

TokoBli E-Commerce

Campaign Study & Customer Spending Potential

Week 2 & 3 - Statistic & Spreadsheet

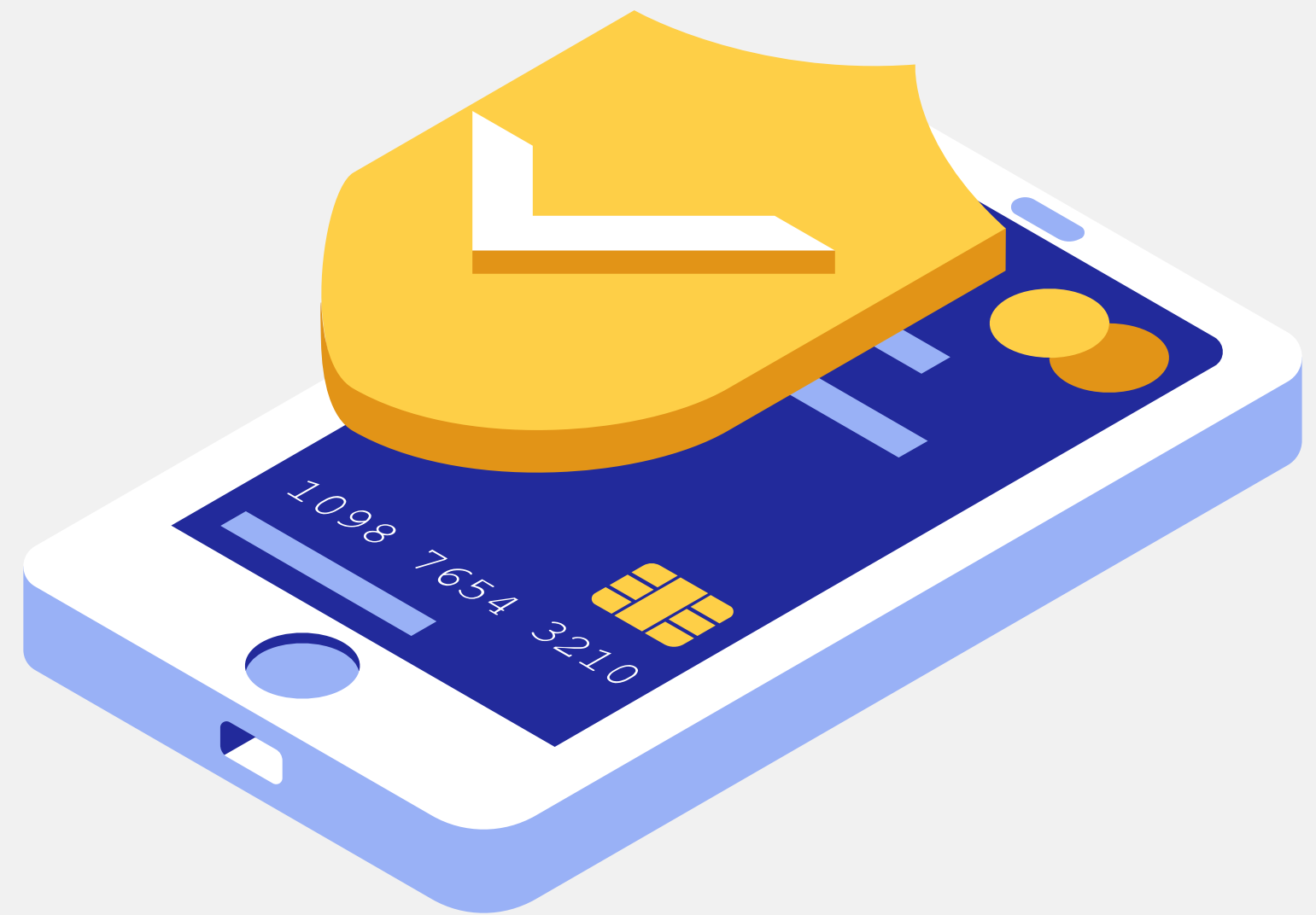
FSDA - Barcelona - Team 1 - Michael

Table of Contents

- Business Understanding
- Data Cleaning
- Descriptive Analytics
- Exploratory Data Analytics
- Hypothesis Testing
- Linear Regressions
- Business Recommendations

Business Background

As a member of the Data Analyst team at TokoBli, we were tasked by the business team to evaluate and derive insights from three recent campaigns, each executed under different scenarios (A, B, and C). Our objective is to analyze the metrics from these campaigns to identify the most effective strategy. The ultimate goal is to optimize revenue and transaction volume while minimizing campaign expenditure.



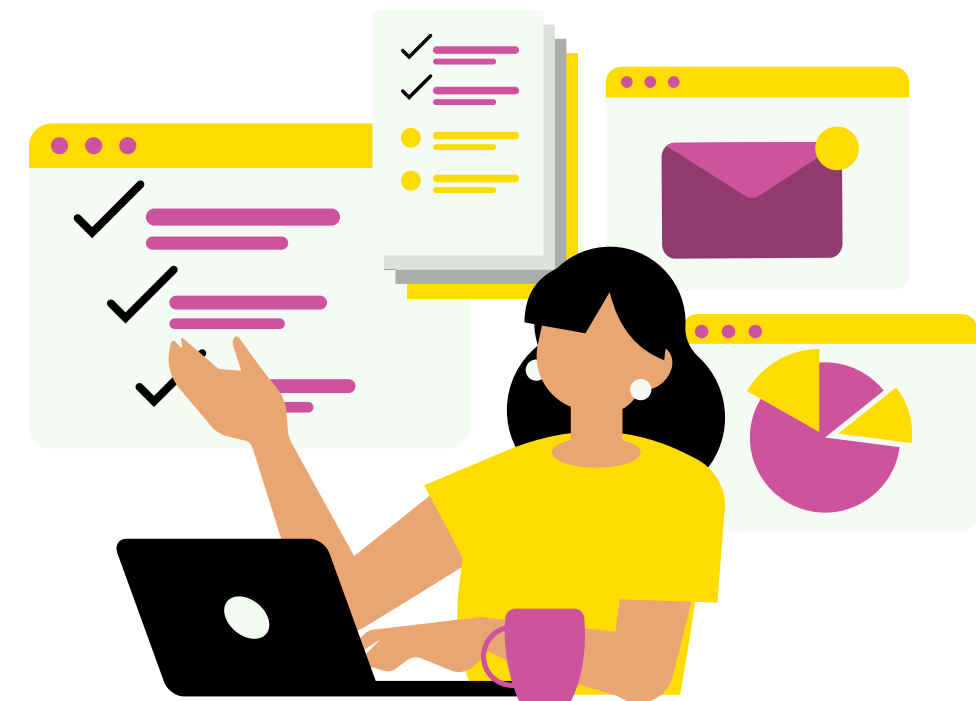
Evaluate Previous Campaigns & Find Insights

To analyze the wholeness of the campaigns conducted, find insights about what have been the outputs from those programs



Determine Most Effective Campaign Strategy

Finalizing what campaign scenario should be made to gain optimal revenue and less budget



Business Understanding

- Considering the data that we have, how do we think the data will help us answer the business problem ?



Business Problem : Customer Behavior

We can find out about the most popular product category.

Analyzed features : Category Name, Price, Quantity, Discount, Total Revenue

Business Problem : Most Effective Campaign

How much campaign budget is spent, and how much revenue is generated for each campaign scenarios.

Analyzed features : Discount, Shipping Cost, Price, Quantity, Total Revenue, Period

Business Understanding

- What Statistics are best to answer these business problems ?



Business Problem : Customer Behavior

Proposed Statistic 1 : SUM & Mean. Provide insights into specific trends or preferences. Example : most popular product category.

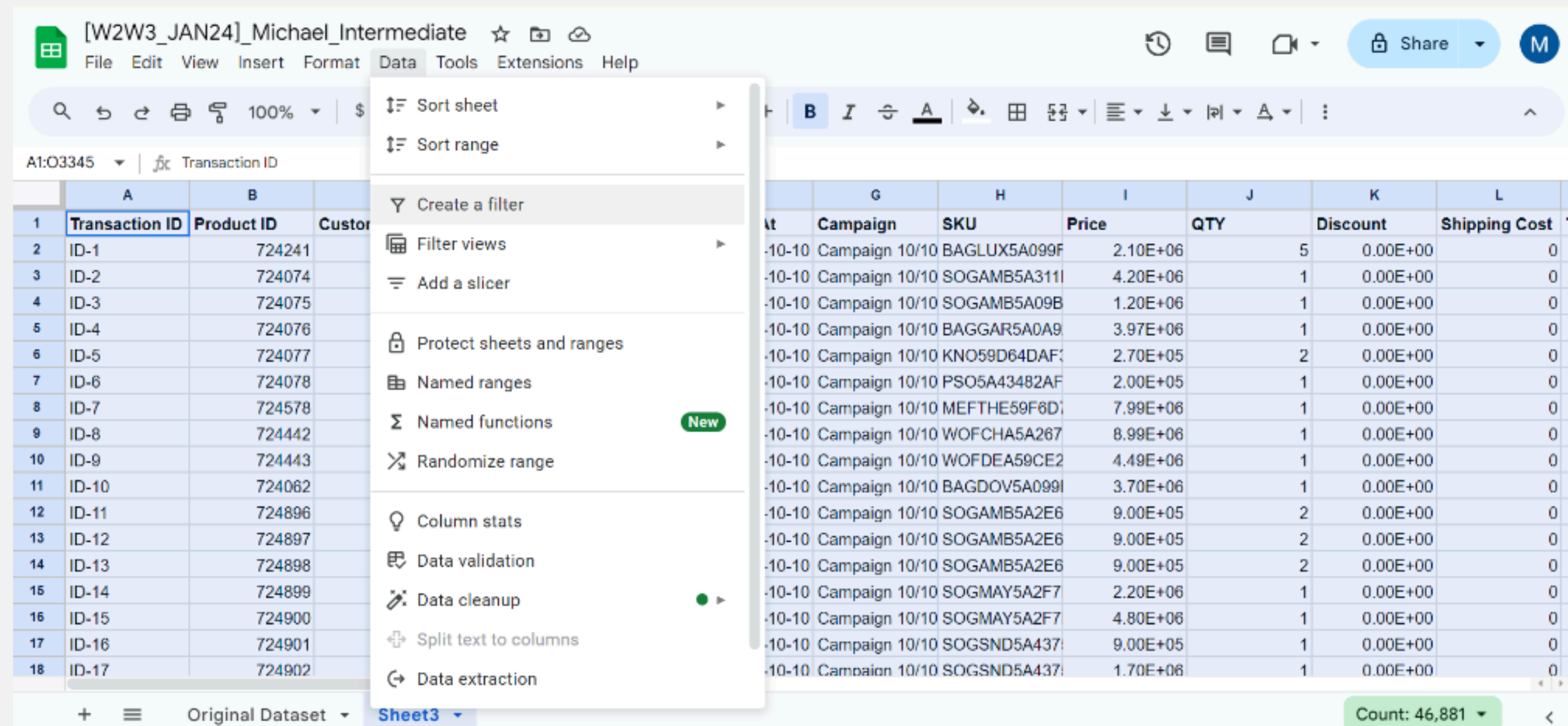
Proposed Statistic 2 : Standard Deviation, Mean, Median, Quartile, Skewness. Example : By examining Quartile and IQR, we can detect outliers that might indicates products that performed extremely well, or the other way around

Business Problem : Most Effective Campaign

Proposed Statistic : Sum, Average, Total Revenue, Total Discount. The results of these statistics help us compare revenue and campaign cost results for each scenario. We can finalized which scenario has the most profitability potential.

First Thing First

Use Filter to help us find missing / blank data



[W2W3_JAN24]_Michael_Intermediate

File Edit View Insert Format Data Tools Extensions Help

100% | \$

A1:O3345 | Transaction ID

	A	B	
1	Transaction ID	Product ID	Customer
2	ID-1	724241	
3	ID-2	724074	
4	ID-3	724075	
5	ID-4	724076	
6	ID-5	724077	
7	ID-6	724078	
8	ID-7	724578	
9	ID-8	724442	
10	ID-9	724443	
11	ID-10	724062	
12	ID-11	724896	
13	ID-12	724897	
14	ID-13	724898	
15	ID-14	724899	
16	ID-15	724900	
17	ID-16	724901	
18	ID-17	724902	

Count: 46,881

Data Cleaning

Easily spot blanks and missing rows by using filter for each column / feature

Remove Irrelevant Columns

Streamlining our works by setting aside unused features / columns

- Gender → There will be no gender behavioral analysis
- Status → All transactions assumed to be completed
- Created at → All transaction happened on the same date depending on the campaign scenario
- SKU → Unused data
- Shipping Costs → All shipping cost are 0 according to provided data
- Payment Method → There will be no payment behavioral analysis

Impute Missing Data

Some missing data can be filled by using reverse maths

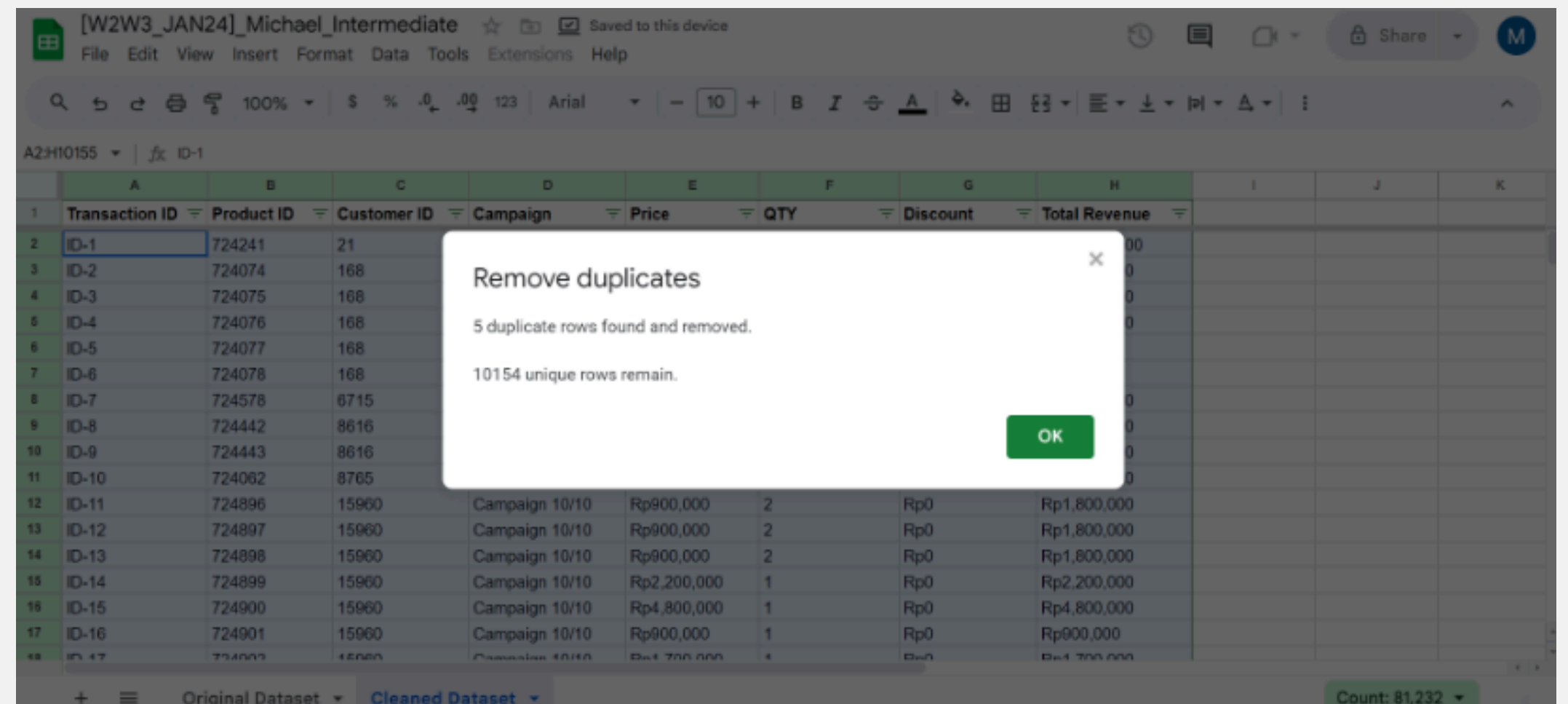
	A	B	C	D	E	F	G	H
1	Transaction	Product ID	Customer ID	Campaign	Price	QTY	Discount	Total Revenue
148	ID-147	725300	84580	Campaign 10/10		1	Rp0	Rp2,500,000

- Example : we can calculate missing value of **Price** by looking over **Discount**, **QTY**, and **Total Revenue** value, assuming the **Price** with "x" variable : **Total Revenue = (x * QTY) - Discount**
- Then : **x = Total Revenue / QTY = 2.500.000**

Remove Duplicated Rows

To raise the accuracy of this analysis, we'll remove duplicated data

The most easy possible way to find duplicates is to look into **Transaction ID**. It would not make any sense for a transaction ID to appear twice on a dataset, every rows of this variable has to be unique, therefore removing duplicates must be applied on duplicated transaction ID rows.



Converting Datatypes

Data Cleaning

Change data type of columns that is not suitable

	A	B	C	D	E	F	G	H
1	Transaction	Product ID	Customer ID	Campaign	Price	QTY	Discount	Total Revenue
2	ID-1	724241	21	Campaign 10/10	2.10E+06	5	0.00E+00	10,500,000
3	ID-2	724074	168	Campaign 10/10	4.20E+06	1	0.00E+00	4,200,000
4	ID-3	724075	168	Campaign 10/10	1.20E+06	1	0.00E+00	1,200,000
5	ID-4	724076	168	Campaign 10/10	3.97E+06	1	0.00E+00	3,968,300
6	ID-5	724077	168	Campaign 10/10	2.70E+05	2	0.00E+00	540,000
7	ID-6	724078	168	Campaign 10/10	2.00E+05	1	0.00E+00	200,000
8	ID-7	724578	6715	Campaign 10/10	7.99E+06	1	0.00E+00	7,990,000
9	ID-8	724442	8616	Campaign 10/10	8.99E+06	1	0.00E+00	8,990,000
10	ID-9	724443	8616	Campaign 10/10	4.49E+06	1	0.00E+00	4,490,000
11	ID-10	724062	8765	Campaign 10/10	3.70E+06	1	0.00E+00	3,700,000
12	ID-11	724896	15960	Campaign 10/10	9.00E+05	2	0.00E+00	1,800,000
13	ID-12	724897	15960	Campaign 10/10	9.00E+05	2	0.00E+00	1,800,000
14	ID-13	724898	15960	Campaign 10/10	9.00E+05	2	0.00E+00	1,800,000
15	ID-14	724899	15960	Campaign 10/10	2.20E+06	1	0.00E+00	2,200,000
16	ID-15	724900	15960	Campaign 10/10	4.80E+06	1	0.00E+00	4,800,000
17	ID-16	724901	15960	Campaign 10/10	9.00E+05	1	0.00E+00	900,000
18	ID-17	724902	15960	Campaign 10/10	1.70E+06	1	0.00E+00	1,700,000



	A	B	C	D	E	F	G	H
1	Transaction	Product ID	Customer ID	Campaign	Price	QTY	Discount	Total Revenue
2	ID-1	724241	21	Campaign 10/10	Rp2,100,000	5	Rp0	Rp10,500,000
3	ID-2	724074	168	Campaign 10/10	Rp4,200,000	1	Rp0	Rp4,200,000
4	ID-3	724075	168	Campaign 10/10	Rp1,200,000	1	Rp0	Rp1,200,000
5	ID-4	724076	168	Campaign 10/10	Rp3,970,000	1	Rp0	Rp3,968,300
6	ID-5	724077	168	Campaign 10/10	Rp270,000	2	Rp0	Rp540,000
7	ID-6	724078	168	Campaign 10/10	Rp200,000	1	Rp0	Rp200,000
8	ID-7	724578	6715	Campaign 10/10	Rp7,990,000	1	Rp0	Rp7,990,000
9	ID-8	724442	8616	Campaign 10/10	Rp8,990,000	1	Rp0	Rp8,990,000
10	ID-9	724443	8616	Campaign 10/10	Rp4,490,000	1	Rp0	Rp4,490,000
11	ID-10	724062	8765	Campaign 10/10	Rp3,700,000	1	Rp0	Rp3,700,000
12	ID-11	724896	15960	Campaign 10/10	Rp900,000	2	Rp0	Rp1,800,000
13	ID-12	724897	15960	Campaign 10/10	Rp900,000	2	Rp0	Rp1,800,000
14	ID-13	724898	15960	Campaign 10/10	Rp900,000	2	Rp0	Rp1,800,000
15	ID-14	724899	15960	Campaign 10/10	Rp2,200,000	1	Rp0	Rp2,200,000
16	ID-15	724900	15960	Campaign 10/10	Rp4,800,000	1	Rp0	Rp4,800,000
17	ID-16	724901	15960	Campaign 10/10	Rp900,000	1	Rp0	Rp900,000
18	ID-17	724902	15960	Campaign 10/10	Rp1,700,000	1	Rp0	Rp1,700,000

Changing the numbers format of Price, Discount, and Total Revenue to currency (Indonesian Rupiah)

Detecting Outliers

Using IQR (Interquartile Range)

Quartile formula in spreadsheet

`QUARTILE(data, quartile_number)`

$$\text{IQR} = Q3 - Q1$$

Outliers Seeking	
Quartile 1	Rp2,241,000
Quartile 2	Rp4,275,000
Quartile 3	Rp6,987,500
IQR	$=M5 - M3$

Quartile 1	Rp2,241,000
Quartile 2	Rp4,275,000
Quartile 3	Rp6,987,500
IQR	Rp4,746,500
Lower Outlier Lim	$=M3 - 1.5 * M6$

$$\text{Lower Outlier Limit} = Q1 - 1.5 * \text{IQR}$$

Quartile 1	Rp2,241,000	
Quartile 2	Rp4,275,000	
Quartile 3	Rp6,987,500	
IQR	Rp4,746,500	
Lower Outlier Limit	-Rp4,878,750	Rp0
Upper Outlier Lim	$=M5 + 1.5 * M6$	

$$\text{Upper Outlier Limit} = Q3 + 1.5 * \text{IQR}$$

Detecting Outliers

Using IQR (Interquartile Range)

We **will remove** outliers from Total Revenue column. Reasoning : **Data for each campaign should be compared evenly.** Based on the IQR method of pinpointing outliers limit, we remove the value that exceed **Rp. 14.107.250**. As for the lower outlier limit, negative value is irrational, therefore we immediately set it to **Rp. 0**

Outliers Seeking		
Quartile 1	Rp2,241,000	
Quartile 2	Rp4,275,000	
Quartile 3	Rp6,987,500	
IQR	Rp4,746,500	
Lower Outlier Limit	-Rp4,878,750	Rp0
Upper Outlier Limit	Rp14,107,250	

Detecting Outliers

Data Cleaning

Using IQR (Interquartile Range)

Identifying outliers with IF function

	G	H	I	J	K	L	M	N	
1	Discount	Total Revenue	Category Name	Outliers ?					
2	Rp0	Rp10,500,000	Beauty & Grooming	=if(H2>\$M\$8,"YES",IF(H2<\$N\$7,"YES","NO"))		Seeking			
3	Rp0	Rp4,200,000	Soghaat	+ Add new function Ctrl + Alt + N		1	Rp2,241,000		
4	Rp0	Rp1,200,000	Soghaat	NO		Quartile 2	Rp4,275,000		
5	Rp0	Rp3,968,300	Beauty & Grooming	NO		Quartile 3	Rp6,987,500		
6	Rp0	Rp540,000	Superstore	NO		IQR	Rp4,746,500		
7	Rp0	Rp200,000	Superstore	NO		Lower Outlier Limit	-Rp4,878,750	Rp0	
8	Rp0	Rp7,990,000	Men's Fashion	NO		Upper Outlier Limit	Rp14,107,250		
9	Rp0	Rp8,990,000	Women's Fashion	NO					

Detecting Outliers

Data Cleaning

Using IQR (Interquartile Range)

Outliers found

	G	H	I	J	
1	Discount	Total Revenue	Category Name	Outliers ?	
274	Rp0	Rp78,400,000	Mobiles & Tablets	YES	
305	Rp0	Rp69,900,000	Entertainment	YES	
329	Rp0	Rp64,000,000	Beauty & Grooming	YES	
333	Rp210,000	Rp50,190,000	Beauty & Grooming	YES	
336	Rp210,000	Rp48,090,000	Beauty & Grooming	YES	
343	Rp210,000	Rp16,590,000	Beauty & Grooming	YES	
360	Rp750,000	Rp74,250,000	Superstore	YES	
361	Rp750,000	Rp74,250,000	Superstore	YES	
419	Rp0	Rp21,000,000	Beauty & Grooming	YES	
441	Rp844,000	Rp83,556,000	Superstore	YES	
818	Rp0	Rp65,100,000	Beauty & Grooming	YES	
9020	Rp0	Rp69,800,000	Men's Fashion	YES	

Statistical Measurement

Create Statistical measurement on important column to know the data distribution

Descriptive Analytics formulas in Spreadsheet

Statistical Measure	Formula
Count	"=COUNT(A1:A100)
Minimum	"=MIN(A1:A100)
Maximum	"=MAX(A1:A100)
Mean (Average)	"=AVERAGE(A1:A100)
Median	"=MEDIAN(A1:A100)
Mode	"=MODE.SNGL(A1:A100)
Q1 (First Quartile)	"=QUARTILE.INC(range, 1)
Q3 (Third Quartile)	"=QUARTILE.INC(range, 3)
Range	"=MAX(A1:A100) - MIN(A1:A100)
Variance	"=VAR(A1:A100)
Standard Deviation	"=STDEV(A1:A100)
Skewness	"=SKEW(A1:A100)

Statistical Measurement

Variable : Discount



Count	10142
Minimum	Rp0
Maximum	Rp900,000
Mean	Rp81,976
Median	Rp0
Mode	Rp0
First Quartile (Q1)	Rp0
Third Quartile (Q3)	Rp21,000
Range	Rp900,000
Variance	35014647140
Std. Deviation	187122.0114
Skewness	2.503916954

- There is an extreme difference between mean & median, this indicates there are outliers to the data distribution where the right tail of the data distribution is longer
- This supported with positive skewness value
- Mode is "Rp. 0" where most products are not discounted, and our customers still prefer to buy them

Statistical Measurement

Variable : Price



Count	10142
Minimum	Rp50,000
Maximum	Rp9,000,000
Mean	Rp4,112,004
Median	Rp3,890,000
Mode	Rp7,200,000
First Quartile (Q1)	Rp1,990,000
Third Quartile (Q3)	Rp6,290,000
Range	Rp8,950,000
Variance	6030363021384
Std. Deviation	2455679.747
Skewness	0.2243167219

- There is a slight difference between mean & median, this indicates there are outliers to the data distribution where the left tail of the data distribution is longer
- This supported with positive skewness value
- Gap between minimum and maximum Total Revenue is quite extreme, meaning there must be outliers that need to be removed to even campaign data for further analysis.

Statistical Measurement

Variable : Quantity



Count	10142
Minimum	1
Maximum	30
Mean	1.263163084
Median	1
Mode	1
First Quartile (Q1)	1
Third Quartile (Q3)	1
Range	29
Variance	1.00351225
Std. Deviation	1.001754586
Skewness	17.0149086

- There is a slight difference between mean & median, this indicates there are outliers to the data distribution where the right tail of the data distribution is longer
- This supported with positive skewness value
- Mode & median has the same value which is "1". Meaning most of customers purchased only 1 item, and certainly there are outliers

Statistical Measurement

Variable : Total Revenue



Count	10142
Minimum	Rp45,000
Maximum	Rp13,940,000
Mean	Rp4,562,944
Median	Rp4,275,000
Mode	Rp7,200,000
First Quartile (Q1)	Rp2,241,000
Third Quartile (Q3)	Rp6,980,000
Range	Rp13,895,000
Variance	7140180293777
Std. Deviation	2672111.58
Skewness	0.3731026644

- There is a slight difference between mean & median, this indicates there are outliers to the data distribution where the left tail of the data distribution is longer
- This supported with positive skewness value
- Gap between minimum and maximum Total Revenue is quite extreme, meaning there must be outliers that need to be removed to even campaign data for further analysis.

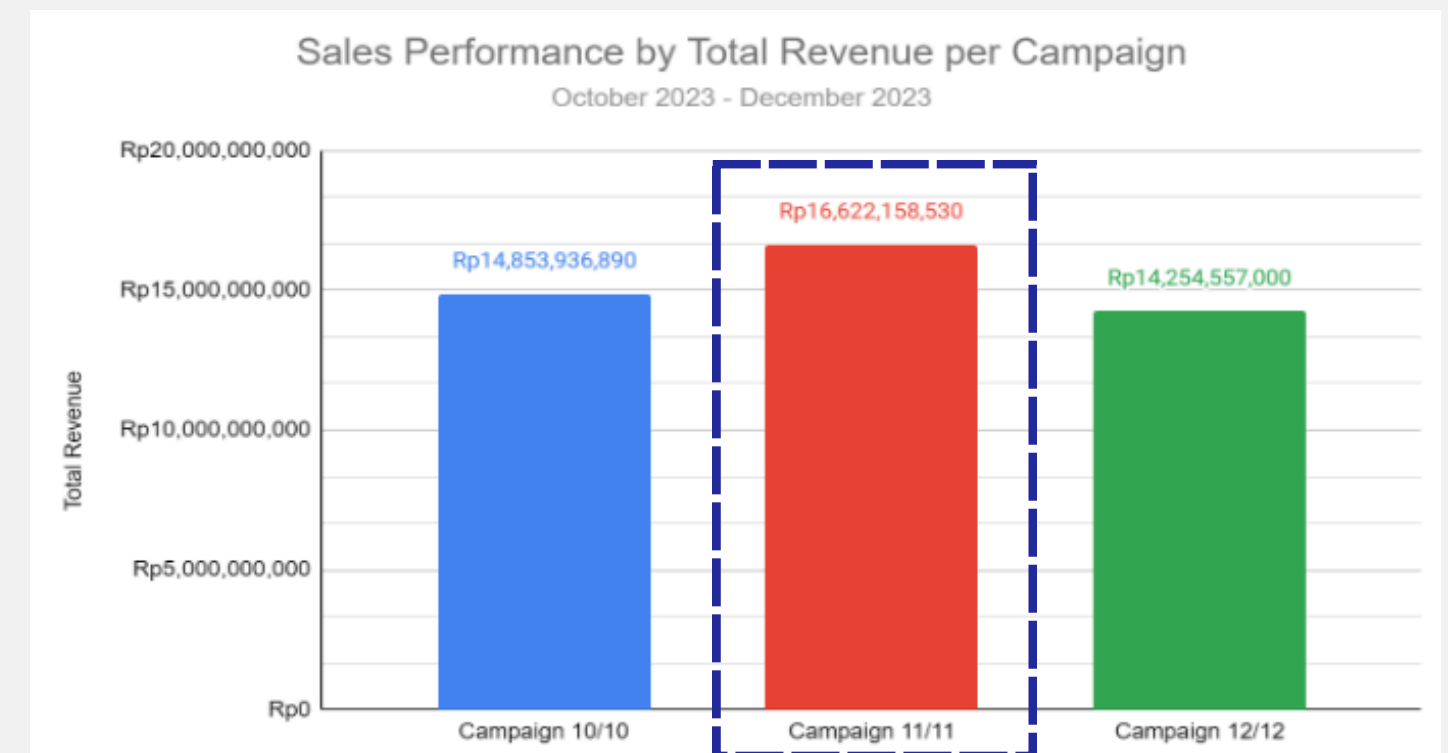
Exploratory Data Analytics

Is there a difference in sales performance between the three campaign periods?

Sales Performance by Total Revenue, Total Transaction, Total Customer, & QTY of Product Sold				
Campaign	Total Revenue	Total Transaction	Total Customer	QTY of Product Sold
Campaign 10/10	Rp14,853,936,890	3329	1498	4211
Campaign 11/11	Rp16,622,158,530	3399	1567	4115
Campaign 12/12	Rp14,254,557,000	3271	1411	4278
Grand Total	Rp45,730,652,420	9999	4242	12604

Above table, describes our sales performance by number of revenue, transaction, customer, and quantity of product sold, however, in this assignment we will focus on the metrics of revenue to do the hypothesis testing.

Campaign	Total Revenue
Campaign 10/10	Rp14,853,936,890
Campaign 11/11	Rp16,622,158,530
Campaign 12/12	Rp14,254,557,000
Grand Total	Rp45,730,652,420



- Campaign 11 / 11 got the highest revenue
- Campaign 12 / 12 got the lowest

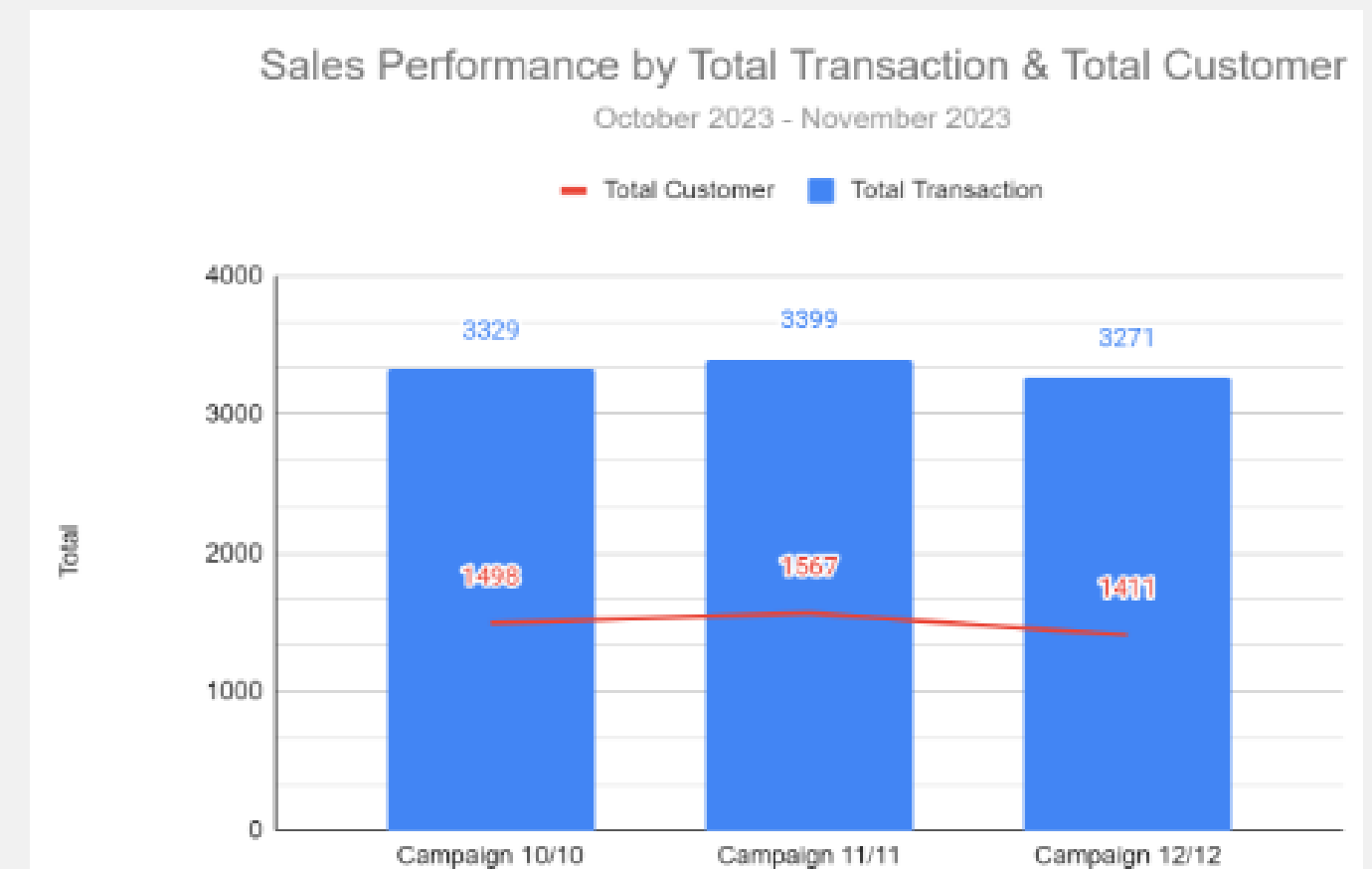
Exploratory Data Analytics

Is there a difference in sales performance between the three campaign periods?

Sales Performance by Total Revenue, Total Transaction, Total Customer, & QTY of Product Sold				
Campaign	Total Revenue	Total Transaction	Total Customer	QTY of Product Sold
Campaign 10/10	Rp14,853,936,890	3329	1498	4211
Campaign 11/11	Rp16,622,158,530	3399	1567	4115
Campaign 12/12	Rp14,254,557,000	3271	1411	4278
Grand Total	Rp45,730,652,420	9999	4242	12604

Above table, describes our sales performance by number of revenue, transaction, customer, and quantity of product sold, however, in this assignment we will focus on the metrics of revenue to do the hypothesis testing.

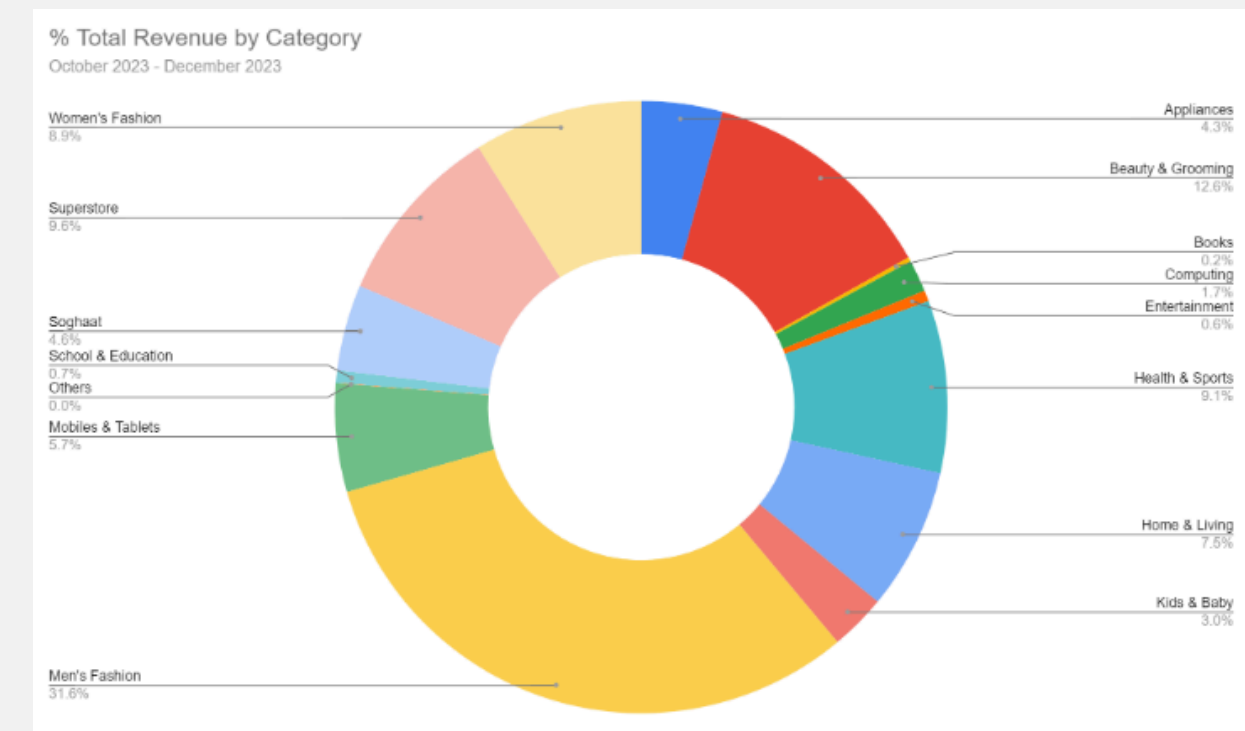
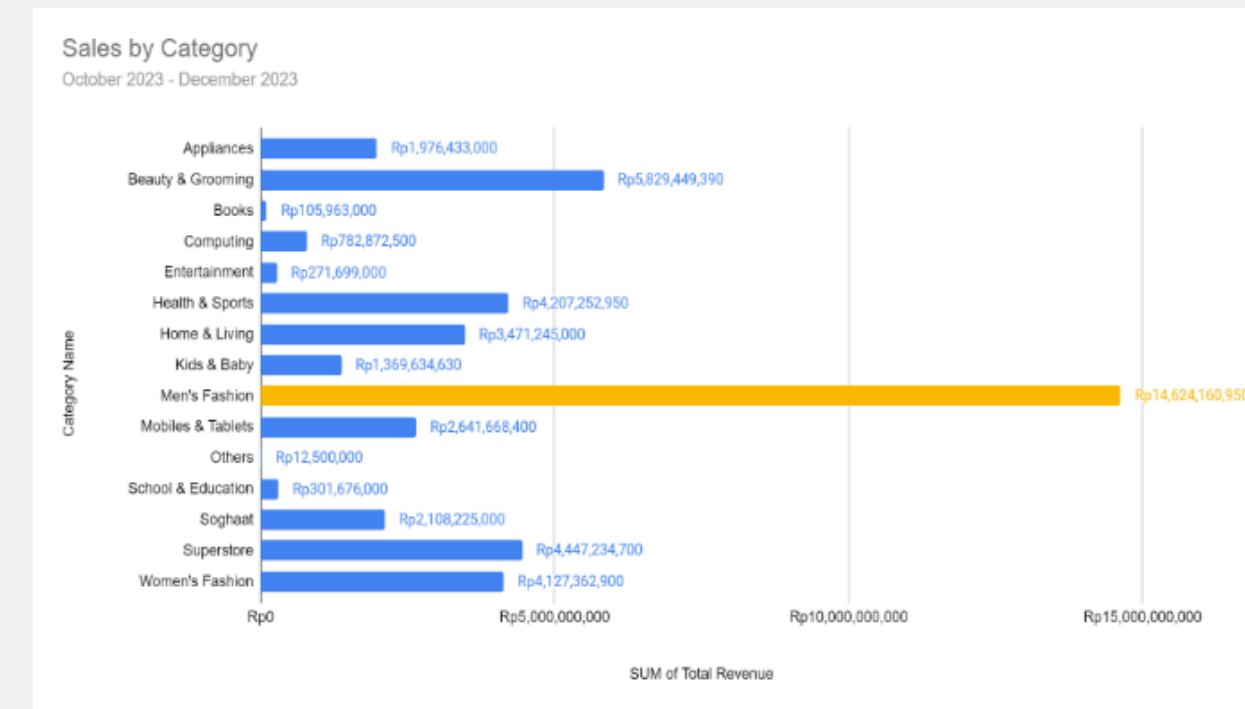
Indicators / Campaign	Campaign 10/10	Campaign 11/11	Campaign 12/12
Total Transaction	3329	3399	3271
Total Customer	1498	1567	1411



- Campaign 11 / 11 had the most transaction
- Campaign 11 / 11 also had the most cust.

Exploratory Data Analytics

What product category is the best seller and produces the highest sales performance from the three campaign periods?

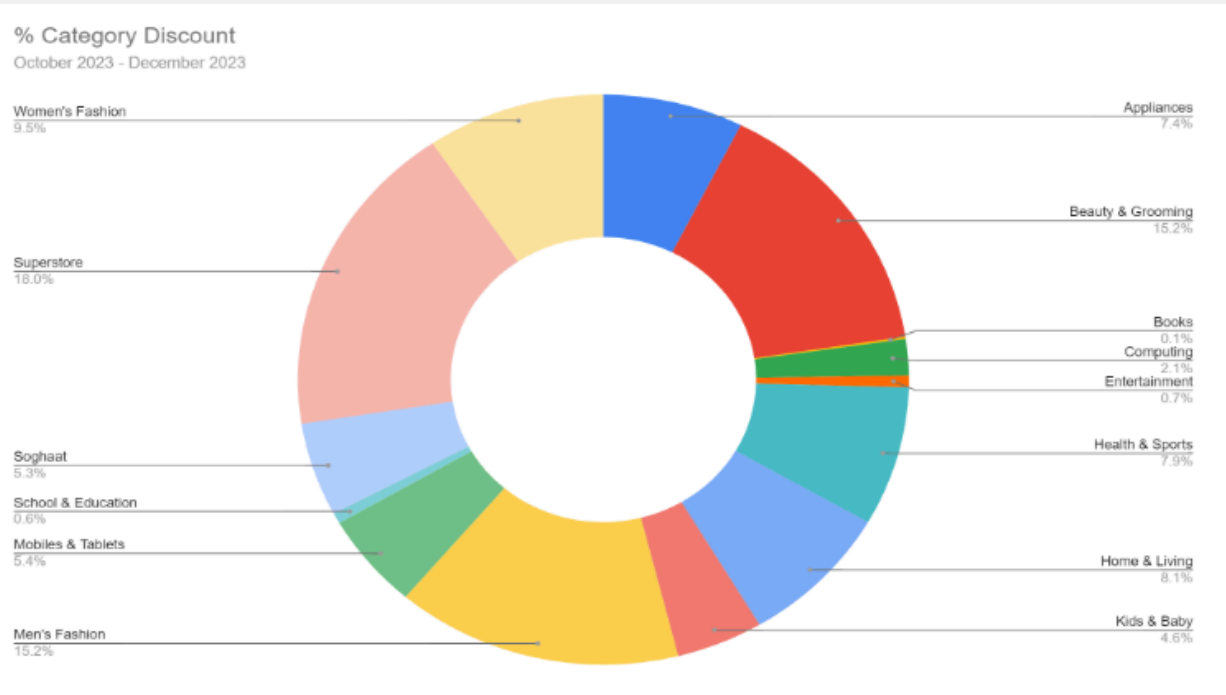
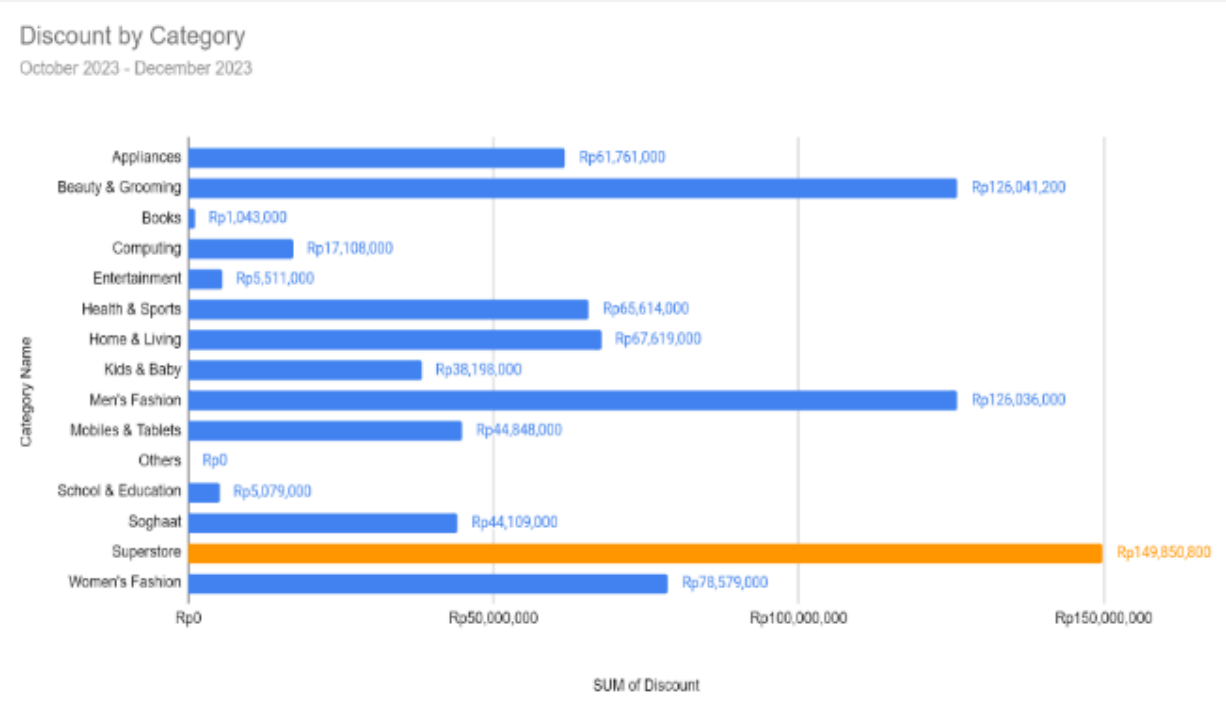


- There are 15 product categories in overall
- Men's fashion is the best performing products from the three previous campaigns (31,6% of revenue)

Exploratory Data Analytics

What product category received the highest discount campaign budget from the three campaign periods?

TOTAL Discount by Category		
Category Name	SUM of Discount	% of Discount
Appliances	Rp61,761,000	7.43%
Beauty & Grooming	Rp126,041,200	15.16%
Books	Rp1,043,000	0.13%
Computing	Rp17,108,000	2.06%
Entertainment	Rp5,511,000	0.66%
Health & Sports	Rp65,614,000	7.89%
Home & Living	Rp67,619,000	8.13%
Kids & Baby	Rp38,198,000	4.59%
Men's Fashion	Rp126,036,000	15.16%
Mobiles & Tablets	Rp44,848,000	5.39%
Others	Rp0	0.00%
School & Education	Rp5,079,000	0.61%
Soghaat	Rp44,109,000	5.31%
Superstore	Rp149,850,800	18.02%
Women's Fashion	Rp78,579,000	9.45%
Grand Total	Rp831,397,000	100.00%



- There are 15 product categories in overall
- Superstore is the category with the highest campaign budget (18,02% of total campaign budget)

Exploratory Data Analytics

Is there a difference in discount campaign budget between the three campaign periods?

Discount Campaign Budget Comparison		
Campaign	SUM of Discount	% of Discount
Campaign 10/10	Rp275,925,900	33.19%
Campaign 11/11	Rp302,636,200	36.40%
Campaign 12/12	Rp252,834,900	30.41%
Grand Total	Rp831,397,000	100.00%



- Campaign 11 / 11 got the highest discount campaign budget
- and campaign 12 / 12 got the lowest.

Exploratory Data Analytics

Some Recommendations

- In order to find the most effective campaign budget distribution for each category, we need to find income ratio value
- Where **income = revenue - campaign discount**
- And **income ratio = income / revenue**
- Find the lowest income ratio to get to know the most ineffective campaign budget distribution.

Category Name	Income each Campaign		
	Campaign 10 / 10	Campaign 11/ 11	Campaign 12 / 12
Appliances	Rp539,422,000	Rp673,024,000	Rp702,226,000
Beauty & Grooming	Rp2,865,622,540	Rp986,282,100	Rp1,851,503,550
Books	Rp5,970,000	Rp13,683,000	Rp85,267,000
Computing	Rp278,646,000	Rp153,040,000	Rp334,078,500
Entertainment	Rp57,084,000	Rp70,710,000	Rp138,394,000
Health & Sports	Rp1,062,763,350	Rp1,051,594,000	Rp2,027,281,600
Home & Living	Rp1,133,846,000	Rp952,689,600	Rp1,317,090,400
Kids & Baby	Rp555,179,000	Rp368,090,930	Rp408,166,700
Men's Fashion	Rp3,337,400,000	Rp8,186,944,600	Rp2,973,780,350
Mobiles & Tablets	Rp990,114,400	Rp623,654,500	Rp983,051,500
School & Education	Rp 171,215,000	Rp32,622,000	Rp92,760,000
Soghat	Rp1,095,039,200	Rp355,386,000	Rp613,690,800
Superstore	Rp1,082,749,500	Rp1,979,450,000	Rp1,235,184,400
Women's Fashion	Rp1,390,460,000	Rp872,351,600	Rp1,785,972,300

Exploratory Data Analytics

Some Recommendations

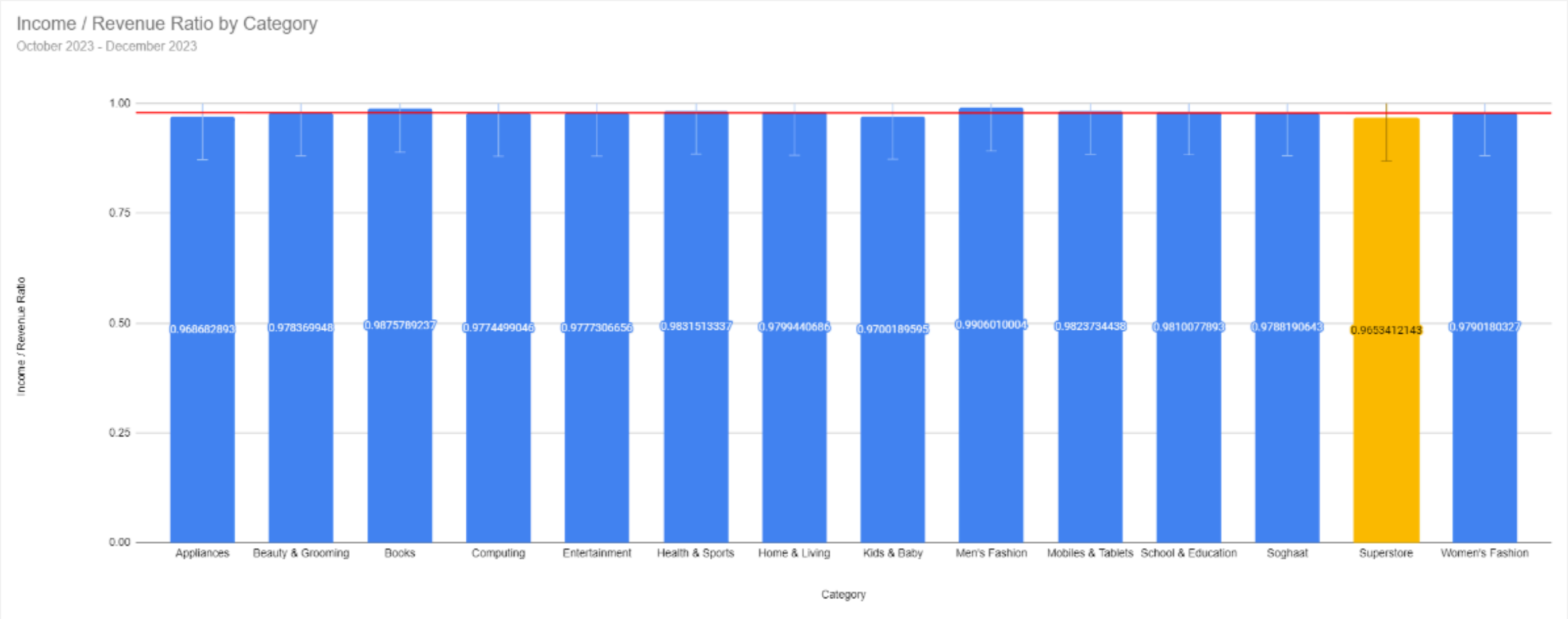
Category Name	Income / Revenue Ratio each Campaign				MIN
	Campaign 10 / 10	Campaign 11/ 11	Campaign 12 / 12	Average	
Appliances	0.968810391	0.9588301834	0.9784081045	0.968882893	0.9653412143
Beauty & Grooming	0.9763254191	0.9758499711	0.9829344537	0.978369948	
Books	1	0.9699439994	0.9927927718	0.9875789237	
Computing	0.9774891868	0.9744667303	0.9803937968	0.9774499048	
Entertainment	0.9735980352	0.9747725393	0.9848214222	0.9777306656	
Health & Sports	0.9818388652	0.9790539358	0.9885612001	0.9831513337	
Home & Living	0.98504081	0.9717013251	0.9830900707	0.9799440688	
Kids & Baby	0.9851005011	0.9470040297	0.9779523476	0.9700189595	
Men's Fashion	0.9899856429	0.9928228138	0.9889945444	0.9906010004	
Mobiles & Tablets	0.9836870527	0.9776321142	0.9858211646	0.9823734438	
School & Education	0.9904664387	0.9822944896	0.9702624395	0.9810077893	
Soghaat	0.9802985838	0.9794889038	0.9766697053	0.9788190643	
Superstore	0.9615401523	0.9702092411	0.9642742496	0.9653412143	
Women's Fashion	0.9867506901	0.9667953213	0.9835080868	0.9790180327	

- **Income ratio = income / revenue**
- Find the lowest income ratio to get to know the most ineffective campaign budget distribution.
- Category : superstore, got the lowest income / revenue ratio, this state that this category had the most inefficient campaign budget

***Other categories to be considered about campaign budget re-adjustment : Appliances (0.96), Kids & Baby (0.97)**

Exploratory Data Analytics

Some Recommendations



CONCLUSION : Overall, the 11 / 11 campaign is the most profitable scenario amongst other, however, it is highly recommended that the business team consider to make some adjustment to the campaign budget distribution, mainly for “superstore” product category. This action might benefits the company even more.

Hypothesis 1	H0 :	Total Revenue Campaign 10 / 10 = 11 / 11	reject H0	
	H1 :	Total Revenue Campaign 10 / 10 <> 11 / 11		
Alpha	5%			
Variance Check	10 / 10	6994163512214		Different variance value
	11 / 11	6957578288904		
Statistical Test	T Test: Two-Sample Assuming Unequal Variances			
		Variable 1	Variable 2	
	Mean	4461981.643	4890308.482	Campaign 11 / 11 performed better in sales performance
	Variance	6.99E+12	6.96E+12	
	Observations	3329	3399	Very tolerrable
	Hypothesized Mean Difference	0		
	df	6722		
	t Stat	-6.650581727		
	P(T<=t) one-tail	1.57E-11		
	t Critical one-tail	1.645080282		
	P(T<=t) two-tail	3.15E-11	<0.05	
	t Critical two-tail	1.960316904		
Conclusion	Business Conclusion	There is a significant difference in Total Revenue of Campaign 10 / 10 compared to Campaign 11/ 11		
	Statistic Conclusion	From the T - test result, we can conclude that the average revenue of campaign 10 / 10 is lower		
		than campaign 11/ 11, also by looking at the number of P - value, we can describe		
		that the difference of the total revenue between those two campaigns is significant		

Hypothesis Testing

Is there any significant difference between Campaign 10 / 10 and Campaign 11 / 11 Total Revenue?

From the process of hypothesis testing using XLminer tools, we can conclude that :

- **Reject H0** (the difference in revenue is significant)
- P value is less than 5%
- Mean of campaign 11 / 11 is greater than campaign 10 / 10

Hypothesis 2	H0 :	Total Revenue Campaign 10 / 10 = 12 / 12	accept H0	
	H1 :	Total Reveneue Campaign 10 / 10 <> 12 / 12		
Alpha	5%			
Variance Check	10 / 10	6994163512214		
	12 / 12	7300112459108		
Statistical Test	T Test: Two-Sample Assuming Unequal Variances			
		Variable 1	Variable 2	
	Mean	4461981.643	4335466.315	there is high significancy between these campaign
	Variance	6.99E+12	7.30E+12	
	Observations	3329	3414	Very tolerrable
	Hypothesized Mean Difference	0		
	df	6740		
	t Stat	1.943111553		
	P(T<=t) one-tail	0.026022133		
	t Critical one-tail	1.645079677		
	P(T<=t) two-tail	0.052044267	>0.05	
	t Critical two-tail	1.960315961		
Conclusion	Business Conclusion	The difference of revenue between campaign 10 / 10 and campaign 12 / 12 is not significance		
	Statistic Conclusion	From the T - test result, we can conclude that the average revenue of campaign 10 / 10 is slightly		
		higher than campaign 12/ 12, also by looking at the number of P - value, we can describe		
		that the difference of the total revenue between those two campaigns is not significanet (>5%)		

Hypothesis Testing

Is there any significant difference between Campaign 10 / 10 and Campaign 12 / 12 Total Revenue?

From the process of hypothesis testing using XLminer tools, we can conclude that :

- **Accept H0** (the difference in revenue is not significant)
- P value is more than 5%
- Mean of campaign 10 / 10 is greater than campaign 12 / 12

Hypothesis 3	H0 :	Total Revenue Campaign 11 / 11 = 12 / 12	reject H0	
	H1 :	Total Revenu Campaign 11 / 11 <> 12 / 12		
Alpha	5%			
Variance Check	11 / 11	6957578288904		
	12 / 12	7300112459108		
Statistical Test	T Test: Two-Sample Assuming Unequal Variances			
		Variable 1	Variable 2	
	Mean	4890308.482	4335466.315	Campaign 11 / 11 performed better in sales performance
	Variance	6.96E+12	7.30E+12	
	Observations	3399	3414	Very tolerrable
	Hypothesized Mean Difference	0		
	df	6808		
	t Stat	8.57648819		
	P(T<=t) one-tail	5.99E-18		
	t Critical one-tail	1.645077418		
	P(T<=t) two-tail	1.20E-17	<0.05	
	t Critical two-tail	1.960312444		
Conclusion	Business Conclusion	There is a significant difference in Total Revenue of Campaign 11 / 11 compared to Campaign 12/ 12		
	Statistic Conclusion	From the T - test result, we can conclude that the average revenue of campaign 12 / 12 is lower		
		than campaign 11 / 11, also by looking at the number of P - value, we can describe		
		that the difference of the total revenue between those two campaigns is significant		

Hypothesis Testing

Is there any significant difference between Campaign 11 / 11 and Campaign 12 / 12 Total Revenue?

From the process of hypothesis testing using XLminer tools, we can conclude that :

- **Reject H0** (the difference in revenue is significant)
- P value is less than 5%
- Mean of campaign 11 / 11 is greater than campaign 12 / 12

Conclusions

- Campaign 11 / 11 produces the highest average revenue amongst other campaigns, therefore it causes significant difference with the rest of comparisons, this also statistically explained with the number of P - value and Mean.
- The number of observations between each campaign is also very tolerable and comparable, this pictures that the number of customer in every campaign is very consistent.

Business Recommendations

Seems that **campaign 11 /11 produces the most revenue in sales**. It is highly recommended that the business team **re - use** this campaign scenario with some improvements like **increasing campaign discount distribution to item's category that performed well** and make some **reducement on category that less efficient** on income ratio.

Unlocking Customer
Spending Potential :

Factors Influencing
Purchase Behavior in
TokoBli E-commerce



XLMiner Analysis ToolPak

Histogram

Linear Regression

Input Y Range:

Input X Range:

☐

Labels

☐

Constant Is Zero

☒

Confidence Level:

95

%

Output Range:

☐

Residuals

☐

Residual Plots

☐

Standardized Residuals☐☐

OK

Checking Correlation Between Variables

To understand the data in general, among the variables **Age, Income, Tenure, Avg. Session Time, Total Promo, and Bounce Rate**, which variables have a strong relationship with each other?

Before diving into correlations checking, First thing first : We check the variable significance first, this can be described with the **P - Value** obtained from the linear regression generated by XLminer.

Checking Correlation Between Variables

Regression								
SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.9639907184							
R Square	0.9292781052							
Adjusted R Square	0.9278249156							
Standard Error	1384406.116							
Observations	299							
ANOVA								
	df	SS	MS	F	Significance F			
Regression	6	7.35363E+15	1.2256E+15	639.4747754	0			
Residual	292	5596414460075	1916580294546					
Total	298	7.91327E+15						
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95%	Upper 95%
Intercept	1947879.706	594065.0494	3.27889969	0.001168198274	778687.603	3117071.81	778687.603	3117071.81
Age	30555.49209	16729.0723	1.826490527	0.06879730844	-2369.352173	63480.33635	-2369.352173	63480.33635
Income	0.03582464067	0.01111730128	3.222422401	0.00141493494	0.0139444427	0.0577048386	0.0139444427	0.0577048386
Tenure	64158.04723	14092.06659	4.55277775	0.0000077803638	36423.15019	91892.94428	36423.15019	91892.94428
Avg. Session Time	6286.982772	456.2684441	13.77913124	0	5388.991109	7184.974436	5388.991109	7184.974436
Total Promo (%)	133215.6971	13692.77277	9.728905847	0	106266.6587	160184.7354	106266.6587	160184.7354
Bounce Rate	-50349.9385	6451.813391	-7.803997954	0	-63047.89015	-37651.98686	-63047.89015	-37651.98686

Only "Age" Variable has low significancy, identified with the P value number > 5%, therefore we remove "Age" variable for the next Iteration.

Checking Correlation Between Variables

After we remove the “Age” variable, All other variables are **significant**, therefore, we check the **correlation** between these variables.

Regression Statistics								
Multiple R	0.9835715414							
R Square	0.9284701154							
Adjusted R Square	0.9272494689							
Standard Error	1389914.044							
Observations	299							
ANOVA								
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>			
Regression	5	7.34723E+15	1.46945E+15	760.6380056	0			
Residual	293	566035287520739	1931861049559					
Total	298	7.91327E+15						
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	2590241.49	480680.5048	5.388896784	0.0000001480639	1644217.366	3536265.614	1644217.366	3536265.614
Income	0.0368431741	0.01114748175	3.305067003	0.001067575078	0.01490388896	0.05878245925	0.01490388896	0.05878245925
Tenure	64576.21329	14146.26528	4.564894832	0.000007363181736	36735.0421	92417.38449	36735.0421	92417.38449
Avg. Session Time	6341.901705	457.0879684	13.87457589	0	5442.309896	7241.493514	5442.309896	7241.493514
Total Promo (%)	136147.7235	13652.45109	9.972401481	0	109278.4248	163017.0223	109278.4248	163017.0223
Bounce Rate	-49209.12635	6447.057484	-7.63280403	0	-61897.53757	-36520.71513	-61897.53757	-36520.71513

Checking Correlation Between Variables

Checking Variables Correlation					
	<i>Income</i>	<i>Tenure</i>	<i>Avg. Session Time</i>	<i>Total Promo (%)</i>	<i>Bounce Rate</i>
Income	1				
Tenure	0.3789467125	1			
Avg. Session Time	0.4836920601	0.6853319563	1		
Total Promo (%)	0.4456727121	0.6213744703	0.7768861588	1	
Bounce Rate	-0.4483756005	-0.5702766089	-0.7873056093	-0.7941434615	1

We set the maximum correlation value is up to 0.8, therefore all these variables are considered **not really correlated**, finally we can make regression formula.

But if we want to look deeper at the numbers, Bounce Rate & Total Promo has the most strong correlation with each other, because it's value is the closest to 1 or -1

XLMiner Analysis ToolPak

Histogram

Linear Regression

Input Y Range:

Input X Range:

☐

Labels

☐

Constant is Zero

☒

Confidence Level: 95 %

Output Range:

☐

Residuals

☐

Residual Plots

☐

Standardized Residuals☐☐

OK

Linear Regression

What variable has the significant impact on the Yearly Spending ?

In order to find the variable with most significant impact, we need to perform linear regression first, this also can be done with XLminer, which the output we want to know is the **P - value**. **Coefficient** value is also mandatory for the regression formula.

y = Yearly Spending

$$y = 2590241.49 + 0.037 * \text{Income} + 64576.2133 * \text{Tenure} + 6341.9017 * \text{Avg. Session Time} + 136147.7235 * \text{Total Promo} - 49209.1264 * \text{Bounce Rate}.$$

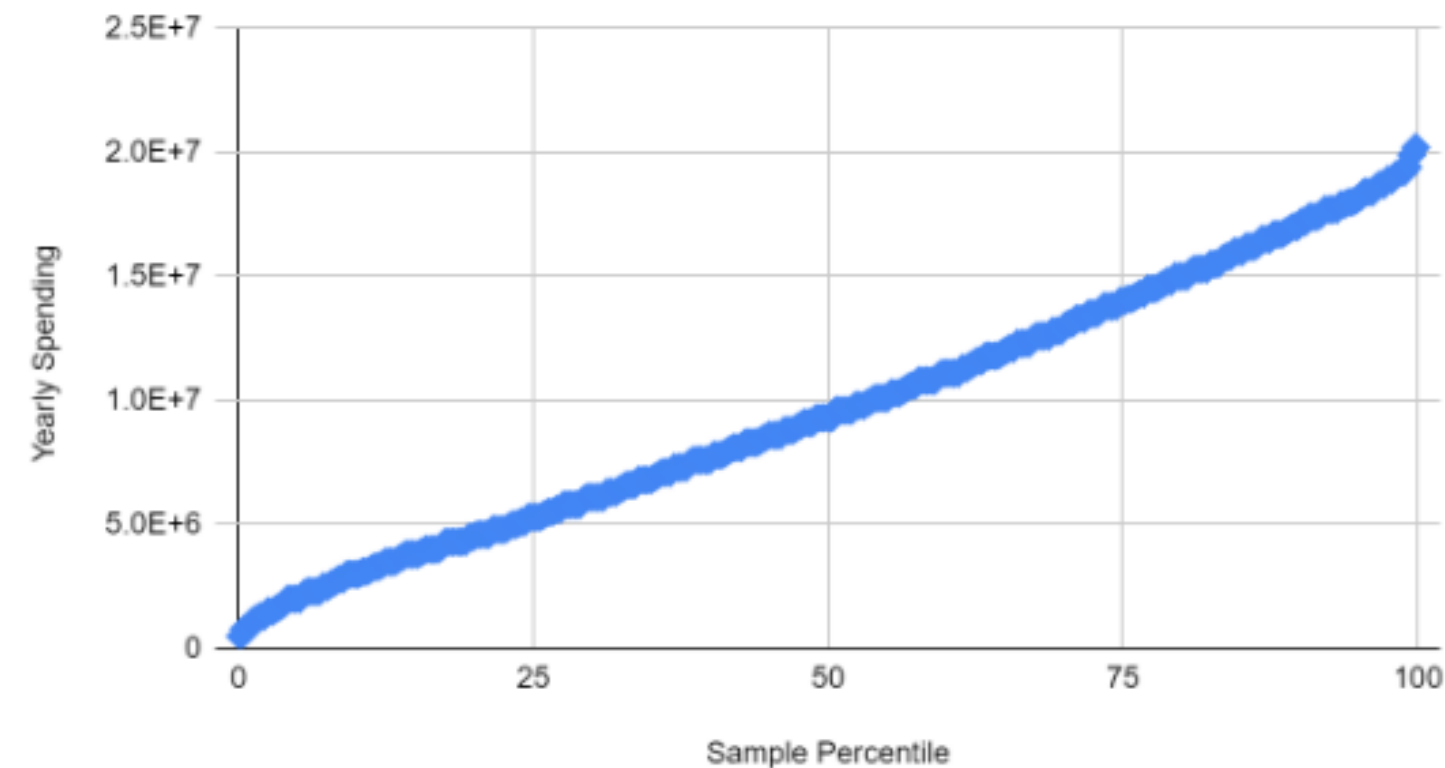
After ensuring that all independent variables are **not strongly correlated** and **significantly impacting the dependant variable** (in this case is yearly spending), we can form the linear regression formula using **coefficients** generated. It is also important to check our assumption of linear regression, namely checking **homocedasticity**, and **residual distribution**

Linear Regression

Variables	Coefficients
Intercept	2590241.49
Income	0.0368431741
Tenure	64576.21329
Avg. Session Time	6341.901705
Total Promo (%)	136147.7235
Bounce Rate	-49209.12635

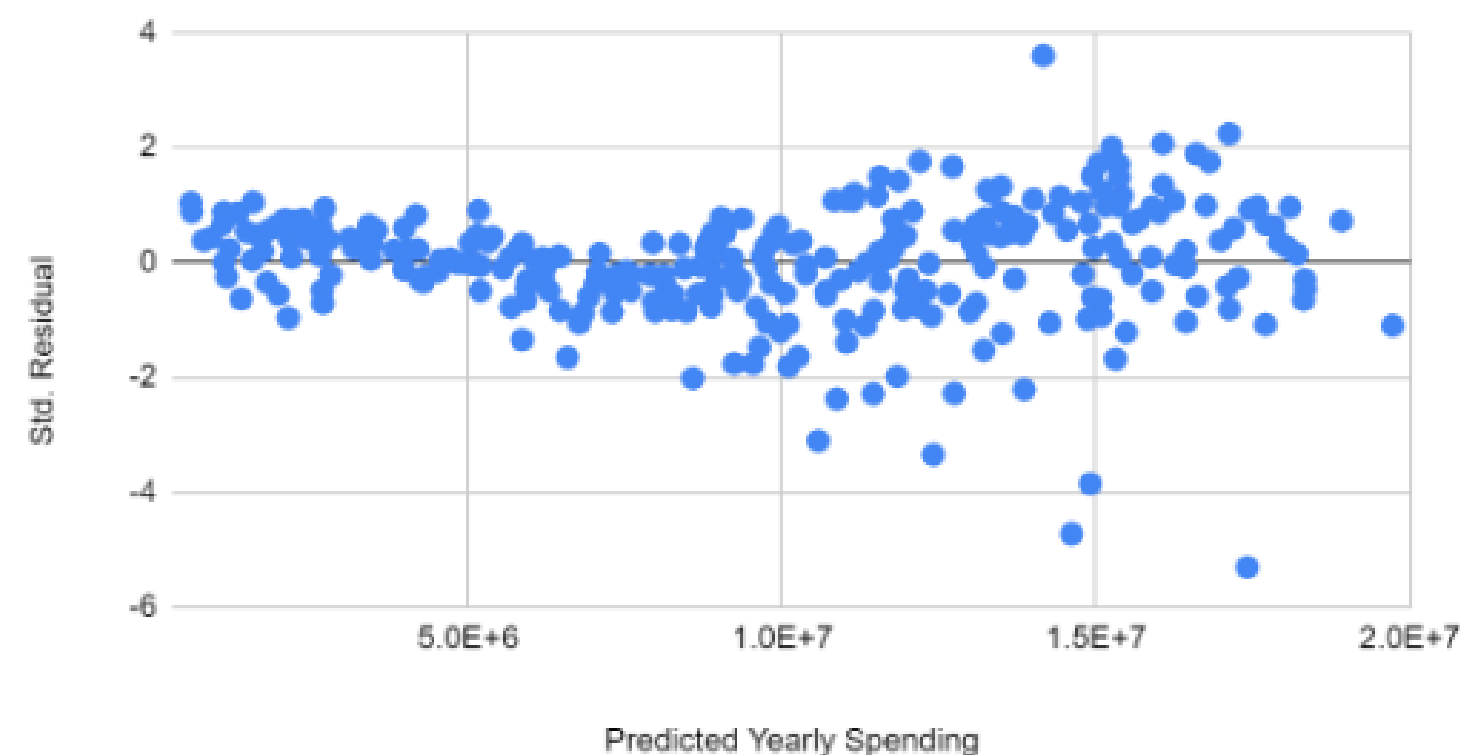
Linear Regression

Normal Probability Plot



Residual distribution can be checked through **Q - Q plot / normal probability plot**, which is the **dependent value (y) vs Sample Percentile**, if the plot is visually linear, then our assumption is probably right

Std. Residual VS Predicted Yearly Spending



Homoscedasticity, can be checked by **plotting standard residual value vs predicted yearly spending** generated from residual output of XLminer. The more scattered the plot is, the more condition it fits **heterocedasticity**, which is not good in terms of regression analysis.

Linear Regression

As stated before, these **P - values** are the **indicator of the significance of a variable**. The smaller the value, the more significance it gives.

On this term, **Avg. Session Time, Total Promo, and Bounce Rate P- Value are zero**, which indicates strong significance to the dependent variable, which is yearly spend.

Variable	P-value
Intercept	0.0000001460639
Income	0.001067575078
Tenure	0.000007363181736
Avg. Session Time	0
Total Promo (%)	0
Bounce Rate	0

Linear Regression

Mr. Faiz Fahmi as the Product Manager Team **wants to predict potential new customer who on-boarded about 1 month ago.** Therefore, he asked the Data Analyst team to predict the potential yearly spending based on the following characteristics:

Age	Income	Tenure	Avg. Session Time	Total Promo	Bounce Rate
20	20,000,000	1	1000	20	30

In order to predict this potential customer yearly spending, we must input above variable characteristics to the regression we formulated before.

New Customer to Predict :					
Age	Income	Tenure	Avg. Session Time	Total Promo	Bounce Rate
20	20000000	1	1000	20	30
y	2590241.49 + 0.037 * Income + 64576.2133 * Tenure + 6341.9017 * Avg. Session Time + 136147.7235 * Total Promo - 49209.1264 * Bounce Rate				
Yearly Spending	Rp10,980,264				
Income * Coefficient	736863.4821				
Tenure * Coefficient	64576.21329				
Avg. Session Time * Coefficient	6341901.705				
Total Promo * Coefficient	2722954.47				
Bounce Rate * Coefficient	-1476273.791				

Predicted Yearly Spending for this potential customer is :

Rp. 10.980.264



Some Recommendations

Top three variables that are strongly & significantly impacting the yearly spending of customers are :

- Avg. Session Time
- Total Promo
- Bounce Rate

- In order to increase yearly spending for customers, the company is highly recommended to **provide more promos** for the customers.
- On the other hand, **bounce rate are negatively affecting yearly spending**, therefore we should **analyze further to find the root cause of why bounce rate happens, and how to reduce them.**
- Avg. Session time, might indicates the relevancy of the **products with the customers**, assuming that we know which products are best selling, we can **convert more discount budget to those products**

Thank you !

Do you have
any questions?

Send it to us! We hope you
learned something new.

