FUTURE ECOMMERCE DATASET

KATARINA NIMAS KUSUMAWATI | FUTURE BATCH V
TRACK DATA

OUTLINE

01 Data Overview

02 OLTP

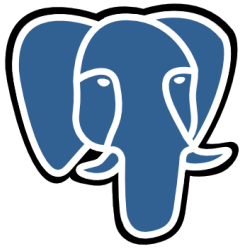
03 Data Warehouse

04-08 Business Questions

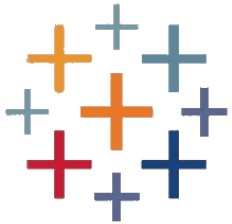
09 Supervised Learning

10 Unsupervised Learning

TOOLS



PostgreSQL



+ab|eau

talend

01

DATA OVERVIEW

Feedback Dataset

- This dataset contains data about customer feedback.
- Total row: 100000

Name	Description
feedback_id	Feedback form unique identifier
order_id	Unique order identifier
feedback_score	Rating
feedback_form_sent_date	Timestamp that survey was sent to the customer
feedback_answer_date	Feedback answered timestamp

Order Dataset

- This dataset contains orders received.
- Total row: 99441
- Total order_id: 99441

Name	Description
order_id	Unique identifier of the order
user_name	Key to the user dataset
order_status	Order status
order_date	Purchase timestamp
order_approved_date	Shows the payment approval timestamp
pickup_date	Timestamp when it was handled to the logistic partner
delivered_date	Actual order delivery date to the customer
estimated_time_delivery	Estimated delivery date that was informed to customer at the purchase moment

Order Item Dataset

- This dataset contains data about the items purchased on each order.
- Total row: 112650

Name	Description
order_id	Unique identifier of the order
order_item_id	Unique identifier of the order item
product_id	Product unique identifier
seller_id	Seller unique identifier
pickup_limit_date	Shows the seller limit date for handling the order over to the logistic partner
price	Item price
shipping_cost	Shipping cost

Payment Dataset

- This dataset includes order payment options.
- Total row: 103886

Name	Description
order_id	Unique identifier of the order
payment_sequential	Unique identifier of the payment
payment_type	Method of payment chosen by the customer
payment_installments	Number of installments chosen by the customer
payment_value	Total order amount

Products Dataset

- This dataset includes data about the product sold.
- Total row: 32951
- Total product_id: 32951

Name	Description
product_id	Unique product identifier
product_category	Root category of product
product_name_length	Number of characters of the product name
product_description_length	Number of characters of the product descripton
product_photos_qty	Number of product photos
product_weight_g	Product weight measure in grams
product_length_cm	Product length measure in centimeters
product_height_cm	Product height measure in centimeters
product_width_cm	Product width measure in centimeters

Seller Dataset

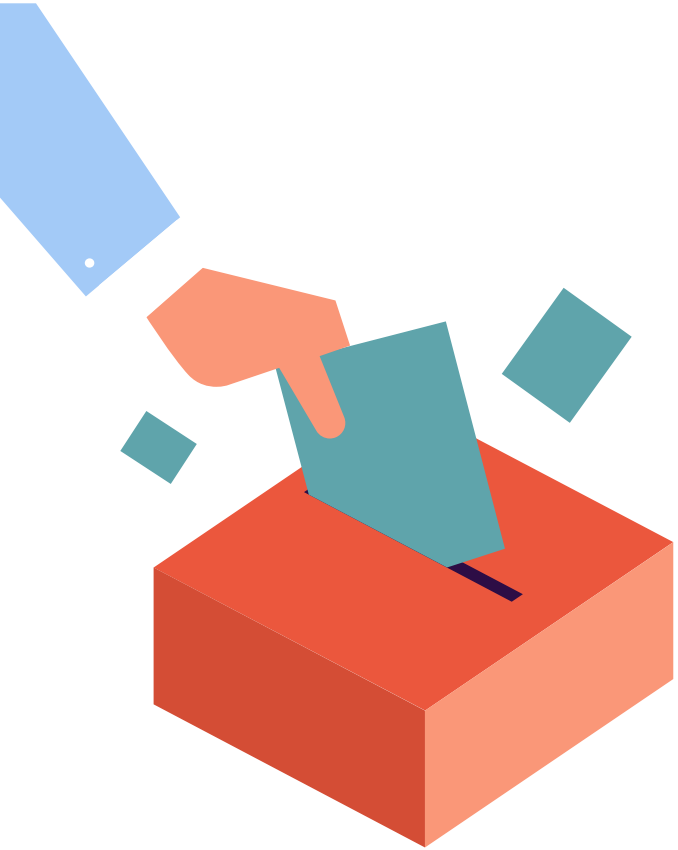
- This dataset includes data about sellers who fulfill orders.
- Total row: 3095
- Total seller_id: 3095

Name	Description
seller_id	Seller unique identifier
seller_zip_code	Zip code
seller_city	City
seller_state	Province

User Dataset

- This dataset includes information about the customer and the location of the customer.
- Total row: 99441
- Total user_id: 96096

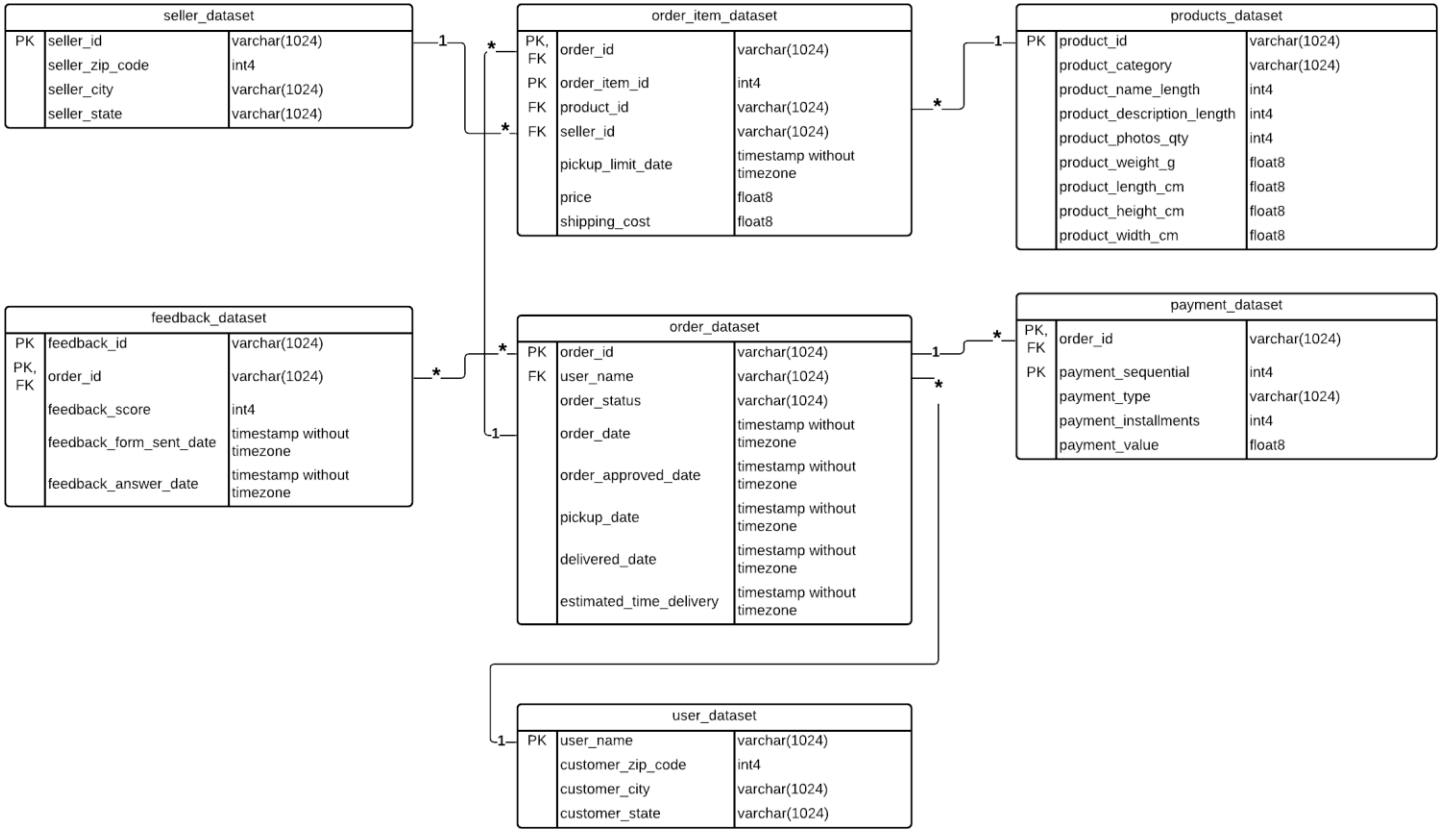
Name	Description
user_name	Unique identifier of a user
customer_zip_code	Zip code
customer_city	City
customer_state	Privince



02

OLTP

ERD OLTP


seller_dataset
PK seller_id varchar(1024)
seller_zip_code int4
seller_city varchar(1024)
seller_state varchar(1024)
order_item_dataset
PK, FK order_id varchar(1024)
PK order_item_id int4
FK product_id varchar(1024)
FK seller_id varchar(1024)
pickup_limit_date timestamp without timezone
price float8
shipping_cost float8
products_dataset
PK product_id varchar(1024)
product_category varchar(1024)
product_name_length int4
product_description_length int4
product_photos_qty int4
product_weight_g float8
product_length_cm float8
product_height_cm float8
product_width_cm float8
feedback_dataset
PK feedback_id varchar(1024)
PK, FK order_id varchar(1024)
feedback_score int4
feedback_form_sent_date timestamp without timezone
feedback_answer_date timestamp without timezone
order_dataset
PK order_id varchar(1024)
FK user_name varchar(1024)
order_status varchar(1024)
order_date timestamp without timezone
order_approved_date timestamp without timezone
pickup_date timestamp without timezone
delivered_date timestamp without timezone
estimated_time_delivery timestamp without timezone
payment_dataset
PK, FK order_id varchar(1024)
PK payment_sequential int4
payment_type varchar(1024)
payment_installments int4
payment_value float8
user_dataset
PK user_name varchar(1024)
customer_zip_code int4
customer_city varchar(1024)
customer_state varchar(1024)

03

DATA WAREHOUSE

DimFeedback

- **Remove Invalid Value**

feedback_dataset to order_dataset is many-to-many. To create a feedback dimension (DimFeedback), feedback_dataset is ranked based on the newest feedback_answer_date to the oldest and only kept the newest feedback_answer_date.

feedback_id	order_id	feedback_score	feedback_form_sent_date	feedback_answer_date	rank
XX	AA	5	12:00:00	12:00:00	2
YY	AA	3	13:00:00	13:00:00	1
YY	BB	3	13:00:00	13:00:00	1

feedback_id	order_id	feedback_score	feedback_form_sent_date	feedback_answer_date
YY	AA	3	13:00:00	13:00:00
YY	BB	3	13:00:00	13:00:00

DimFeedback (cont)

- **Drop order_id**

So to join the fact, only rely on feedback_id.

feedback_id	order_id	feedback_score	feedback_form_sent_date	feedback_answer_date
YY	AA	3	13:00:00	13:00:00
YY	BB	3	13:00:00	13:00:00

feedback_id	feedback_score	feedback_form_sent_date	feedback_answer_date
YY	3	13:00:00	13:00:00
YY	3	13:00:00	13:00:00

- **Remove Duplicate Value**

DimFeedback (cont)

- **Add Surrogate Key**

Added surrogate key with data type int.

feedback_id	feedback_score	feedback_form_sent_date	feedback_answer_date
YY	3	13:00:00	13:00:00

feedback_key	feedback_id	feedback_score	feedback_form_sent_date	feedback_answer_date
1	YY	3	13:00:00	13:00:00

DimCustomer

- **Remove Invalid Value**

In the user_dataset, 1 username has many addresses so user_name is not unique. To create the customer dimension (DimCustomer), query rank the user_dataset, and ranked based on customer_zip_code, customer_state, customer_city

user_name	customer_zip_code	customer_city	customer_state	rank
AA	140	SURABAYA	JAWA TIMUR	1
BB	140	SURABAYA	JAWA TIMUR	3
BB	150	DENPASAR	BALI	2
BB	160	SOLO	JAWA TENGAH	1

user_name	customer_zip_code	customer_city	customer_state
AA	140	SURABAYA	JAWA TIMUR
BB	160	SOLO	JAWA TENGAH

DimCustomer (cont)

- **Uppercase**

Uppercase on customer_state and customer_city. Most of them have been uppercase from the beginning, but to prevent any values that have not been uppercase.

user_name	customer_zip_code	customer_city	customer_state
AA	140	surabaya	jawa timur
BB	160	SOLO	JAWA TENGAH

user_name	customer_zip_code	customer_city	customer_state
AA	140	SURABAYA	JAWA TIMUR
BB	160	SOLO	JAWA TENGAH

DimCustomer (cont)

- **Add Surrogate Key**

Added a surrogate key with data type int.

user_name	customer_zip_code	customer_city	customer_state
AA	140	SURABAYA	JAWA TIMUR
BB	160	SOLO	JAWA TENGAH

customer_key	user_name	customer_zip_code	customer_city	customer_state
1	AA	140	SURABAYA	JAWA TIMUR
2	BB	160	SOLO	JAWA TENGAH

DimProducts

- **Fix Wrong Attribute Naming**

In the DimProducts table, there is a typo in the table name.
Correction:
product_name_lenght → product_name_length
product_description_lenght → product_description_length

- **Replace “_” with “ “ in product_category Value**

Replace “_” with “ “ in product_category value.
Example: bed_bath_table → bed bath table

DimProducts (cont)

- **Add Row for Order that doesn’t have product_id**

product_id = 0
product_category = UNKNOWN
product_name_length = 0
product_description_length = 0
product_photos_qty = 0
product_weight_g = 0
product_length_cm = 0
product_height_cm = 0
product_width_cm = 0

- **Sort product_id**

Sort product_id. This is to make it easier to assign surrogate key where the surrogate key is 0 for product_id which null in fact. So that product_id 0 can exist in the first row and be given a surrogate key 0 as well.

DimProducts (cont)

- **Add Surrogate key**

Surrogate key are given using data type int.

- **Uppercase product_category**

Uppercase the product_category.

- **Fill NULL with 0**

Fill 0 in the product_weight_g, product_length_cm, product_height_cm, and product_width_cm column which have NULL values.

- **Make product_volume_cm3 column**

Create product_volume_cm3 column by multiplying product_length_cm, product_height_cm, and product_width_cm.

DimSeller

- **Uppercase**

Uppercase on seller_state and seller_city. Most of them have been uppercase from the beginning, but to prevent any values that have not been uppercase.

seller_id	seller_zip_ code	seller_city	seller_state
AA	140	surabaya	jawa timur
BB	160	SOLO	JAWA TENGAH

seller_id	seller_zip_ code	seller_city	seller_state
AA	140	SURABAYA	JAWA TIMUR
BB	160	SOLO	JAWA TENGAH

- **Add Row for Order that doesn’t have seller_id**

seller_id = 0
seller_zip_code = 0
seller_city = UNKNOWN
seller_state = UNKNOWN

DimSeller (cont)

- **Sort seller_id**

This is to make it easier to assign surrogate key where the surrogate key is 0 for seller_id which null in fact. So that seller_id 0 can exist in the first row and be given a surrogate key 0 as well.

- **Add Surrogate Key**

Added surrogate key with data type int.

seller_id	seller_zip_code	seller_city	seller_state
AA	140	SURABAYA	JAWA TIMUR
BB	160	SOLO	JAWA TENGAH

seller_key	seller_id	seller_zip_code	seller_city	seller_state
1	AA	140	SURABAYA	JAWA TIMUR
2	BB	160	SOLO	JAWA TENGAH

DimDate

- **Generate Date**

Generate date with the format yyyy-MM-dd (example: 2021-07-30).

- **Make Date Column**

Column	Description
date_id	Unique identifier for date (2021-07-30)
DayNumberOfWeek	The week on that date (Week 1-4)
DayNumberOfMonth	The day of the month (1-31)
DayNumberOfYear	The day of the year (1-366 atau 1-365)
MonthYear	The month with data type varchar (January, February, etc)
MonthNumberOfYear	The month with data type int (1-12)
CalendarQuarter	The quarter on that date (2016-2018)
CalendarYear	The year on that date (2016-2020)
CalendarSemester	The semester on that date (1-2)
DayName	Name of the day (Monday-Sunday)

DimDate (cont)

- **Uppercase**

Uppercase column with data type varchar in MonthYear and DayName column

- **Add Surrogate Key**

Add surrogate key with the format yyyyMMdd.
Example: 2021-07-30 → 20210730

DimTime

- **Generate Time**

Generate date with the format HH:mm:ss (example: 07:00:00).

- **Make Time Column**

Column	Description
time_id	Unique identifier for time (07:00:00)
hour	Hour (0-23)
minute	Minute (0-59)
second	Second (0:59)
meridiem	AM/PM
time_of_day	MORNING, AFTERNOON, EVENING

DimTime (cont)

- **Add Surrogate Key**

Add surrogate key

time_key	time_id	hour	minute	second	meridiem	time_of_day
0	00:00:00	0	0	0	AM	MORNING
1	00:00:01	0	0	1	AM	MORNING
2	00:00:02	0	0	2	AM	MORNING

FactOrder

- **Outer Join order_dataset and order_item_dataset**

Outer join between order_dataset and order_item_dataset.

- **Left Join payment_dataset**

In payment_dataset, there are several process carried out in it.

order_id	payment_sequential	payment_type	payment_installments	payment_value
AA	1	credit_card	1	100000
AA	2	debit_card	1	100000

FactOrder (cont)

order_id	count_blipay	total_blipay	count_credit_card	total_credit_card	count_debit_card	total_debit_card	count_voucher	total_voucher	count_not_defined	total_not_defined	payment_sequential	payment_installments	total_payment_value
AA	0	0	1	100000	1	100000	0	0	0	0	1	2	200000

FactOrder (cont)

Name	Aggregate
count_blipay	Sum
total_blipay	Sum
count_credit_card	Sum
total_credit_card	Sum
count_debit_card	Sum
total_debit_card	Sum
count_voucher	Sum
total_voucher	Sum
count_not_defined	Sum
total_not_defined	Sum
payment_sequential	Max
payment_installements	Sum
total_payment_value	Sum

FactOrder (cont)

- **Insert feedback_id in FactOrder**

In FactOrder there is no feedback_id yet. So that DimFeedback can be joined with FactOrder is insert the feedback_id.

- **Make total_payment column**

This column is price+shipping cost. This is necessary because the granularity used is per order item, so a total payment with granularity per order item is required.

- **Handling NULL**

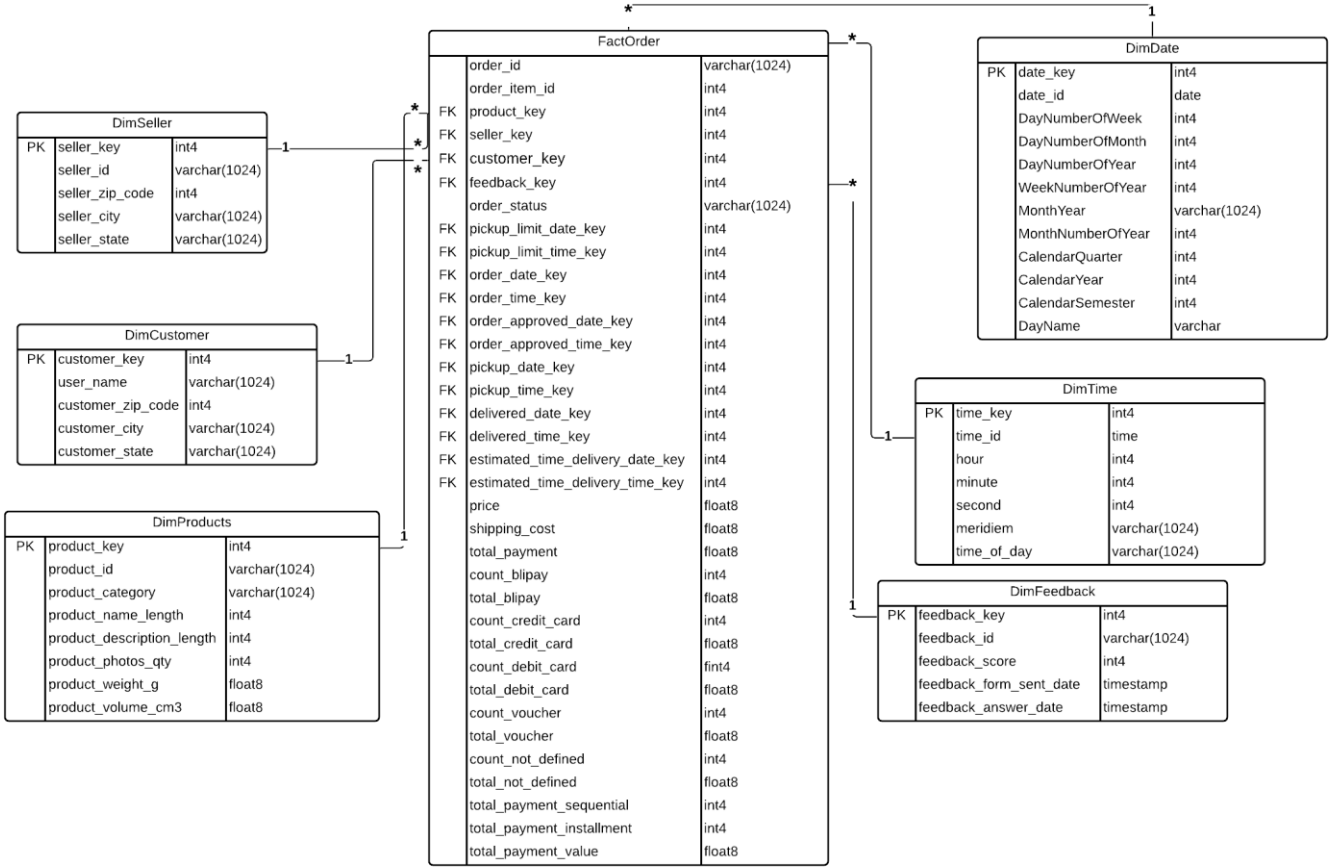
Replace NULL with 0 in the order_item_id, product_id, seller_id, price, shipping_cost, and all payment columns.

- **Uppercase**

Uppercase on column order_status.

- **Replace Natural Key into Surrogate Key**

ERD DATAWAREHOUSE



04

BUSINESS QUESTIONS 1

BUSINESS QUESTIONS 1

How are payment types, product categories, and sales used in each region?

BACKGROUND

- Reviewing the distribution of data from the customer side
- So we can find out:

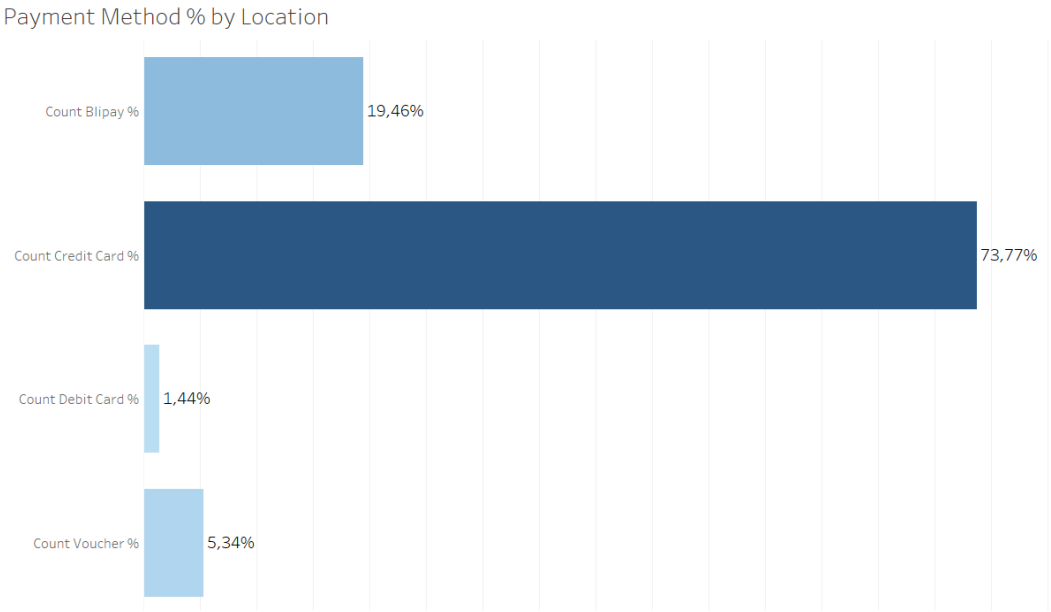
1. What are the most frequently used payment methods
2. What product categories are often purchased and make a big profit
3. Which provinces are quite consumptive

PURPOSE

Knowing the initial description of the data so we can dig deeper about the things that need to be asked in data.

HYPOTHESIS 1: PAYMENT METHODS DONE IN EACH REGION HAVE DIFFERENT TRENDS

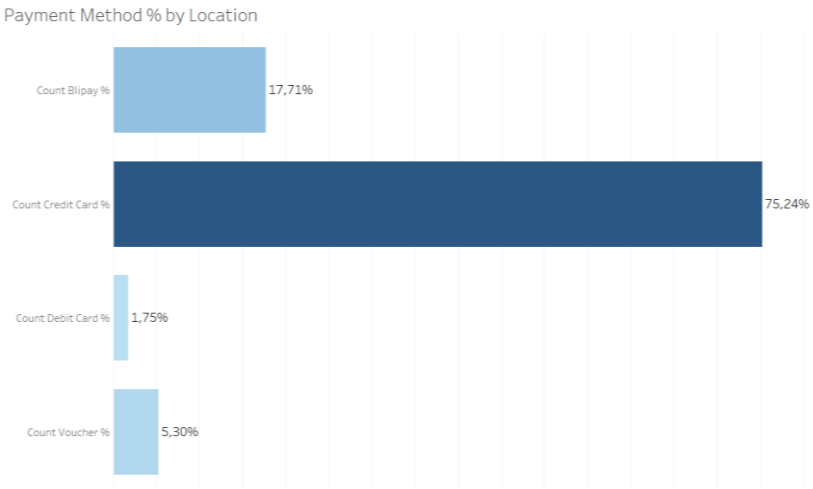
All Province



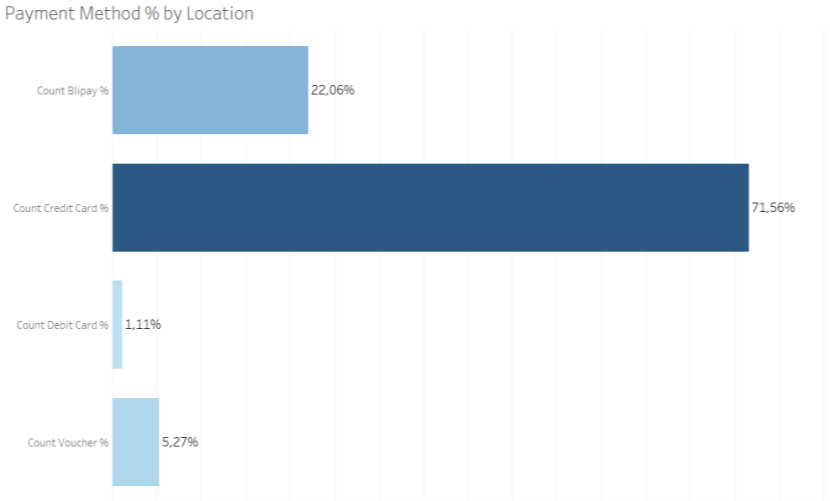
The most **widely used** payment method is a **credit card** with a total transaction of **73,77%** from all transactions.

HYPOTHESIS 1: PAYMENT METHODS DONE IN EACH REGION HAVE DIFFERENT TRENDS

Banten



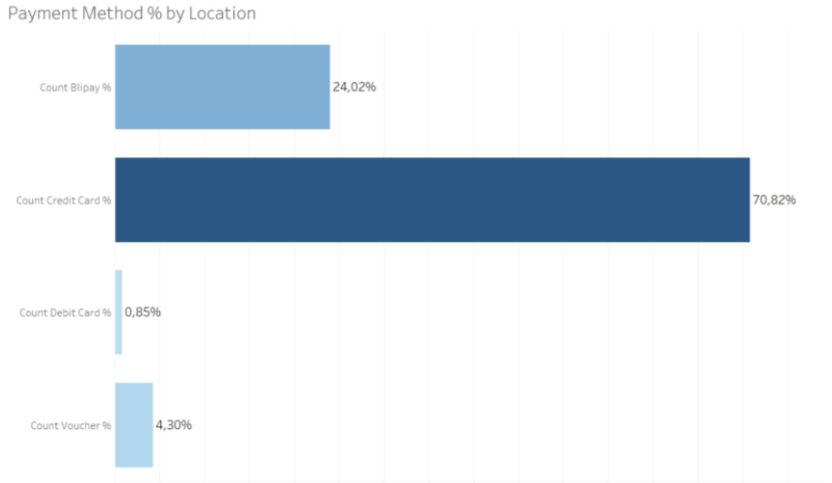
North Sumatera



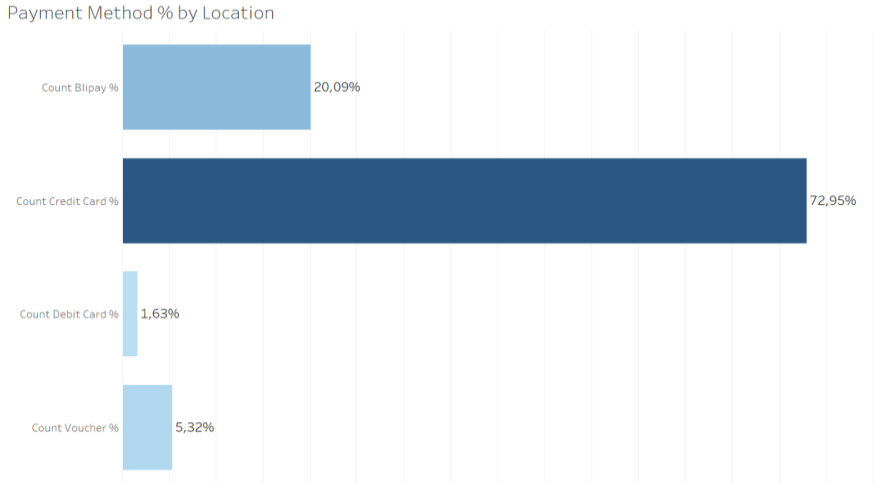
1. The most **widely used** payment method in **Banten province** is **credit card** with a total of **75,24%** transactions.
2. The most **widely used** payment method in **North Sumatera province** is **credit card** with a total of **71,56%** transactions.

HYPOTHESIS 1: PAYMENT METHODS DONE IN EACH REGION HAVE DIFFERENT TRENDS

East Kalimantan



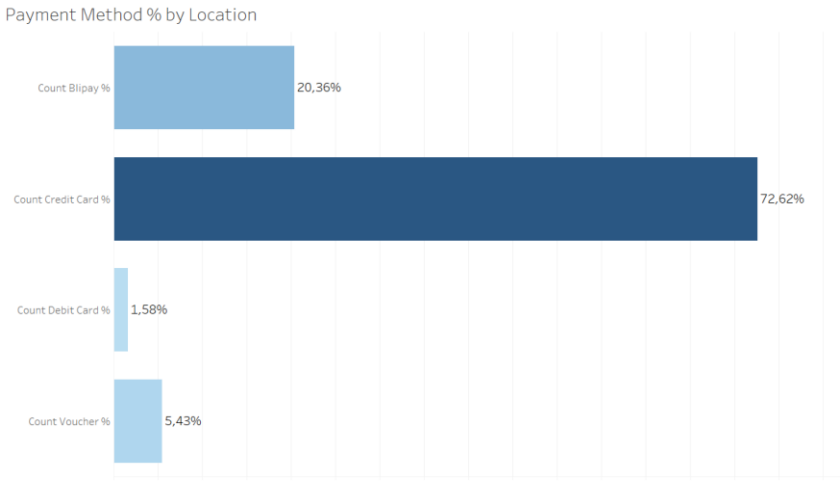
South Sulawesi



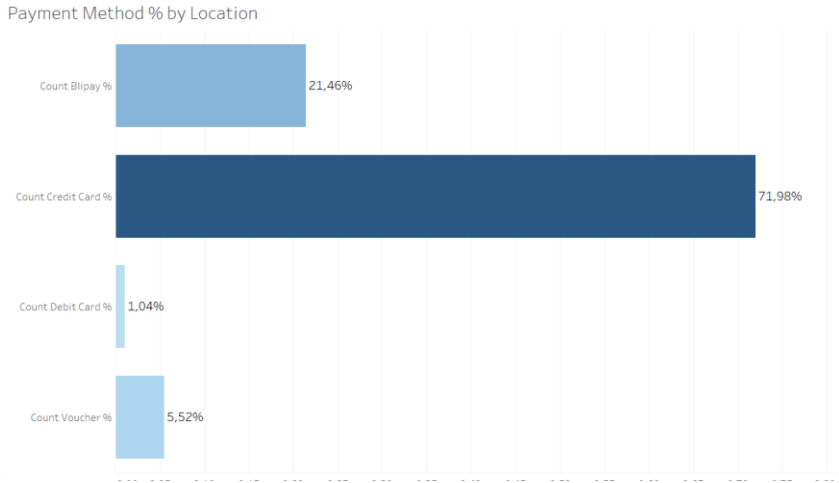
1. The most **widely used** payment method in **East Kalimantan province** is **credit cards** with a total of **70,82%** transactions.
2. The most **widely used** payment method in **South Sulawesi province** is **credit cards** with a total of **72,95%** transactions.

HYPOTHESIS 1: PAYMENT METHODS DONE IN EACH REGION HAVE DIFFERENT TRENDS

West Nusa Tenggara

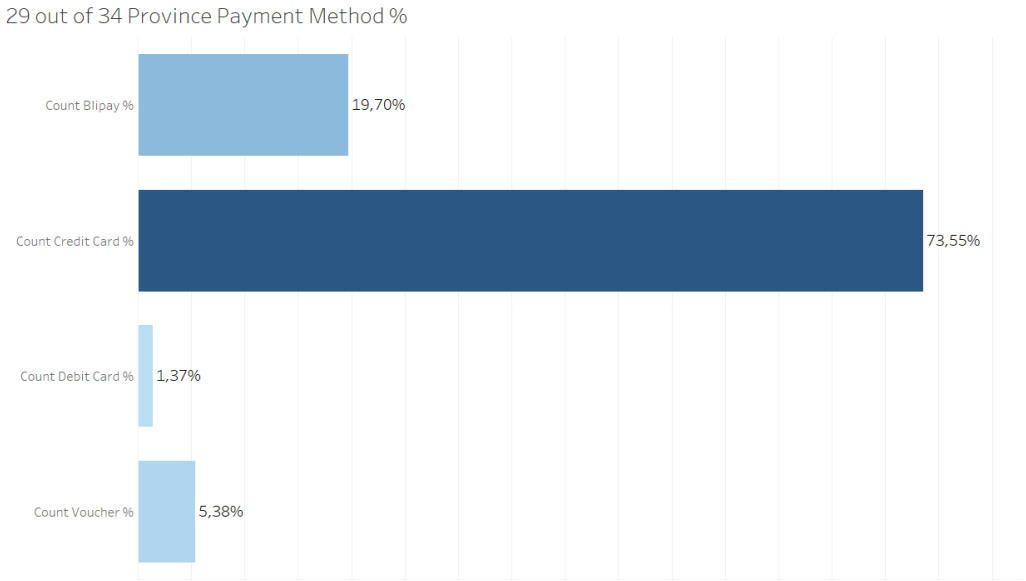


Papua



1. The most **widely used** payment method in the **West Nusa Tenggara province** is **credit cards** with a total of **72,62%** transactions.
2. The most **widely used** payment method in **Papua province** is **credit cards** with a total of **71,98%** transactions.

HYPOTHESIS 1: PAYMENT METHODS DONE IN EACH REGION HAVE DIFFERENT TRENDS

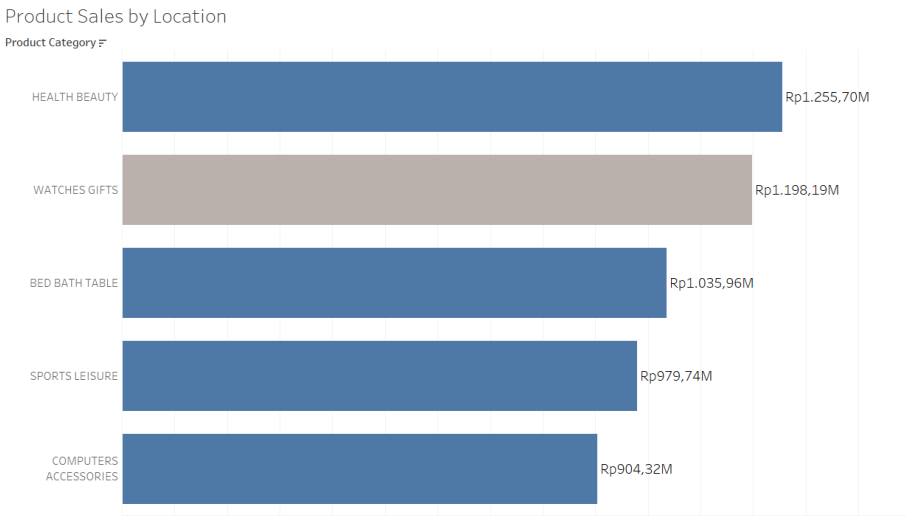


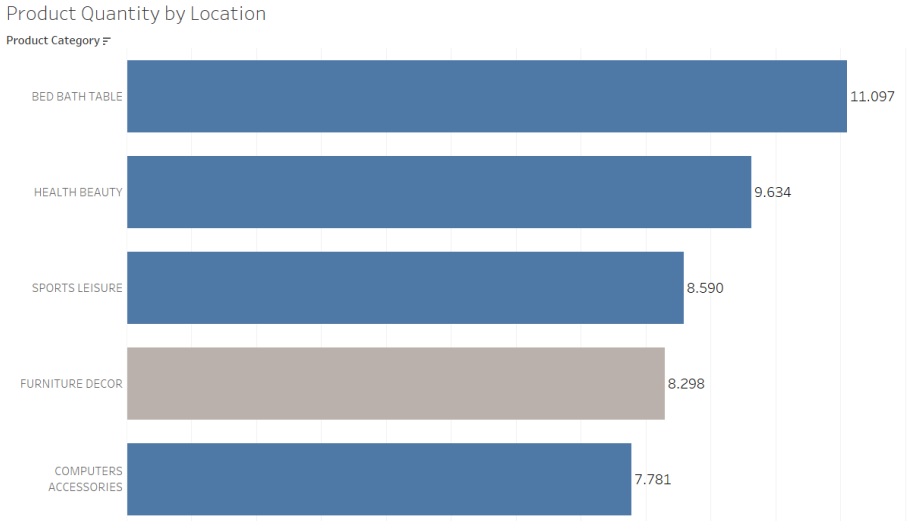
- The most **widely used** payment method in **29 of 34 provinces** is **credit cards** with a total of **73,55%** transactions.
- From **all provinces, 6 regions, and 29 of 34 provinces**, it was found that the most frequently used payment method was **credit cards**.
- For the next step, it will be explored more deeply why this phenomenon can occur in **business question no 2**

HYPOTHESIS 2: CONSUMPTION NEEDS ARE THE MOST SELLING PRODUCT CATEGORIES IN EVERY REGION

All Province

Category products that are in the top 5 highest selling and most sold are marked with a blue bar



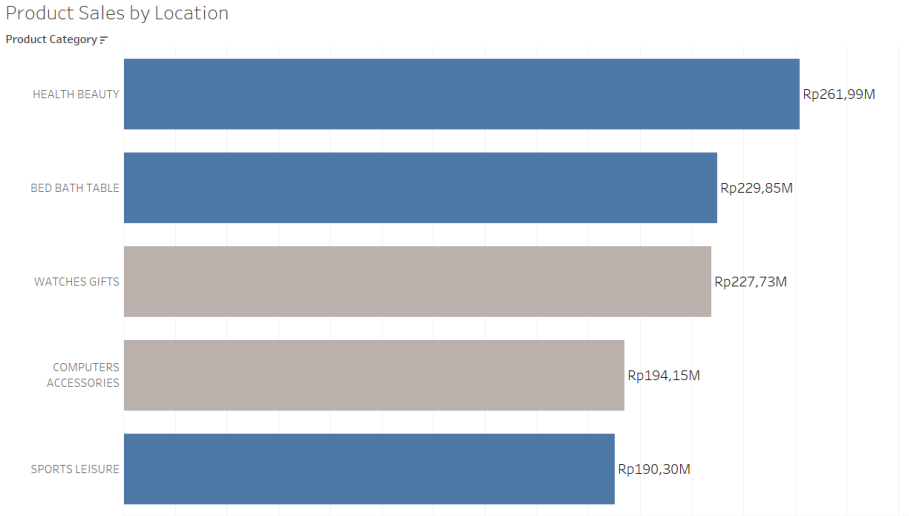


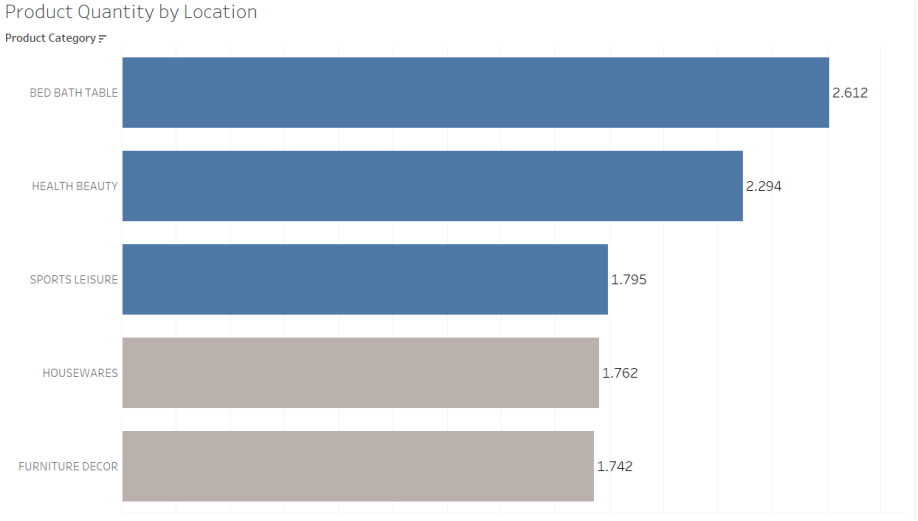
Product categories that are **frequently purchased** and have the **highest overall sales** include **bed bath tables, health beauty, computers accessories,** and **sports leisure**.

HYPOTHESIS 2: CONSUMPTION NEEDS ARE THE MOST SELLING PRODUCT CATEGORIES IN EVERY REGION

Banten

Category products that are in the top 5 highest selling and most sold are marked with a blue bar



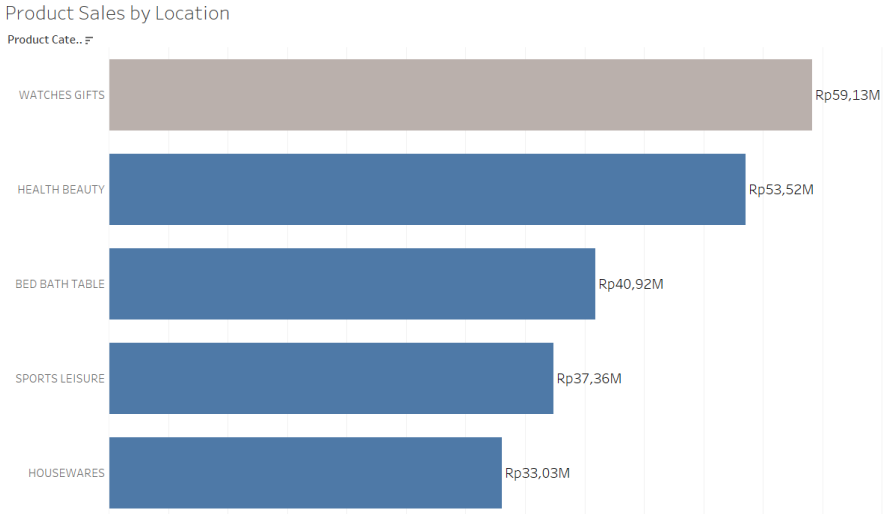


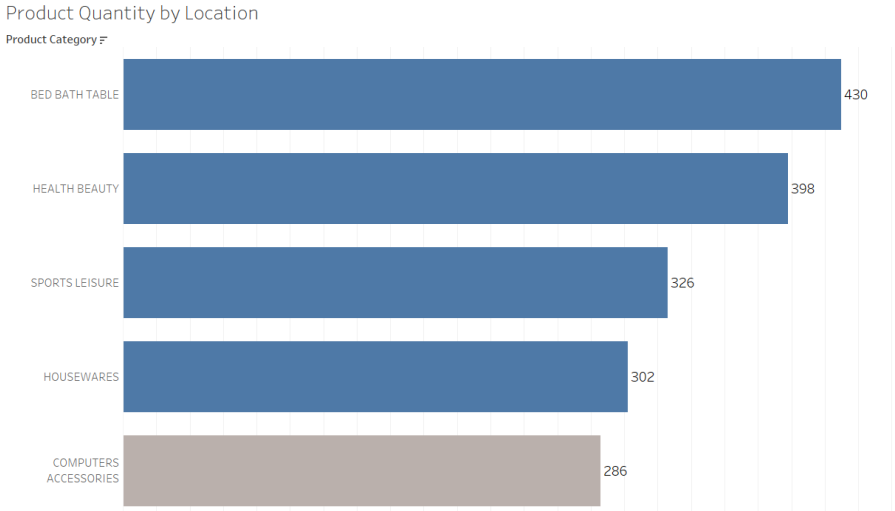
Product categories that are **frequently purchased** and have the **highest sales** in **Banten Province** include **bed bath tables, health beauty,** and **sports leisure**.

HYPOTHESIS 2: CONSUMPTION NEEDS ARE THE MOST SELLING PRODUCT CATEGORIES IN EVERY REGION

North Sumatera

Category products that are in the top 5 highest selling and most sold are marked with a blue bar



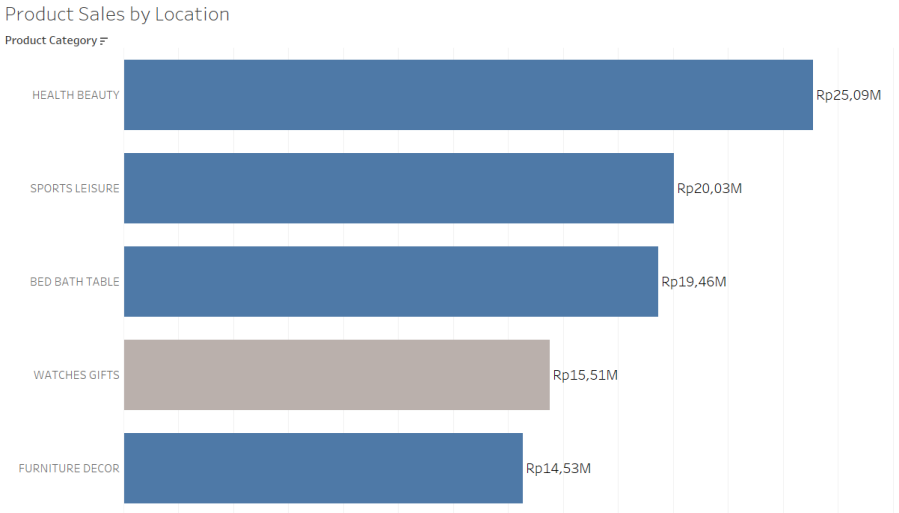


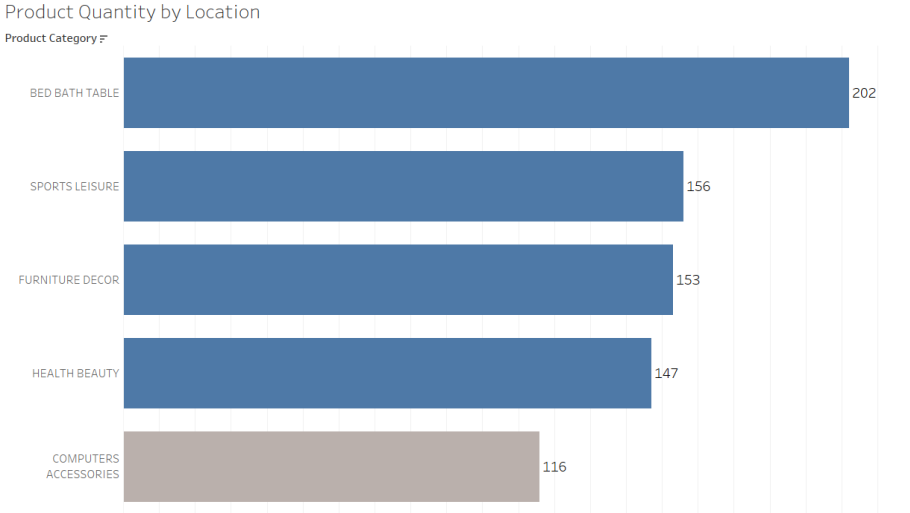
Product categories that are **frequently purchased** and have the **highest sales** in **North Sumatera Province** include **bed bath tables, health beauty, housewares,** and **sports leisure**.

HYPOTHESIS 2: CONSUMPTION NEEDS ARE THE MOST SELLING PRODUCT CATEGORIES IN EVERY REGION

East Kalimantan

Category products that are in the top 5 highest selling and most sold are marked with a blue bar



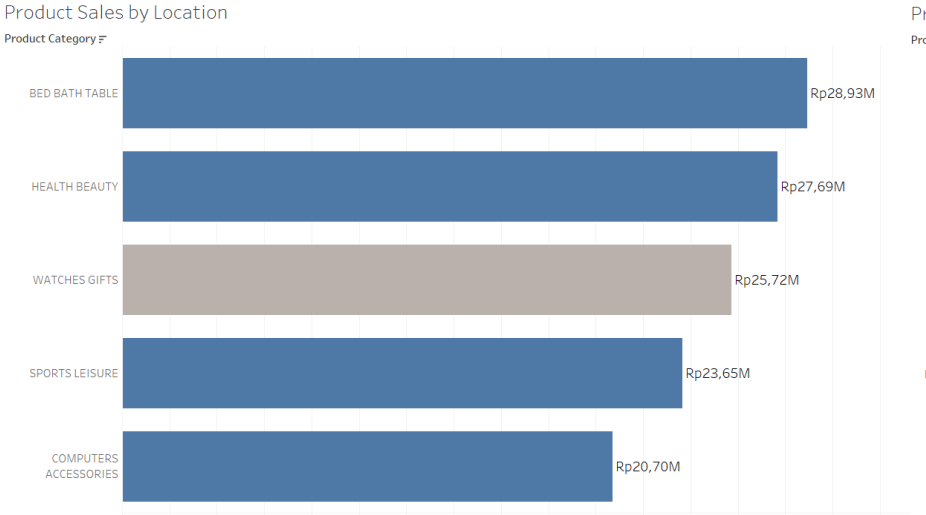


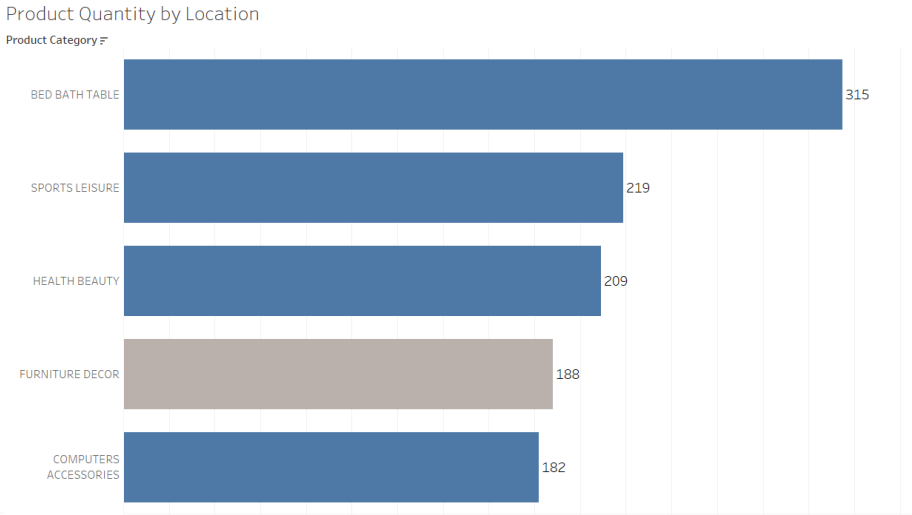
Product categories that are **frequently purchased** and have the **highest sales** in **East Kalimantan Province** include **bed bath tables, health beauty, furniture decor,** and **sports leisure**.

HYPOTHESIS 2: CONSUMPTION NEEDS ARE THE MOST SELLING PRODUCT CATEGORIES IN EVERY REGION

South Sulawesi

Category products that are in the top 5 highest selling and most sold are marked with a blue bar



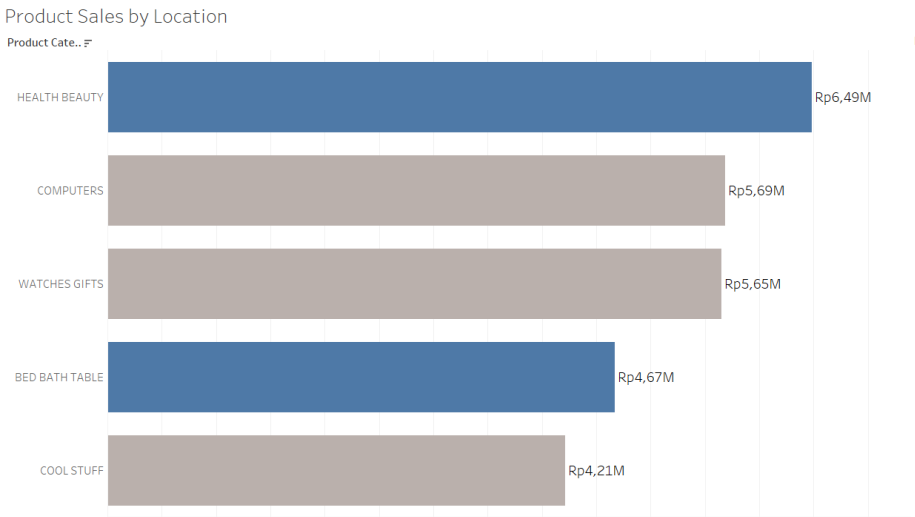


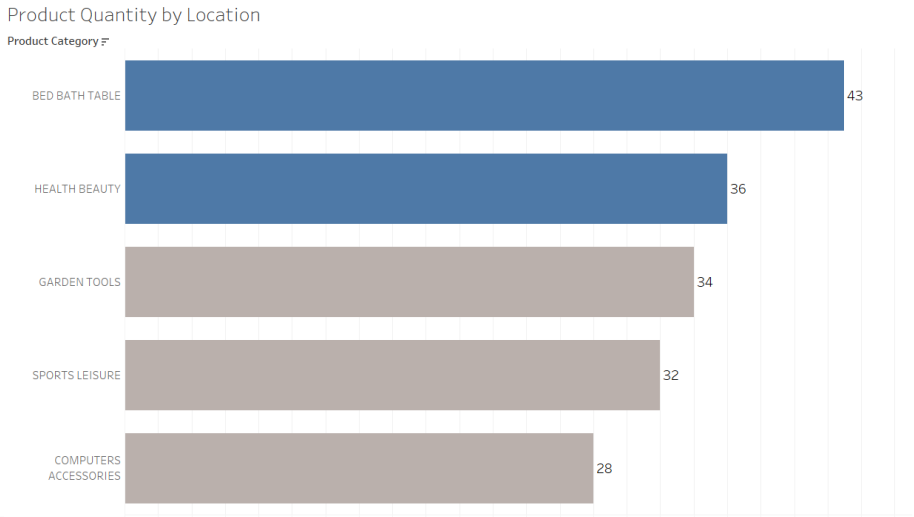
Product categories that are **frequently purchased** and have the **highest sales** in **South Sulawesi Province** include **bed bath tables, health beauty, computer accessories,** and **sports leisure**.

HYPOTHESIS 2: CONSUMPTION NEEDS ARE THE MOST SELLING PRODUCT CATEGORIES IN EVERY REGION

West Nusa Tenggara

Category products that are in the top 5 highest selling and most sold are marked with a blue bar



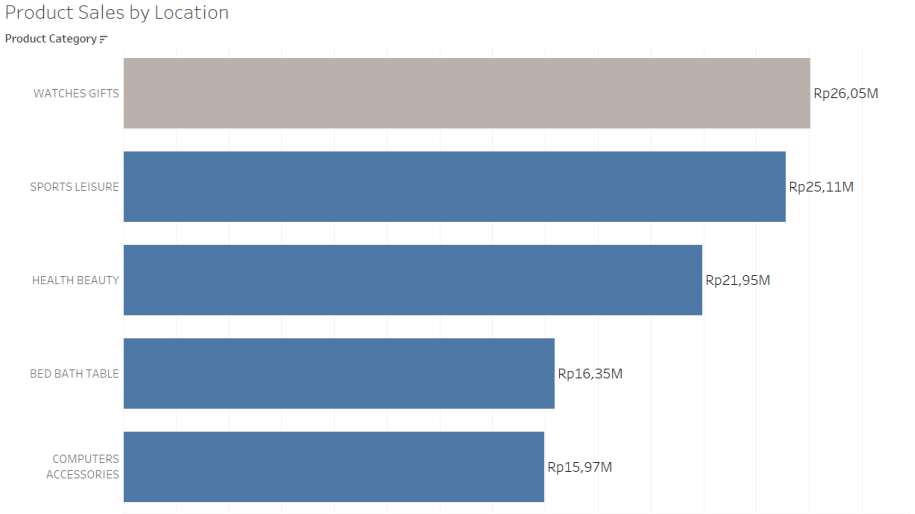


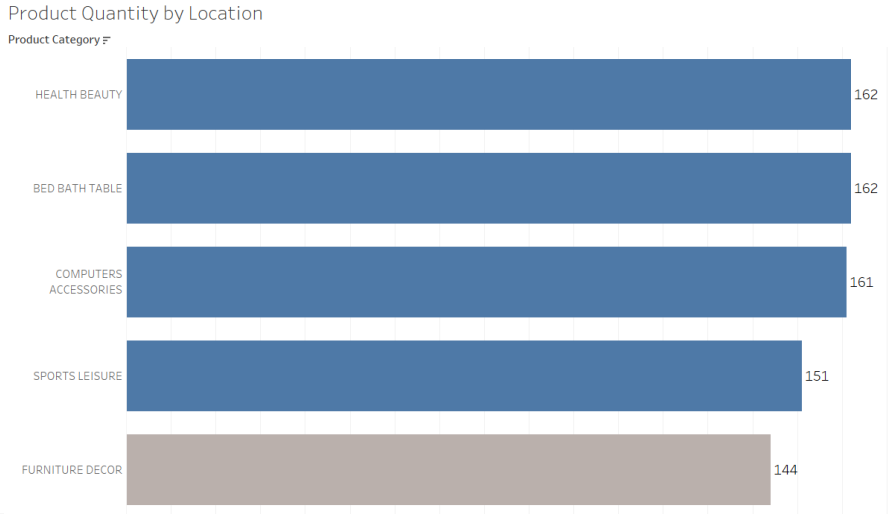
Product categories that are **frequently purchased** and have the **highest sales** in **West Nusa Tenggara Province** include **bed bath tables** and **health beauty**.

HYPOTHESIS 2: CONSUMPTION NEEDS ARE THE MOST SELLING PRODUCT CATEGORIES IN EVERY REGION

Papua

Category products that are in the top 5 highest selling and most sold are marked with a blue bar



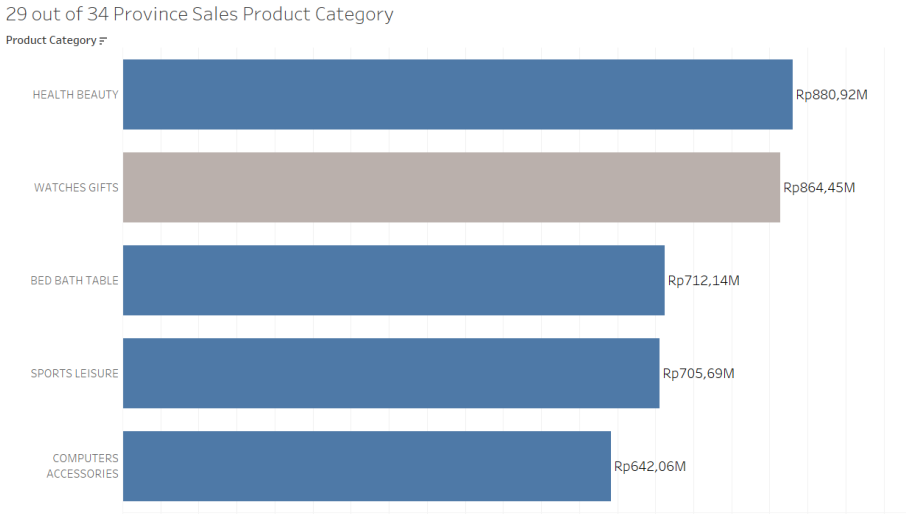


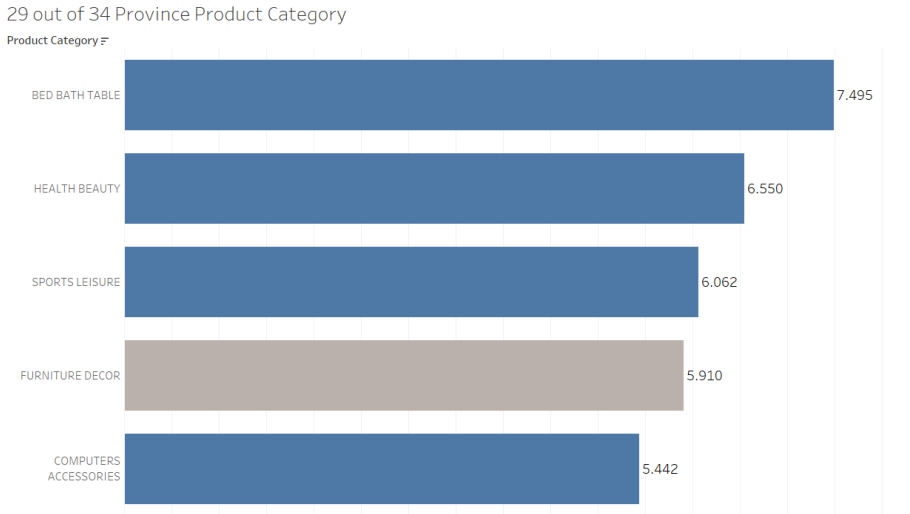
Product categories that are **frequently purchased** and have the **highest sales** in **Papua Province** include **bed bath tables, health beauty, computer accessories,** and **sports leisure.**

HYPOTHESIS 2: CONSUMPTION NEEDS ARE THE MOST SELLING PRODUCT CATEGORIES IN EVERY REGION

29 of 34 Province

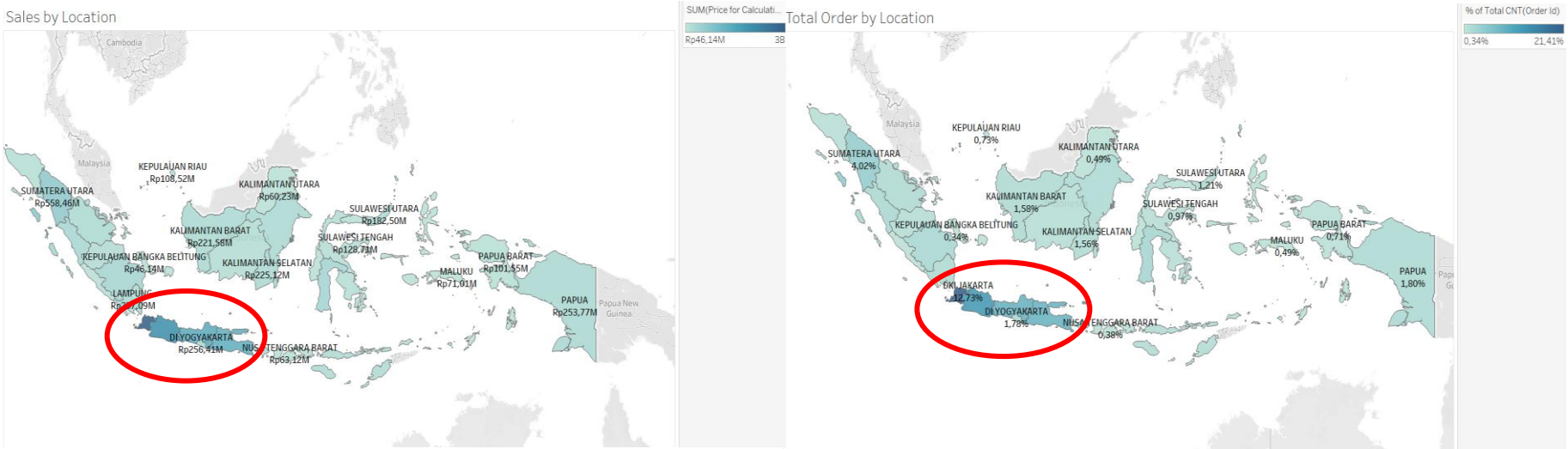
Category products that are in the top 5 highest selling and most sold are marked with a blue bar





- In **29 out of 34 provinces**, it was found that the product categories that were **frequently purchased** and have the **highest sales** were **bed bath tables, health beauty, sports leisure**, and **computers accessories**.
- The product categories with the **highest sales** and **most purchased** are **bed bath tables** and **health beauty**. So there are categories of goods that are often purchased apart from consumables such as **sports leisure** and **computer accessories**.
- For the next step, it will be explored more deeply why this phenomenon can occur in **business question no 3**

HYPOTHESIS 3: JAVA ISLAND IS THE REGION WITH THE MOST BUYING OF GOODS ONLINE

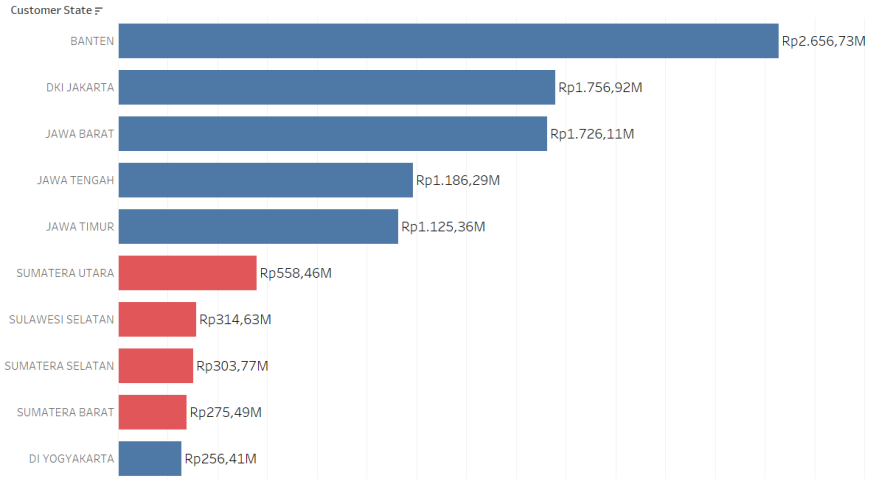


Overall, almost all provinces on the island of Java have a fairly **high tendency** to **purchase goods** and have **high sales** compared to other islands in Indonesia (the darker the color of the province, the higher the sales/purchase of goods in that province).

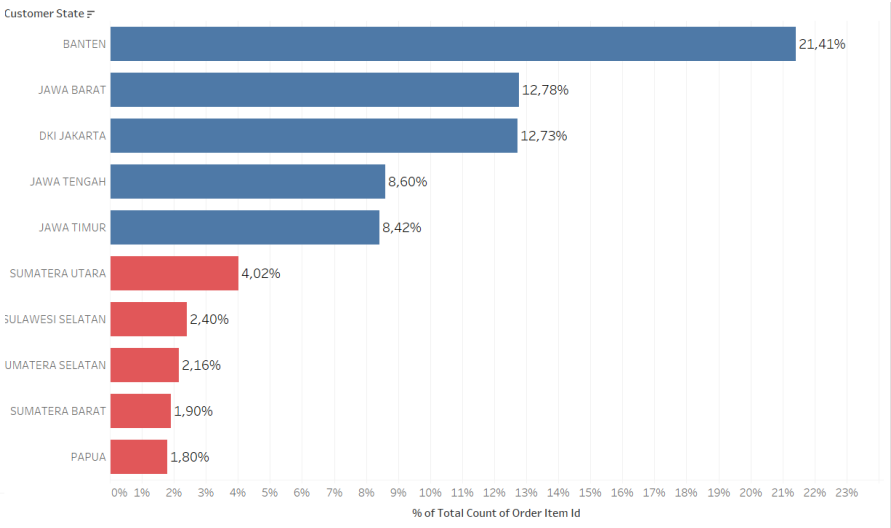
HYPOTHESIS 3: JAVA ISLAND IS THE REGION WITH THE MOST BUYING OF GOODS ONLINE

The area with the blue graph is Java Island and the red graph is outside Java Island.

Sales by Location



Order by Location



- **5 of the 6 provinces** on the **Java Island** are included in the top 10 provinces that **have lot of total purchases of goods** and **all provinces** on the Java Island are included in the top 10 provinces with the **highest sales** in Indonesia.
- For the next step, I will find out more about why this phenomenon can occur in **business question 2**.

05

BUSINESS QUESTIONS 2

BUSINESS QUESTIONS 2

Is the distribution of payment methods evenly distributed between Java and other islands?

BACKGROUND

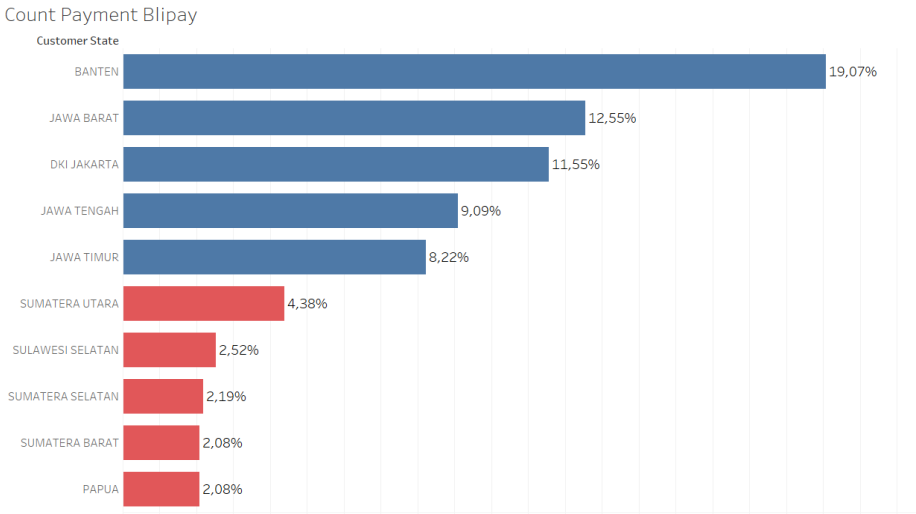
I want to explore the phenomena that occur in business question 1 **(Most sales** are only on **Java Island** and the use of **credit cards** is **very high** compared to the other payment methods)

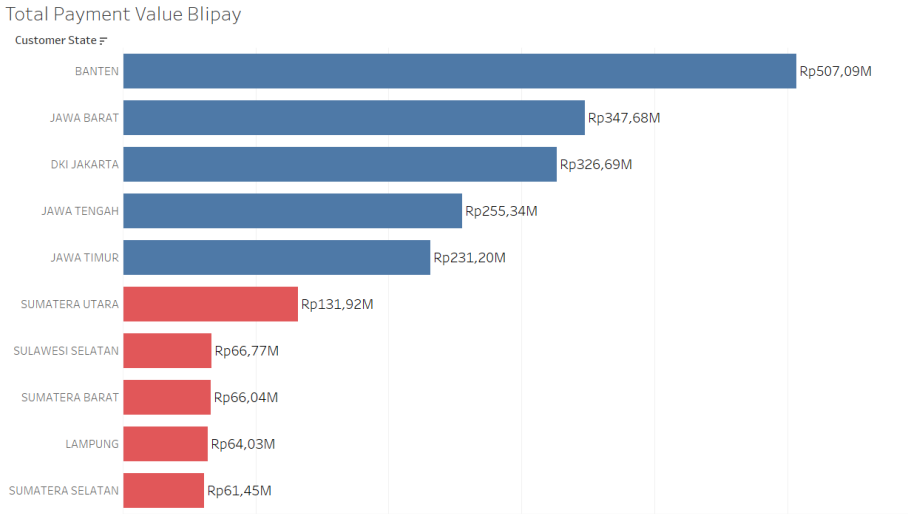
PURPOSE

Knowing the distribution of payment methods in each region whether it is evenly distributed or not

HYPOTHESIS: PAYMENT METHODS ARE VERY DIFFERENT IN EVERY REGION

The area with the blue graph is Java Island and the red graph is outside Java Island.

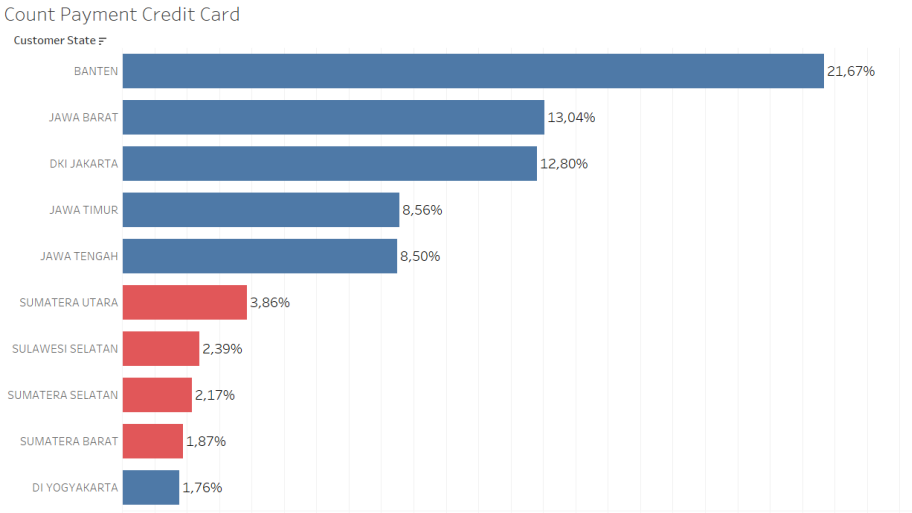


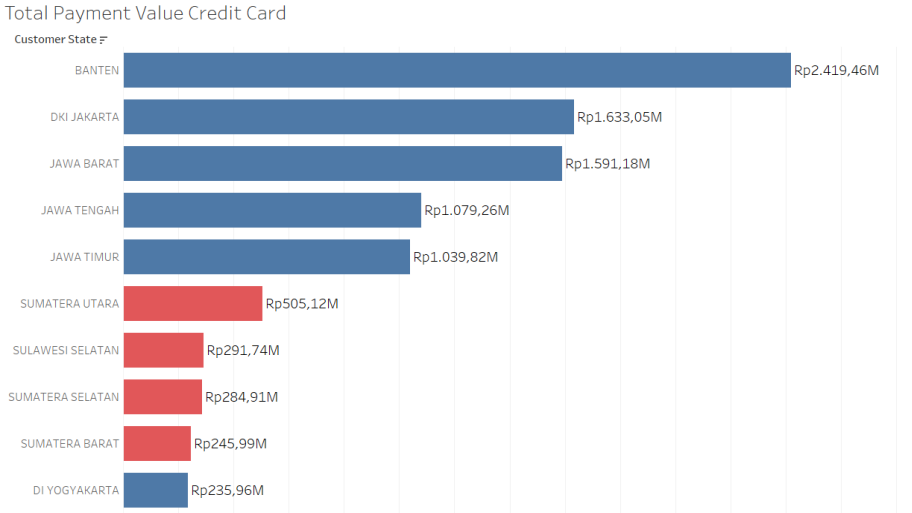


1. The area with the most transactions using **Blipay** is the **Banten Province** with a total of **19.07%** of the total Blipay transactions in Indonesia. Here it can be seen that the **Java Island** dominates in terms of transactions using Blipay.
2. The area with the largest total nominal transaction using **Blipay** is **Banten Province** with a total of **Rp 507,09M** transactions. Here it can be seen that the **Java Island** dominates in terms of total transactions using Blipay.

HYPOTHESIS: PAYMENT METHODS ARE VERY DIFFERENT IN EVERY REGION

The area with the blue graph is Java Island and the red graph is outside Java Island.

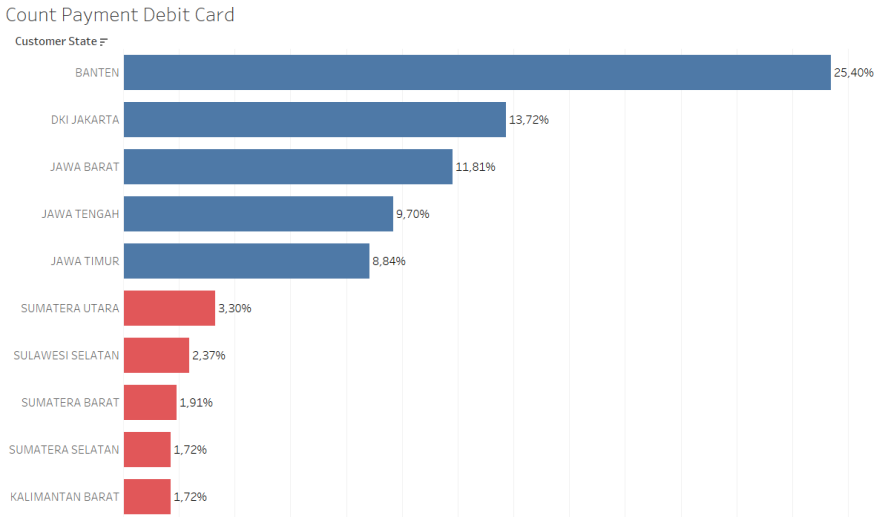


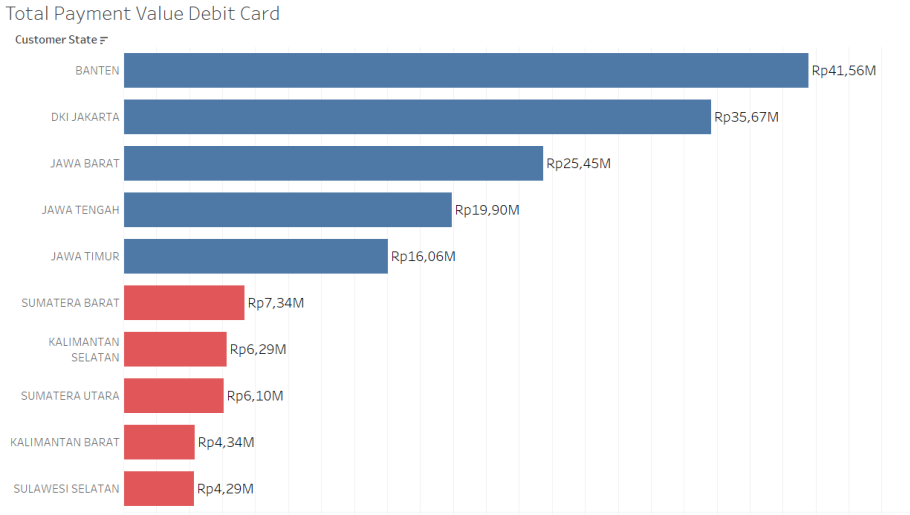


1. The area with the most transactions using **credit card** is the **Banten Province** with a total of **21.67%** of the total credit card transactions in Indonesia. Here it can be seen that the **Java Island** is **quite dominant** (in **fact all provinces in Java are in the top 10**) in terms of transactions using credit cards.
2. The area with the largest total nominal transaction using a **credit card** is the **Banten Province** with a total of **Rp 2.419,46M** transactions. Here it can be seen that Java is quite dominating (in fact **all provinces in Java are in the top 10**) in terms of total transactions using credit cards.

HYPOTHESIS: PAYMENT METHODS ARE VERY DIFFERENT IN EVERY REGION

The area with the blue graph is Java Island and the red graph is outside Java Island.

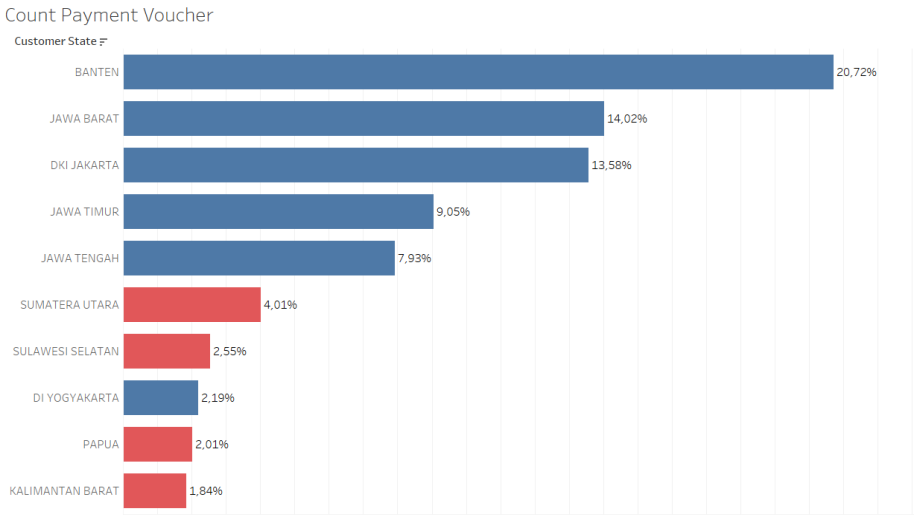


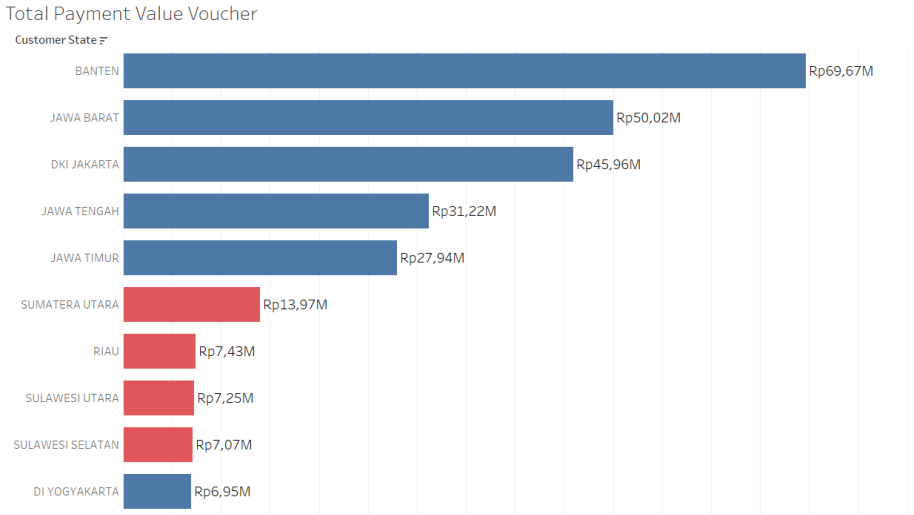


1. The area with the most transactions using **debit card** is the **Banten Province** with a total of **25.40%** of the total debit card transactions in Indonesia. Here it can be seen that the **Java Island** dominates in terms of transactions using debit cards.
2. The area that has the largest total nominal transaction using a **debit card** is the **Banten Province** with a total of **Rp 41,56M** transactions. Here it can be seen that the **Java Island** dominates in terms of total transactions using debit cards.

HYPOTHESIS: PAYMENT METHODS ARE VERY DIFFERENT IN EVERY REGION

The area with the blue graph is Java Island and the red graph is outside Java Island.





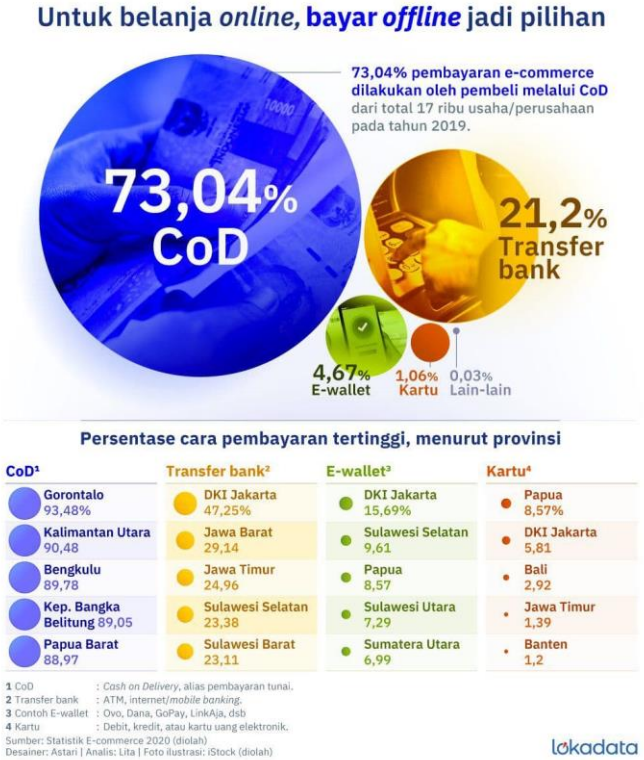
1. The area with the most transactions using **vouchers** is the **Banten Province** with a total of **20.72%** of the total voucher transactions in Indonesia. Here it can be seen that Java is quite dominating (in fact **all provinces in Java are in the top 10**) in terms of transactions using vouchers.
2. The region that has the largest total nominal transaction using **vouchers** is **Banten Province** with a total of **Rp 69,67M** transactions. Here it can be seen that Java is quite dominating (in fact **all provinces in Java are in the top 10**) in terms of total transactions using vouchers.

WHAT ARE THE MOST USED PAYMENT METHOD TO BUY ONLINE?

(Asosiasi Penyelenggara Jasa Internet Indonesia, 2019-2020)

In a survey conducted by **APJII**, the top 6 **most frequently used** online payment methods to purchase online was **COD** as much as **27.5%.**

WHAT ARE THE MOST FREQUENTLY USED PAYMENT METHODS IN EACH REGION IN INDONESIA?

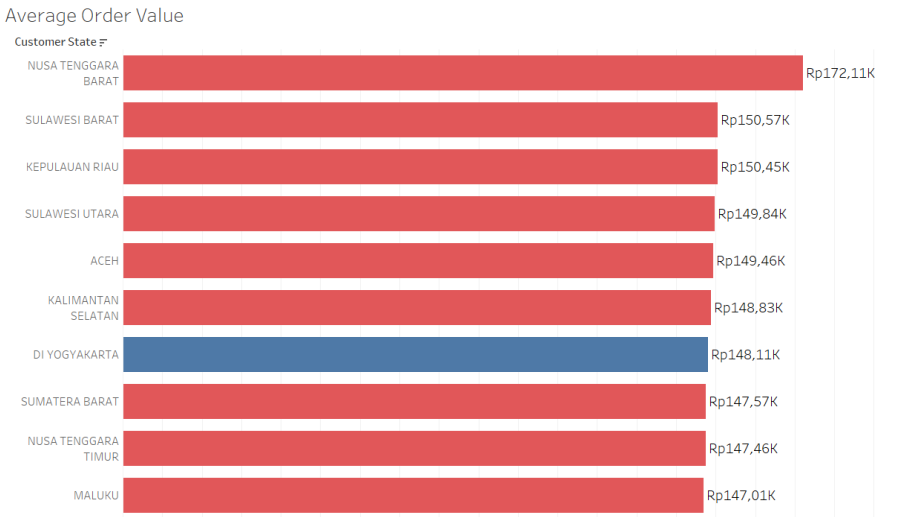


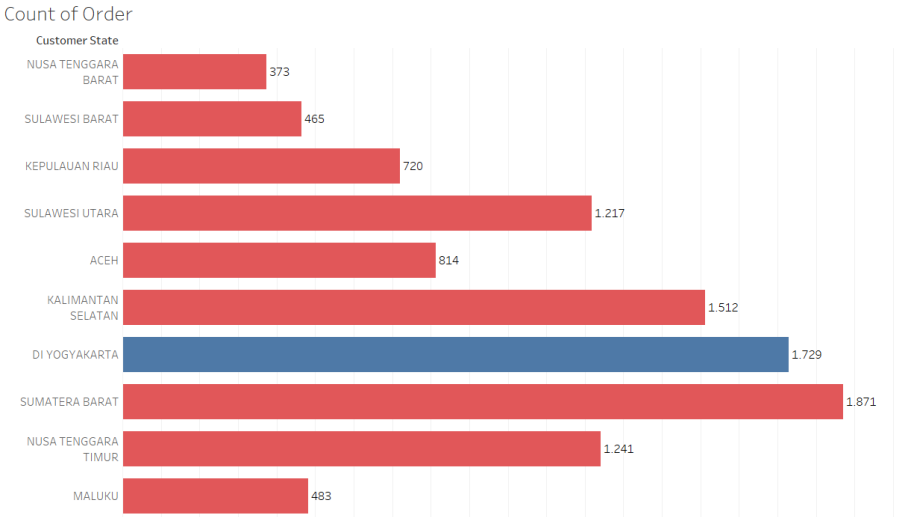
In a survey from **lokadata**, **73.04%** of e-commerce payments were made by buyers via **COD**. The area that did the most COD are from **outside of Java Island**, one of them was Gorontalo at 93.48%.

Survey Lokadata Mengenai Metode Pembayaran yang Paling Diminati (Lita, 2020)

AVERAGE ORDER VALUE

The area with the blue graph is Java Island and the red graph is outside Java Island.





- In the following visualization, the regions that tend to have **the most Average Order Value** are **regions outside Java** where **West Nusa Tenggara** ranks **first** wherein one order they spend a budget of around **Rp. 172,11K.**
- This means that even **outside of Java**, they spend quite a **lot of money** during online shopping even there are **only fewer orders**. The **addition** of **payment methods** is expected to **reach markets outside** Java to **increase sales**.

CONCLUSION

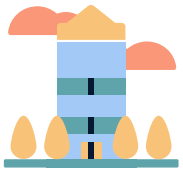
1. Customers on the **Java Island dominate transactions** on each payment method provided.

2. **Banten Province always ranks first** in the most total transactions.

3. The most frequently used payment methods **outside Java** are **COD**.

4. Although areas outside Java tend to have fewer orders, in one order **they spend quite a lot of money**.

SUGGESTION



Based on surveys from external sources, I recommend adding a payment method in the form of **COD**. Users are expected to have many choices of payment methods to increase user flexibility in making payments.

By implementing payment methods that are widely used outside Java, it is expected to be able to **embrace markets outside Java**.

06

BUSINESS QUESTIONS 3

BUSINESS QUESTIONS 3

What is the trend of the 5 best-selling product categories in each year and how are the sales of these product categories?

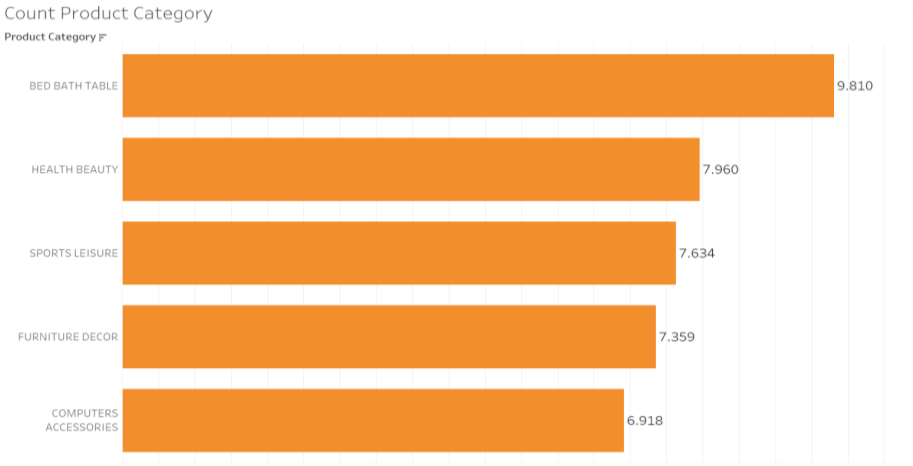
BACKGROUND

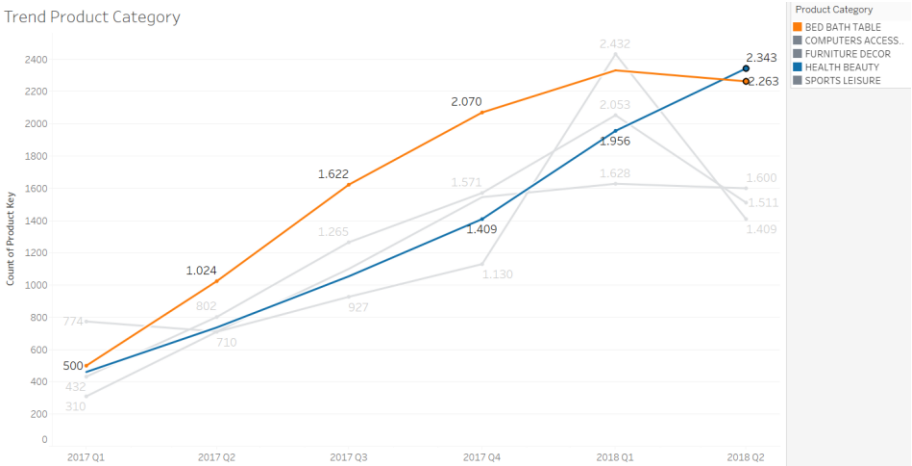
In business question 1, it is known that the **product categories** that are **most frequently purchased** and **have the highest sales** are **bed bath tables, health beauty, sports leisure**, and **computers accessories**. Here I want to find out how **the trend of the existing product categories**? Is the trend up or down?

PURPOSE

Knowing what is the trend of product categories in Blibli data so that product categories that have an increasing trend can keep in stock.

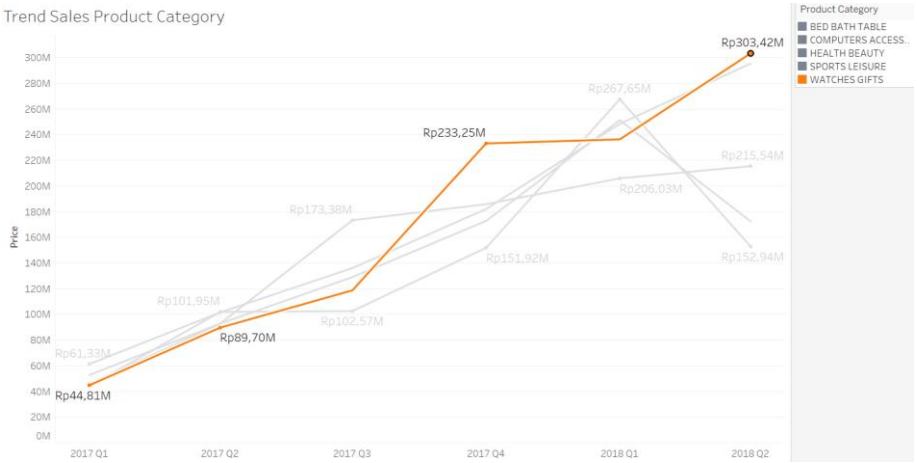
MOST BUYED PRODUCT CATEGORIES IN 2017-2018





1. The **5 most frequently purchased product categories** during the **2017Q1-2018Q2** were **bed bath tables, health beauty, sports leisure, furniture decor**, and **computers accessories**.
2. Products in the **bed bath table** and **health beauty** categories experienced an **increase** in the number of products purchased, although sales of bed bath tables decreased slightly in the second quarter.

PRODUCT CATEGORIES WITH THE HIGHEST SALES IN 2017-2018



1. The **5 product categories** with the highest sales during the **2017Q1-2018Q2** were **watches gifts, health beauty, bed bath tables, sports leisure,** and **computers accessories**.
2. Products in the watch gifts category have always experienced an **increase in sales** throughout **2017Q1 to 2018Q2**.

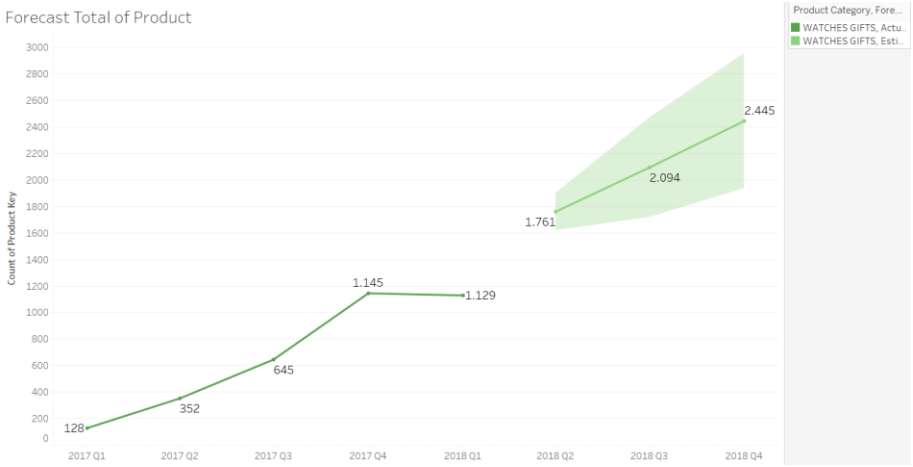
SUGGESTIONS

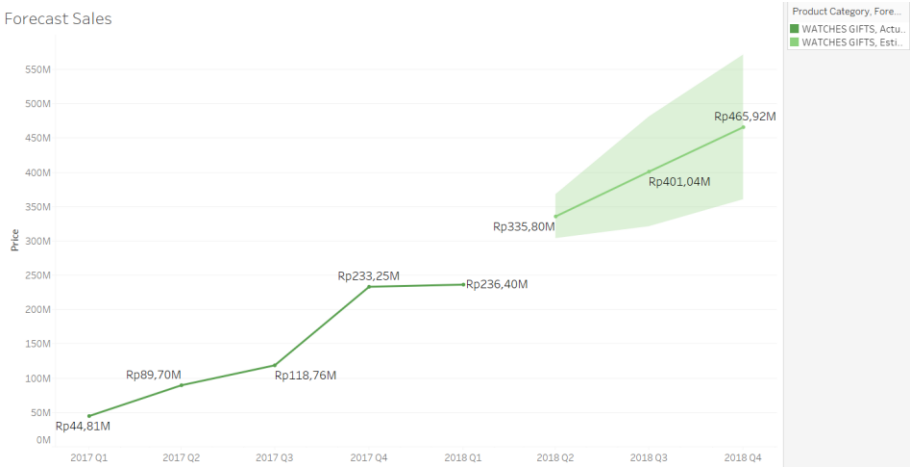
- **Bed bath tables** and **health beauty** have quite high sales and are necessities that quickly run out. So doing a fairly massive sale like giving some promo can **increase sales**.
- **Watches gifts** are a category product whose sales are quite high, but because watch gifts are items that can last for a long time, it is necessary to review the trend in the future, whether it will continue to rise.
- The following data can be used to perform **predictive analysis** to predict how the trend will be in the following years.

PREDICTIVE ANALYSIS USING TABLEAU

Predictive analysis was performed using Tableau. Products in the **Bed bath Table** and **Health Beauty** categories are predicted to **increase** in sales until the end of 2018.

PREDICTIVE ANALYSIS USING TABLEAU





Predictive analysis on **Watches Gifts** category products using Tableau is also predicted to **increase** sales until the end of 2018.

SUMMARY

1. **Bed bath table** and **health beauty** are the most frequently purchased category products

2. **Watches gifts** are the products that have the highest sales

3. **Bed bath tables, health beauty**, and **watches gifts** have pretty **good trend** predictions so they need to be kept in stock

07

BUSINESS QUESTIONS 4

BUSINESS QUESTIONS 4

Is the product delivered on time? If not, does late delivery affect customer satisfaction?

BACKGROUND

Reported from the article Aulia, et al with respondents are students of the Faculty of Economics, Islamic University of Borneo Muhammad Arsyad Al-Banjari (UNISKA) Banjarmasin with a total of **100 respondents**, in which the article contains a **service quality factor** as one of the determinants of **someone buying goods in e-commerce** or no. Based on the survey results, **68%** of respondents **agree** with the services provided by the seller. One of the things that can be done to **improve the quality of customer satisfaction** is **on-time delivery**.

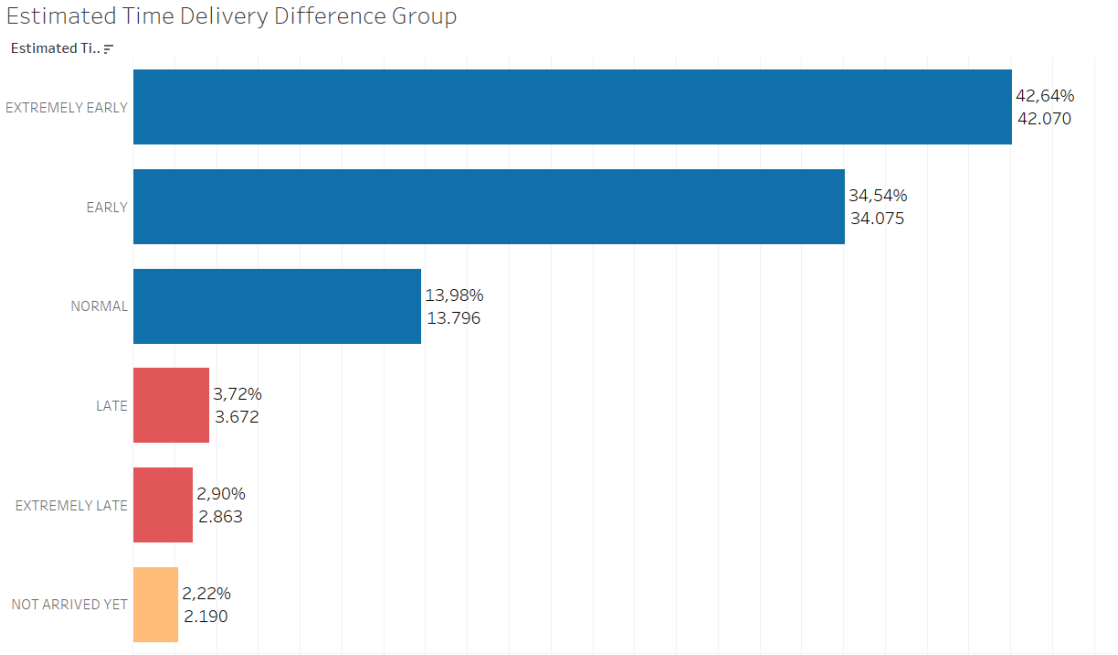
PURPOSE

Knowing whether the product sent to the customer has arrived on time and how the customer responds in the form of a rating.

ESTIMATED TIME DELIVERY

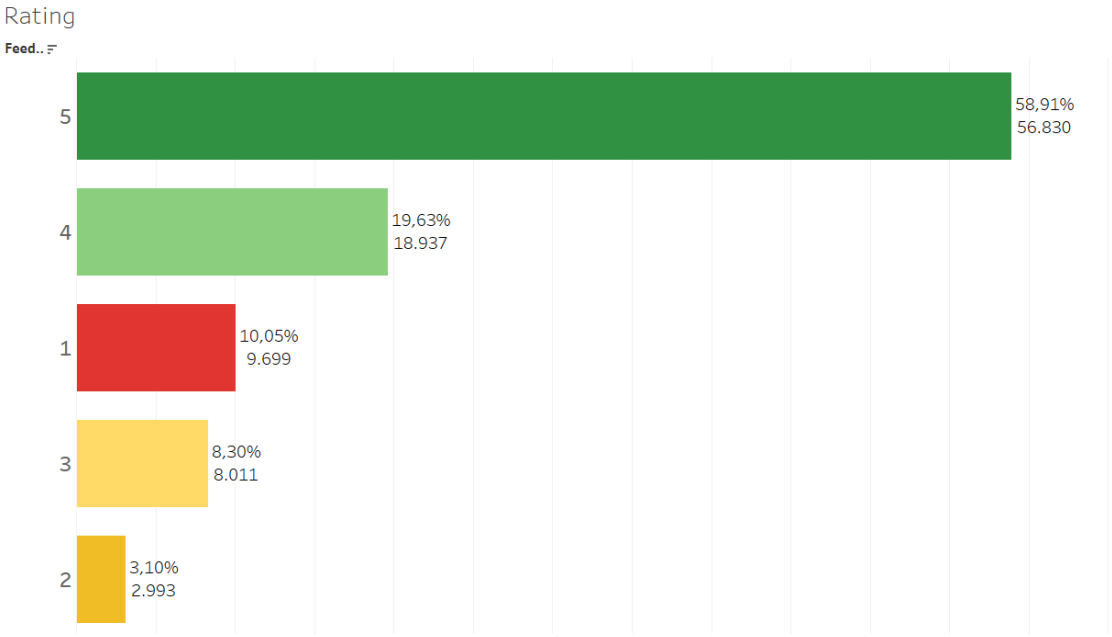
- If an order **doesn't have a delivered date key** then the order **has not arrived.**
- If the order is delivered **more than 2 weeks to 2 weeks than estimated**, the order is classified as **extremely early**.
- If the order is delivered **less than 2 weeks to 1 week earlier** than estimated, the order is classified as **early**.
- If the order is delivered **less than 1 week earlier to the D day** of the estimate, the order is classified as **normal (on time)**.
- If the order is delivered **1 day to 1 week late** from the estimate, the order is classified as **late**.
- If the order is delivered **more than 1 week later** than estimated, the order is classified as **extremely late.**

HYPOTHESIS 1: 80% ORDERS SENT ON TIME



Most orders were shipped **very earlier** than the estimated, **42.070** orders. Only **13,98%** of orders that **arrived on time** (13.796 out of 98.666). This phenomenon is quite **strange** because it means the distance between the estimated delivery time and the actual delivery time is **still too far**.

OVERALL RATING

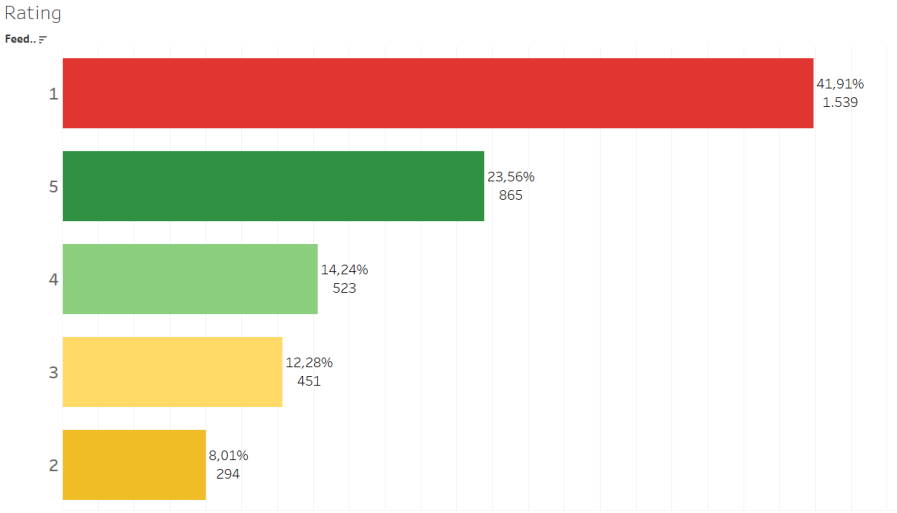


Most orders get a **perfect rating** (5), which is **58.91%** of the total orders.

ORDER STATUS

Most of the orders were sent to customers, which was **97.78%** (96.478) of the total orders.

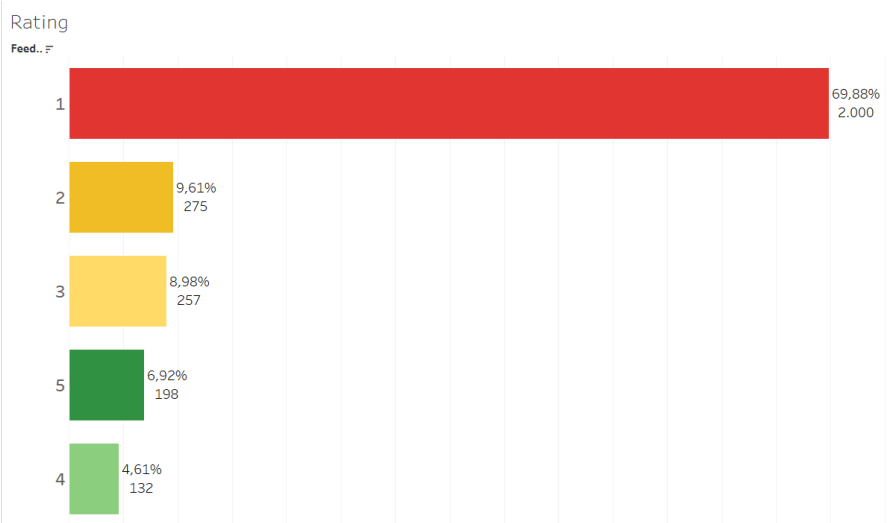
HYPOTHESIS 2: LATE DELIVERY WILL AFFECT CUSTOMER SATISFACTION



KEPULAUAN BANGKA BELITUNG	SUMATERA UTARA	
	SUMATERA SELATAN	
	SUMATERA BARAT	
	SULAWESI UTARA	
	SULAWESI TENGGARA	
	SULAWESI TENGAH	
	SULAWESI SELATAN	
	SULAWESI BARAT	
	RIAU	
	PAPUA BARAT	
	PAPUA	
	NUSA TENGGARA TIMUR	
	MALUKU UTARA	50,00%
	MALUKU	
	LAMPUNG	

1. Late orders that get a rating of **1** are **41.91%** of the total late orders.
2. The highest percentage of orders that are late orders from the **Kepulauan Bangka Belitung** - **North Maluku**. With a delay percentage of **50%** of the total orders sent.

HYPOTHESIS 2: LATE DELIVERY WILL AFFECT CUSTOMER SATISFACTION



Seller State	Customer State	EXTREMELY LATE
ACEH	BALI	
	JAMBI	100,00%
	SULAWESI BARAT	
	SULAWESI TENGGARA	14,29%
KALIMANTAN UTARA	BALI	100,00%
	JAMBI	
	SULAWESI BARAT	
	SULAWESI TENGGARA	
KEPULAUAN BANGKA BELITUNG	BALI	
	JAMBI	
	SULAWESI BARAT	
	SULAWESI TENGGARA	100,00%
PAPUA BARAT	BALI	25,00%
	JAMBI	
	SULAWESI BARAT	100,00%
	SULAWESI TENGGARA	

1. Extremely late orders that get a rating of **1** are **69.88%** of the total very late orders.
2. The highest percentage of orders that are extremely late orders from **Aceh-Jambi, North Kalimantan-Bali, Kepulauan Bangka Belitung - Southeast Sulawesi**, and **West Papua-West Sulawesi** where all orders are delivered extremely late.

CONCLUSION

1. A total of **42,64%** (42.070) orders arrived extremely early.

2. Only **13,98%** of orders arrived on time (13.796 out of 98.666 orders)

3. A total of **97,78%** (96.478) orders were successfully delivered.

4. Most orders get a **perfect rating** (5), which is **58.91%** of the total orders.

5. For **late orders**, most of the **rating** is **1** with a total of **41.91%**.

6. For orders that are **extremely late**, the highest rating obtained is **rating 1** with a total of **69.88%**, so **orders that arrive late and very late affect customer satisfaction**.

7. Orders that are **late** and **extremely late** happened in **regions outside Java**.

SUGGESTION

1. Adjust the estimated delivery time so that the distance is not too far from the original delivery date.

2. Provide notification to customers that the ordered goods are expected to arrive late.

08

BUSINESS QUESTIONS 5

BUSINESS QUESTIONS 5

How is the customer recency, frequency, and monetary?

BACKGROUND

Customer segmentation is the practice of **dividing a customer** base into groups of individuals that are **similar** in specific ways. One of the techniques of customer segmentation is **RFM** (recency, frequency, monetary). RFM segmentation allows marketers to **target specific clusters of customers** with communications that are much **more relevant to their behavior**.

PURPOSE

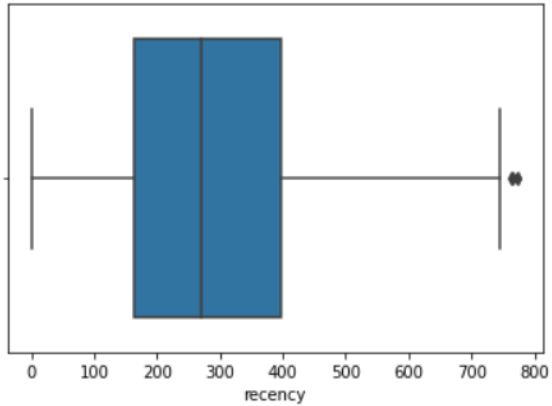
Create effective allocation of marketing resources and the maximization of cross and up-selling opportunities

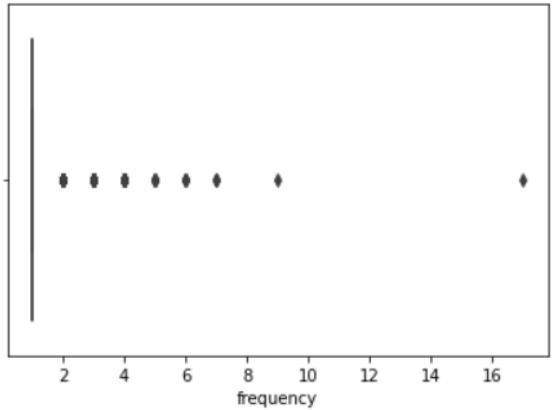
RFM

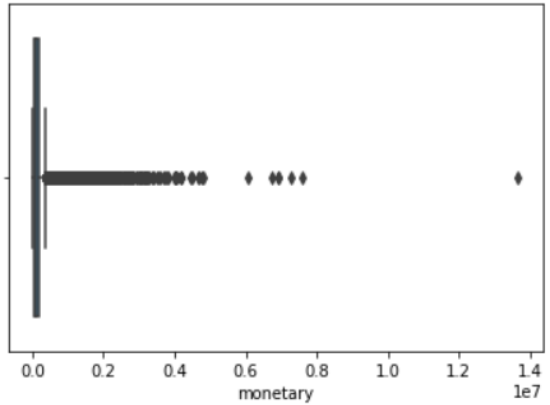
- **Build RFM**

recency	the maximum time an order occurs overall – time_limit each customer + 1
frequency	Counting the number of orders per customer using order_id
monetary	Total payment

HANDLING OUTLIERS







	recency	frequency	monetary
Minimum limit	-184.5	1	-117426.25
Maximum limit	747.5	1	362083.75
Outliers percentage under the minimum limit	0.0%	0.0%	0.0 %
Outliers percentage under the maximum limit	0.004162504162504162 %	3.118886067518628 %	7.950713899179952 %

CHECK FOR QUANTILE

	recency	R
Q1	recency ≤ 165	4
Q2	165 < recency ≤ 270	3
Q3	270 < recency ≤ 399	2
Q4	recency > 399	1

frequency	Count user_name	F
1	93099	1
2	2745	2
3	203	3
4	30	4
5	8	4
6	6	4
7	3	4
9	1	4
17	1	4

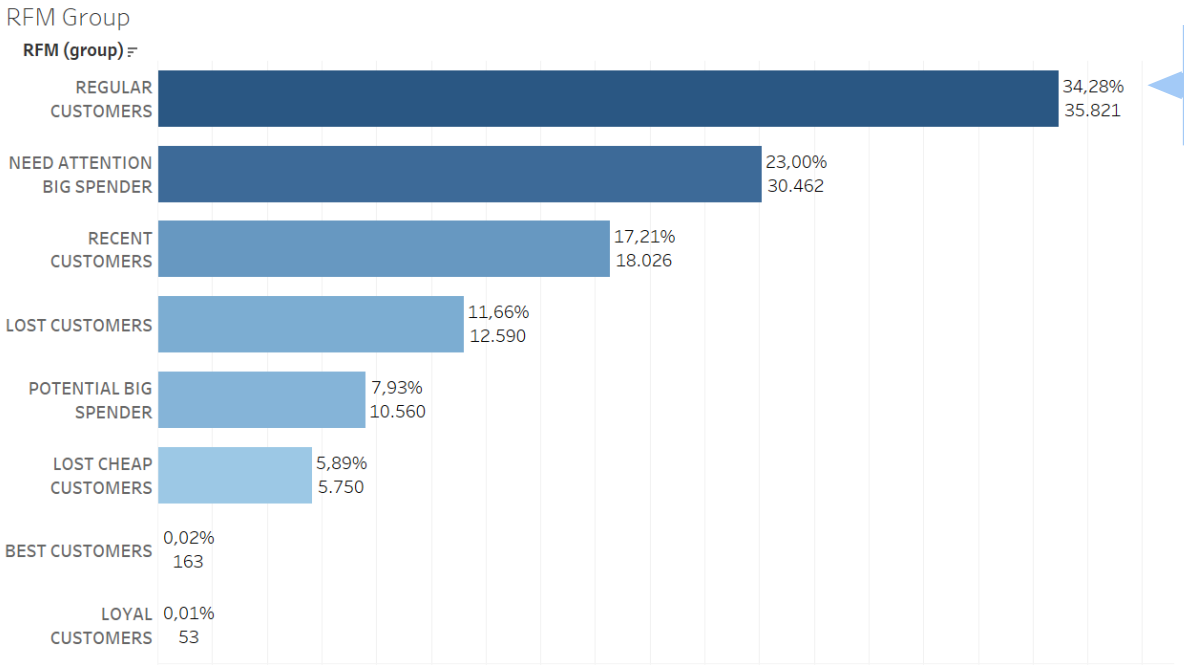
	monetary	M
Q1	monetary ≤ 58980	1
Q2	58980 < monetary ≤ 98840	2
Q3	98840 < monetary ≤159252.5	3
Q4	monetary > 159252.5	4

This is a limit in the formation of recency and monetary. Recency, frequency, and monetary will be divided into 4 parts.

RFM SEGMENTATION

Customer	RFM Score
Best Customers	444
Loyal Customers	x4x
Potential Big Spender	4x4
Need Attention Big Spender	xx4
Recent Customer	4xx
Lost Cheap Customers	111
Lost Customers	1xx
Regular Customers	Others

RFM GROUP



A customer group that has many customers is a regular customer

#1 REGULAR CUSTOMERS

F X M

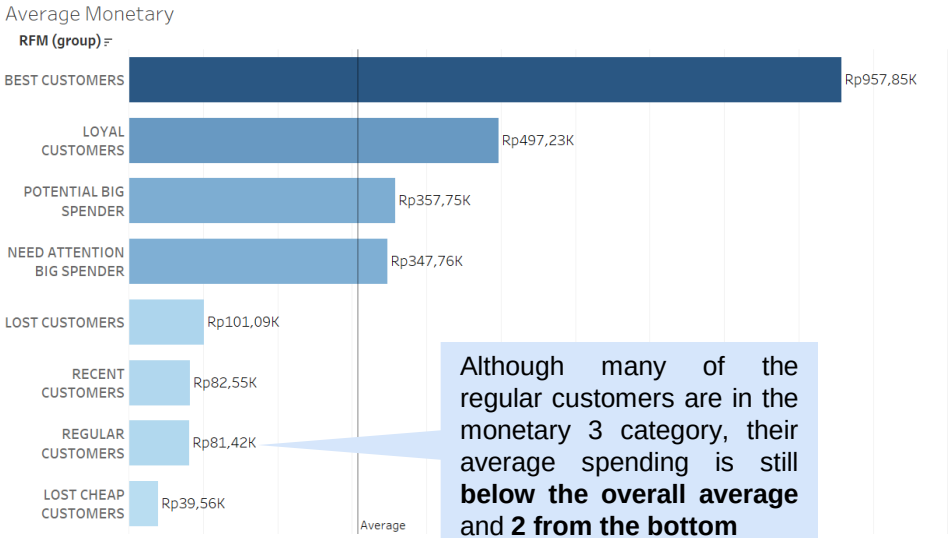
F	M 1	2	3
1	31,07%	32,39%	33,81%
2	0,15%	0,68%	1,82%
3		0,01%	0,08%

Many customers spend a lot of money but have a very small frequency.

R X F

Most of them have a distant last shopping time (9 months – 1 year ago)

F	R 2	3
1	49,31%	47,96%
2	1,39%	1,26%
3	0,03%	0,05%



Although many of the regular customers are in the monetary 3 category, their average spending is still **below the overall average** and **2 from the bottom**

- Make limited-time offers.
- Send personalised emails.
- Look back at marketing strategy on 9 months - 1 year ago.
- Offer personalized recommendations

#2 LOST CUSTOMERS

Most customers spend quite a bit of money and very little shopping frequency

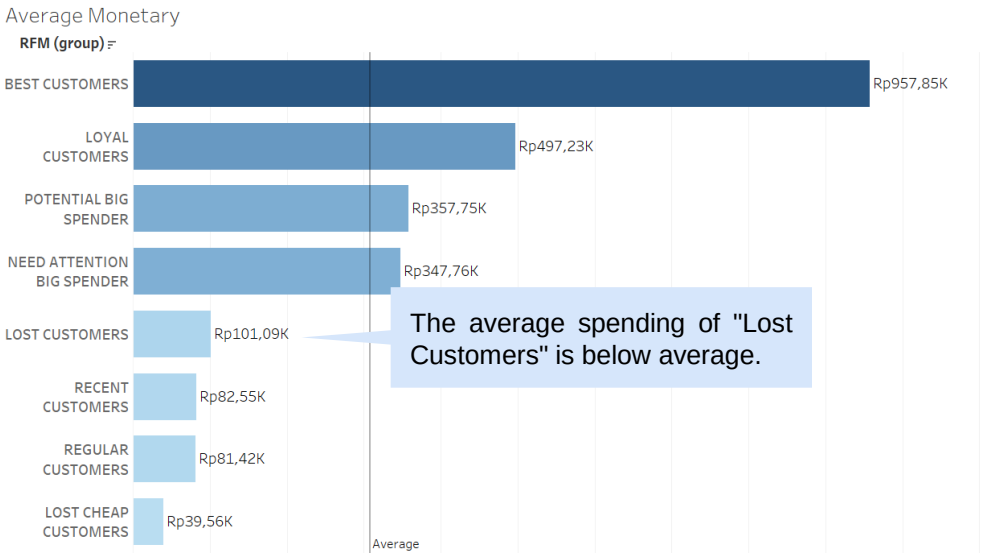
FXM

F	M 1	2	3
1		48,31%	48,11%
2	0,25%	0,98%	2,13%
3	0,05%	0,05%	0,12%

RXF

F	R 1
1	96,42%
2	3,37%
3	0,21%

Some customers have not shopped for more than 1 year and only one time



The average spending of "Lost Customers" is below average.

- Send personalised emails.
- It is better to ignore it because attracting customers who have not shopped for more than 1 year will require more effort and cost a lot of money.

#3 RECENT CUSTOMERS

FXM

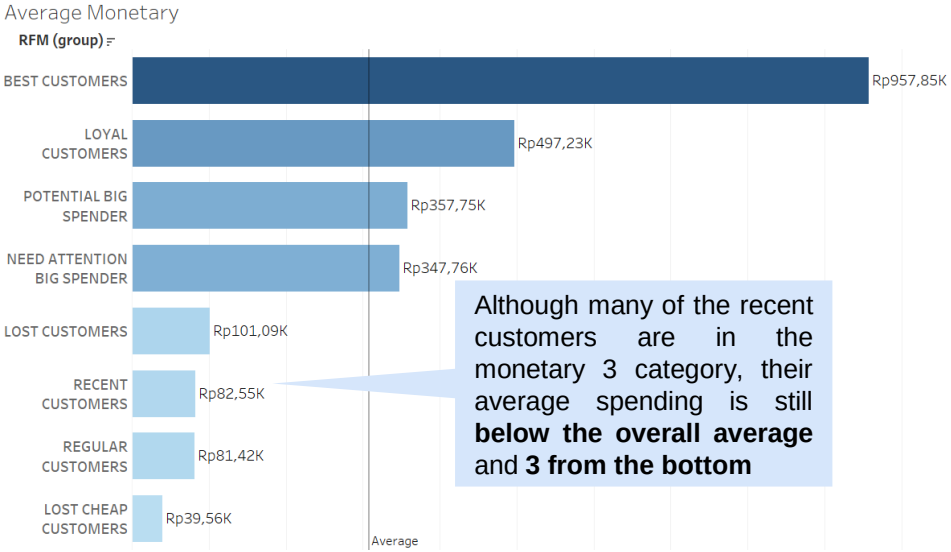
F	M 1	2	3
1	30,99%	31,71%	34,27%
2	0,21%	0,73%	1,91%
3	0,02%	0,07%	0,11%

Many customers spend a lot of money but have a very small frequency.

RXF

F	R 4
1	96,97%
2	2,84%
3	0,19%

Shopping these days but most of them still have less frequency



Although many of the recent customers are in the monetary 3 category, their average spending is still **below the overall average** and **3 from the bottom**

- Gift them discounts/promo.

#4 LOYAL CUSTOMERS

F X M

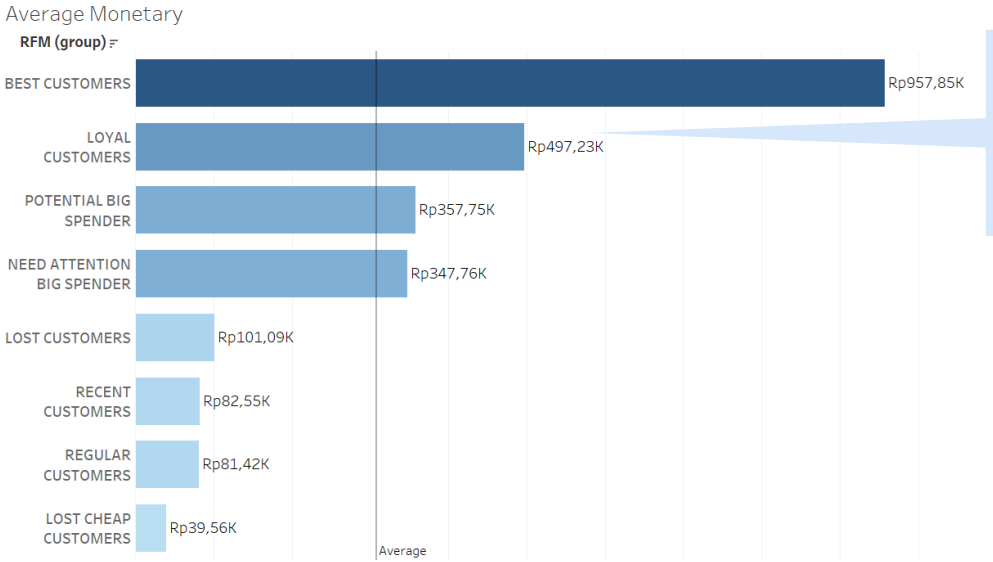
F	M 2	3	4
4	7,55%	18,87%	73,58%

Many customers spend quite a lot of money.

R X F

F	R 1	2
4	43,40%	56,60%

Most of them have a distant last shopping time (9 months – 1 year ago)



Their monetary is very good and **above** average

- Offer personalized recommendations
- Upselling
- Look back at marketing strategy on 9 months - 1 year ago

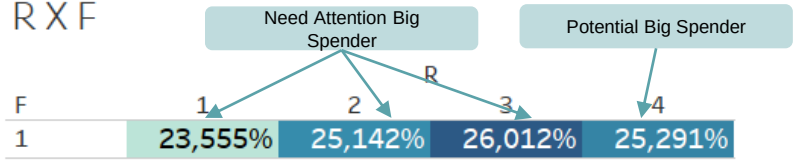
#5 CUSTOMER THAT NEED ATTENTION

F X M

F	M 1	2	3	4
1	19,81%	20,63%	21,46%	31,33%
2	0,11%	0,44%	1,11%	4,14%
3	0,01%	0,02%	0,05%	0,61%
4		0,00%	0,01%	0,27%

Many customers spend quite a lot of money but have a very small frequency.

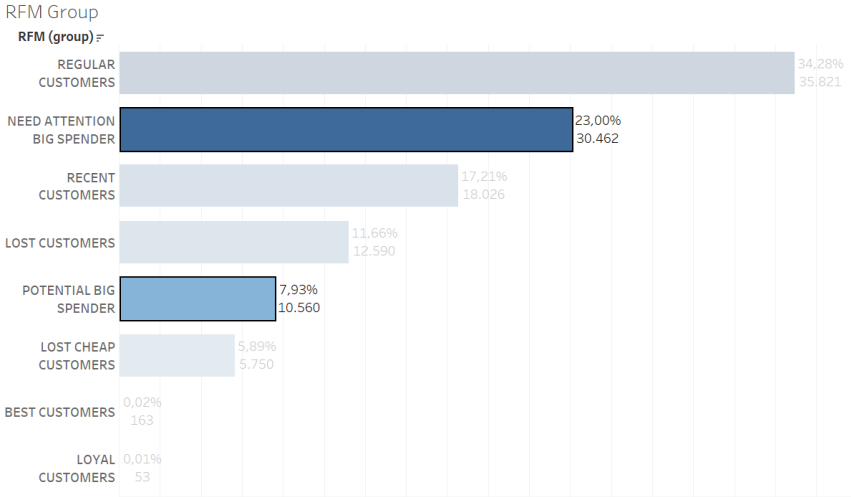
R X F

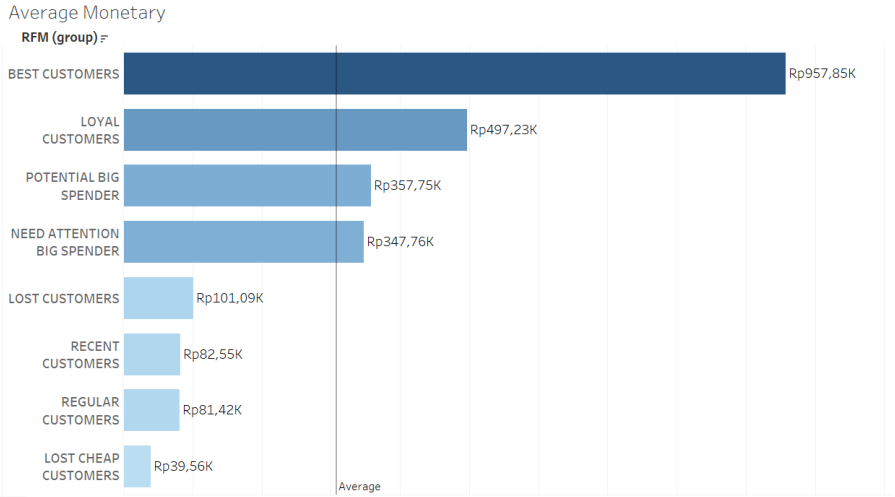


Most of them have a fairly distant last shopping time (5 - 9 months ago)

Group of customers that have highest monetary but less in frequency

WHY POTENTIAL BIG SPENDER & NEED ATTENTION BIG SPENDER?





- There are quite several of them, it will be very profitable if they can level up to become best customers

- Conduct in-depth analysis of events 5 – 9 months ago
- **Potential big spender**: Offer personalized recommendations, encourage them to buy products more frequently so they can level up their member status with so much benefit

- Can shop with a large nominal

- **Need attention big spender**: reach them via email/notifications, make subject lines of emails very personalized, revive their interest by a specific discount on a specific product.

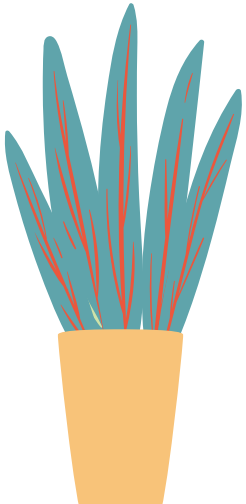
SUGGESTION

Customer	RFM Score	Marketing Tips
Best Customers	444	Give the big bonus points when they shop and special promos for the best customers.
Loyal Customers	x4x	Offer personalized recommendations, upselling
Potential Big Spender	4x4	Offer personalized recommendations, encourage them to buy products more frequently so they can level up their member status with so much benefit.
Need Attention Big Spender	xx4	Reach them via email/notifications, make subject lines of emails very personalized, revive their interest by a specific discount on a specific product.
Recent Customer	4xx	Gift them discounts/promo.
Lost Cheap Customers	111	Ignore.
Lost Customers	1xx	Send personalised emails, it is better to ignore it because attracting customers who have not shopped for more than 1 year will require more effort and cost a lot of money.
Regular Customers	Others	Make limited-time offers, send personalised emails, offer personalized recommendations

09

SUPERVISED LEARNING

"Adjust Estimated Delivery Date"



BACKGROUND

In business questions no 4, it was found that as many as **42.64%** of orders arrived **extremely early**. This amount is greater than the total orders that **arrived on time**, which was **13.98%**. This is not reasonable, therefore the estimated delivery time **needs to be adjusted**.

OBJECTIVES

Adjust the estimated delivery time so that the distance is not too far from the original delivery date

METRICS

- **Directed Error**

Metric that preservers the magnitude of the target and the direction of the errors made.

$$Directed\ error = \hat{y} - y$$

- **Mean Absolute Error**

Average of the absolute difference between the actual and predicted values in the dataset.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}|$$

- **Mean Squared Error (MSE)**

Average the squared difference between original and predicted values in the dataset.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2$$

- **Root Mean Squared Error (RMSE)**

Square root of Mean Squared Error.

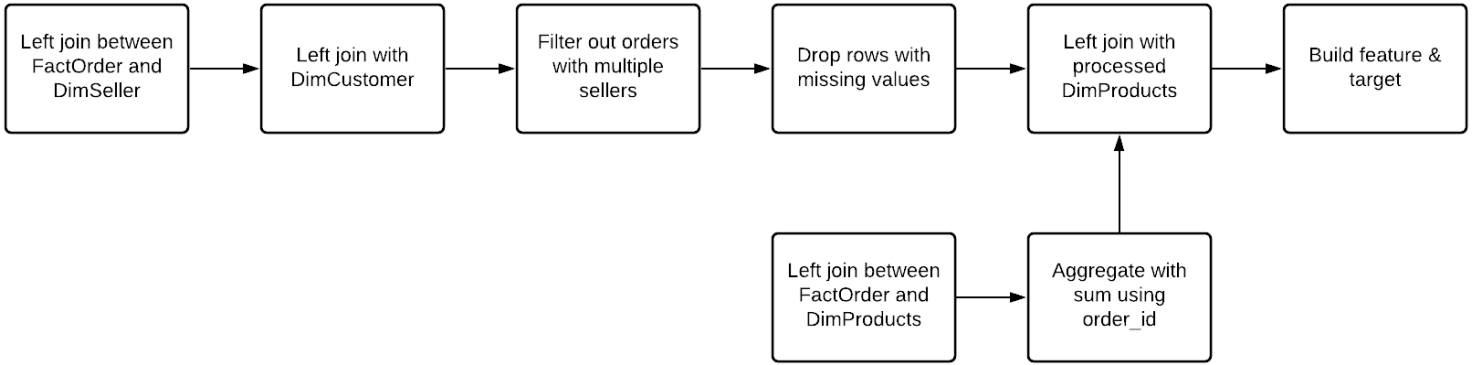
$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2}$$

LOAD DATA

- Data is taken from the data warehouse. Only the following table will be used.

Table Name	Total Row	Total Column
DimCustomers	5	96096
DimSellers	5	3096
FactOrder	35	113425
DimProducts	8	32952

PREPROCESSING DATA



- Left join between FactOrder and DimSeller.
- Left join again with DimCustomer.
- Then filter out orders that have more than 1 seller, because there are orders that have more than one seller and are located in different locations, but there is only one delivered date. The total order that has one seller in each order is 98,163.
- Drop rows with missing values.
- Left join with processed DimProducts.
- Build feature and target (next slide).

FEATURE

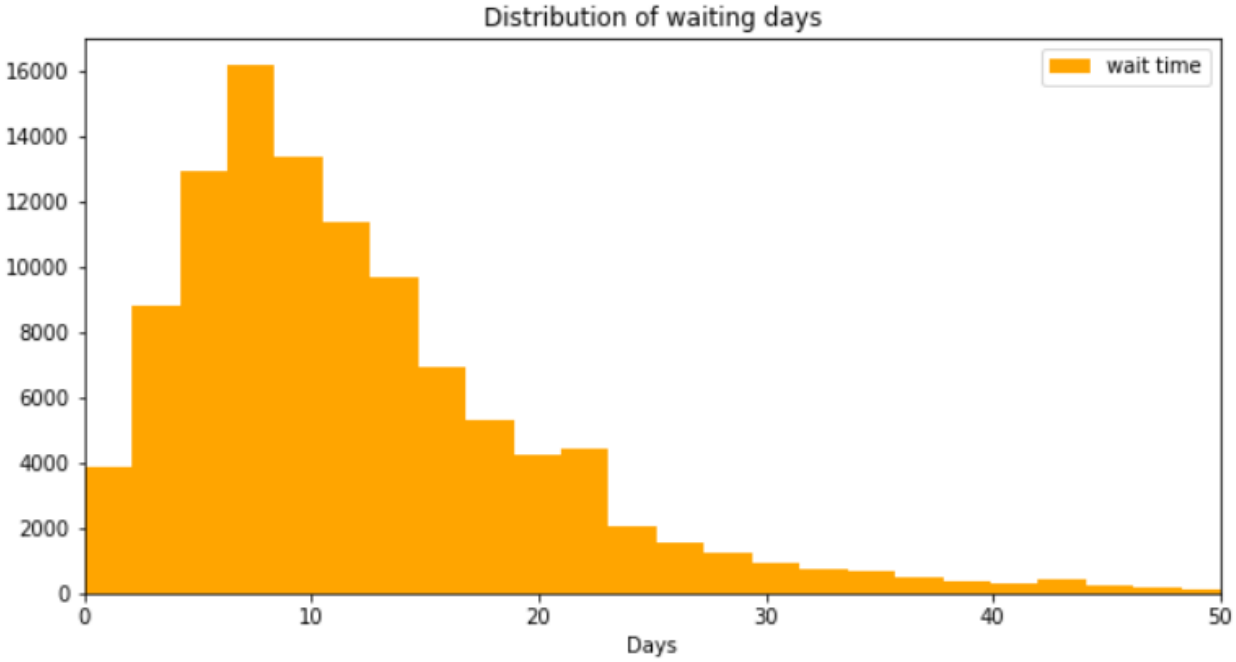
Rows marked in orange are features obtained from feature engineering

Feature	Description
customer_state	Customer province
seller_state	Seller province
product_volume_cm3	Product volume measure in cm3
product_weight_g	Product weight measure in grams
day_of_week	The day of the week based on order date (0-6, 0 = Monday)
month	The month of that date based on order date (1-12, 1 = January)
year	The year of that date based on order date

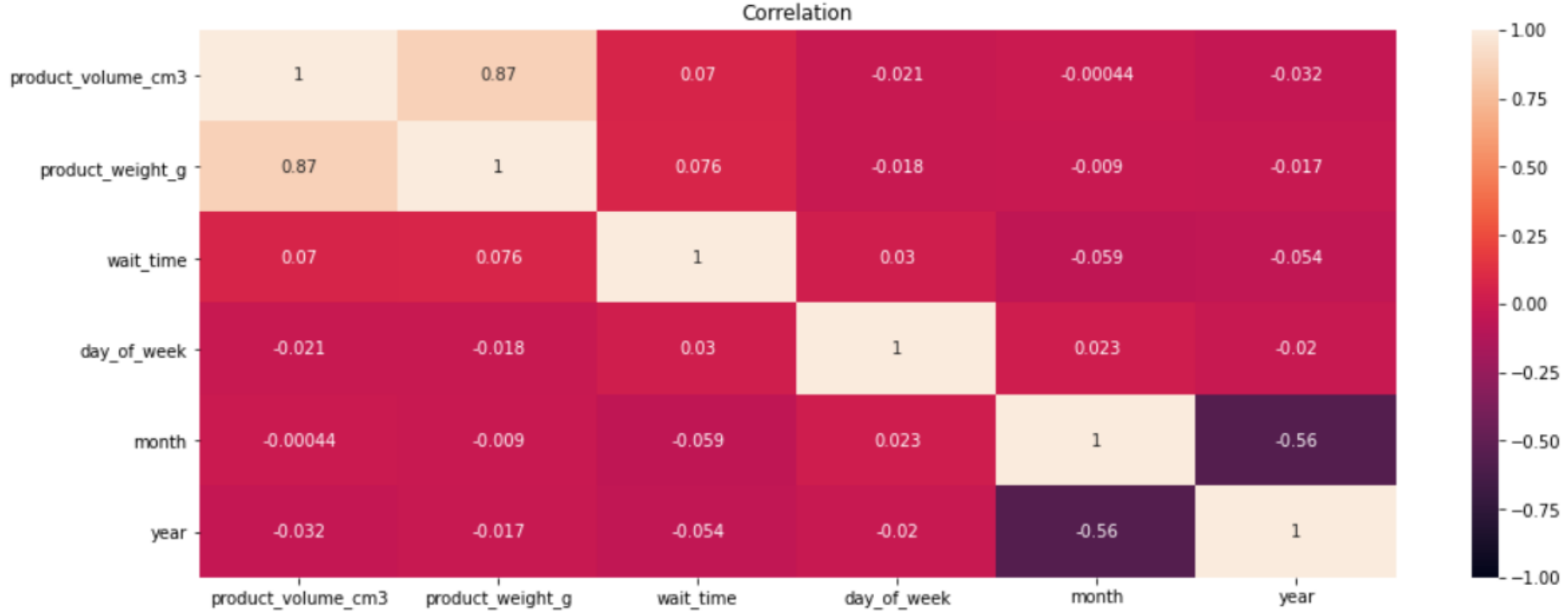
VISUALIZATION OF TARGET (WAIT TIME)

- **Distribution of wait time**

To calculate the wait time is delivered_date – order_date.



CORRELATION BETWEEN TARGET AND FEATURE



PREPOCESSING FEATURES

- **Preprocessing categorical features**

Categorical features include customer_state, seller_state, year, order_month, order_day. Preprocessing is using get_dummies

get_dummies = convert categorical variable into dummy/indicator variables.

- **Preprocessing numerical features**

Numerical features include product_volume_cm3 and product_weight_g. Preprocessing is using MinMaxScaler.

$$X_{sc} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

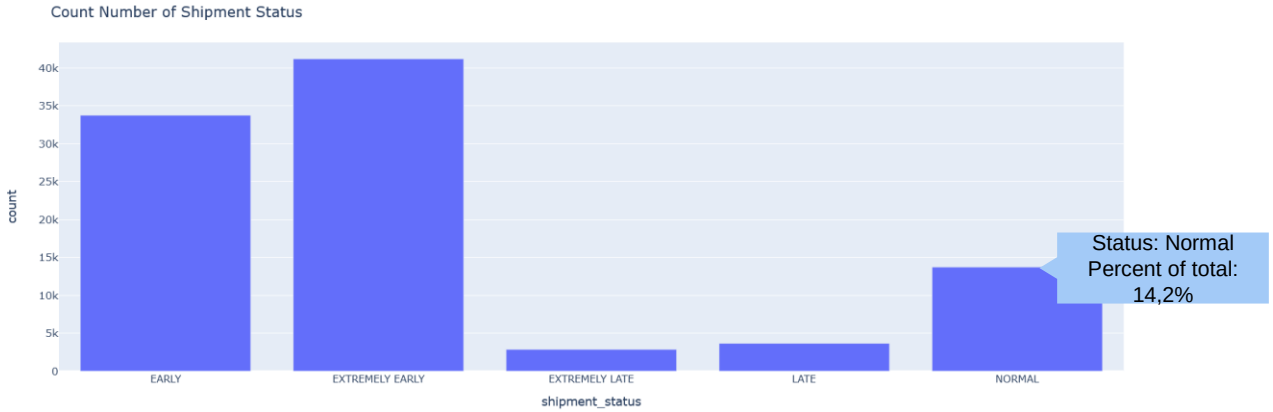
MODEL USING HYPERPARAMETER TUNING

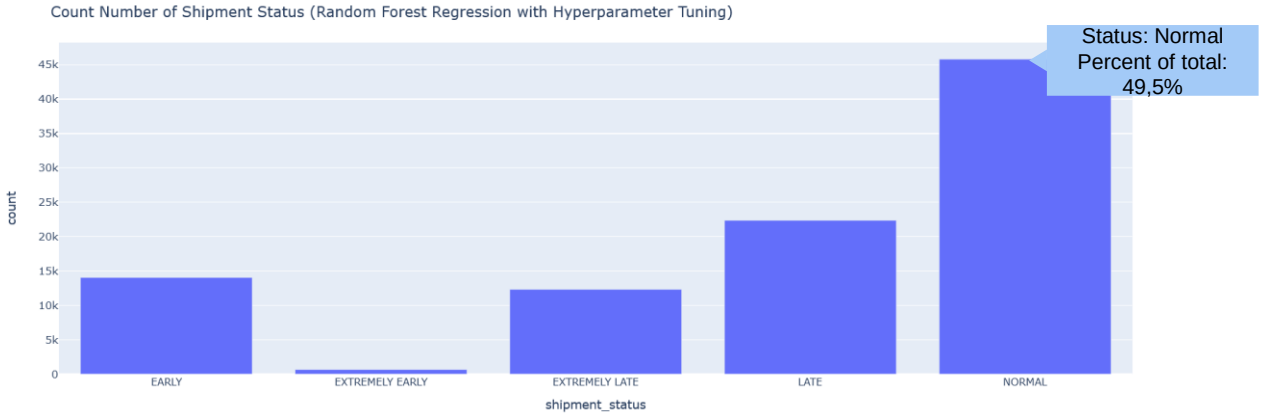
Model	Best Hyperparameter
Linear Regression with Hyperparameter Tuning	• fit_intercept : False • n_jobs : -1 • normalize : True
Ridge Regression with Hyperparameter Tuning	• alpha : 10 • fit_intercept : True • normalize : False • solver : cholesky
Lasso Regression with Hyperparameter Tuning	• alpha : 0.001 • fit_intercept : False • normalize : False • selection : random
Elastic Net Regression with Hyperparameter Tuning	• alpha : 0.0001 • l1_ratio : 0.0
Random Forest Regression with Hyperparameter Tuning	• max_features : sqrt • min_samples_leaf : 5 • n_estimators : 200

RESULT

Model	Directed Error	MSE	RMSE	MAE
Original Data	11.879840871099988	244.66256081730992	15.641693029122836	13.300832064846894
Linear Regression without Hyperparameter Tuning	0.0031657894136623284	79.29071748546892	8.904533535535082	5.826485465679313
Linear Regression with Hyperparameter Tuning	0.0004417485195023224	79.29022344332685	8.904505794446251	5.8258756107117895
Ridge Regression without Hyperparameter Tuning	0.0005377459398175368	79.28917049024821	8.904446669515641	5.825872401637961
Ridge Regression with Hyperparameter Tuning	0.0004407797461583732	79.28282096278585	8.90409012548648	5.825726858073235
Lasso Regression without Hyperparameter Tuning	9.715380220469766e-08	90.71908223786426	9.524656541727069	6.406244791614647
Lasso Regression with Hyperparameter Tuning	-0.0025917988006382907	79.28724122575531	8.904338337336206	5.824885857554923
Elastic Net Regression without Hyperparameter Tuning	6.658209179889073e-06	90.25032408962853	9.500017057333556	6.382150168057148
Elastic Net Regression with Hyperparameter Tuning	0.0004441711442821272	79.28298524834035	8.90409935076762	5.825728790773919
Random Forest Regression without Hyperparameter Tuning	0.1150174566129856	83.63595840140434	9.145269728193059	5.9056959198050745
Random Forest Regression with Hyperparameter Tuning	0.014058656000391708	76.46514378967417	8.744435018323035	5.662922168107006

COUNT NUMBER OF SHIPMENT STATUS DIFFERENCE

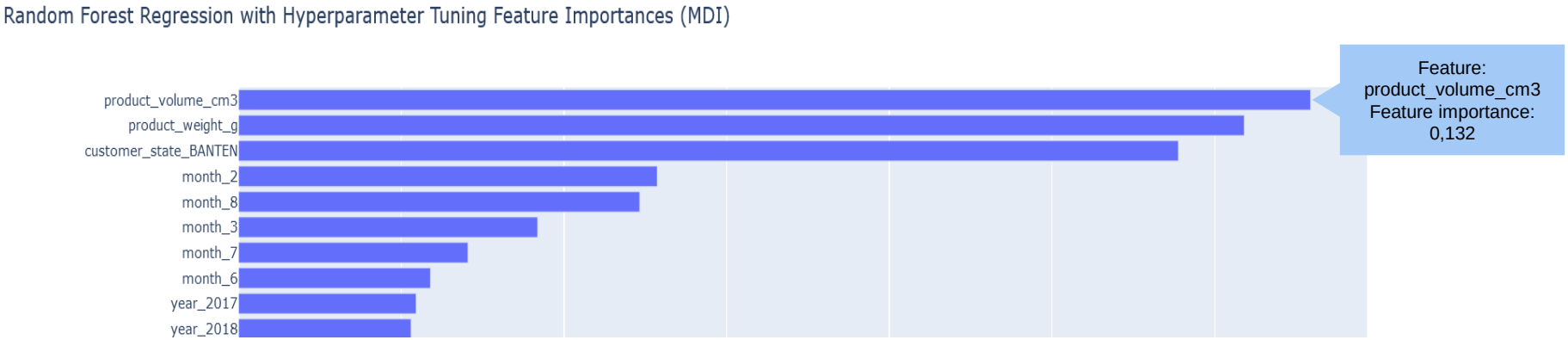




Prior to modelling, many shipments arrived **too soon** than estimated. After modelling, **49,5%** of deliveries arrived **on time**. This means that on the new model, they can estimate arrival times better than the previous model.

FEATURE IMPORTANCE

The **random forest algorithm** can measure the **relative importance** of **each feature** on the prediction. Sklearn provides a tool for this that measures a feature's importance by looking at how much the tree nodes use that feature, is the feature contributes enough. It computes this score automatically for each feature after training and scales the results so the sum of all importance is equal to one.

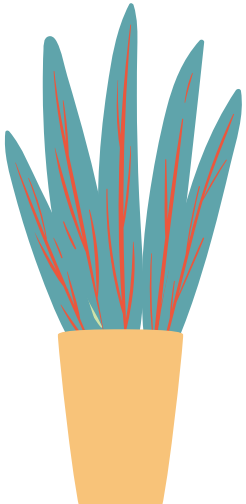


The most **influential feature** in this modelling is **product_volume_cm3** or volume of the product. The feature importance that is much higher than other features.

10

UNSUPERVISED LEARNING

“CUSTOMER SEGMENTATION USING RFM”



BACKGROUND

Customer segmentation is the practice of **dividing a customer** base into groups of individuals that are **similar** in specific ways. One of the techniques of customer segmentation is **RFM** (recency, frequency, monetary). RFM segmentation allows marketers to **target specific clusters of customers** with communications that are much **more relevant to their behavior**.

OBJECTIVES

Create best clustering for customer segmentation.

LOAD DATA

- Data is taken from the data warehouse. Only the following table will be used.

Table Name	Total Row	Total Column
DimCustomers	5	96096
FactOrder	35	113425

RFM

- **Join Between Customers & Orders**

Inner join the customer & fact orders dimension table and only use the user_name, order_date_key, order_id, and total_payment columns

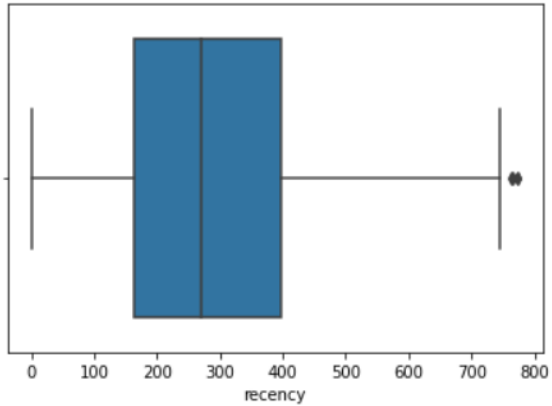
- **Declaring time_limit**

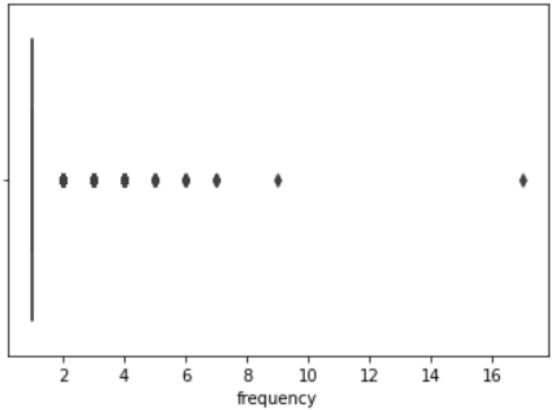
time_limit is the last time an order was placed on each customer plus 1.

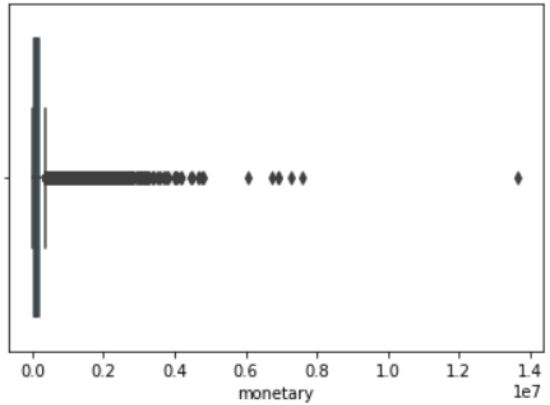
- **Build RFM**

recency	the maximum time an order occurs overall - time_limit each customer
frequency	Counting the number of orders per customer using order_id
monetary	Total payment

HANDLING OUTLIERS







	recency	frequency	monetary
Minimum limit	-184.5	1	-117426.25
Maximum limit	747.5	1	362083.75
Outliers percentage under the minimum limit	0.0%	0.0%	0.0 %
Outliers percentage under the maximum limit	0.004162504162504162 %	3.118886067518628 %	7.950713899179952 %

CHECK FOR QUANTILE

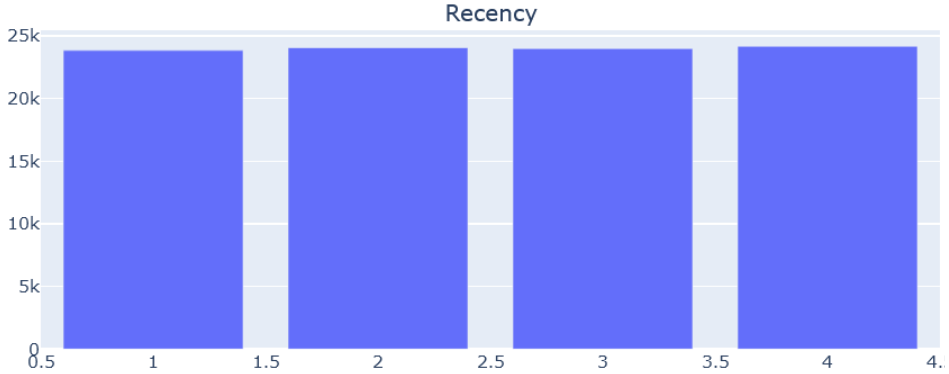
	recency	R
Q1	recency ≤ 165	4
Q2	165 < recency ≤ 270	3
Q3	270 < recency ≤ 399	2
Q4	recency > 399	1

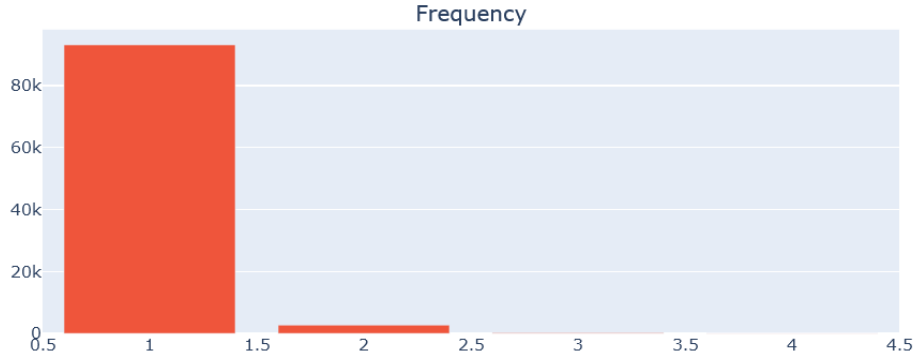
frequency	Count user_name	F
1	93099	1
2	2745	2
3	203	3
4	30	4
5	8	4
6	6	4
7	3	4
9	1	4
17	1	4

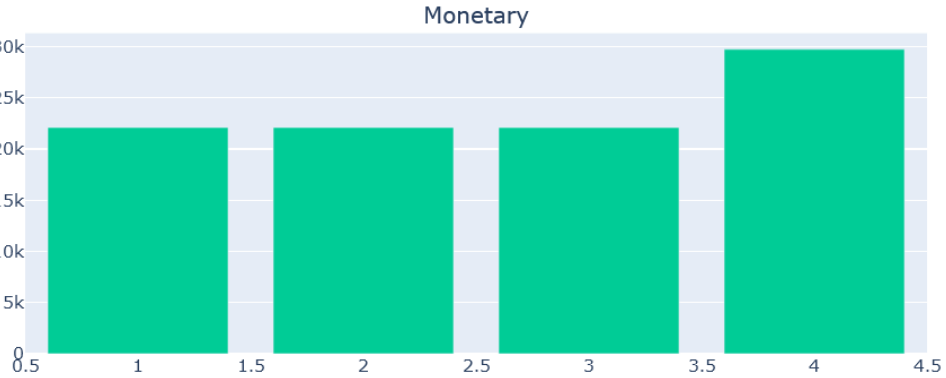
	monetary	M
Q1	monetary ≤ 58980	1
Q2	58980 < monetary ≤ 98840	2
Q3	98840 < monetary ≤159252.5	3
Q4	monetary > 159252.5	4

This is a limit in the formation of recency and monetary. Recency, frequency, and monetary will be divided into 4 parts.

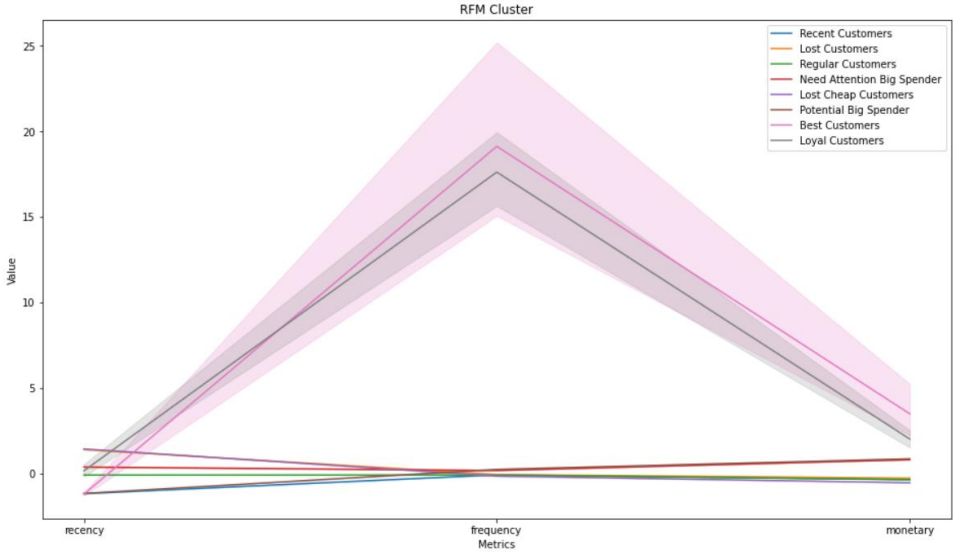
RFM

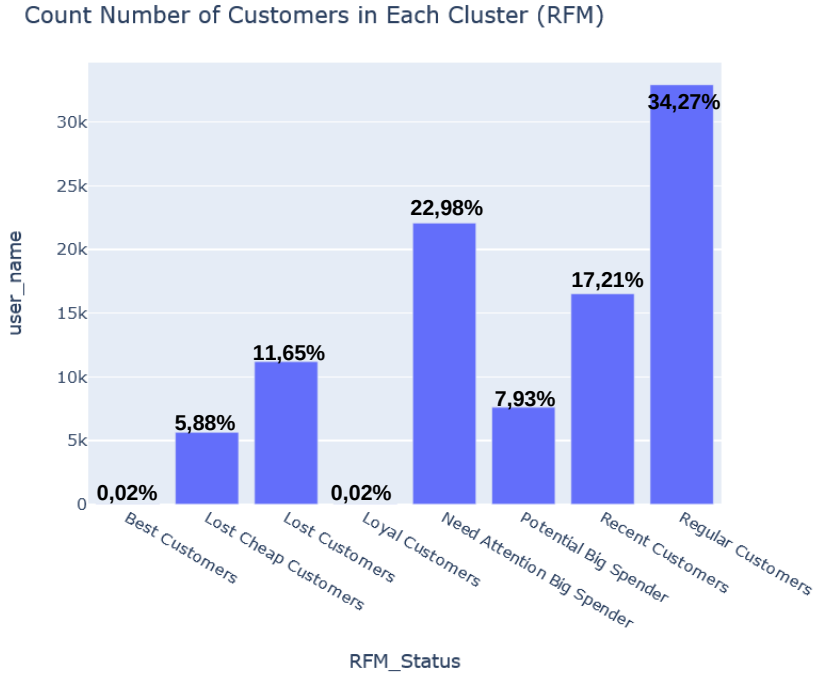

Recency
25k
20k
15k
10k
5k
0
0.5
1
1.5
2
2.5
3
3.5
4


Frequency
80k
60k
40k
20k
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5


Monetary
30k
25k
20k
15k
10k
5k
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5

RFM

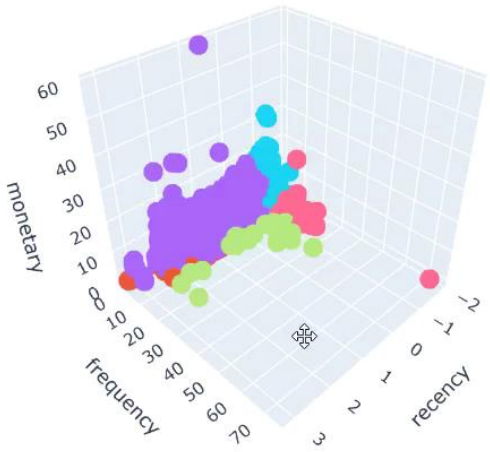

RFM Cluster
Recent Customers
Lost Customers
Regular Customers
Need Attention Big Spender
Lost Cheap Customers
Potential Big Spender
Best Customers
Loyal Customers
Value
25
20
15
10
5
0
recency
frequency
monetary
Metrics

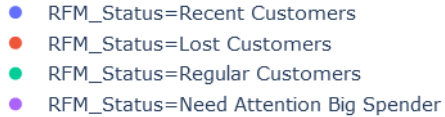

Count Number of Customers in Each Cluster (RFM)
user_name
30k
25k
20k
15k
10k
5k
0
0,02%
5,88%
11,65%
0,02%
22,98%
7,93%
17,21%
34,27%
Best Customers
Lost Cheap Customers
Lost Customers
Loyal Customers
Need Attention Big Spender
Potential Big Spender
Recent Customers
Regular Customers
RFM_Status

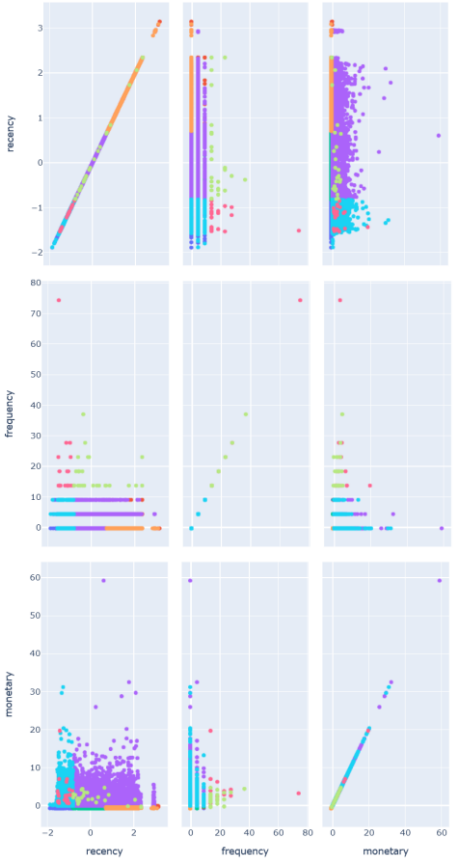
RFM

Advantage:

- Can divide customers based on specific criteria (highest shopping frequency, highest spending, customers who have high frequency, monetary, and high recency)

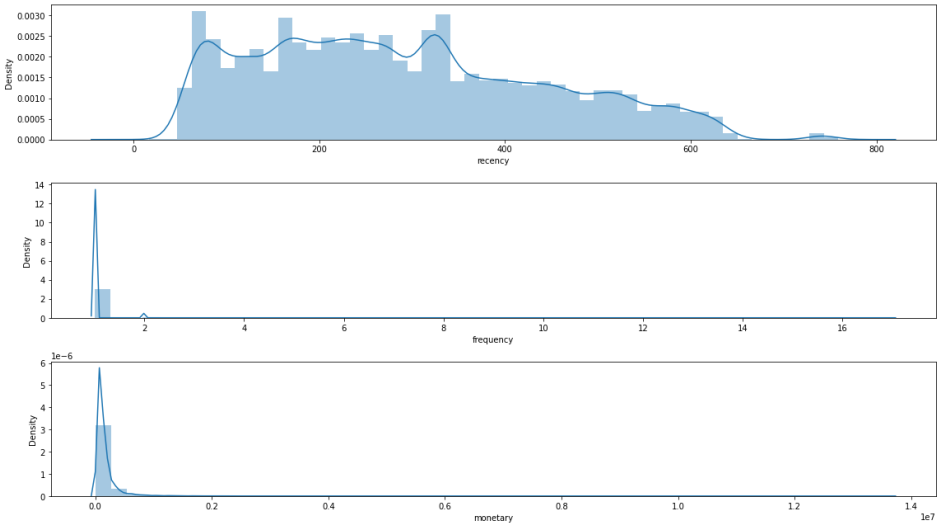




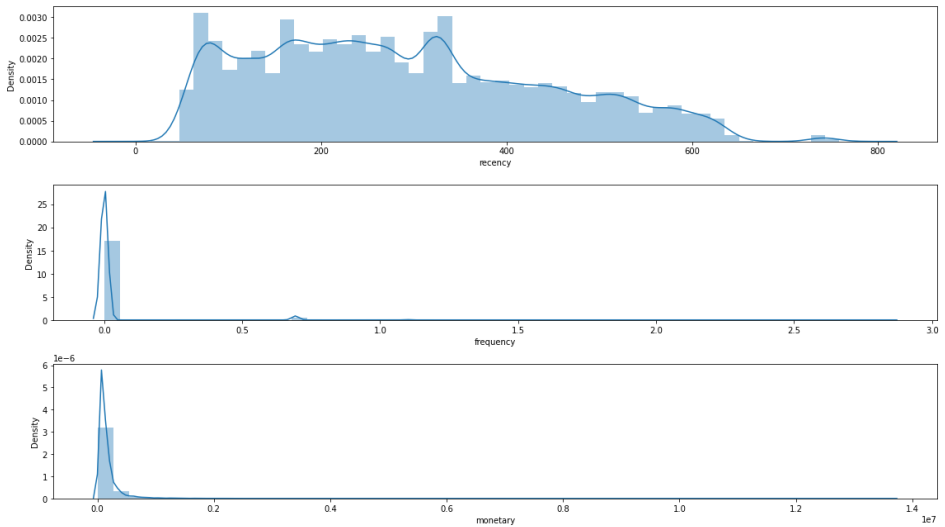


NORMALIZATION (KMEANS, BIRCH, OPTICS)

Normalization is only carried out on **frequency** because the total number of customers in each frequency has a fairly large inequality.



BEFORE NORMALIZATION



AFTER NORMALIZATION

SCALING (KMEANS, BIRCH, OPTICS)

Scaling is using StandardScaler.

Standardization:

$$z = \frac{x - \mu}{\sigma}$$

Mean:

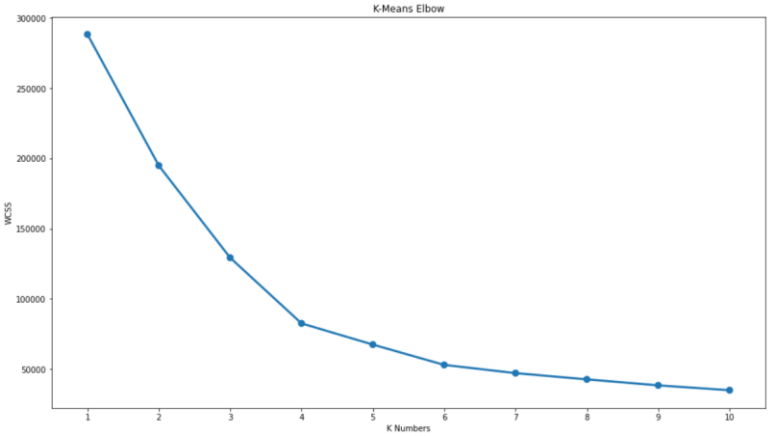
$$\mu = \frac{1}{N}\sum_{i=1}^{N}(x_i)$$

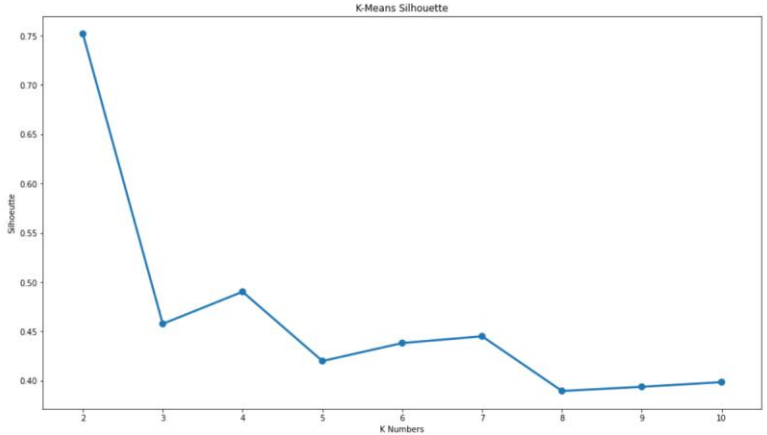
Standard Deviation:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i}^{N}(x_i - \mu)^2}$$

user_name	recency	monetary	frequency
0000366f3b9a7992bf8c76cfdf3221e2	-0.835036	-0.100783	-0.175637
0000b849f77a49e4a4ce2b2a4ca5be3f	-0.815482	-0.604035	-0.175637
0000f46a3911fa3c0805444483337064	1.941708	-0.345060	-0.175637
0000f6ccb0745a6a4b88665a16c9f078	0.533782	-0.531954	-0.175637
0004aac84e0df4da2b147fca70cf8255	0.318682	0.140468	-0.175637

KMEANS MODELLING

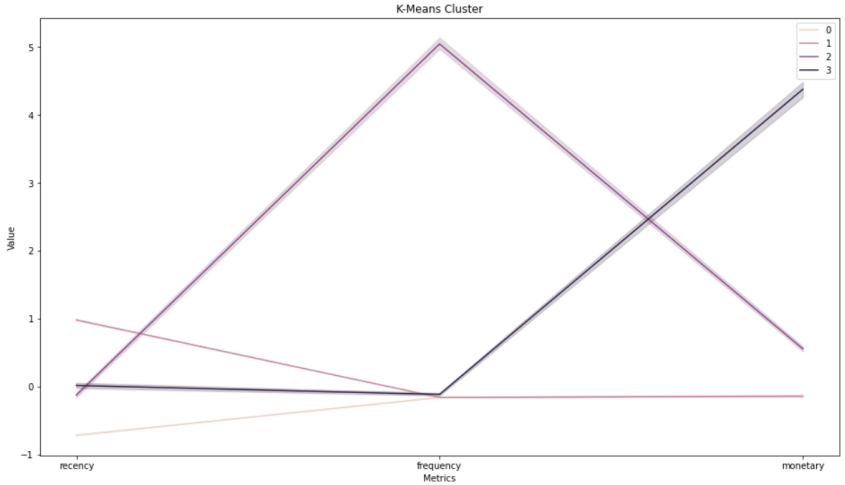


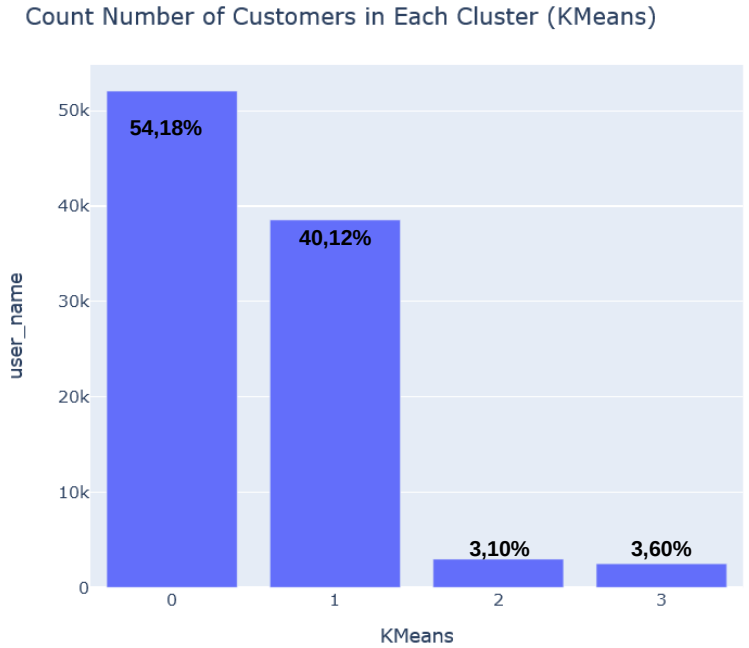


Silhouette Score (max_iter = 100, init= ‘k-means++’)

Cluster	Silhouette Score
2	0.7517797029179484
3	0.45771592654193066
4	0.490269525589124
5	0.42002437486655225
6	0.438253675897778
7	0.4451040664250428
8	0.3896524204148909
9	0.3938681767015192
10	0.3986208012081028

KMEANS MODELLING

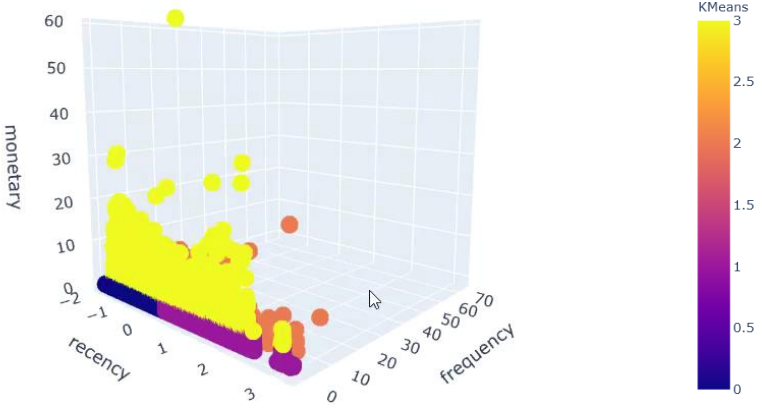


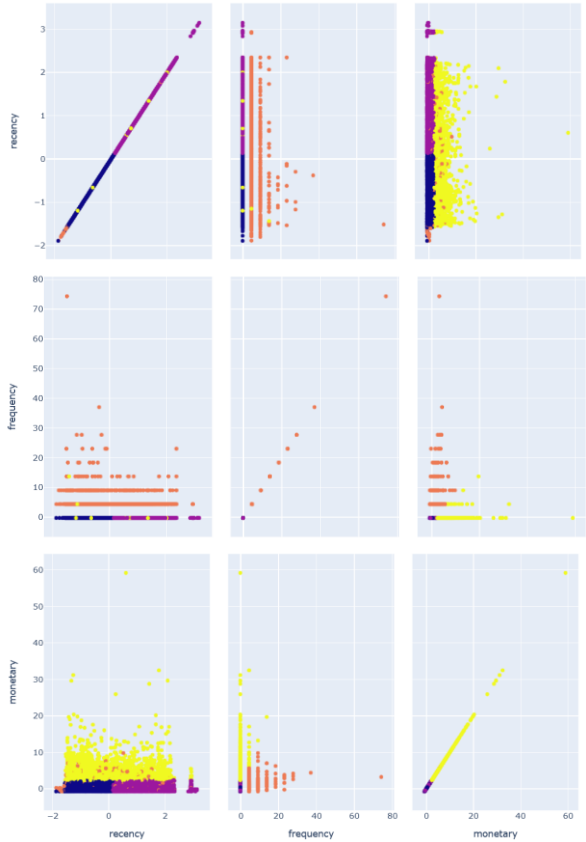


KMEANS MODELLING

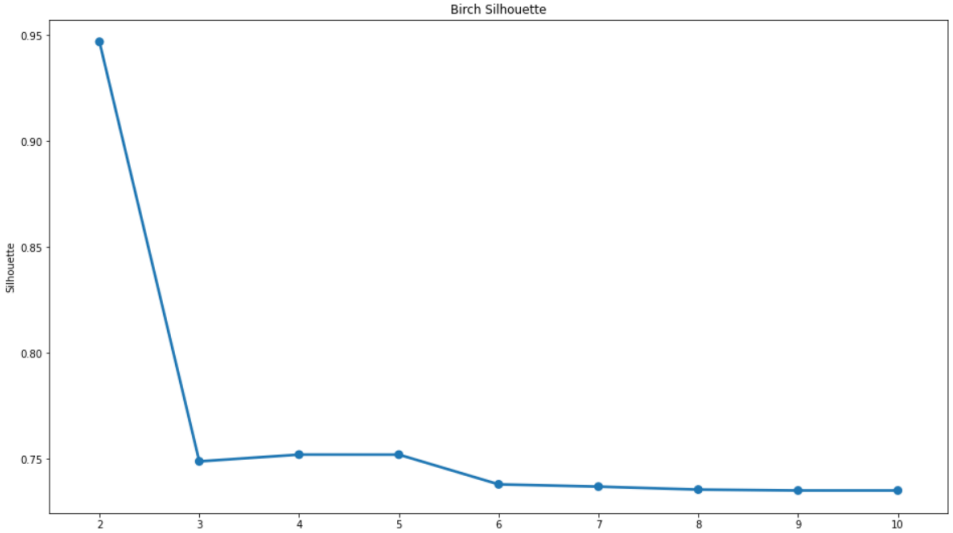
Advantage:

- Can separate customers who have a high shopping frequency, high spending, and high recency





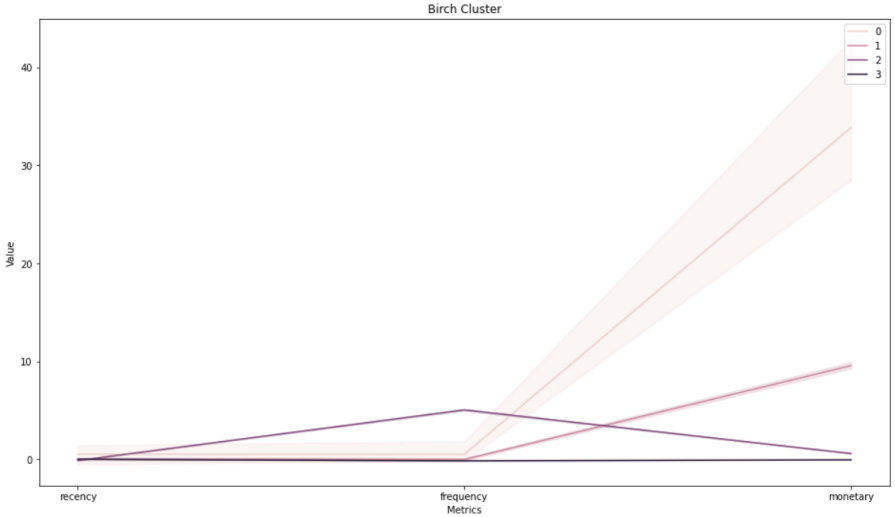
BIRCH MODELLING

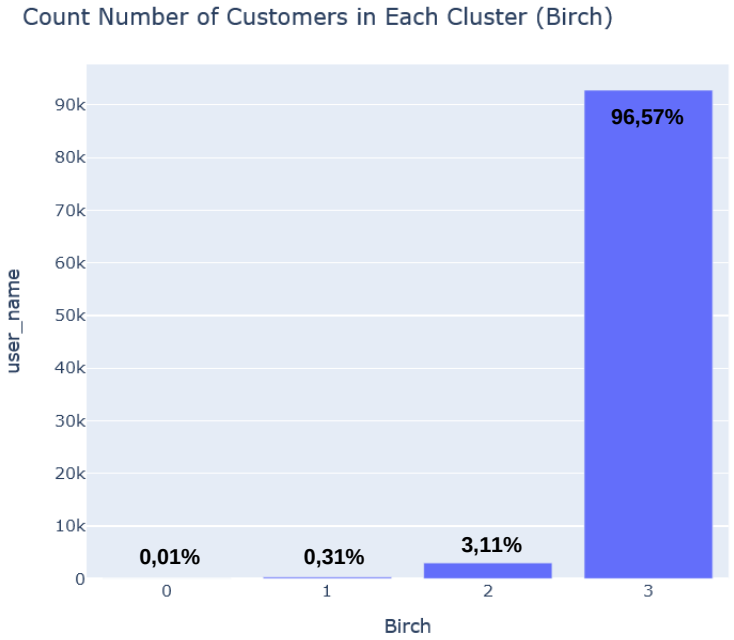


Silhouette Score

Cluster	Silhouette Score
2	0.9468885924086847
3	0.7488208001526748
4	0.7520245513585802
5	0.7520014183081782
6	0.7379752003204659
7	0.7369073103754451
8	0.7355094165562268
9	0.7350791922577496
10	0.7350714636575758

BIRCH MODELLING

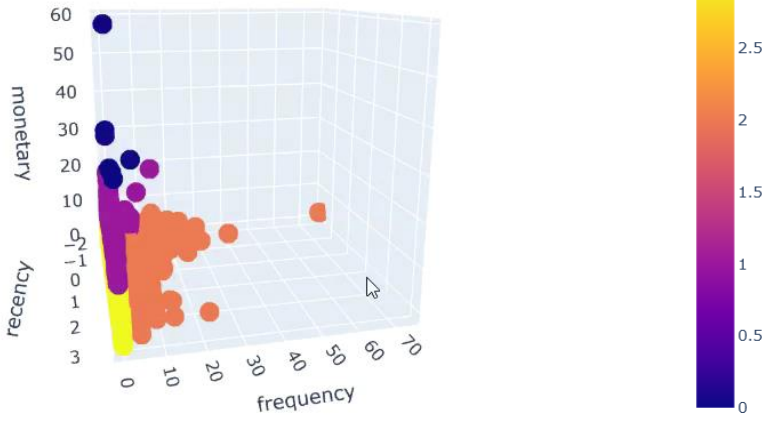


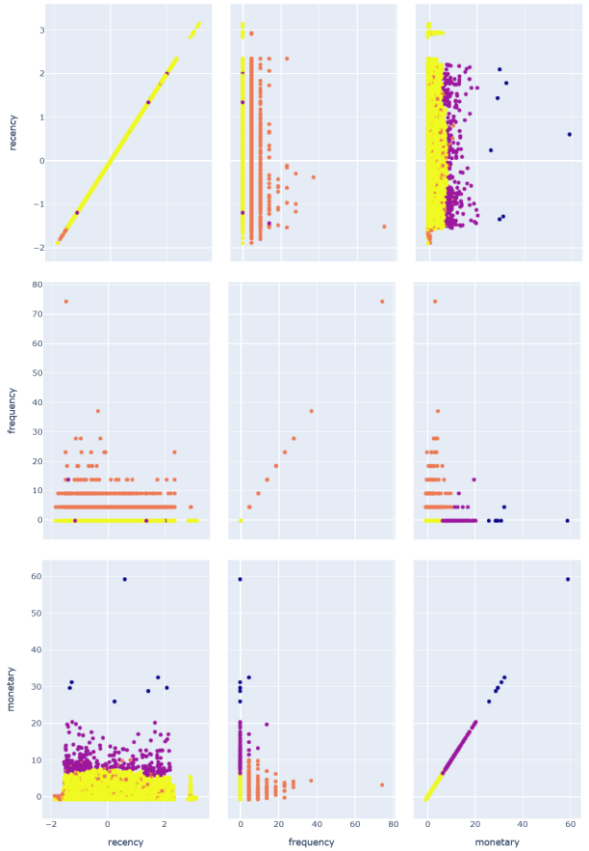


BIRCH MODELLING

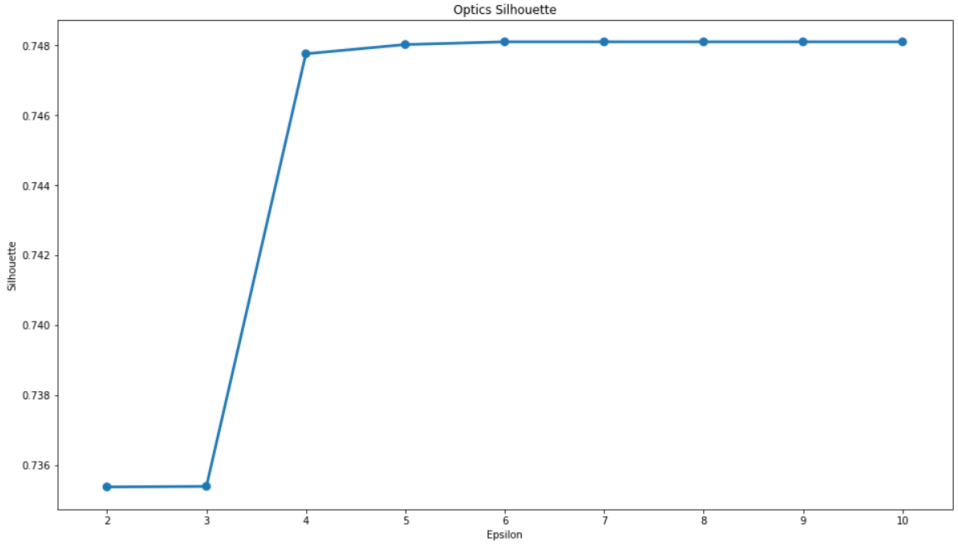
Advantage:

- Can separate customers who have a high shopping frequency, and divide customers up to 3 segments based on high spending





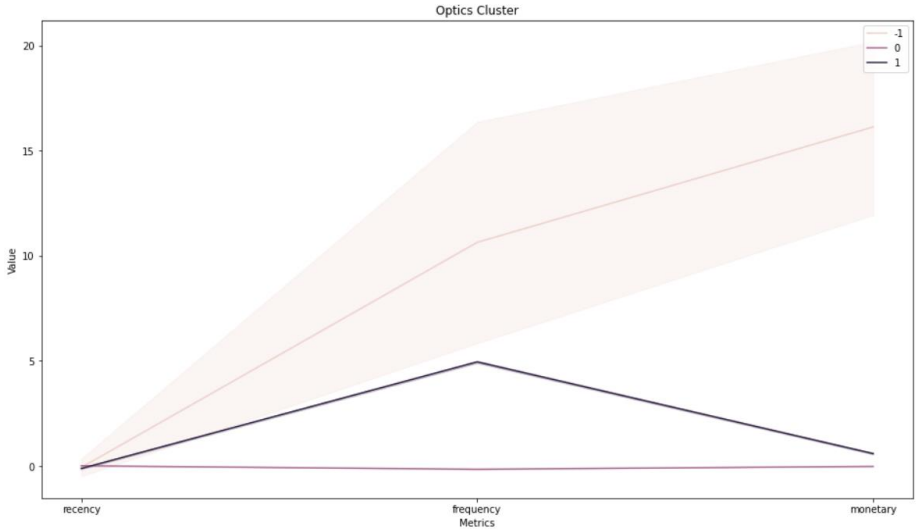
OPTICS MODELLING

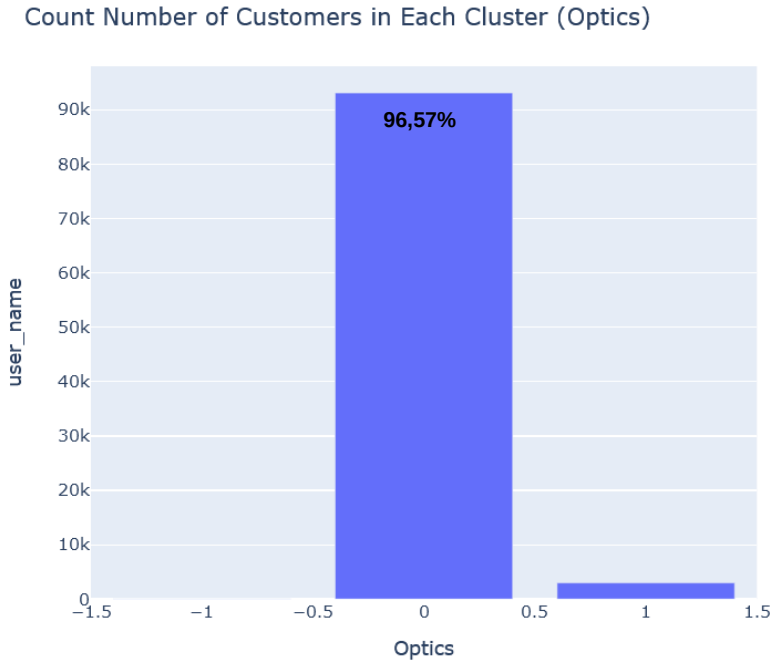


Silhouette Score (min_samples = 300)

Epsilon	Silhouette Score
2	0.7353730783275159
3	0.7353882489934999
4	0.7477505713411909
5	0.7480140604901347
6	0.7480917157564253
7	0.7480917157564253
8	0.7480917157564253
9	0.7480917157564253
10	0.7480917157564253

OPTICS MODELLING

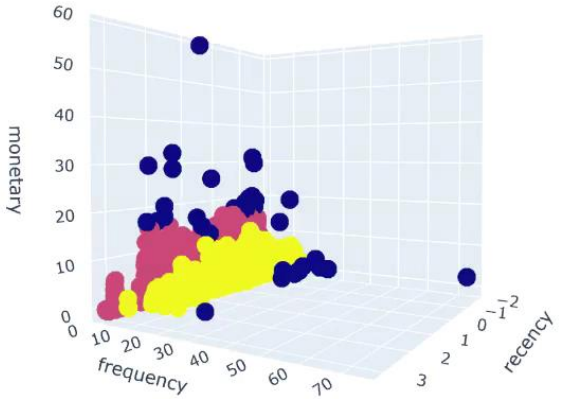


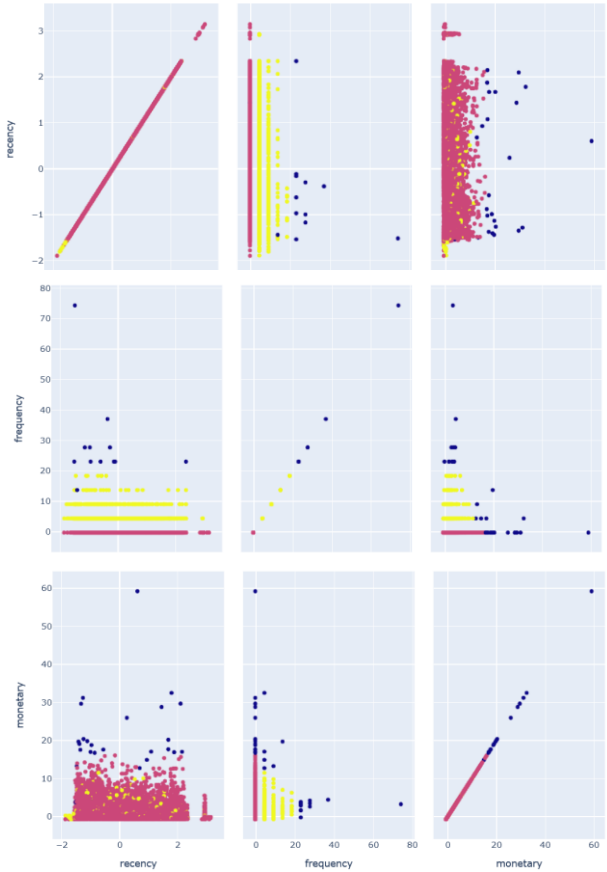


OPTICS MODELLING

Advantage:

- Can separate customers who have the most frequency or monetary





SILHOUETTE SCORE SUMMARY

Model	Silhouette Score								
	2	3	4	5	6	7	8	9	10
RFM							-0,23060		
K-Means	0,75178	0,45760	0,49026	0,41998	0,43825	0,44808	0,38900	0,40098	0,39105
Birch	0,94689	0,74882	0,75202	0,75200	0,73798	0,73691	0,73551	0,73508	0,73507
Optics	0,73537	0,73539	0,74775	0,74801	0,74809	0,74809	0,74809	0,74809	0,74809

In the table above, it is known that the **highest silhouette score** is **Birch**. In terms of **dividing customer segments**, **RFM** is a good model because it can **adapt to marketing needs**.
However, among k-means, birch, and optics, in my opinion **the best** is **k-means**, because it also divides customers based on **recency** so that from a marketing perspective, customers with low recency, frequency, monetary can be marked not to need to be reached out , but need to look at customer segments that need more attention.
While the distribution on Birch and Optics models is uneven.

THANK YOU

Do you have any questions?