Data Analytics project n°1: Technical file

Student's full Name : Yannick MPOYI

I. Data understanding

I.1. Dataset

File name	Data Analytics Cumulative Project Data
Extension (file format)	Microsoft Excel Worksheet (.xlsx)
Source	Moringa School
Dimension	(16 columns, 700 rows)
	5 categorical + 9 numeric + 2 date format columns
Missing data	22 missing data found in 14 columns

II. Data cleaning

A. Missing value

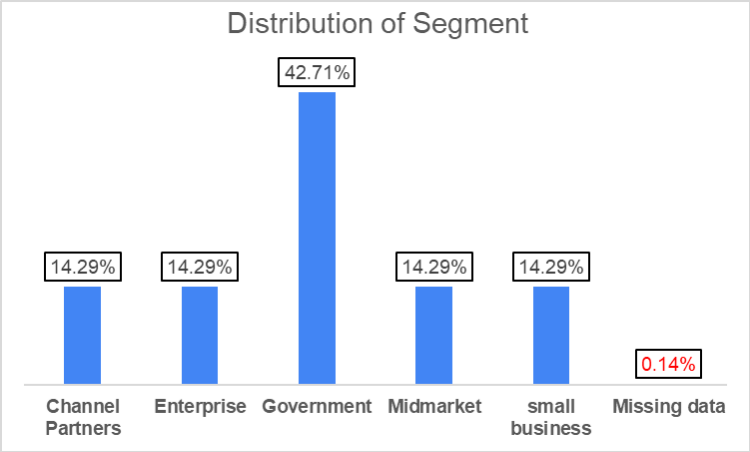
A.1. There are no raw entirely missing data in the dataset

A.2. Missing data by column

To maintain the representativeness of the population data, we will replace the missing values rather than deleting entire rows. Numeric data will be imputed using the median, while categorical data will be imputed using the mode, as the mean can be disproportionately affected by outliers.

A.2.1. Column A (segment) : 1 missing data

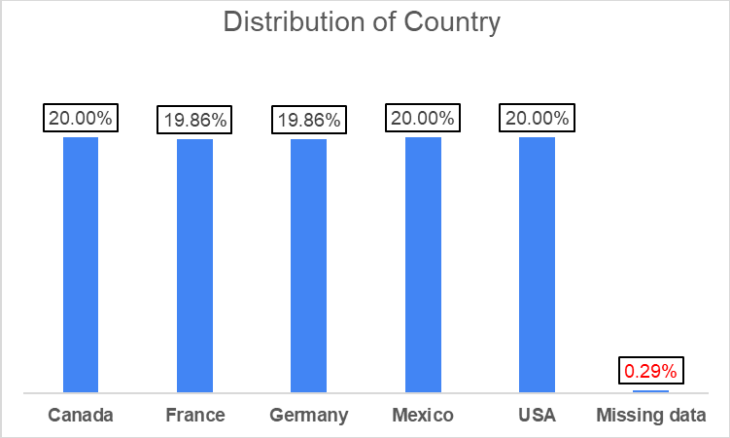
Replaced by mode with the Mode = Government



A.2.2. Column B (country) : 2 missing data

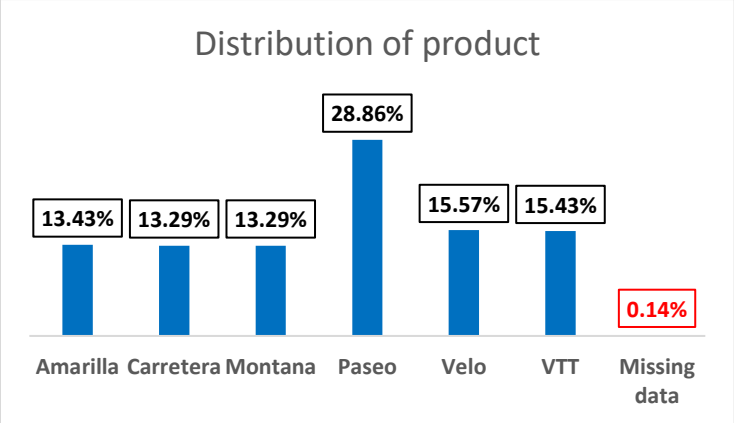
Replaced by mode with the Mode = Canada, Mexico and United States of America (changed to USA) are the modalities that appear the most in our dataset.

Therefore, we will chose randomly by this formula :
=CHOOSE(RANDBETWEEN(1,3),"Canada","Mexico","USA") + copy the same value and paste them as value (special copy).
The random imputation is supported by scientific evidence.

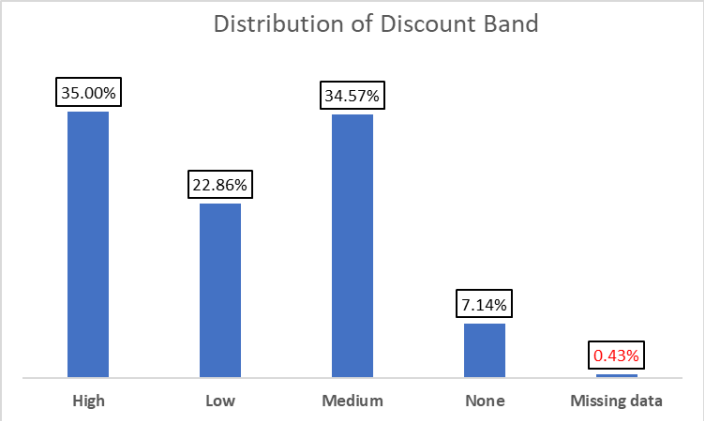


A.2.3. Column C(Product) : one missing data

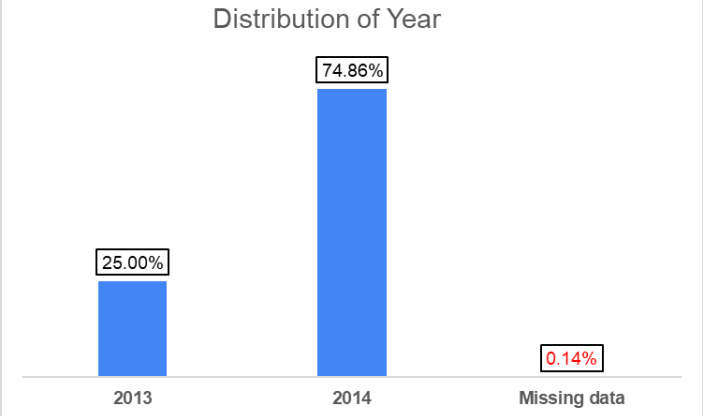
Replaced by mode with the Mode = Paseo



A.2.4. Column D(Discount Band) : 3 missing data
Replaced by mode with the Mode = High



A.2.5. Column P(Year) : 1 missing data
Replaced by mode with the Mode = 2014 (Modal year)



A.2.6. Other columns missing data

Columns & missing data	Description
E(Units Sold) : 1 missing data	replaced by mode with the Median MEDIAN(E2:E701) = 1,540 units sold Column rounded
F(Manufacturing Price) : 1 missing data	replaced by mode with the Median = =MEDIAN(F2:F700)= 10 USD
G(Sale Price) : 3 missing data	replaced by mode with the Median = =MEDIAN(G2:G700)= 20 USD
H(Gross Sales) :3 missing data	replaced by mode with the Median = =MEDIAN(H2:H700)= 38,302.50 USD
I(Discounts) : 1 missing data	replaced by mode with the Median = MEDIAN(I2:I700)= 2,585.25 USD
J(Sales) : 2 missing data	replaced by mode with the Median = MEDIAN(J2:J700)= 35,585.60USD
K(COGS) : 1 missing data	replaced by mode with the Median = MEDIAN(K2:K700)= $22,580.00 USD
L(Profit) Column L(Profit) : 1 missing data	replaced by mode with the Median = MEDIAN(L2:L700)= 9,241.80 USD

B. Checking duplicate

Let's create a new column to be identified as our primary key.
The new olumn will be a concatenation of all existing column
=CONCAT(Table1[@[segment]:[Year]]).
Then, click on Conditional formatting>new rule>Format only unique or duplicate values + choose the red color
>>No duplicated row found or highlighted

Addition information

We did create the column named ''Quarter'' based on the dates data column N (Date) ="Q"&ROUNDUP(MONTH(N2)/3,0) and slide down the formula.