esprit
Se former autrement
ÉCOLE SUPÉRIEURE PRIVÉE D'INGÉNIERIE ET DE TECHNOLOGIES

# OPTIMIZATION OF INTERNATIONAL E-COMMERCE SERVICES THROUGH THE ANALYSIS OF CUSTOMER DATA

## STATISTICS PROJECT

DATA
PULSE

# CONTENT



DATA PULSE

# OUR TEAM MEMBERS

**Neil Monastiri**

**Montassar Maiza**

**Mahdi Fathallah**

**Zeyneb Ben Arab**

**Iness Elahchaichi**


INTRODUCTION
DATA
PULSE

# OPTIMIZING E-COMMERCE SERVICES THROUGH CUSTOMER DATA ANALYSIS

DATA
PULSE

**Objective :** Enhance customer satisfaction and loyalty in international e-commerce by analyzing customer data.

## Key Components

### DataSet

Analysis of 10,999 customer interactions.

### Focus

Concentrating on customer engagement and retention metrics to drive strategic decisions

### Goals

Identify service improvement areas and explore the impact of AI on sales.

### Deliverables

Comprehensive analysis with R scripts, documentation, and a strategic presentation.of 10,999 customer interactions.

# DATA PRE-PROCESSING AND CLEANING

DATA
PULSE

# SNAPSHOT OF DATA CHARACTERISTICS

## Dimension of data

```
> dim(data)
[1] 10999    12
```

## Columns names

```
> names(data)
 [1] "ID"                 "Warehouse_block"    "Mode_of_Shipment"    "Customer_care_calls" "Customer_rating"
 [6] "Cost_of_the_Product" "Prior_purchases"    "Product_importance"  "Gender"              "Discount_offered"
[11] "Weight_in_gms"       "Reached.on.Time_Y.N"
```

## Some rows of data

| | ID | Warehouse_block | Mode_of_Shipment | Customer_care_calls | Customer_rating | Cost_of_the_Product | Prior_purchases | Product_importance | Gender | Discount_offered | Weight_in_gms | Reached.on.Time_Y.N |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | D | Flight | 4 | 2 | 177 | 3 | low | F | 44 | 1233 | 1 |
| 2 | 2 | F | Flight | 4 | 5 | 216 | 2 | low | M | 59 | 3088 | 1 |
| 3 | 3 | A | Flight | 2 | 2 | 183 | 4 | low | M | 48 | 3374 | 1 |
| 4 | 4 | B | Flight | 3 | 3 | 176 | 4 | medium | M | 10 | 1177 | 1 |
| 5 | 5 | C | Flight | 2 | 2 | 184 | 3 | medium | F | 46 | 2484 | 1 |
| 6 | 6 | F | Flight | 3 | 1 | 162 | 3 | medium | F | 12 | 1417 | 1 |
| 7 | 7 | D | Flight | 3 | 4 | 250 | 3 | low | F | 3 | 2371 | 1 |
| 8 | 8 | F | Flight | 4 | 1 | 233 | 2 | low | F | 48 | 2804 | 1 |
| 9 | 9 | A | Flight | 3 | 4 | 150 | 3 | low | F | 11 | 1861 | 1 |
| 10 | 10 | B | Flight | 3 | 2 | 164 | 3 | medium | F | 29 | 1187 | 1 |
| 11 | 11 | C | Flight | 3 | 4 | 189 | 2 | medium | M | 12 | 2888 | 1 |
| 12 | 12 | F | Flight | 4 | 5 | 232 | 3 | medium | F | 32 | 3253 | 1 |
| 13 | 13 | D | Flight | 3 | 5 | 198 | 3 | medium | F | 1 | 3667 | 1 |
| 14 | 14 | F | Flight | 4 | 4 | 275 | 3 | high | M | 29 | 2602 | 1 |
| 15 | 15 | A | Flight | 4 | 3 | 152 | 3 | low | M | 43 | 1009 | 1 |

# Data Frame Structure

```
> str(data)
'data.frame':   10999 obs. of  12 variables:
 $ ID                 : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Warehouse_block    : chr  "D" "F" "A" "B" ...
 $ Mode_of_Shipment   : chr  "Flight" "Flight" "Flight" "Flight" ...
 $ Customer_care_calls: int  4 4 2 3 2 3 3 4 3 3 ...
 $ Customer_rating    : int  2 5 2 3 2 1 4 1 4 2 ...
 $ Cost_of_the_Product: int  177 216 183 176 184 162 250 233 150 164 ...
 $ Prior_purchases    : int  3 2 4 4 3 3 3 2 3 3 ...
 $ Product_importance : chr  "low" "low" "low" "medium" ...
 $ Gender             : chr  "F" "M" "M" "M" ...
 $ Discount_offered   : int  44 59 48 10 46 12 3 48 11 29 ...
 $ Weight_in_gms      : int  1233 3088 3374 1177 2484 1417 2371 2804 1861 1187 ...
 $ Reached.on.Time_Y.N: int  1 1 1 1 1 1 1 1 1 1 ...
```

# Summary of data

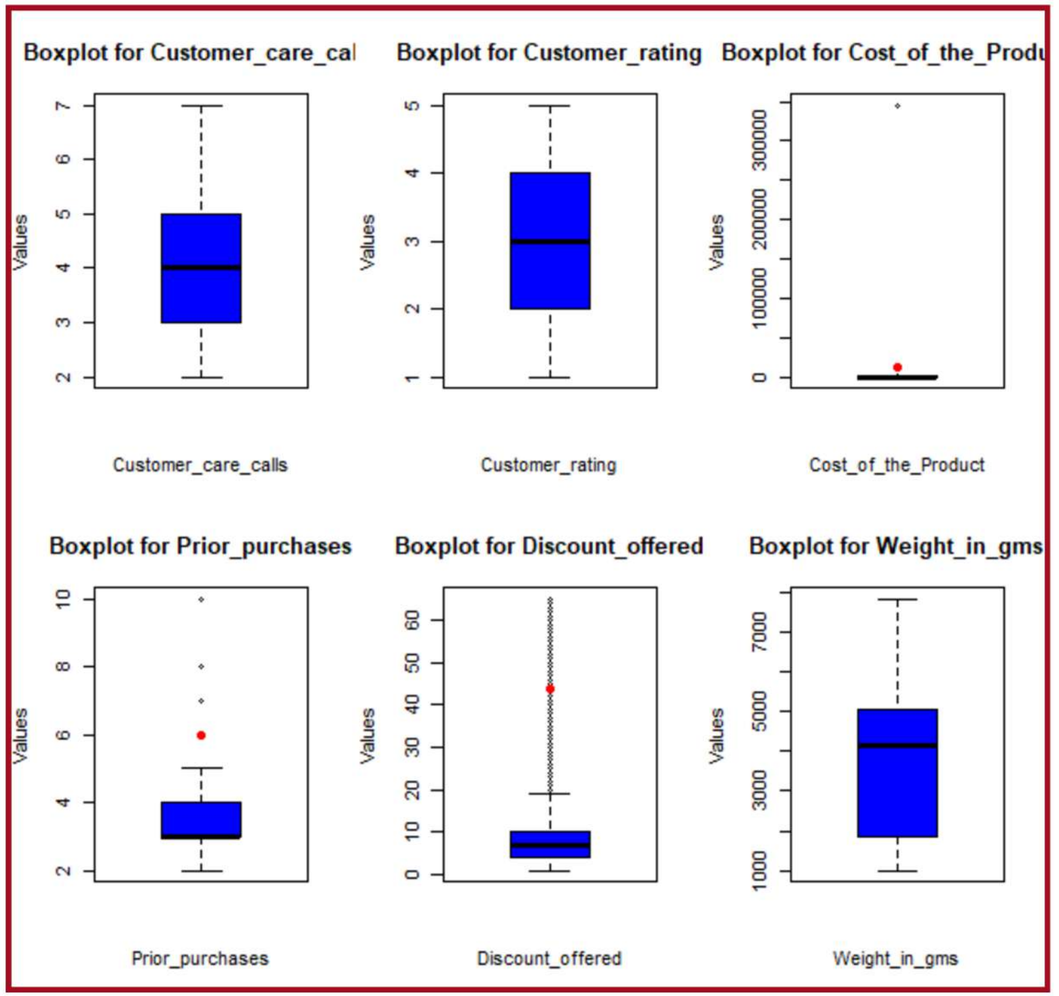
```
> summary(data)
       ID        Warehouse_block    Mode_of_Shipment   Customer_care_calls Customer_rating Cost_of_the_Product
 Min.   :    1   Length:10999       Length:10999       Min.   :2.000       Min.   :1.000   Min.   :    96.0
 1st Qu.: 2750   Class :character   Class :character   1st Qu.:3.000       1st Qu.:2.000   1st Qu.:   170.0
 Median : 5500   Mode  :character   Mode  :character   Median :4.000       Median :3.000   Median :   214.0
 Mean   : 5500                                         Mean   :4.055       Mean   :2.991   Mean   :   242.9
 3rd Qu.: 8250                                         3rd Qu.:5.000       3rd Qu.:4.000   3rd Qu.:   251.0
 Max.   :10999                                         Max.   :7.000       Max.   :5.000   Max.   :345213.0
                                                       NA's   :11                          NA's   :48
 Prior_purchases  Product_importance    Gender           Discount_offered Weight_in_gms  Reached.on.Time_Y.N
 Min.   : 2.000   Length:10999       Length:10999       Min.   : 1.00    Min.   :1001   Min.   :0.0000
 1st Qu.: 3.000   Class :character   Class :character   1st Qu.: 4.00    1st Qu.:1842   1st Qu.:0.0000
 Median : 3.000   Mode  :character   Mode  :character   Median : 7.00    Median :4157   Median :1.0000
 Mean   : 3.568                                        Mean   :13.34    Mean   :3641   Mean   :0.5967
 3rd Qu.: 4.000                                        3rd Qu.:10.00    3rd Qu.:5055   3rd Qu.:1.0000
 Max.   :10.000                                        Max.   :65.00    Max.   :7846   Max.   :1.0000
                                                       NA's   :23       NA's   :56     NA's   :1
```

DATA
PULSE

# OUTLIERS IDENTIFICATION

- The distant values observed in the boxplots of **Discount_offer** are derived from real observations, as a customer's prior purchases and associated discounts can vary.

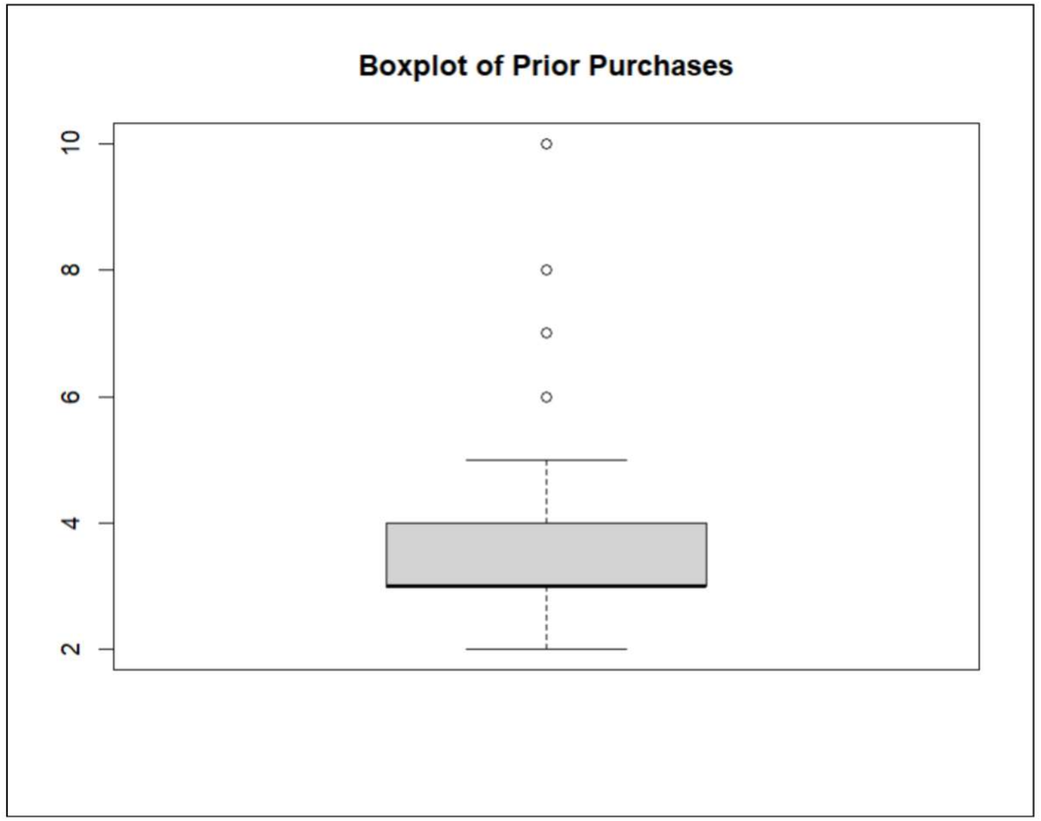
- We identify the presence of outliers in the variables **Cost_of_the_Product**, **Prior_purchases** and as the distant values exhibit abnormal and illogical figures for the price of a product. We will address these outliers for a more comprehensive analysis.




DATA PULSE

# OUTLIERS HANDLING

- Compute First and Third Quartiles (Q1 and Q3):
  - Q1 is the median of the lower half of the dataset.
  - Q3 is the median of the upper half of the dataset.
- Calculate Interquartile Range (IQR):
  - IQR is the difference between Q3 and Q1.
  - IQR=Q3−Q1
- Determine Lower and Upper Bounds:
  - Lower Bound: Q1−1.5×IQR
  - Upper Bound: Q3+1.5×IQR
- Identify Outliers:
  - Any data point that falls below the Lower Bound or above the Upper Bound is considered a potential outlier.




DATA
PULSE

# MISSING VALUES

# IDENTIFICATION

- The observations in rows 314, 320, and 353 represent outlier values of the variable **Cost_of_the_Product**

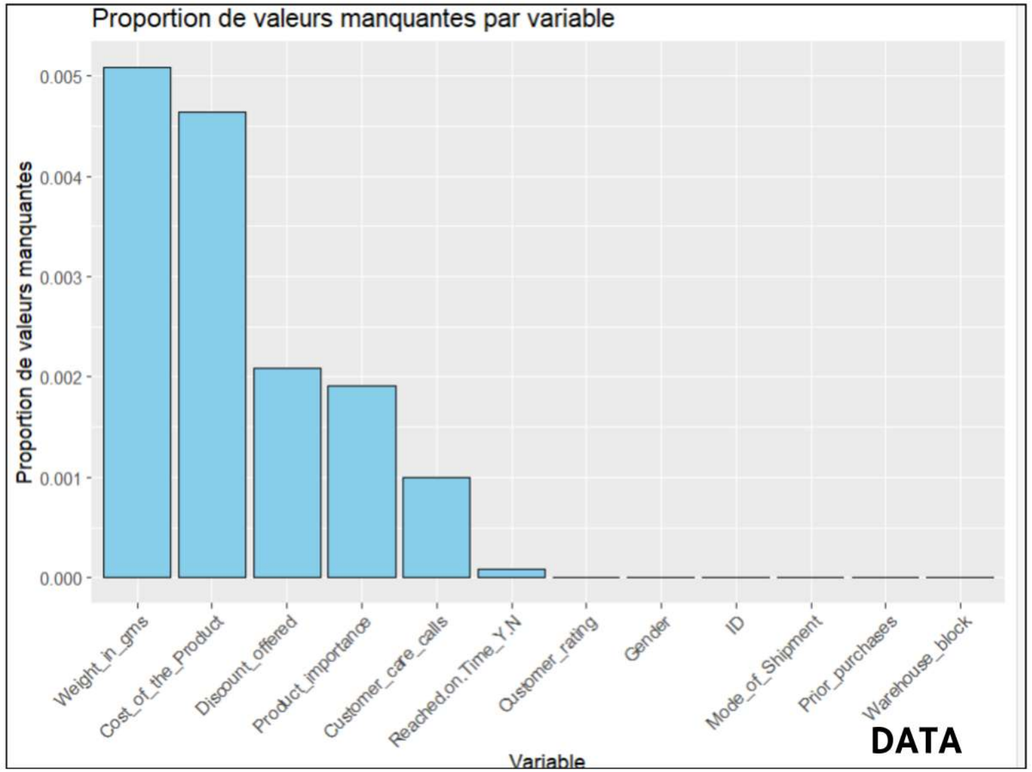
```
[1] 314 320 353
```

- Our dataset contains 163 missing values, including the 3 outlier values and the 21 empty values that we have converted into missing values.

- Number of missing values per variable

```
> colSums(is.na(data1))
                ID     Warehouse_block     Mode_of_Shipment Customer_care_calls
                 0                   0                    0                  11
   Customer_rating Cost_of_the_Product     Prior_purchases  Product_importance
                 0                  51                    0                  21
            Gender     Discount_offered      Weight_in_gms Reached.on.Time_Y.N
                 0                  23                   56                   1
```

- Total number of missing values : `[1] 163`

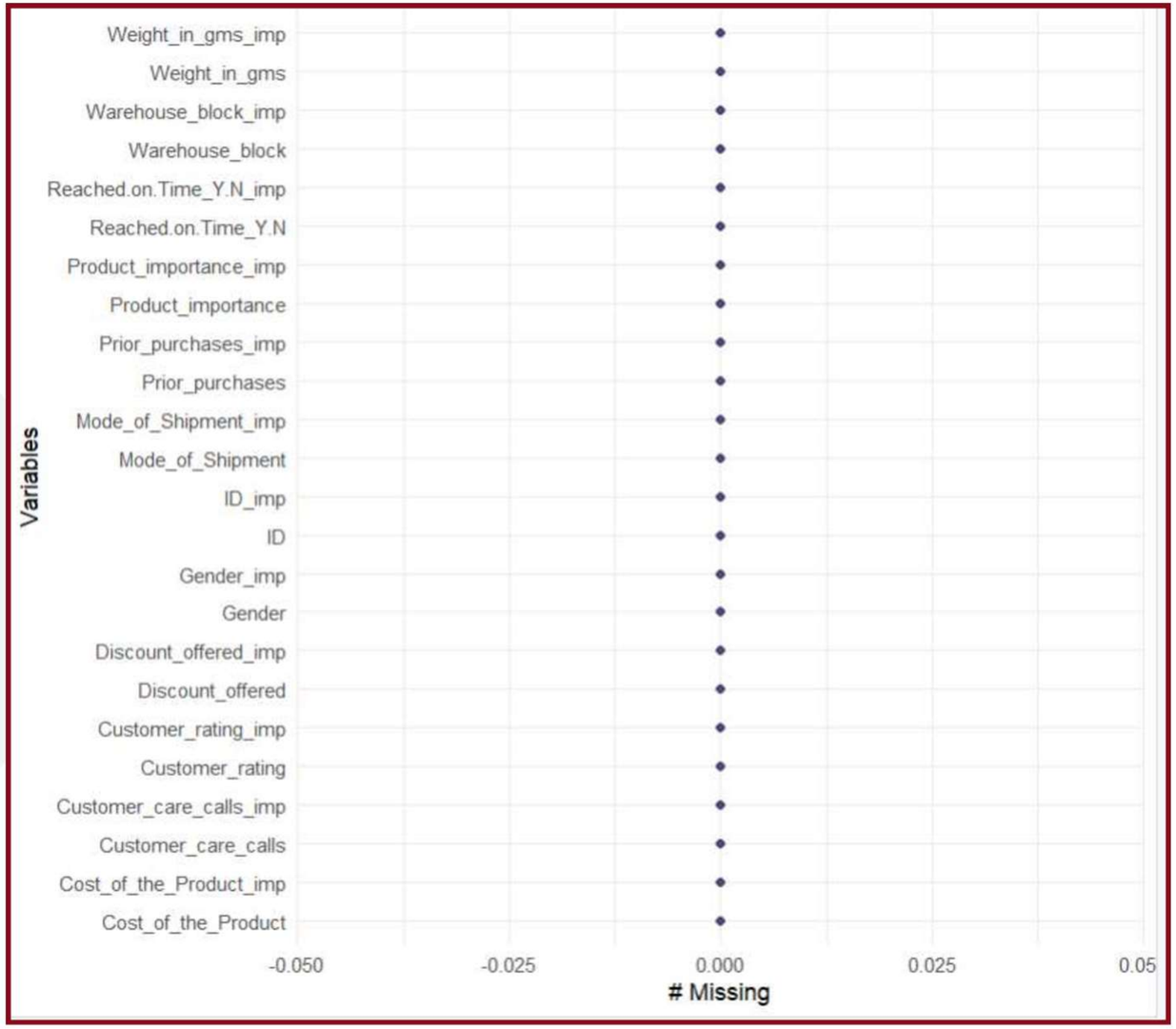



DATA PULSE

# DATA IMPUTATION

- Testing if Missing Values Follow MCAR **(Missing Completely at Random)** Pattern
  If **p-value = 0**, it indicates that the missing values are not MCAR, and deletion is not a suitable approach. Instead, we proceed with imputation using the **KNN** algorithm. This involves replacing missing values with the values of the k nearest neighbors to the missing data point.

- There are no more missing values in the dataset.




DATA PULSE

# DATA TRANSFORMATION

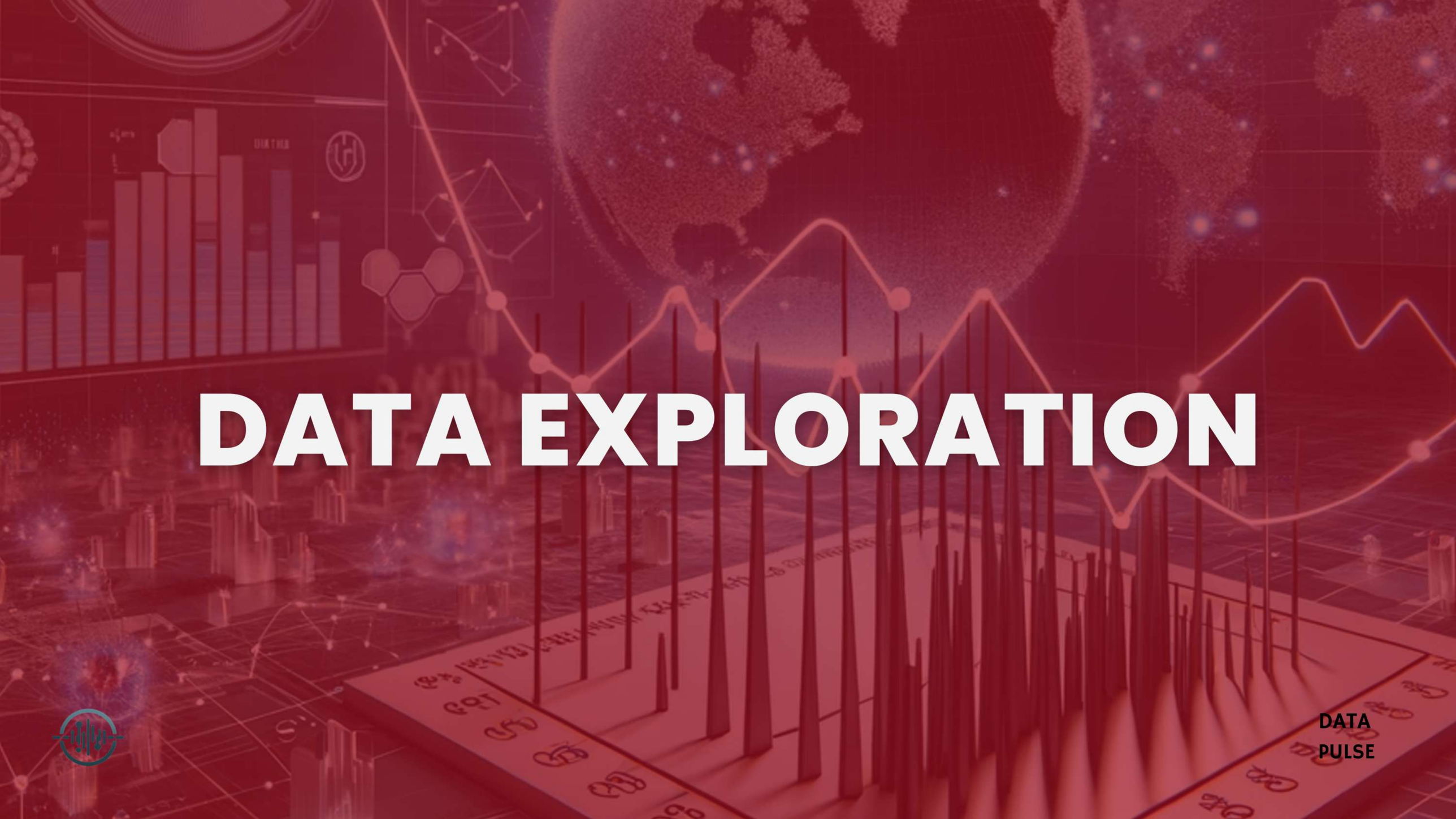
## ENCODING :

```
'data.frame':   10999 obs. of  9 variables:
 $ Warehouse_block    : num  4 5 1 2 3 5 4 5 1 2 ...
 $ Mode_of_Shipment   : num  1 1 1 1 1 1 1 1 1 1 ...
 $ Customer_care_calls: int  4 4 2 3 2 3 3 4 3 3 ...
 $ Customer_rating    : int  2 5 2 3 2 1 4 1 4 2 ...
 $ Cost_of_the_Product: int  177 216 183 176 184 162 250 233 150 164 ...
 $ Gender             : num  1 2 2 2 1 1 1 1 1 1 ...
 $ Discount_offered   : int  44 59 48 10 46 12 3 48 11 29 ...
 $ Weight_in_gms      : int  1233 3088 3374 1177 2484 1417 2371 2804 1861 1187
...
 $ Reached.on.Time_Y.N: int  1 1 1 1 1 1 1 1 1 1 ...
```

## NORMALIZATION :

Data: Anor

| | Warehouse_block | Mode_of_Shipment | Customer_care_calls | Customer_rating | Cost_of_the_Product | Gender |
|---|---|---|---|---|---|---|
| 1 | 0.4471689 | -2.0040666 | -0.04747132 | -0.700723240 | -0.69025180 | -0.9917154 |
| 2 | 1.1179832 | -2.0040666 | -0.04747132 | 1.421513085 | 0.12213187 | 1.0082622 |
| 3 | -1.5652740 | -2.0040666 | -1.79960901 | -0.700723240 | -0.56526970 | 1.0082622 |
| 4 | -0.8944597 | -2.0040666 | -0.92354017 | 0.006688868 | -0.71108215 | 1.0082622 |
| 5 | -0.2236454 | -2.0040666 | -1.79960901 | -0.700723240 | -0.54443935 | -0.9917154 |
| 6 | 1.1179832 | -2.0040666 | -0.92354017 | -1.408135349 | -1.00270706 | -0.9917154 |
| 7 | 0.4471689 | -2.0040666 | -0.92354017 | 0.714100977 | 0.83036380 | -0.9917154 |
| 8 | 1.1179832 | -2.0040666 | -0.04747132 | -1.408135349 | 0.47624784 | -0.9917154 |
| 9 | -1.5652740 | -2.0040666 | -0.92354017 | 0.714100977 | -1.25267127 | -0.9917154 |
| 10 | -0.8944597 | -2.0040666 | -0.92354017 | -0.700723240 | -0.96104636 | -0.9917154 |




DATA
PULSE

# NORMALITY TEST

- This presentation will use the Shapiro-Wilk test to assess the normality of distributions in three dataset subsets. We'll briefly introduce the test's significance, discuss why we divided the dataset, and highlight unique subset characteristics. The core will focus on the test methodology, providing detailed results analysis with interpretations for each subset. Overall, the goal is to offer insights into dataset distribution patterns and enhance understanding of its statistical properties.

```
Shapiro-Wilk test for Warehouse_block : p-value = 8.486695e-52
Shapiro-Wilk test for Mode_of_Shipment : p-value = 2.086795e-68
Shapiro-Wilk test for Customer_care_calls : p-value = 5.957023e-45
Shapiro-Wilk test for Customer_rating : p-value = 5.207002e-47
Shapiro-Wilk test for Cost_of_the_Product : p-value = 5.33167e-27
Shapiro-Wilk test for Gender : p-value = 2.159003e-68
Shapiro-Wilk test for Discount_offered : p-value = 1.091122e-43
Shapiro-Wilk test for Weight_in_gms : p-value = 4.889119e-37
Shapiro-Wilk test for Reached.on.Time_Y.N : p-value = 2.376541e-79
```

```
Shapiro-Wilk test for Warehouse_block : p-value = 8.440301e-52
Shapiro-Wilk test for Mode_of_Shipment : p-value = 1.360966e-68
Shapiro-Wilk test for Customer_care_calls : p-value = 1.951173e-41
Shapiro-Wilk test for Customer_rating : p-value = 5.692495e-47
Shapiro-Wilk test for Cost_of_the_Product : p-value = 2.939574e-28
Shapiro-Wilk test for Gender : p-value = 2.100796e-68
Shapiro-Wilk test for Discount_offered : p-value = 3.303085e-38
Shapiro-Wilk test for Weight_in_gms : p-value = 1.646199e-55
Shapiro-Wilk test for Reached.on.Time_Y.N : p-value = 1.228065e-68
```

```
+ }
Shapiro-Wilk test for Warehouse_block : p-value = 1.364151e-46
Shapiro-Wilk test for Mode_of_Shipment : p-value = 3.190352e-62
Shapiro-Wilk test for Customer_care_calls : p-value = 2.153764e-36
Shapiro-Wilk test for Customer_rating : p-value = 1.603903e-42
Shapiro-Wilk test for Cost_of_the_Product : p-value = 4.714173e-26
Shapiro-Wilk test for Gender : p-value = 3.323549e-62
Shapiro-Wilk test for Discount_offered : p-value = 2.361516e-35
Shapiro-Wilk test for Weight_in_gms : p-value = 1.629186e-48
Shapiro-Wilk test for Reached.on.Time_Y.N : p-value = 1.211783e-62
>
```

DATA PULSE

# UNIVARIATE ANALYSIS

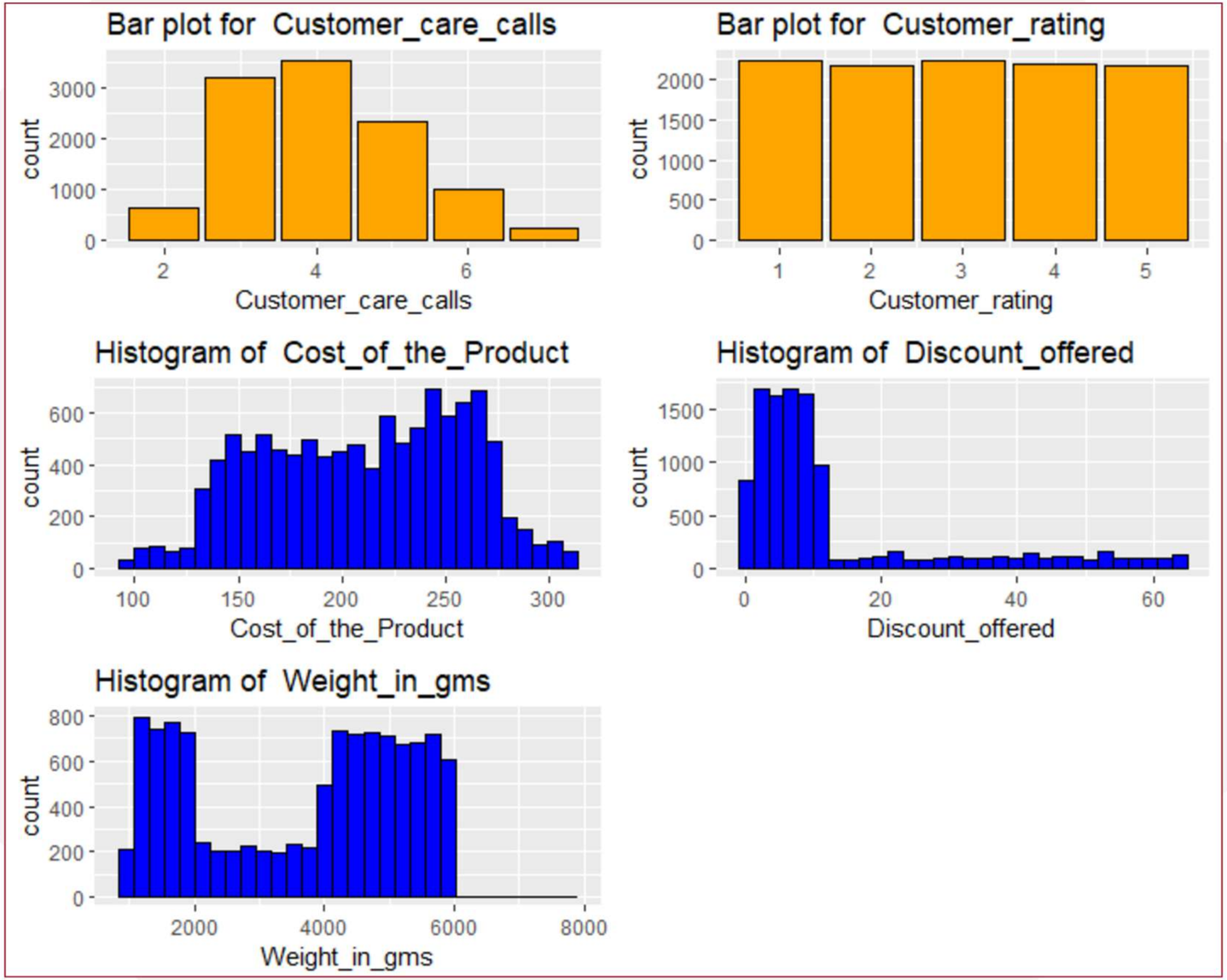
## QUANTITATIVE DATA

This slide presents a univariate statistical analysis of customer data variables. It includes measures of central tendency such as mean and median, measures of spread like standard deviation. Such analysis is crucial for understanding the individual characteristics of each variable, which can inform subsequent analytics steps and decision-making processes.

```
[1] "descriptive statistics for  Customer_care_calls :"
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.000   3.000   4.000   4.054   5.000   7.000
[1] "descriptive statistics for  Customer_rating :"
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   2.000   3.000   2.991   4.000   5.000
[1] "descriptive statistics for  Cost_of_the_Product :"
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   96.0   170.0   214.0   210.1   251.0   310.0
[1] "descriptive statistics for  Discount_offered :"
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00    4.00    7.00   13.38   10.00   65.00
[1] "descriptive statistics for  Weight_in_gms :"
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1001    1840    4149    3633    5048    7846
```

DATA PULSE

# VISUALISATION



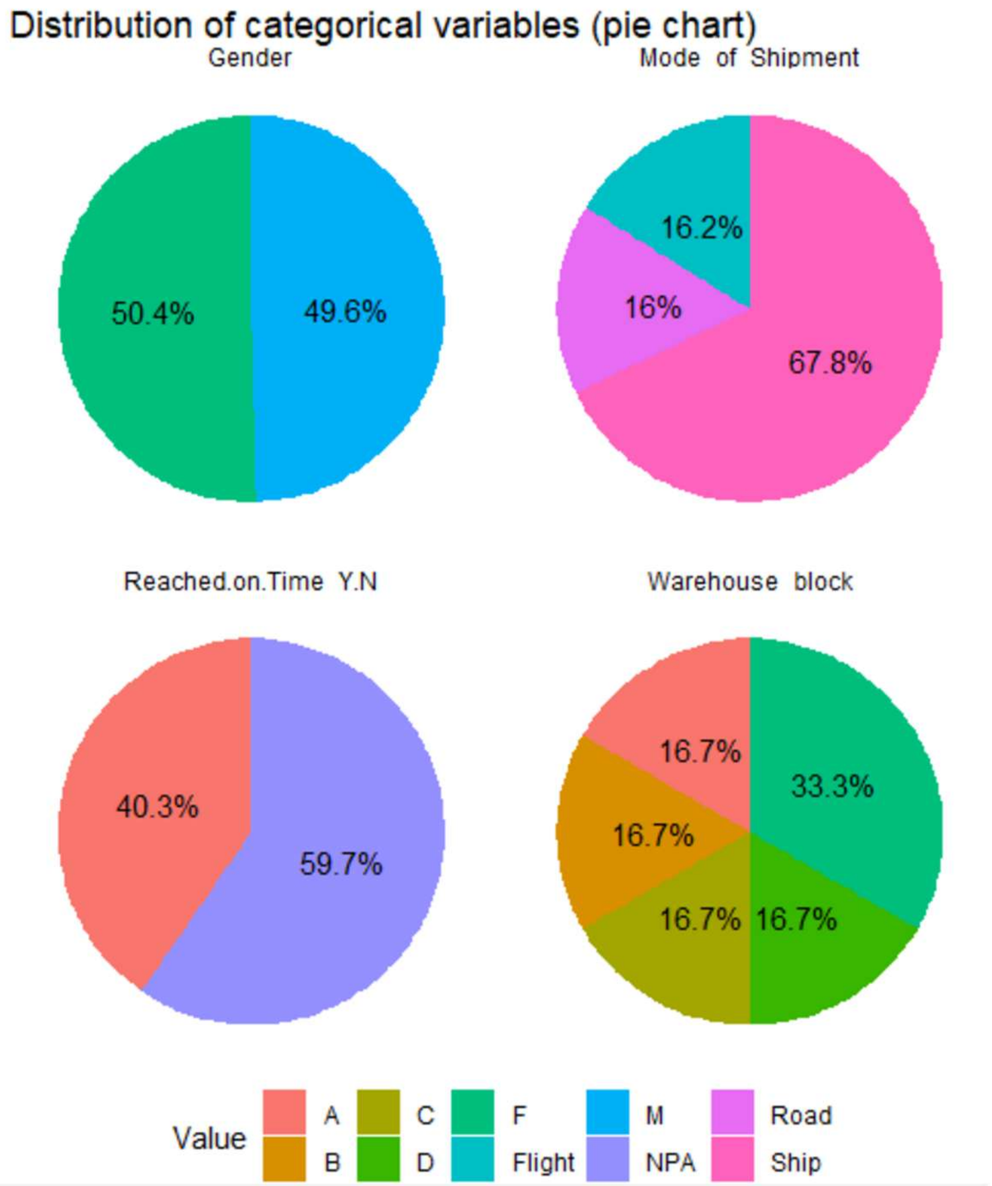

DATA
PULSE

# DISTRIBUTIONS OF KEY METRICS

```
> summary(impdata)
 Warehouse_block    Mode_of_Shipment   Customer_care_calls Customer_rating
 Length:10999       Length:10999       Min.   :2.000       Min.   :1.000
 Class :character   Class :character   1st Qu.:3.000       1st Qu.:2.000
 Mode  :character   Mode  :character   Median :4.000       Median :3.000
                                       Mean   :4.054       Mean   :2.991
                                       3rd Qu.:5.000       3rd Qu.:4.000
                                       Max.   :7.000       Max.   :5.000
 Cost_of_the_Product    Gender            Discount_offered Weight_in_gms  Reached.on.Time_Y.N
 Min.   : 96.0       Length:10999       Min.   : 1.00    Min.   :1001   Length:10999
 1st Qu.:170.0       Class :character   1st Qu.: 4.00    1st Qu.:1840   Class :character
 Median :214.0       Mode  :character   Median : 7.00    Median :4149   Mode  :character
 Mean   :210.1                          Mean   :13.38    Mean   :3633
 3rd Qu.:251.0                          3rd Qu.:10.00    3rd Qu.:5048
 Max.   :310.0                          Max.   :65.00    Max.   :7846
 ifelse(Reached.on.Time_Y.N == 1, "NPA", "A") Reached.on.Time_Y.N == 0
 Length:10999                                 Mode:logical
 Class :character                             TRUE:10999
 Mode  :character
```

DATA
PULSE

# DISTRIBUTIONS OF KEY METRICS




DATA PULSE

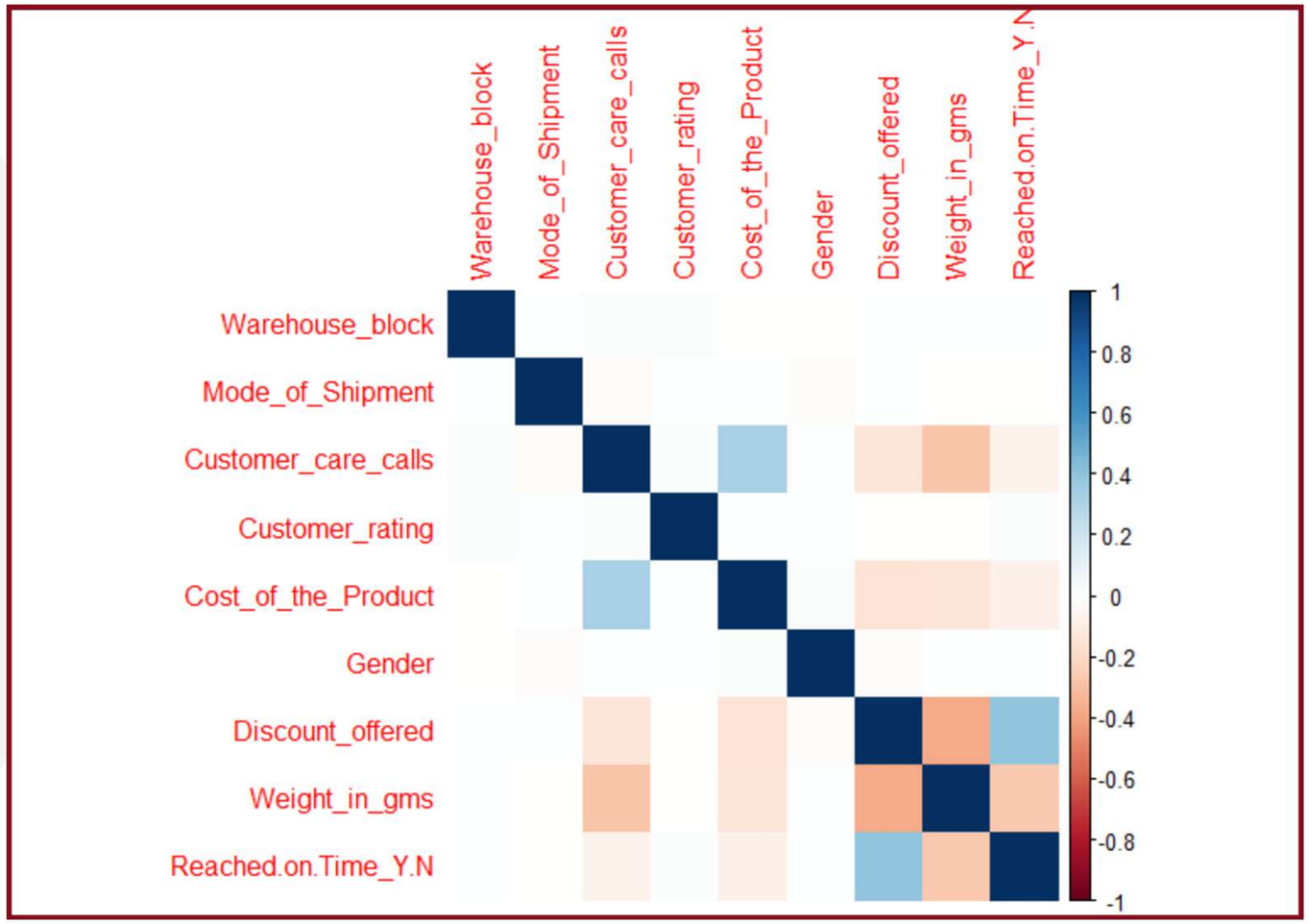
# FREQUENCY OF QUALITATIVE DATA

```
[1] "Frequencies for Mode_of_Shipment :"

Flight   Road   Ship
  1777   1760   7462
[1] "Frequencies for Warehouse_block :"

   A    B    C    D    F
1833 1833 1833 1834 3666
[1] "Frequencies for Gender :"

   F    M
5545 5454
[1] "Frequencies for Reached.on.Time_Y.N :"

    A
10999
```

DATA
PULSE

# BIVARIATE ANALYSIS

- Visualisation : the relationships between pairs of variables using a correlation matrix. Darker colors represent stronger correlations, whether positive (blue) or negative (red). This analysis helps identify which factors are related and can potentially influence each other, providing insights that inform more complex models and decision-making processes



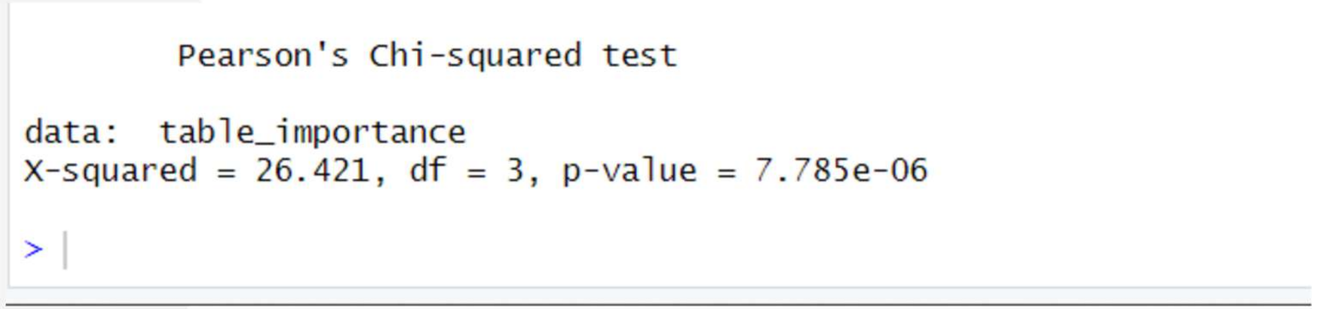

DATA PULSE

# BIVARIATE ANALYSIS

**‘Product Importance’ vs ‘On-Time Delivery’:**

Our Chi-squared test results suggest that there is a significant association between the importance of a product and whether it’s delivered on time or late:

- p-value = 7.785e-06: The p-value is much less than 0.05, indicating strong evidence against the null hypothesis, so we reject it. The two variables are significantly correlated.




**Example result of Chi-squared test**

# BIVARIATE ANALYSIS

**‘Product Importance’ vs ‘On-Time Delivery’:**

Our Chi-squared test results suggest that there is a significant association between the importance of a product and whether it’s delivered on time or late:

- p-value = 7.785e-06: The p-value is much less than 0.05, indicating strong evidence against the null hypothesis, so we reject it. The two variables are significantly correlated.

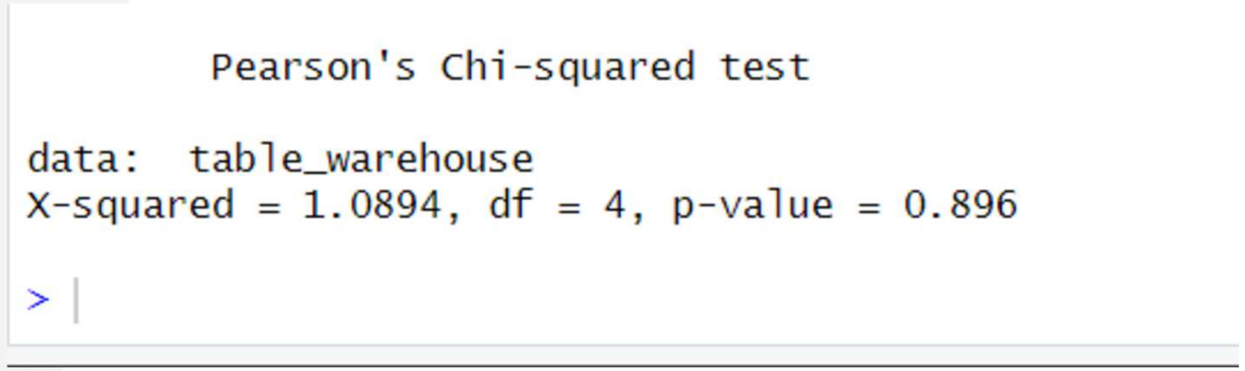

DATA PULSE

# BIVARIATE ANALYSIS

**‘Warehouse Block’ vs ‘On-Time Delivery’:**

Based on our Chi-squared test, there seems to be no significant association between the ‘Warehouse Block’ and ‘On-Time Delivery’.

- p-value = 0.896: The p-value is greater than 0.05. We fail to reject the null hypothesis that there is no association.




**Example result of Chi-squared test**

# BIVARIATE ANALYSIS

**‘Warehouse Block’ vs ‘On-Time Delivery’:**

Based on our Chi-squared test, there seems to be no significant association between the ‘Warehouse Block’ and ‘On-Time Delivery’.

- p-value = 0.896: The p-value is greater than 0.05. We fail to reject the null hypothesis that there is no association.


DATA PULSE

# BIVARIATE ANALYSIS

**‘Customer_rating’ vs ‘Prior_purchases’:**

based on this Pearson’s correlation test, there seems to be a very weak, non-significant positive correlation between ‘Customer_rating’ and ‘Prior_purchases’:

- p-value = 0.1669: The p-value is greater than 0.05. We fail to reject the null hypothesis that there is no correlation.

```
        Pearson's product-moment correlation

data:  data3$Customer_rating and data3$Prior_purchases
t = 1.3822, df = 10997, p-value = 0.1669
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.005510714  0.031860287
sample estimates:
       cor
0.01317939
```

**Example result of Pearson correlation test**

# NON-PARAMETRIC TEST

DATA
PULSE

# BIVARIATE ANALYSIS

**Weight_in_gms vs reached_On_Time:**

Our Wilcoxon test results suggest that there is a significant difference in the weights of items that reached on time and those that did not. Heavier items are more likely to not reach on time:

- p-value < 2.2e-16: The p-value is less than 0.05.

```
> print(result_wilcoxon_weight)

        Wilcoxon rank sum test with continuity correction

data:  on_time and not_on_time
W = 19115459, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0

>
```

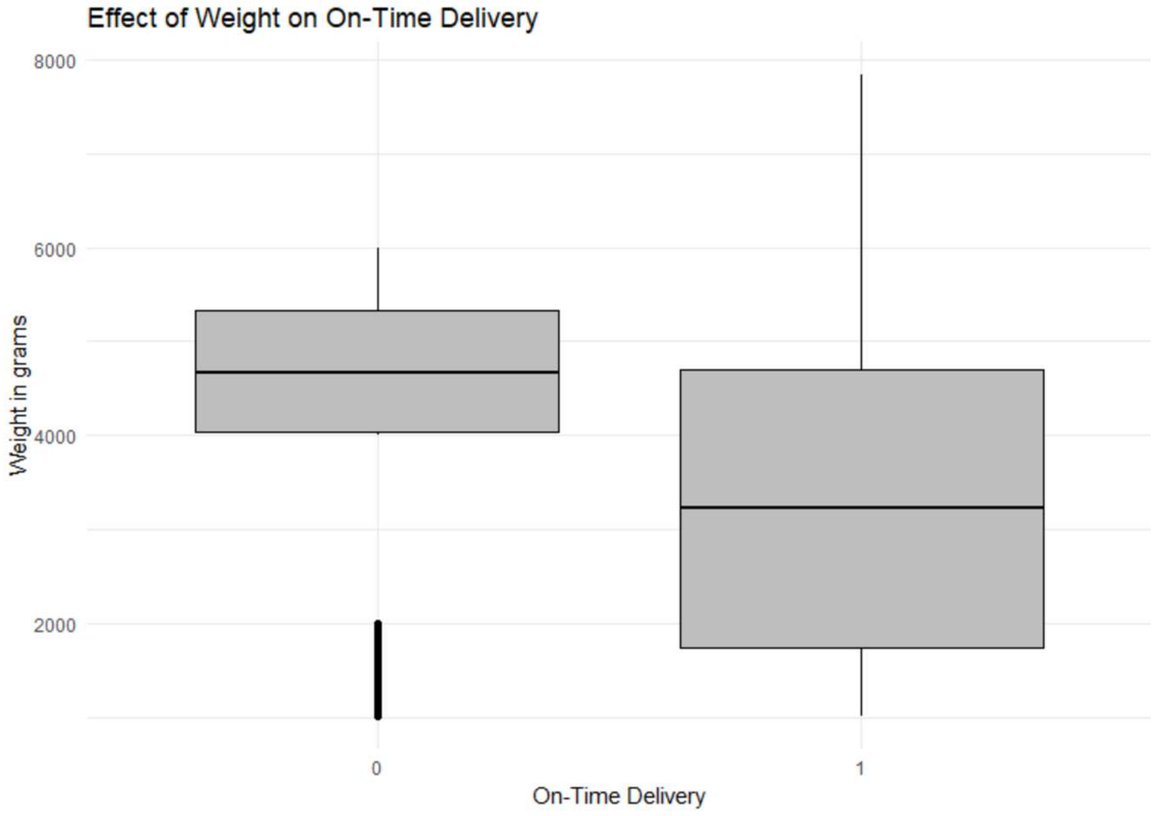
**Example result of Wilc test**

# BIVARIATE ANALYSIS

**Weight_in_gms vs reached_On_Time:**

Our Wilcoxon test results suggest that there is a significant difference in the weights of items that reached on time and those that did not. Heavier items are more likely to not reach on time:

- p-value < 2.2e-16: The p-value is less than 0.05.

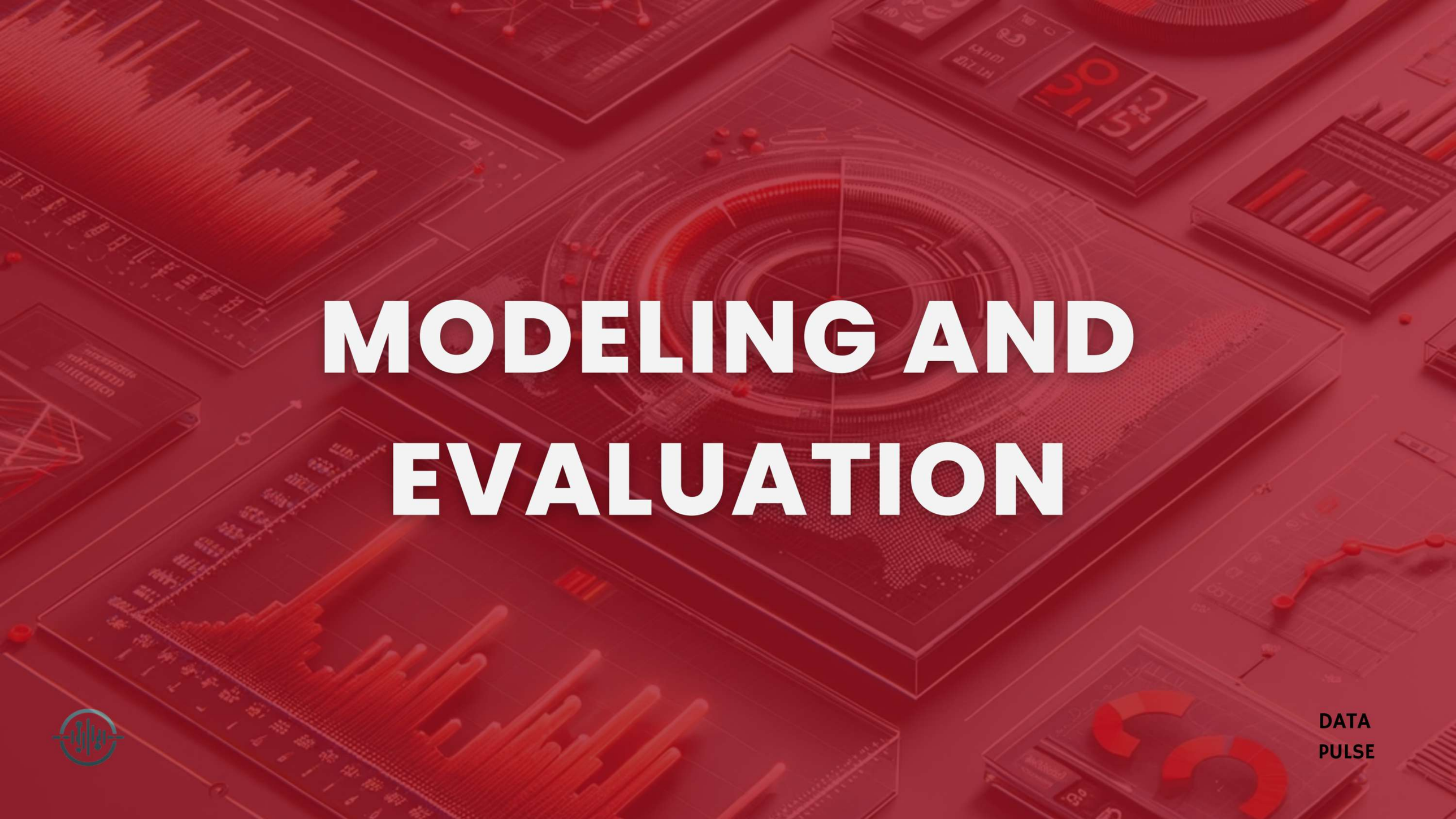



DATA PULSE

# BIVARIATE ANALYSIS

- **Discout_Offered vs reached_On_Time:**


DATA PULSE




DATA
PULSE

# THE DATASET THAT WE WILL MODEL AND EVALUATE

RStudio Source Editor

df

Filter

| | Warehouse_block | Mode_of_Shipment | Customer_care_calls | Customer_rating | Cost_of_the_Product | Gender | Discount_offered | Weight_in_gms | Reached.on.Time_Y.N | Product_importance | Prior_purchases |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.4471689 | -2.0040666 | -0.04747132 | -0.700723240 | -0.69025180 | -0.9917154 | 1.89140728 | -1.46860103 | 0.8221003 | low | 3 |
| 2 | 1.1179832 | -2.0040666 | -0.04747132 | 1.421513085 | 0.12213187 | 1.0082622 | 2.81775413 | -0.33368149 | 0.8221003 | low | 2 |
| 3 | -1.5652740 | -2.0040666 | -1.79960901 | -0.700723240 | -0.56526970 | 1.0082622 | 2.13843311 | -0.15870199 | 0.8221003 | low | 4 |
| 4 | -0.8944597 | -2.0040666 | -0.92354017 | 0.006688868 | -0.71108215 | 1.0082622 | -0.20831224 | -1.50286275 | 0.8221003 | medium | 4 |
| 5 | -0.2236454 | -2.0040666 | -1.79960901 | -0.700723240 | -0.54443935 | -0.9917154 | 2.01492020 | -0.70321864 | 0.8221003 | medium | 3 |
| 6 | 1.1179832 | -2.0040666 | -0.92354017 | -1.408135349 | -1.00270706 | -0.9917154 | -0.08479932 | -1.35602680 | 0.8221003 | medium | 3 |
| 7 | 0.4471689 | -2.0040666 | -0.92354017 | 0.714100977 | 0.83036380 | -0.9917154 | -0.64060743 | -0.77235390 | 0.8221003 | low | 3 |
| 8 | 1.1179832 | -2.0040666 | -0.04747132 | -1.408135349 | 0.47624784 | -0.9917154 | 2.13843311 | -0.50743737 | 0.8221003 | low | 2 |
| 9 | -1.5652740 | -2.0040666 | -0.92354017 | 0.714100977 | -1.25267127 | -0.9917154 | -0.14655578 | -1.08438029 | 0.8221003 | low | 3 |
| 10 | -0.8944597 | -2.0040666 | -0.92354017 | -0.700723240 | -0.96104636 | -0.9917154 | 0.96506044 | -1.49674459 | 0.8221003 | medium | 3 |
| 11 | -0.2236454 | -2.0040666 | -0.92354017 | 0.714100977 | -0.44028759 | 1.0082622 | -0.08479932 | -0.45604479 | 0.8221003 | medium | 2 |
| 12 | 1.1179832 | -2.0040666 | -0.04747132 | 1.421513085 | 0.45541749 | -0.9917154 | 1.15032981 | -0.23273178 | 0.8221003 | medium | 3 |
| 13 | 0.4471689 | -2.0040666 | -0.92354017 | 1.421513085 | -0.25281444 | -0.9917154 | -0.76412035 | 0.02056024 | 0.8221003 | medium | 3 |
| 14 | 1.1179832 | -2.0040666 | -0.04747132 | 0.714100977 | 1.35112257 | 1.0082622 | 0.96506044 | -0.63102430 | 0.8221003 | high | 3 |
| 15 | -1.5652740 | -2.0040666 | -0.04747132 | 0.006688868 | -1.21101057 | 1.0082622 | 1.82965083 | -1.60564792 | 0.8221003 | low | 3 |
| 16 | -0.8944597 | -2.0040666 | -0.04747132 | 0.006688868 | 0.35126573 | -0.9917154 | 1.95316374 | -0.56678357 | 0.8221003 | low | 3 |
| 17 | -0.2236454 | -2.0040666 | -0.92354017 | 0.714100977 | -1.39848373 | -0.9917154 | -0.45533806 | -1.49246187 | 0.8221003 | medium | 2 |
| 18 | 1.1179832 | 0.6383127 | 0.82859753 | 1.421513085 | 0.35126573 | 1.0082622 | 1.39735563 | 0.19492793 | 0.8221003 | medium | 3 |
| 19 | 0.4471689 | 0.6383127 | 0.82859753 | 1.421513085 | 0.60122994 | 1.0082622 | 0.28573942 | -0.69648866 | 0.8221003 | high | 3 |

Showing 1 to 20 of 10,999 entries, 11 total columns

DATA PULSE

# LINEAR REAGRESSION

- We applied linear regression to predict prior purchases using predictors like warehouse block and mode of shipment, without any transformations, to establish a straightforward linear relationship between these factors and the purchasing behavior.

```
Call:
lm(formula = Prior_purchases ~ Warehouse_block + Mode_of_Shipment +
    Customer_care_calls + Customer_rating + Cost_of_the_Product +
    Gender + Discount_offered + Weight_in_gms + Reached.on.Time_Y.N,
    data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-2.4290 -0.8304 -0.3486  0.6154  7.1540

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         3.5675971  0.0140225 254.419  < 2e-16 ***
Warehouse_block    -0.0066063  0.0140284  -0.471    0.638
Mode_of_Shipment    0.0008803  0.0140290   0.063    0.950
Customer_care_calls 0.1417601  0.0157114   9.023  < 2e-16 ***
Customer_rating     0.0175953  0.0140270   1.254    0.210
Cost_of_the_Product 0.0709812  0.0149930   4.734 2.23e-06 ***
Gender             -0.0166775  0.0140283  -1.189    0.235
Discount_offered   -0.1791737  0.0166267 -10.776  < 2e-16 ***
Weight_in_gms      -0.2962114  0.0164979 -17.954  < 2e-16 ***
Reached.on.Time_Y.N -0.0782931  0.0154877  -5.055 4.37e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.471 on 10989 degrees of freedom
Multiple R-squared:  0.06819,   Adjusted R-squared:  0.06742
F-statistic: 89.35 on 9 and 10989 DF,  p-value: < 2.2e-16

>
```

DATA PULSE

# ELIMINATION OF UNSGNIFICANT VARIABLES

- We utilized the Akaike Information Criterion (AIC) to assess model quality, specifically focusing on non-significant variables. Notably, an increase in AIC values indicated a better fit, prompting attention to variables that contributed meaningfully to the model. This approach enhances our model selection process by emphasizing the importance of relevant variables, leading to more refined and accurate analyses.

```
Call:
lm(formula = Prior_purchases ~ -Warehouse_block - Mode_of_Shipment +
    Customer_care_calls - Customer_rating - Cost_of_the_Product -
    Gender + Discount_offered + Weight_in_gms + Reached.on.Time_Y.N,
    data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-2.4886 -0.8498 -0.3425  0.6219  7.2388

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         3.56760    0.01404 254.182  < 2e-16 ***
Customer_care_calls 0.16101    0.01518  10.604  < 2e-16 ***
Discount_offered   -0.18881    0.01651 -11.434  < 2e-16 ***
Weight_in_gms      -0.30464    0.01642 -18.556  < 2e-16 ***
Reached.on.Time_Y.N -0.08062    0.01549  -5.204 1.98e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.472 on 10994 degrees of freedom
Multiple R-squared:  0.06602,   Adjusted R-squared:  0.06568
F-statistic: 194.3 on 4 and 10994 DF,  p-value: < 2.2e-16

>
```

```
> AIC(model5)
[1] 39725.71
> AIC(reg_multi)
[1] 39710.19
```

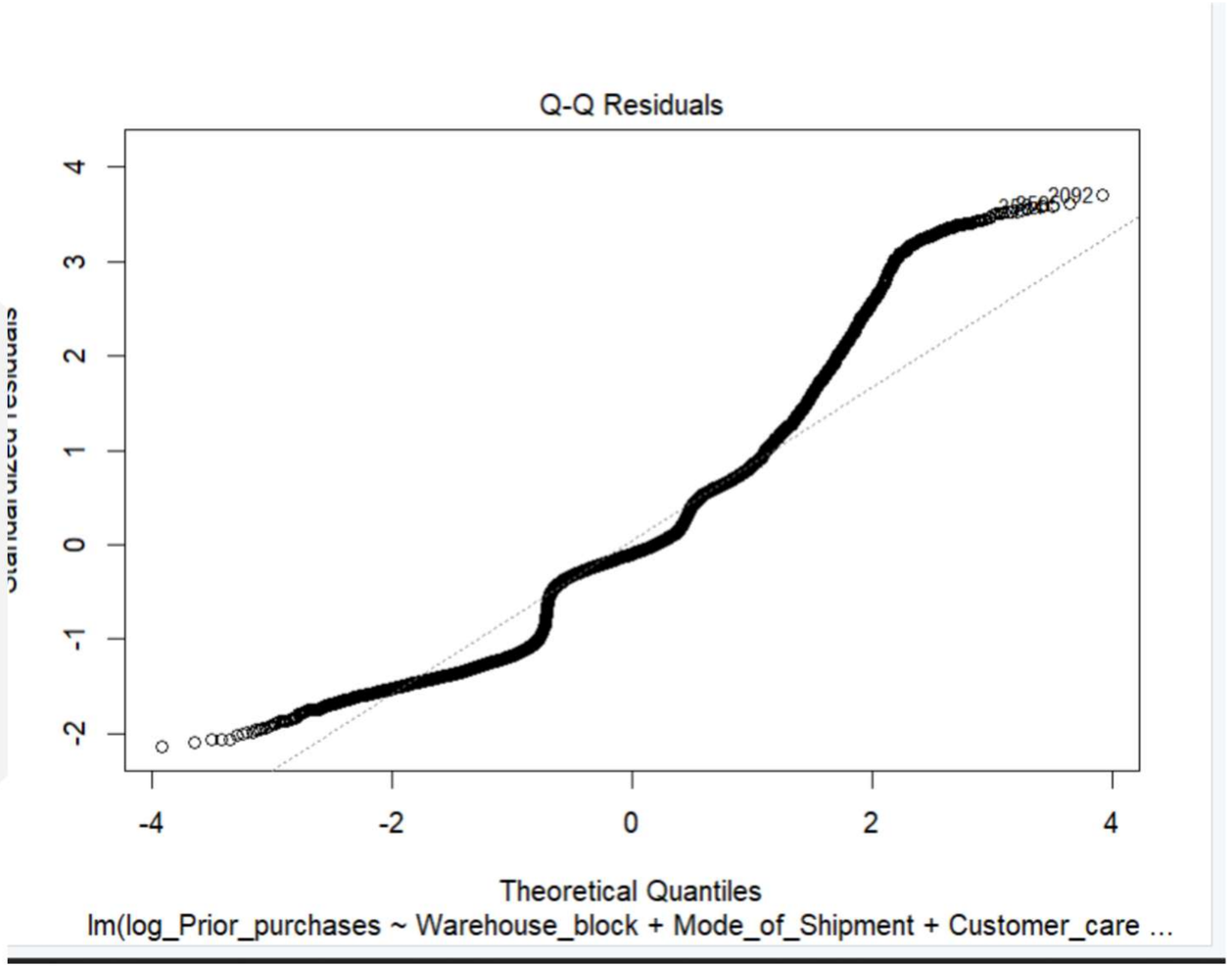
# TARGET CHANGMENT

- To improve our regression model, we transformed the target variable by taking its logarithm. This step often helps in stabilizing variance, normalizing the distribution, and making the relationships more linear, thereby enhancing the model's performance.

```
Call:
lm(formula = log_Prior_purchases ~ Warehouse_block - Mode_of_Shipment +
    Customer_care_calls - Customer_rating - Cost_of_the_Product -
    Gender + Discount_offered + Weight_in_gms + Reached.on.Time_Y.N,
    data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-0.78519 -0.19466 -0.03182  0.21417  1.35455

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         1.195669   0.003431 348.460  < 2e-16 ***
Warehouse_block    -0.002418   0.003432  -0.705    0.481
Customer_care_calls 0.053649   0.003713  14.448  < 2e-16 ***
Discount_offered   -0.056329   0.004037 -13.952  < 2e-16 ***
Weight_in_gms      -0.092391   0.004014 -23.015  < 2e-16 ***
Reached.on.Time_Y.N -0.023187   0.003786  -6.124 9.46e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3599 on 10993 degrees of freedom
Multiple R-squared:  0.1025,    Adjusted R-squared:  0.1021
F-statistic: 251.1 on 5 and 10993 DF,  p-value: < 2.2e-16
```

# RESIDUALS

- This slide, titled "RESIDUALS," presents a Q-Q plot to assess the normality of residuals from a linear model. The plot compares standardized residuals to a normal distribution, where data points deviating from the diagonal line indicate non-normality.



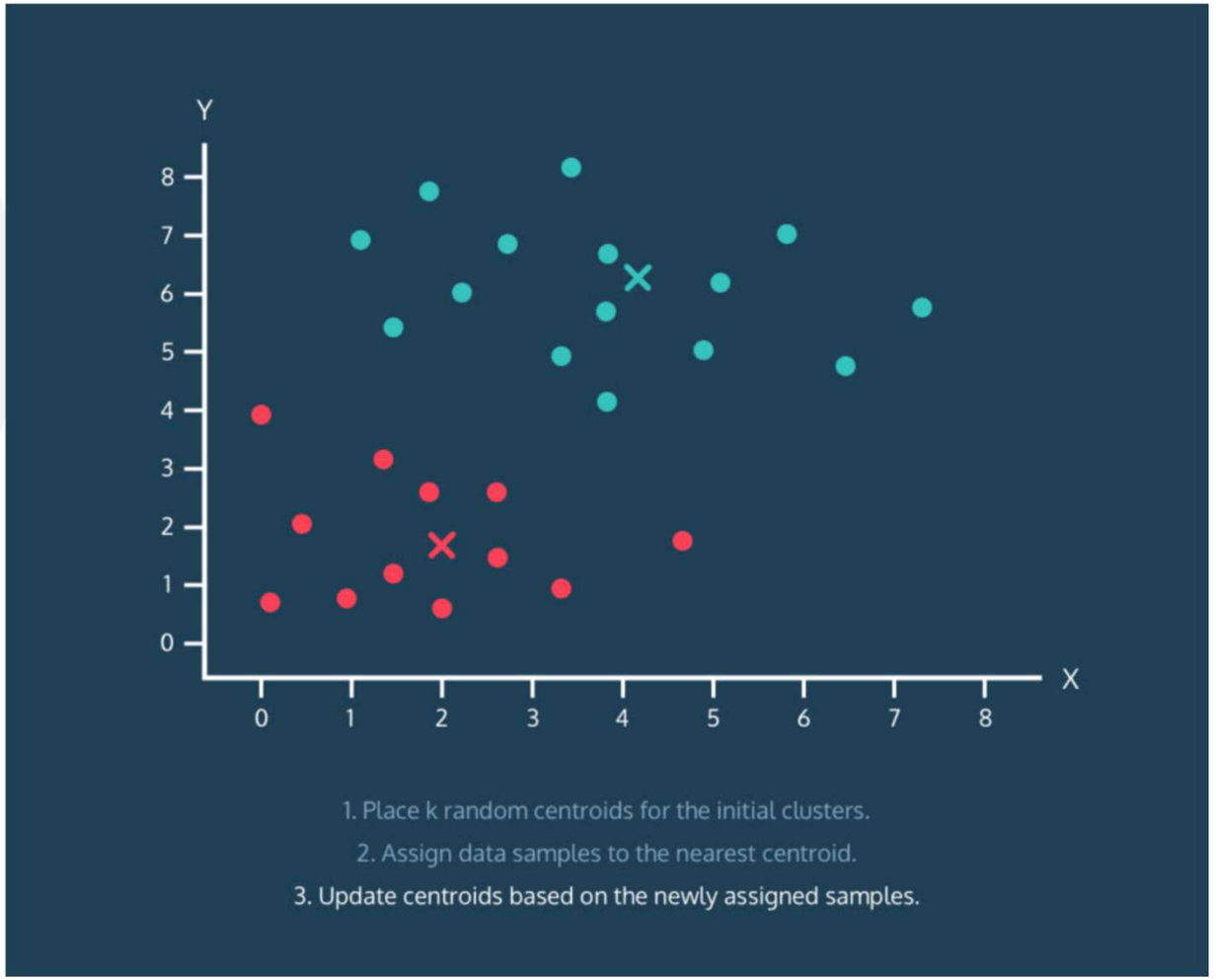

DATA PULSE


DATA
PULSE

# UNDERSTANDING CLUSTERING

- Clustering is a fundamental technique in unsupervised machine learning used to group similar data points together.

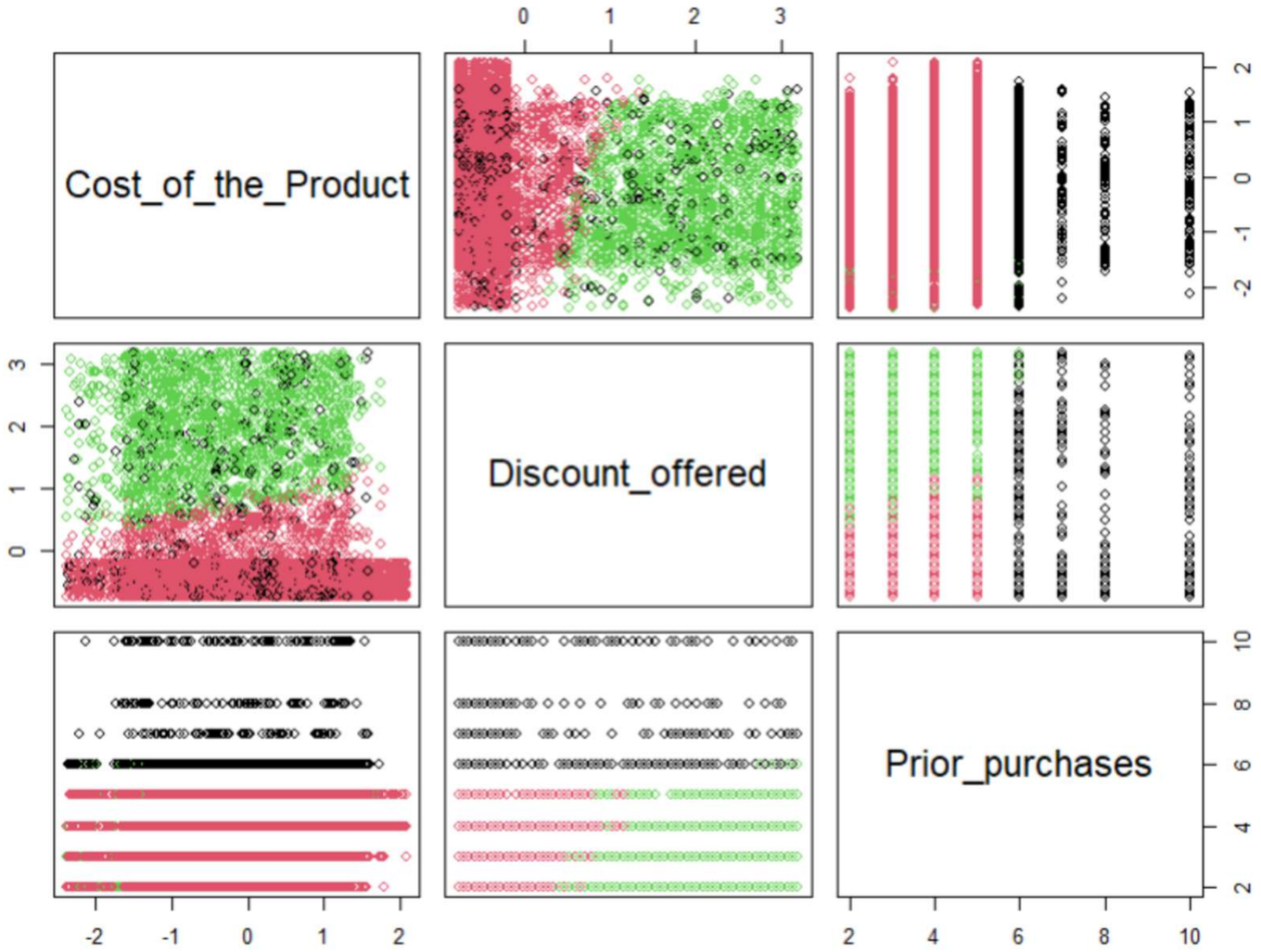
- K-means is a clustering algorithm used to partition data into distinct groups (clusters).




DATA
PULSE

# K-MEANS

- In our analysis, we employed the K-means clustering algorithm, resulting in the identification of three distinct clusters within the dataset. These clusters represent groupings of data points that share similar characteristics, allowing us to uncover meaningful patterns and insights. This clustering approach facilitates a clearer understanding of the inherent structures within the data, aiding in more targeted and nuanced data interpretation.



# GENERATIVE ARTIFICIAL INTELLIGENCE FOR SALES

DATA
PULSE

# IMPACT OF GENERATIVE AI ON SALES

| BENIFITS | | CHALLENGES |
|---|---|---|
| Personalized Customer Interactions | | Data Privacy and Security |
| Enhanced Data Analysis | | Accuracy and Relevance |
| Automated Content Creation | | Integration with Existing Systems |
| Streamlined Sales Processes | | Reliance on Data Quality |

DATA PULSE

# TRANSFORMATION OF SALES STRATEGY VIA AI

Predictive Analytics:

- AI can predict which leads are most likely to convert, allowing sales teams to prioritize their efforts.

Dynamic Pricing:

- AI can dynamically adjust pricing based on market conditions, competition, and customer profiles to maximize profits and sales volume.

Sales Forecasting:

- More accurate sales forecasts can be generated, helping in inventory management and financial planning.

Customer Insights:

- Deep learning can reveal complex patterns in customer behavior, leading to more effective sales strategies.

DATA PULSE

# CONCLUSION

DATA
PULSE

Optimized e-commerce via statistical & AI analysis

Enhanced customer satisfaction & loyalty

Data-driven sales strategy transformation

Predictive modeling for improved decision-making

AI-facilitated personalized customer interactions

Sales team empowerment & operational efficiency

Ready for continuous improvement & AI integration

# THANK YOU

## FOR YOUR ATTENTION

DATA
PULSE