***Sales Analysis of Local Ramen Shop***

Section 52745 | Group Number 30 | Fidel Otero (fco229) | Ryan Mehendale (ram3273) 10/31/2023

In this phase, you will do a systematic analysis of your project. Here are the steps of your project:

1. Read the project description and create a goal for your project. Articulate that goal in a short paragraph.

1. Familiarize yourself with the data set. Here are questions that I want you to answer:

* What is the size of the dataset (how many rows and columns)?
  + 22323 rows × 13 columns
* What are the column names?
  + Order\_Number, Order\_Type, Order\_Date, Order\_Sub\_Total, Order\_Sales\_Tax, Order\_Total, Payment\_Type, Product\_Name, Category, Exchange\_Void\_or\_Return, Total\_Sales\_Price, Total\_Product\_Tax, Product\_Quantity
* Classify each column as - categorical, ordinal, discrete, or continuous.
  + Order\_Number : Categorical, Order\_Type : Categorical, Order\_Date: Continuous, Order\_Sub\_Total: Continuous, Order\_Sales\_Tax: Continuous, Order\_Total: Continuous, Payment\_Type: Categorical, Product\_Name: Categorical Category: Categorical, Exchange\_Void\_or\_Return: Categorical, Total\_Sales\_Price: Continuous, Total\_Product\_Tax: Continuous, Product\_Quantity: Discrete
* For categorical data, find the discrete values.
  + Order\_Number: 8789 different unique order numbers
  + Order\_Type: 'To Go', 'Eat In', 'Doordash', 'Uber', 'Grubhub'
  + Product\_Name: 267 different product names
  + Exchange\_Void\_or\_Return: 0 and 1
  + Category: 'Ramen', 'Specials', 'Appetizers', 'Food', 'Drinks', nan, 'Special', 'Extras', 'Sazan @ Daiboku', 'HAPPY HOUR', 'Online Food', 'Online Drinks', 'Gift', 'Event'
* For categorical data, find the distribution of the discrete values. In particular, identify the mode.
  + Modes for all categorical data:
    - Order\_Number: 228348 - 30 occurrences
    - Order\_Type: ‘To Go’ – 11035
    - Product\_Name: ‘Shoyu’ – 3247
    - Exchange\_Void\_or\_Return: 0 – 21532
    - Category: ‘Food’ - 13964
* For quantitative data (discrete or continuous), find the range, median, mean, and standard deviation. What are the units? Do this for columns where it makes sense. If the column has social security numbers, none of the statistics mentioned is relevant.
  + Order\_Sub\_Total
    - count 22323.000000 mean 37.054772 std 29.548151 min 0.000000 25% 18.000000 50% 32.000000 75% 48.000000 max 759.000000
  + Order\_Sales\_Tax
    - count 22323.000000 mean 2.968230 std 2.430112 min 0.000000 25% 1.400000 50% 2.600000 75% 3.880000 max 62.620000
  + Order\_Total
    - count 22323.000000 mean 38.779181 std 31.725330 min 0.000000 25% 17.940000 50% 33.560000 75% 50.340000 max 821.620000
  + Total\_Sales\_Price
    - count 22323.000000 mean 10.554211 std 8.551853 min 0.000000 25% 4.500000 50% 10.000000 75% 14.500000 max 759.000000
  + Total\_Product\_Tax:
    - count 22323.000000 mean 0.884989 std 0.732687 min 0.000000 25% 0.330000 50% 0.820000 75% 1.360000 max 62.620000
  + Product\_Quantity:
    - count 22323.000000 mean 1.068808 std 0.336416 min 1.000000 25% 1.000000 50% 1.000000 75% 1.000000 max 10.000000

2. In this step, you will have to do data wrangling.

* Find all missing data.
* If it makes sense, replace missing categorical data with mode and replace quantitative data with mean.
* You may remove a row with missing data. Make a record of it with your reasoning for removing it.
* Make sure that all the data is within the expected range. For example, you should not have negative salaries.
* Make sure that all the data is in the correct format.
* Check for duplicates and remove them if it makes sense.

3. Do an initial statistical analysis of your dataset.

* Check for correlation among columns. Find the Pearson correlation coefficient among quantitative data.
* If the correlation coefficient is high, check for the best fit - linear, quadratic, or cubic.

**Goal (or Thesis)**

**Description of the Dataset**

* **Size of the dataset**
* **Column Names**
* **Classification of each column**
* **List of Discrete Values for Categorical Variables**
* **Distribution of Discrete Variables**
* **Distribution of Quantitative Variables**

**Data Wrangling**

* **List missing data**
* **Replacement of missing data (if any)**
* **Deletion of rows (if any) and reasons**
* **Check that the data is in the correct range**
* **Check that the data is in the correct format.**
* **Check for duplicates and remove them (if any)**

**Preliminary Statistical Analysis**

* **Determination of the Pearson correlation coefficient among quantitative data**
* **Determination of the best fit among highly correlated quantities - linear, quadratic, or cubic**

**Narration**

* **Present in a bulleted form your narration**