**Data Cleaning (chipotle.tsv):**

**Columns:**

* Order Id
* Quantity
* Item Name
* Choice Description
* Item Price

In this task we are working on tsv file so, before diving into the questions, here's a brief introduction to working with TSV files in **Pandas**:

A TSV (Tab-Separated Values) file is a plain text file where each line represents a data record, and values are separated by tabs. In Pandas, you can use the **read\_csv** function with the sep parameter set to **'\t'** to read TSV files.

**Optional: Converting TSV to CSV in pandas:**

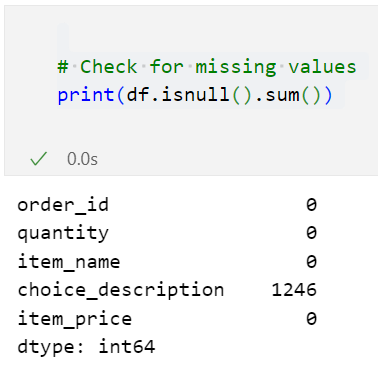
If needed, you can convert a TSV file to CSV using the to\_csv function with the sep parameter set to ','. Now, let's proceed with the questions for data cleaning using the Chipotle dataset:

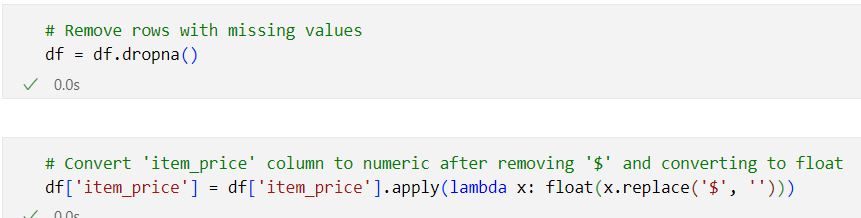
**Reading TSV File:**

Use the **pd.read\_csv** function with sep='\t' to read the Chipotle TSV file into a pandas DataFrame.

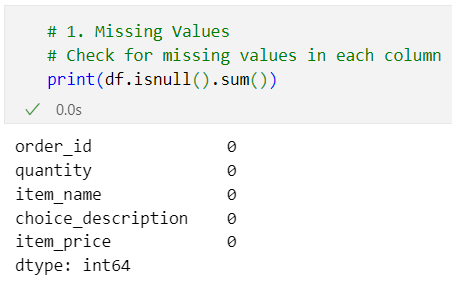
**1. Missing Values:**

**Question:** Check for missing values in each column (Order ID, Quantity, Item Name, Choice Description, Item Price). How should missing values be handled?



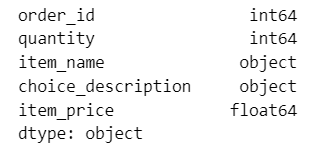


After Cleaning:



**2. Data Types:**

**Question:** Verify the data types of each column. Do they align with their expected types, and should any adjustments be made?



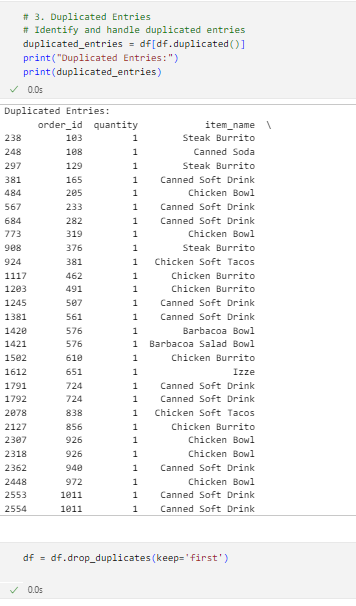
It looks like the data types for each column align with their expected types:

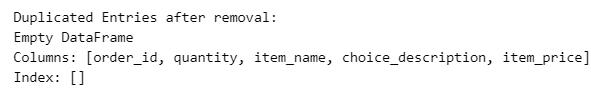
* order\_id: **int64**
* quantity: **int64**
* item\_name: **object**
* choice\_description: **object**
* item\_price: **float64**

Since the data types seem appropriate, there's no immediate need for adjustments in this regard. The integer types (**int64**) are suitable for order\_id and quantity, while the object type is used for textual data (item\_name and choice\_description), and float64 is appropriate for item\_price.

**3. Duplicated Entries:**

**Question:** Identify and handle duplicated entries in the dataset. How might duplicates impact analysis, and what is the appropriate action?

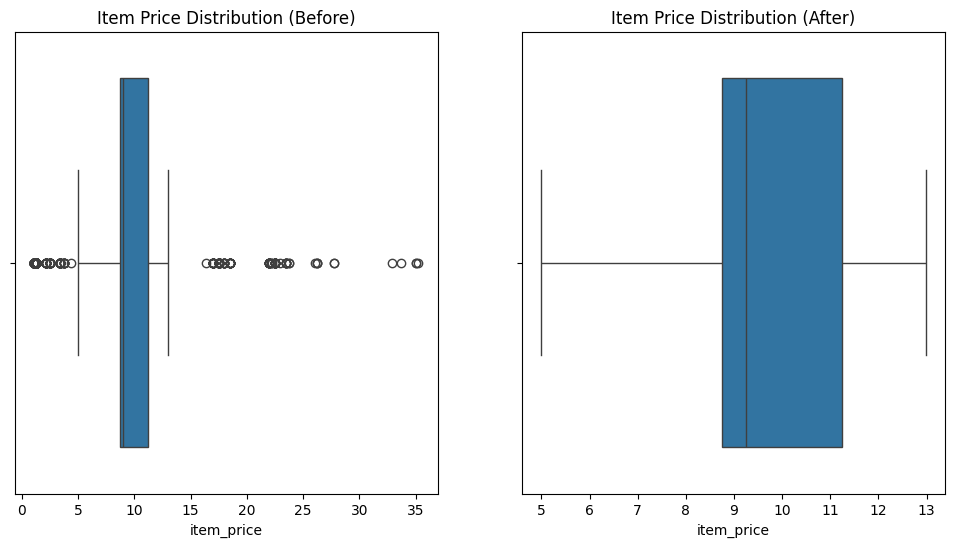
****

****

**4. Quantity and Item Price:**

**Question:** Examine the Quantity and Item Price columns. Are there any inconsistencies or anomalies that need correction?

we successfully removed the item price outliers using the IQR method.



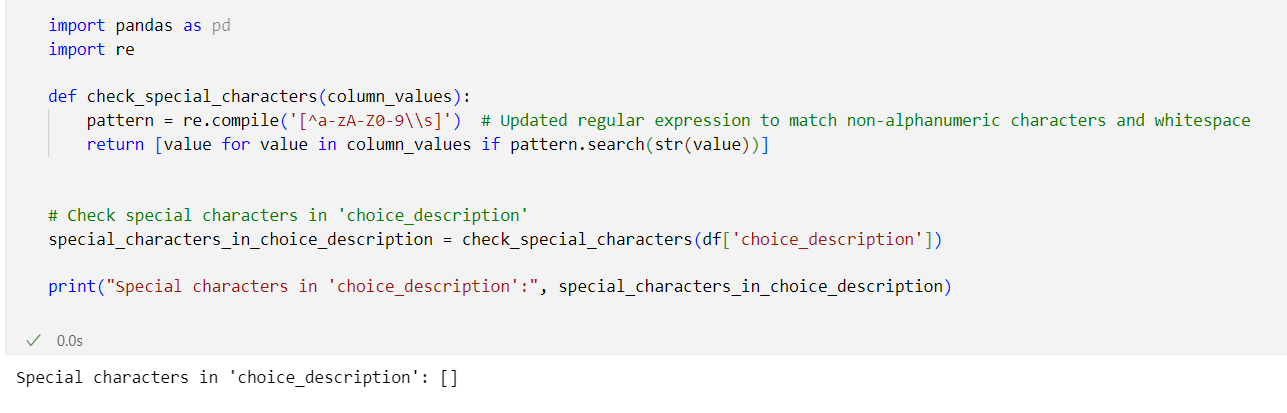
**5. Choice Description:**

**Question:** Analyze the Choice Description column. How should choices be handled, especially when there are multiple descriptions for a single item?

**6. Handling Special Characters:**

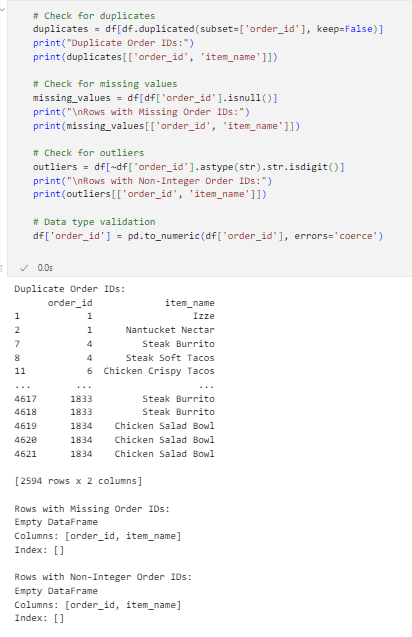
**Question:** Check for special characters in text-based columns (e.g., Item Name, Choice Description). How can these be addressed for consistency?





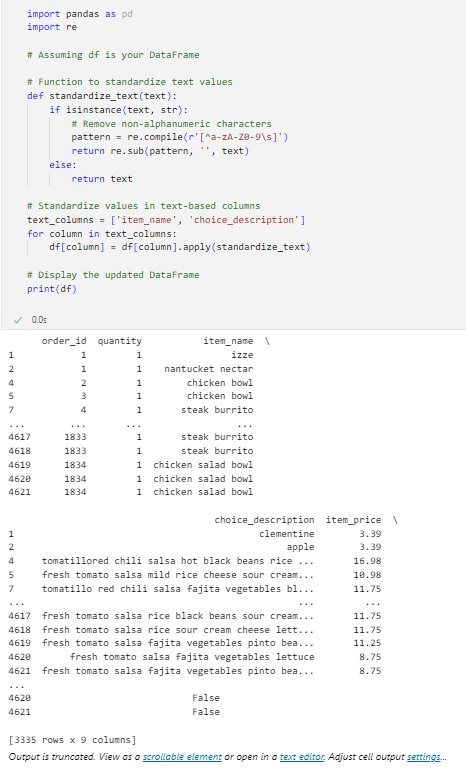
**7. Order Id Integrity:**

**Question:** Cross-reference the Order ID column for integrity. Are there any irregularities or patterns that need validation?

  
It appears that there are duplicate Order IDs in your dataset. The identified duplicates show the same Order ID associated with different item names. This suggests that there might be multiple items associated with a single order. It's not uncommon for an order to have multiple items associated with it, so having duplicate Order IDs may be valid.

**8. Item Name Standardization:**

**Question:** Standardize the Item Name column. Are there variations that can be unified for better analysis?

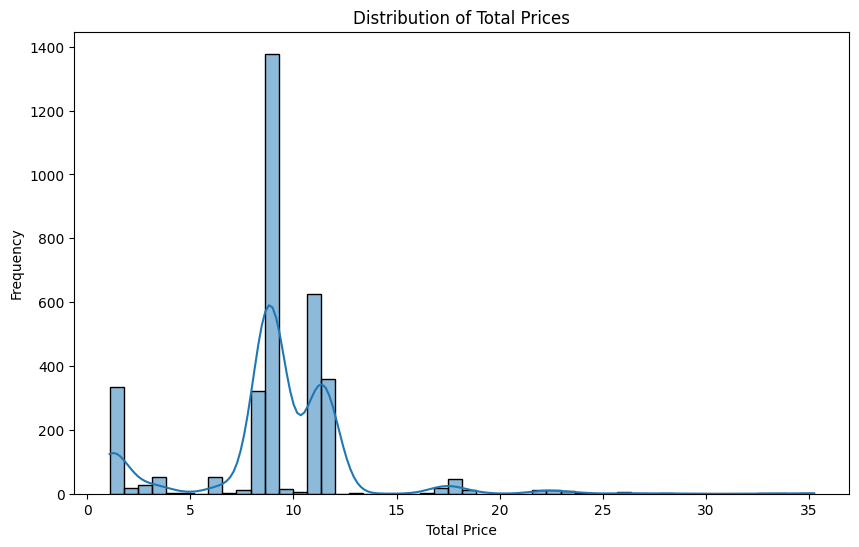


The output provided indicates that the 'item\_name' and 'choice\_description' columns have been standardized. The 'item\_name' column has been transformed to lowercase, and non-alphanumeric characters have been removed. Similarly, the 'choice\_description' column has been processed to remove special characters, convert to lowercase, and unify the format.

The boolean column 'choice\_description\_has\_special\_chars' indicates whether the standardized 'choice\_description' column contains any special characters. In this case, it seems there are no special characters present in the standardized 'choice\_description' column.

**9. Quantity and Price Relationships:**

**Question:** Investigate the relationships between Quantity and Item Price. Are there cases where adjustments need to be made for accurate analysis?



The summary statistics for the total prices provide useful information about the distribution of values:

* Count: 3335 items
* Mean: Approximately 9.04
* Standard Deviation: Approximately 3.78
* Minimum: 1.09
* 25th Percentile (Q1): 8.69
* Median (50th Percentile): 8.99
* 75th Percentile (Q3): 11.25
* Maximum: 35.25

The distribution is right-skewed, as indicated by the higher mean compared to the median. The majority of total prices fall within the range of 8.69 to 11.25, with some items having higher prices (up to a maximum of 35.25).

**10. Data Integrity Check:**

**Question:** Perform a data integrity check by ensuring that quantities and prices align with the corresponding items and descriptions.



If the checks for total prices and choice descriptions did not print anything, it indicates that there are no discrepancies found in the calculated total prices and the alignment of choice descriptions with item names. This is a positive outcome, suggesting that the data is consistent and aligns with expectations in terms of quantities, prices, and item descriptions.

**11. Converting to CSV:**

**Optional Question:** If needed, convert the cleaned dataset to a CSV file using the to\_csv function with sep=','.



**12. Handling Categorical Data:**

**Question:** For categorical columns (e.g., Item Name), consider encoding or transforming them into a format suitable for analysis.

**13. Consistent Quantity and Price Units:**

**Question:** Ensure consistency in units for Quantity and Item Price. Should any conversions or adjustments be made for uniform analysis?

No …

These questions guide you through the process of cleaning and preparing the Chipotle dataset using pure pandas, addressing various aspects of data quality and consistency. Feel free to explore the dataset.