**CLEANING AND TRANSFORMING THE DATASET USING POWER QUERY**

## **Introduction**

Source File: malltransactions\_mockdataset.csv

### Description of the file

• A mocked data of well-known malls in the capital region of the Philippines

• Created from mockaroo.com

### Dimensions

* transaction\_id
* date
* mall
* city
* gender
* age
* product category
* price
* quantity
* discount
* payment

### Business Requirements

* All Ages less than 21 or 60 and above will have 20% discount
* For undeclared ages, replaced it with average age per gender and per mall.
* For undeclared gender, replace it with “Undeclared”

### Business Notes

* There are data entry mistakes on discounts and payment

## **Steps In Cleaning and Transforming the Dataset**

1. Open file in Power Query
2. Do an Exploratory Data Analysis of the file
   1. Check for error/null values on each column
   2. Check for misspelled / case inconsistency in the values
   3. Check for Outliers
3. Clean all necessary columns
   1. Correct misspelled / case inconsistency values
   2. Filter or Replace Outliers with standard values
   3. Replace or Remove error/null values
   4. Check if all columns have the right data type and format
4. Save and Close file to be used on Reports/Creating Dashboard

## Step 1: Open file in Power Query

* I use Excel to use the built-in Power Query tool.
* I clicked the Transform Data for the next Step.

A screenshot of a computer

Description automatically generated

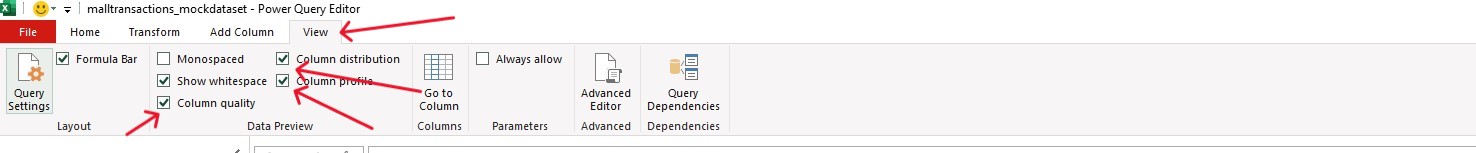
A screenshot of a computer

Description automatically generated

## Step 2: Do an Exploratory Data Analysis (EDA)

#### ***On View Ribbon, check the following*:**

* Column Profile
* Column Quality
* Column Distribution



This is used for EDA, checking all the needed things for cleaning and transforming.

### **Check all column profiles, column quality, column distribution of each column**

#### **Gender column**

* Gender column have EMPTY cells which is 22%.

A screenshot of a computer

Description automatically generated

#### **Age column**

* Age column has null values and is about 17%.

A screenshot of a computer

Description automatically generated

#### **Discount column**

* Discount column has negative values and an Outlier. Data should be in percentage format.

A screenshot of a computer

Description automatically generated

#### **Payment column**

* Payment column has misspelled / case inconsistency in its values A screenshot of a computer

  Description automatically generated

## Step 3: Clean all necessary columns

#### **Payment Column**

* For Payment Column, I trim, clean, and lower case the values of the column
* Replace ewallet to e-wallet, card-debit to debit card, coins to cash, card-credit to credit card for standardization
* Only 4 distinct values remain

A screenshot of a computer

Description automatically generated

#### **Gender Column**

* For Gender Column, I replaced blank values to undeclared values

A screenshot of a computer

Description automatically generated

#### **Discount Column**

* For Discount Column, I remove all negative values by using the absolute value feature
* Replace the Outlier 2 to .2

A screenshot of a computer

Description automatically generated

* Created a new column “discount\_new” that will change values based on the age (All Ages less than 21 or 60 and above will have 20% discount)
  + Temporarily, I replace all null values with -1 on the Age column to avoid Errors after manipulation

A screenshot of a computer

Description automatically generated

* + I use conditional column for manipulation for the discount\_new column

A screenshot of a computer

Description automatically generated

* + Replace the -1 to Null again and delete the old column.
  + Replace the format to percentage

A screenshot of a computer

Description automatically generated

#### **Age Column**

* For Age Column, I need to get the average age per gender per mall.
  + I duplicate the query, remove other columns except Age, Mall, and Gender

A screenshot of a computer

Description automatically generated

* + I use the Group BY feature to get the average age

A screenshot of a computer

Description automatically generated

* + Replace the *null* to the average age between Female and Male of WalterMart

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

* + This will be the reference table for replacing the null values.
* I’ll create a new query and create a custom function that will replace the null values based on the mall and the gender
  + On Advanced Editor, Put this code

A screenshot of a computer

Description automatically generated

* + Rename the function avgAge\_function

A screenshot of a computer

Description automatically generated

* + On the main table, add column by custom column and put this code:

A screenshot of a computer

Description automatically generated

* + Remove the old age column and change format to whole Number, rename the column to age

## Step 4: Save and Close file to be used on Reports/Creating Dashboard

* Close and Load the file. Save the file for Reports/ Creating Dashboard