REPORT

PRODUCT SALES ANALYSIS

DATACAMP

REQUEST REPORT

You written report should include written text summaries and graphics of the following:

Data validation:

- Describe validation and cleaning steps for every column in the data

Exploratory Analysis to answer the customer questions ensuring you include:

- Two different types of graphics showing single variables only
- At least one graphic showing two or more variables
- Description of your findings

Definition of a metric for the business to monitor

- How should the business monitor what they want to achieve?
- Estimate the initial value(s) for the metric based on the current data?

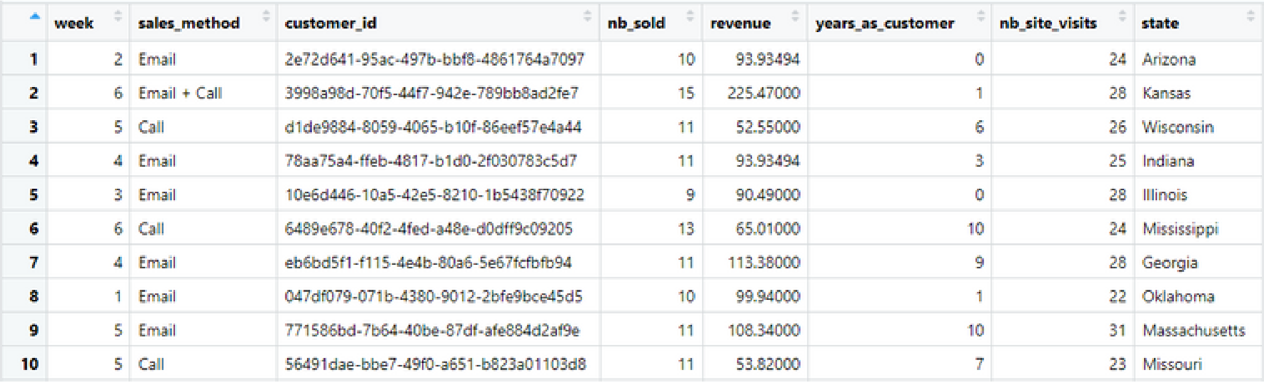
Final summary including recommendations that the business should undertake

Report by khanhhuyenthai

OVERVIEW **DATASET**

The original dataset contains:
product_sales**:** 15,000 rows and 8 columns. In dataset product_sales, there are 8 attributes, including:

week: week sale was made
sales_method: Selling method for customers (Email, Call, Email + Call)
customer_id: unique identifier for the customer
nb_sold: number of new products sold
revenue: revenue from the sales
years_as_customer: number of years customer has been buying from us (company founded in 1984)
nb_site_visits: number of times the customer has visited our website in the last 6 month
state: location of the customer i.e. where orders are shipped



	week	sales_method	customer_id	nb_sold	revenue	years_as_customer	nb_site_visits	state
1	2	Email	2e72d641-95ac-497b-bbf8-4861764a7097	10	93.93494	0	24	Arizona
2	6	Email + Call	3998a98d-70f5-44f7-942e-789bb8ad2fe7	15	225.47000	1	28	Kansas
3	5	Call	d1de9884-8059-4065-b10f-86eef57e4a44	11	52.55000	6	26	Wisconsin
4	4	Email	78aa75a4-ffeb-4817-b1d0-2f030783c5d7	11	93.93494	3	25	Indiana
5	3	Email	10e6d446-10a5-42e5-8210-1b5438f70922	9	90.49000	0	28	Illinois
6	6	Call	6489e678-40f2-4fed-a48e-d0dff9c09205	13	65.01000	10	24	Mississippi
7	4	Email	eb6bd5f1-f115-4e4b-80a6-5e67fcfbfb94	11	113.38000	9	28	Georgia
8	1	Email	047df079-071b-4380-9012-2bfe9bce45d5	10	99.94000	1	22	Oklahoma
9	5	Email	771586bd-7b64-40be-87df-afe884d2af9e	11	108.34000	10	31	Massachusetts
10	5	Call	56491dae-bbe7-49f0-a651-b823a01103d8	11	53.82000	7	23	Missouri

Figure 1.1: Fisrt 10 rows and several colomns of dataset product_sales

DATA VALIDATION

Before delving into the analysis process, data pre-processing steps, including cleaning and transformation, should be performed.

Firstly, check for missing values in the 'product_sales' dataset. Only the 'revenue' column has NA values. Replace these NA values with the mean of the 'revenue' column, rounded to two decimal places.

Secondly, the 'sales_method' column contains exceptional values such as 'em + call' and 'email'. Transform these exceptional values to the correct format: 'em + call' = 'Email + Call', and 'email' = 'Email'.

Thirdly, although the company was established in 1984, the 'year_as_customer' column has outlier values like 63 and 47 (which exceed the current year back to 1984). Remove these outlier values from the 'year_as_customer' column.

Dataset product_sales final có 14998 obs of 8 variables:

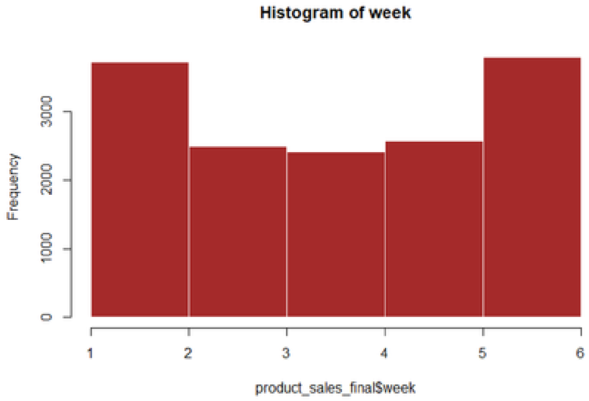
- **week:** ranges from 1 week to 6 weeks.
- **sales_method:** there are 3 sales methods (Email, Call, Email + Call)
- **customer_id:** unique identifier for the customer
- **nb_sold:** ranges from 7 to 16 products.
- **revenue:** ranges from $32.54 to $238.32.
- **years_as_customer:** ranges from 0 to 39 years.
- **nb_site_visits:** ranges from 12 to 41 visits in the last 6 months.
- **state:** 49 location of the customer i.e. where orders are shipped

EXPLORATORY **DATA ANALYSIS**

Week, year_as_customer, sales_method, state

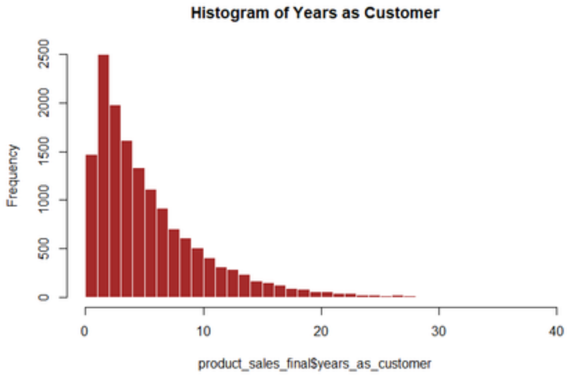
WEEK

From chart 3.1, weeks 1 and 6 are the two weeks with the highest sales quantities since the implementation of the new approaches



YEAR_AS_CUSTOMER

Chart 3.2 skews to the right, mostly focusing on customers with less than 10 years of shopping (new customers).

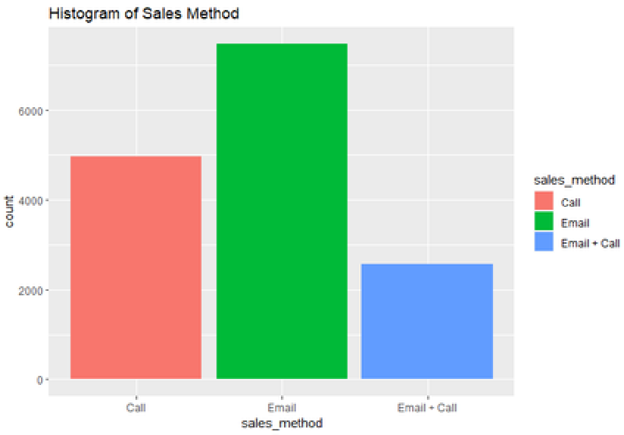


SALES_METHOD

Email: Customers received two emails, one at product launch and another three weeks later, requiring minimal team effort.

Call: Sales team members called customers, averaging thirty minutes per call.

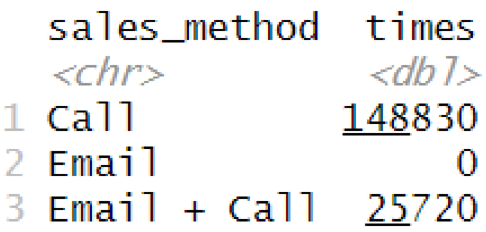
Email + Call: Customers got an initial email and a follow-up call a week later, with minimal email effort and a ten-minute call per customer.



In Chart 3.3, the company primarily utilizes the Email sales method with over 7000 occurrences, followed by Call with approximately 5000, and Email + Call with over 2500 occurrences.

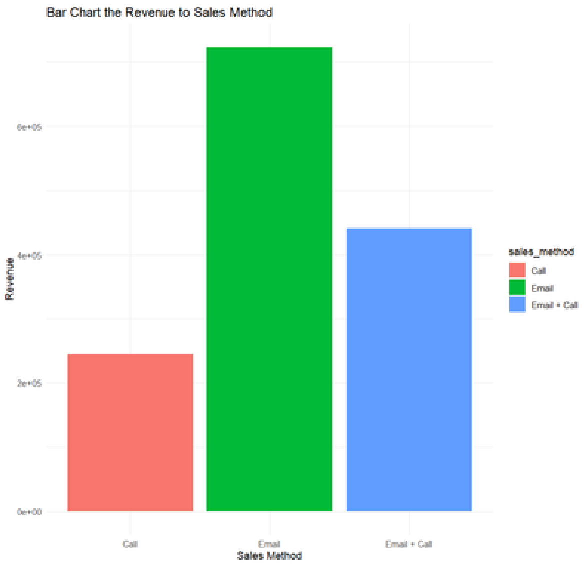
SALES_METHOD (REVENUE AND SALES)

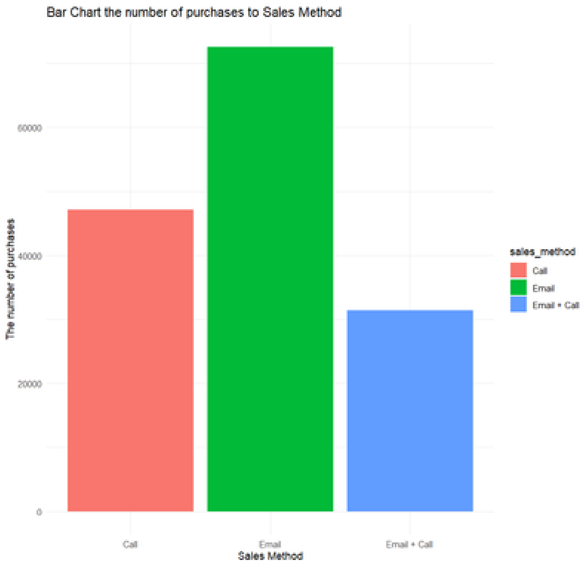
From charts 3.4 and 3.5, it is observed that the sales volume of Call is greater than Email + Call, but the revenue from Call is less than Email + Call. It's noticed that the company invests the most time in Call, but the effectiveness (as measured by revenue) is the lowest.



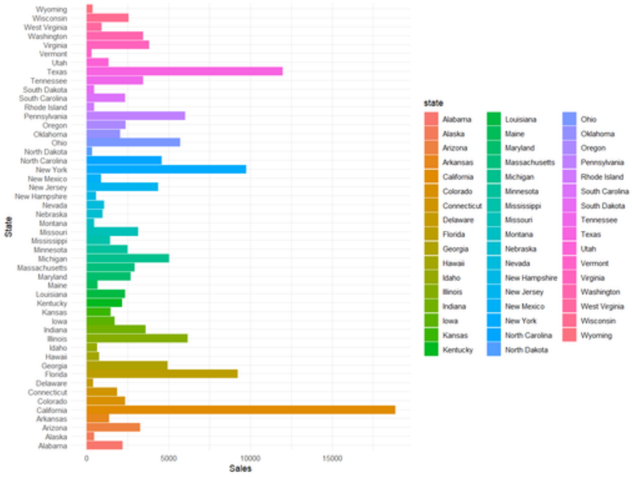
```
  sales_method   times
  <chr>          <dbl>
1 Call          148830
2 Email              0
3 Email + Call   25720
```

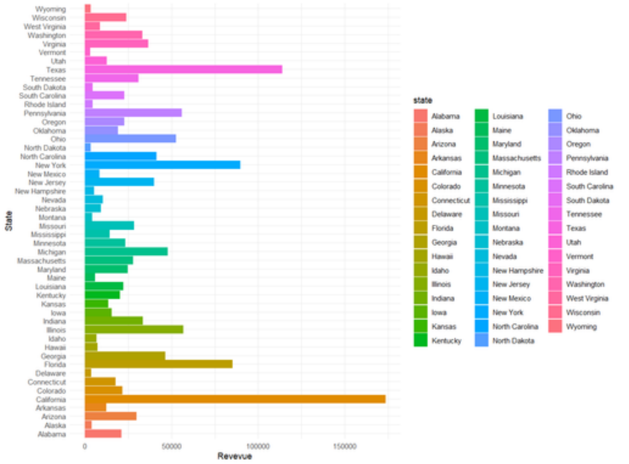
Suggestion: The company may consider eliminating the Call-only sales method.





STATE (REVENUE AND SALES)

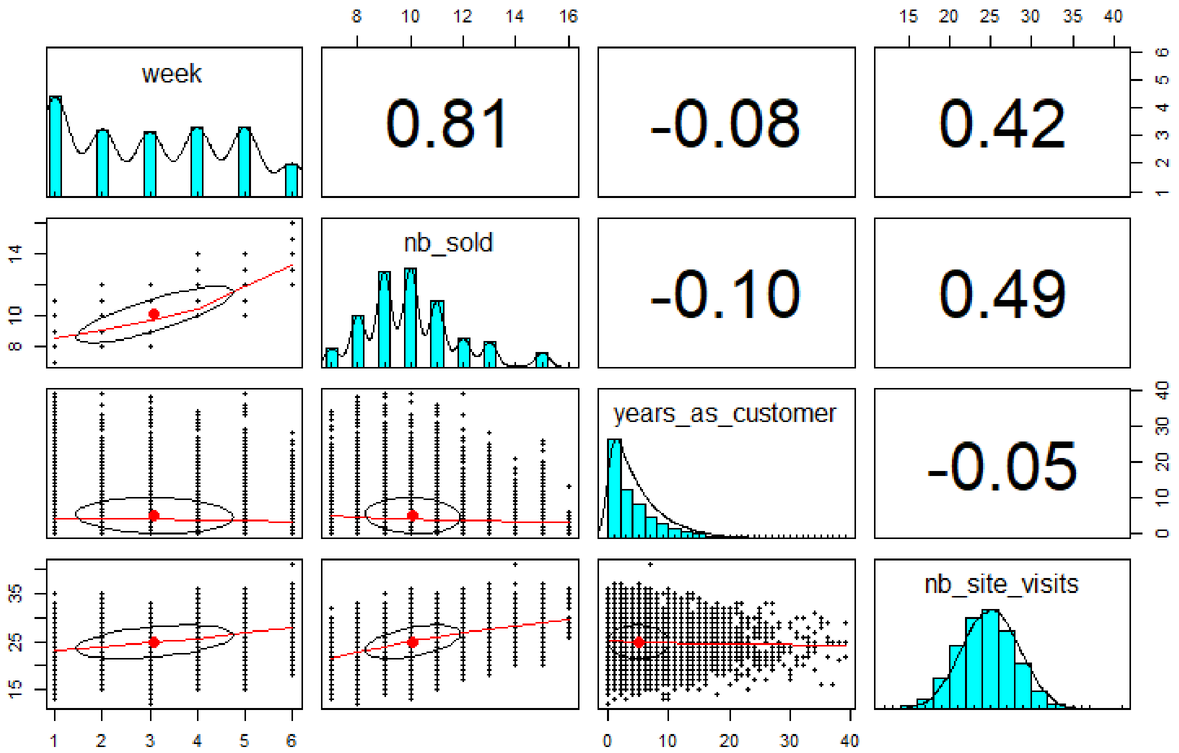




From charts 3.6 and 3.7, it is evident that sales volume is proportional to revenue. The top 4 states with the highest revenue are California, Texas, New York, and Florida.

DIAGNOSTIC **ANALYSIS**

Observation reveals a high correlation among the variables week, nb_sold, and nb_site_visit. Now, considering the use of PCA to analyze and determine if it's possible to reduce the number of variables.

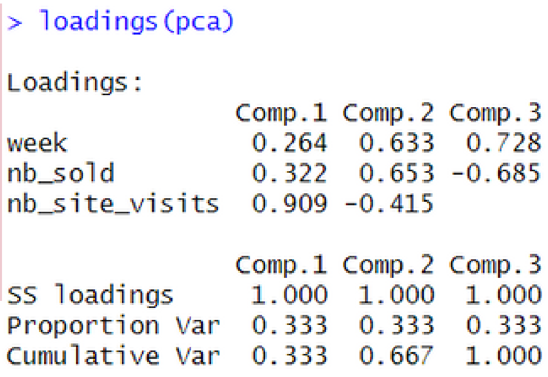


PHÂN TÍCH PCA

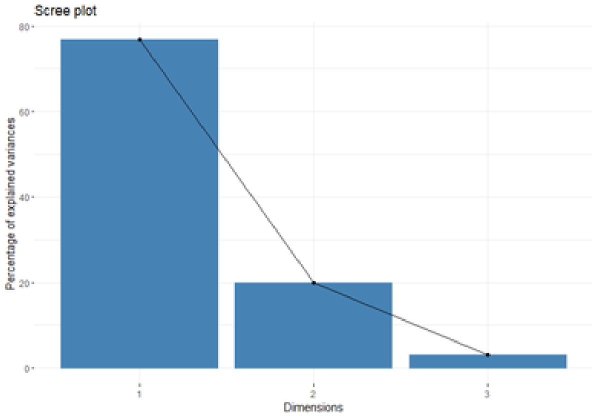
PRINT1 = 0.26*week + 0.32*nb_sold + 0.9*nb_site_visits

PRINT2 = 0.63*week + 0.65*nb_sold - 0.4*nb_site_visits

The contribution of PRINT1 is the highest, explaining nearly 80% of the data. Without reducing the dimensionality, all three variables are kept intact for further cluster analysis.



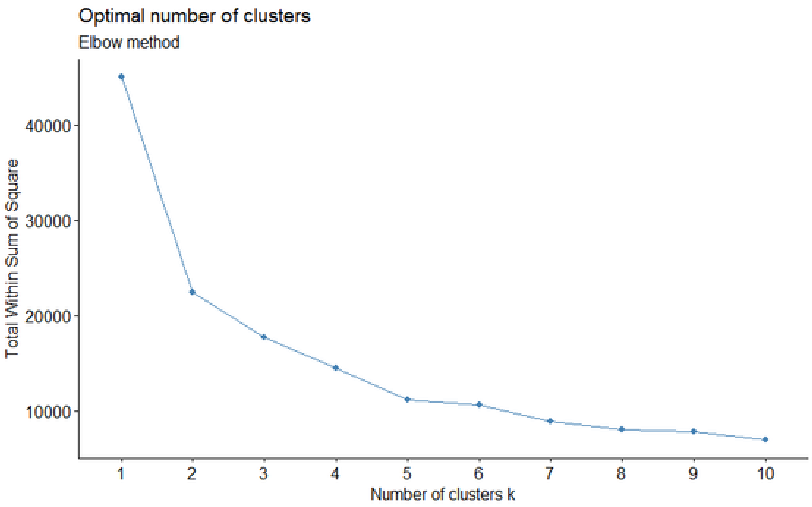
```
> loadings(pca)

Loadings:
               Comp.1 Comp.2 Comp.3
week            0.264  0.633  0.728
nb_sold         0.322  0.653 -0.685
nb_site_visits  0.909 -0.415       

               Comp.1 Comp.2 Comp.3
SS loadings     1.000  1.000  1.000
Proportion Var  0.333  0.333  0.333
Cumulative Var  0.333  0.667  1.000
```



CLUSTER **CUSTOMER**

The Hopkins index is approximately 0.86 (close to 1), meeting the conditions for cluster analysis.

ELBOW METHOD



Using two methods:

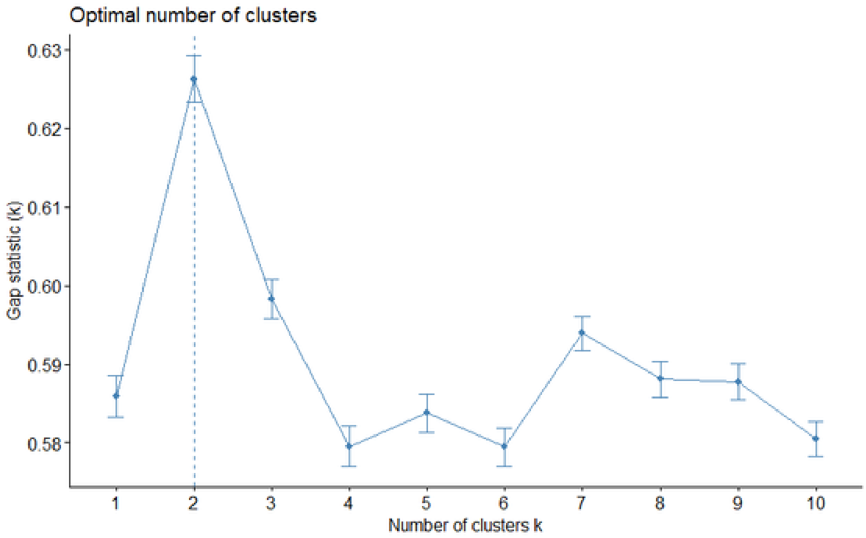
- Elbow Method
- Gap Statistic

to select the number of clusters for K-means analysis.

GAP STATISTIC

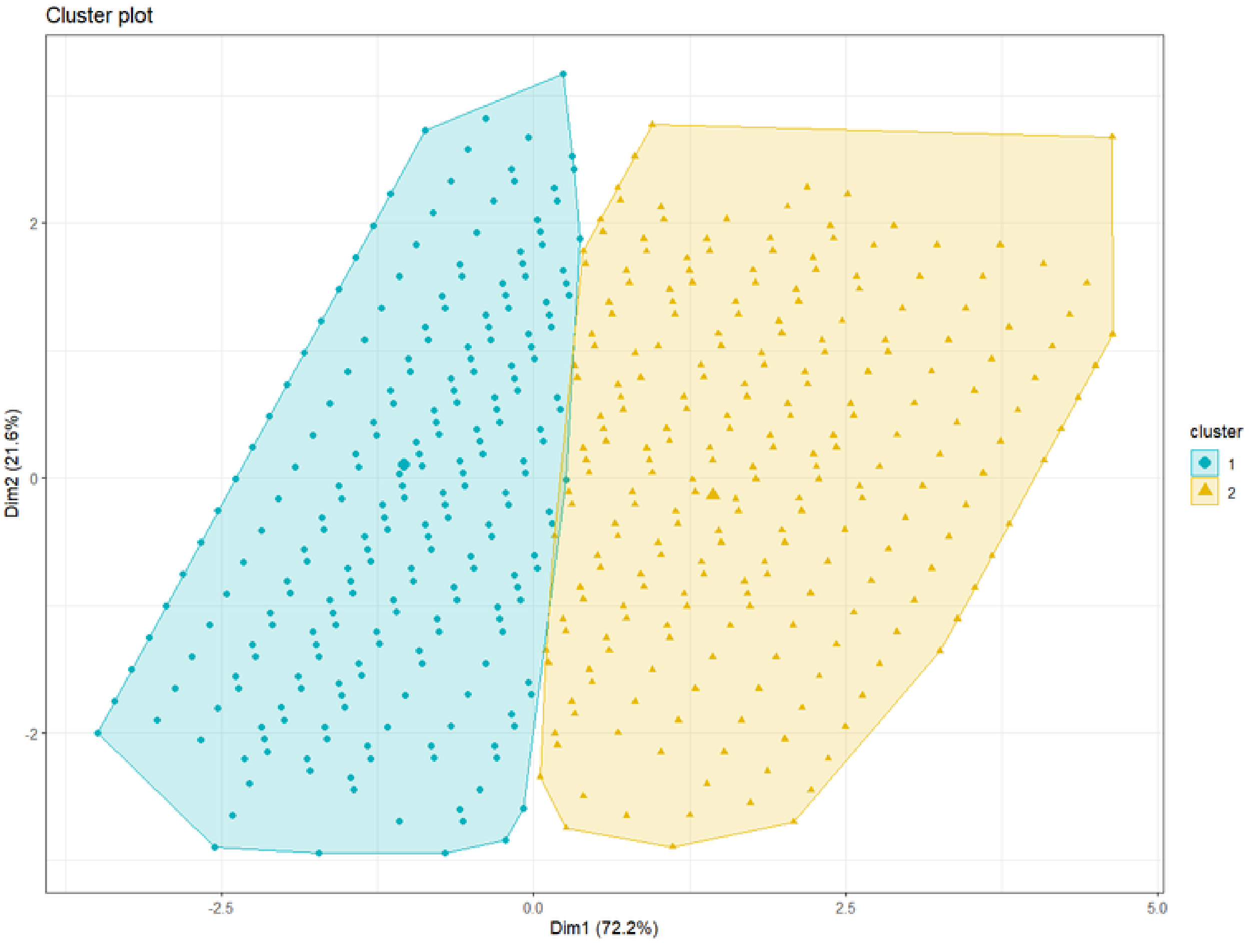
From the two methods, the optimal number of clusters, k, is determined to be 2 (k = 2).

Therefore, customers will be grouped into 2 main clusters.



COMPUTATION **K-MEANS**

Using K-Means with k = 2, we observe that with 2 explanatory dimensions, such as dim1 and dim2, with explanatory values of 72.2% and 21.6%, the variation in the data is predominantly explained by dim1 (72.2%) and to a lesser extent by dim2 (21.6%). In summary, the 2 explanatory dimensions collectively explain 93.8% of the Within-Cluster Sum of Squares (WSS).


Cluster plot
Dim2 (21.6%)
Dim1 (72.2%)
cluster