**Data Cleaning Project Documentation**

**Introduction**

This document outlines the process of cleaning and preparing the Chipotle dataset. The goal is to ensure data quality, resolve inconsistencies, and create a clean dataset for further analysis.

Environment Setup

import pandas as pd

Reading TSV File

# Reading Chipotle TSV File

df = pd.read\_csv('chipotle.tsv', sep='\t')

1. Missing Values

# Checking for Missing Values

missing\_values = df.isnull().sum()

**Handling Missing Values**

# Filling NaN values in 'item\_price' with the mean

df['item\_price'] = pd.to\_numeric(df['item\_price'], errors='coerce')

df['item\_price'] = df['item\_price'].fillna(df['item\_price'].mean())

2. Data Types

# Verifying Data Types

df['item\_price'] = df['item\_price'].str.replace('$', '').astype(float)

3. Duplicated Entries

# Identifying and Handling Duplicated Entries

duplicated\_entries = df.duplicated().sum()

df = df.drop\_duplicates()

4. Quantity and Item Price

# Examining Quantity and Item Price Columns

df['Quantity'] = pd.to\_numeric(df['Quantity'], errors='coerce')

5. Choice Description

# Analyzing the Choice Description Column

df['Primary Choice'] = df['Choice Description'].apply(lambda x: eval(x)[0] if pd.notnull(x) else None)

6. Handling Special Characters

# Removing Special Characters from 'Item Name'

df['Item Name'] = df['Item Name'].str.replace('[^a-zA-Z0-9\s]', '')

7. Order Id Integrity

# Cross-referencing the Order ID Column

unique\_order\_ids = df['Order Id'].unique()

8. Item Name Standardization

# Standardizing Item Names

df['Standardized Item Name'] = df['Item Name'].str.lower()

9. Quantity and Price Relationships

# Investigating Relationships between Quantity and Item Price

negative\_values = df[(df['Quantity'] < 0) | (df['item\_price'] < 0)]

10. Data Integrity Check

# Checking Data Integrity

integrity\_check = df.groupby(['Standardized Item Name', 'Primary Choice']).agg({

'Quantity': ['sum', 'mean'],

'item\_price': ['sum', 'mean']

}).reset\_index()

11. Converting to CSV

# Converting the Cleaned Dataset to CSV

df.to\_csv('cleaned\_chipotle.csv', index=False)

Conclusion

This data cleaning project has addressed missing values, data types, duplicated entries, and other issues in the Chipotle dataset. The resulting cleaned dataset ('cleaned\_chipotle.csv') is now ready for further analysis.