Step 9. Which was the most-ordered item?

```
most_ordered = chipo.groupby('item_name').sum(numeric_only=True).sort_values('quantity', ascending=False).head(1)
print(most_ordered)
```

```
              order_id  quantity
item_name
Chicken Bowl    713926       761
```

Step 10. For the most-ordered item, how many items were ordered?

```
print(most_ordered['quantity'].values[0])
```

```
761
```

Step 11. What was the most ordered item in the choice_description column?

```
most_choice = chipo['choice_description'].value_counts().head(1)
print(most_choice)
```

```
choice_description
[Diet Coke]    134
Name: count, dtype: int64
```

Step 12. How many items were orderd in total?

```
print(chipo['quantity'].sum())
```

```
4972
```

Step 13. Turn the item price into a float

Step 13.a. Check the item price type

```
print(chipo['item_price'].dtype)
```

```
object
```

Step 13.b. Create a lambda function and change the type of item price

```
chipo['item_price'] = chipo['item_price'].apply(lambda x: float(x.replace('$', '')))
```

Step 13.c. Check the item price type

```
print(chipo['item_price'].dtype)
```

```
float64
```

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

```
revenue = (chipo['quantity'] * chipo['item_price']).sum()
print(f"Revenue: ${revenue:.2f}")
```

```
Revenue: $39237.02
```

Step 15. How many orders were made in the period?

```
orders = chipo['order_id'].nunique()
print(f"Total orders: {orders}")
```

```
Total orders: 1834
```

Step 16. What is the average revenue amount per order?

```
# Solution 1
print(f"Average revenue per order: ${revenue / orders:.2f}")
```

```
Average revenue per order: $21.39
```

```
# Solution 2
order_totals = chipo.groupby('order_id').apply(lambda x: (x['quantity'] * x['item_price']).sum())
print(f"Average per order (method 2): ${order_totals.mean():.2f}")
```

```
Average per order (method 2): $21.39
<ipython-input-19-6291081fe1bf>:2: DeprecationWarning: DataFrameGroupBy.apply operated on the grouping columns. This behavior is dep
  order_totals = chipo.groupby('order_id').apply(lambda x: (x['quantity'] * x['item_price']).sum())
```

Step 17. How many different items are sold?

```
print("Unique items sold:", chipo['item_name'].nunique())
```

```
Unique items sold: 50
```