Startup Campus

Kampus Merdeka
INDONESIA JAYA

Startup Campus Final Project

MITIGATE CUSTOMER CHURN

MACHINE LEARNING APPROACH

GROUP ITALY

MENTOR:

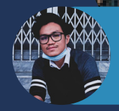
Muhammad Anwar Sanusi

MENTEE:

Faridha Berliana
DS03054

Fitri Ayu L.
DS03060

Fathan Mubina
DS03055



Gidion Buranda
DS03064

Faustina Sari K.
DS03057

Gress Polina S.
DS03065

CONTENTS

BUSINESS UNDERSTANDING

There's an E-commerce company named Fashion Campus that operates in the fashion industry with the "Indonesia Young Urbans" as the target market. By June 2022 this company already has 10,000 active users and reaches 100,000 orders each month.

Problem Statement

The interesting promos that attract new users, caused the rise of non-organic users. They tend to not come back to do any transaction which **increases the churn rate.**

Analytical Approach

Using the machine learning approach to predict possible factors, the analysis goal is to predict and know user churn probability, which is by providing solutions to increase and maintain user retention. We also use RFM Analysis to categorize the customer.

DATA UNDERSTANDING

CUSTOMER

PRODUCT

TRANSACTION

CLICK STREAMS

MERGED DATASET

43 COLUMNS
1,987,222 ROWS

NUMERICAL

Home location latitude
Home location longitude
Shipment location latitude
Shipment location longitude
Promo amount
Shipment fee
Age
Recency

CATEGORICAL

Gender
Device type
Device version
Home location
Home Country
Payment method
Payment status
Promo code
Event name
Created day
Created month
Created year
Created time
Churn

DATA UNDERSTANDING

Customers who had been 30 days not returned to use the application, categorized as churn. The team found customers who Churn over the past few years from 2016 - 2021, have always experienced an increase. From 2020 to 2021, this will be the year with the highest number of customers churning with a total of 4888 customers. The city location of the most churned customers is in Greater Jakarta with a total of 236 customers, followed by Central Java with a total of 131 customers.

Order

Customers made the most orders in 2021

The last churn customer made a purchase in May with a total of 4000 customers

Promo code

Customers use the AZ2022 promo in 2022 and Customers who use promo code AZ2022, mostly make payments by Credit Card with a total of 3,513 customers

Payment method

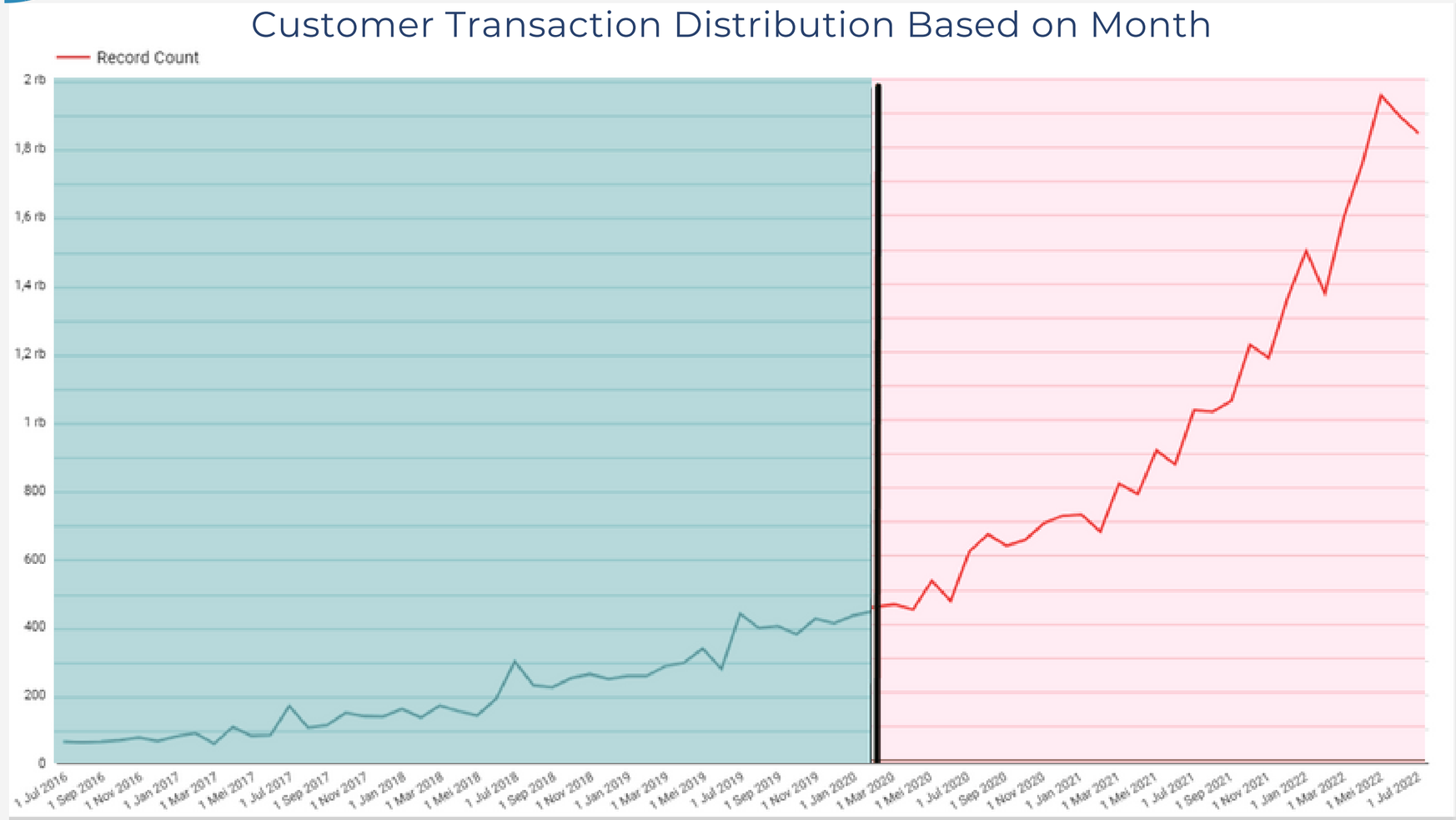
The payment method that is often used by customers is Credit Card with a total of 13,057 and rarely used is LinkAja with a total of 3,272

DATASET DEFINITION

- **Promo amount:** total discount from the promo
- **Shipment fee:** expenses that must be paid to ship a product through the delivery service.
- **Age:** age of customers
- **Recency:** the day lag of the user's last transaction with the current date
- **Home location latitude:** customer location' latitude
- **Home location longitude:** customer location' longitude
- **Shipment location latitude:** seller location' latitude
- **Shipment location longitude:** customer location' longitude
- **Gender:** customer gender
- **Device type:** device type that customer use to order (iOS=1, Android =0)
- **Device version:** device version that customer use to order
- **Home location:** customer location region
- **Home Country:** customer country
- **Payment method:** type of payment customer choose when paid their order
- **Payment status:** customer order payment status (Success=1, Failed=0)
- **Promo code:** promo that customer used to get discount
- **Event name:** events that user clicked while surfing the platform
- **Created day:** name of day customer transaction
- **Created month:** name of month customer transaction
- **Created year:** year of customer transaction
- **Created time:** parts of the day customer transaction
- **Churn: has customer churn?**

EXPLORATORY DATA ANALYSIS & VISUALIZATION

DATA VISUALIZATION

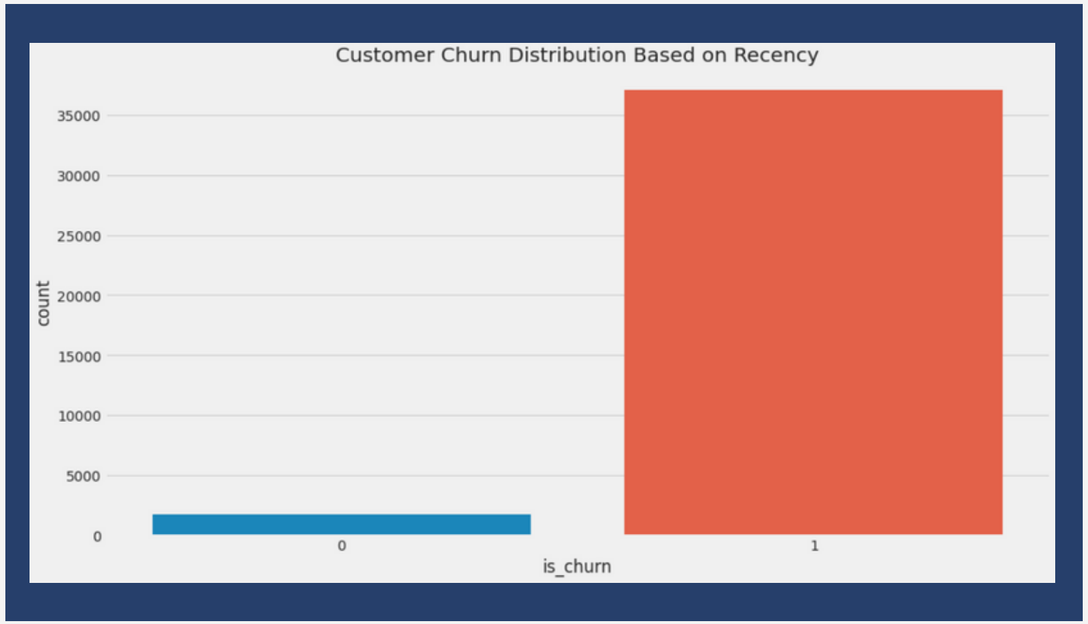


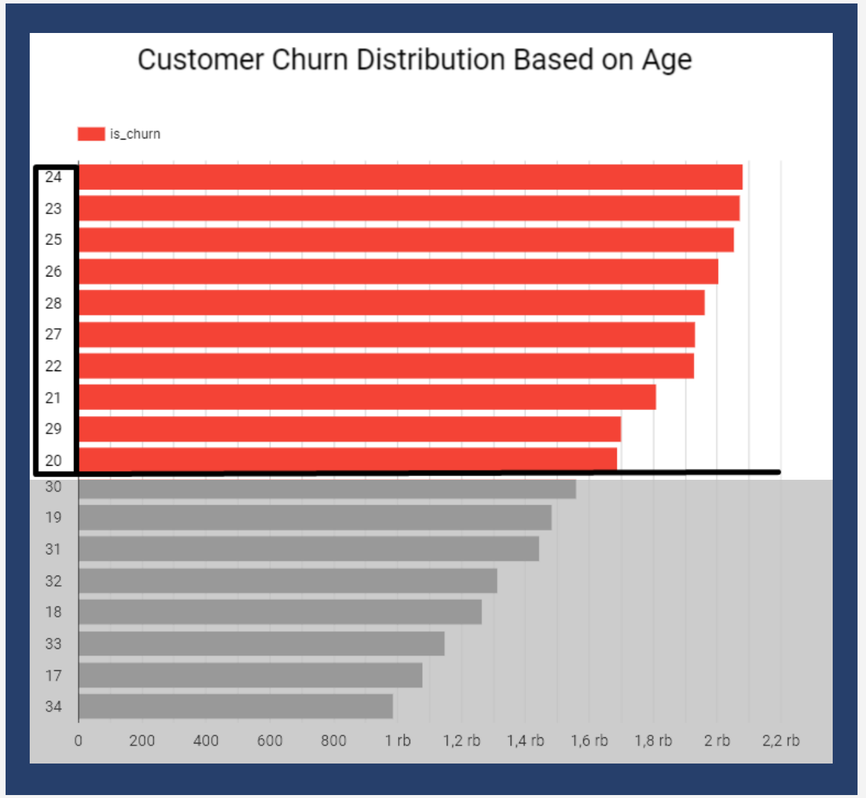
March 2, 2020
Indonesia First Covid Case

From the picture we can see that the transaction rate from customers is increasing from month to month

Also we can see that the transaction rate from customers has increased more after the arrival of Covid in Indonesia

DATA VISUALIZATION





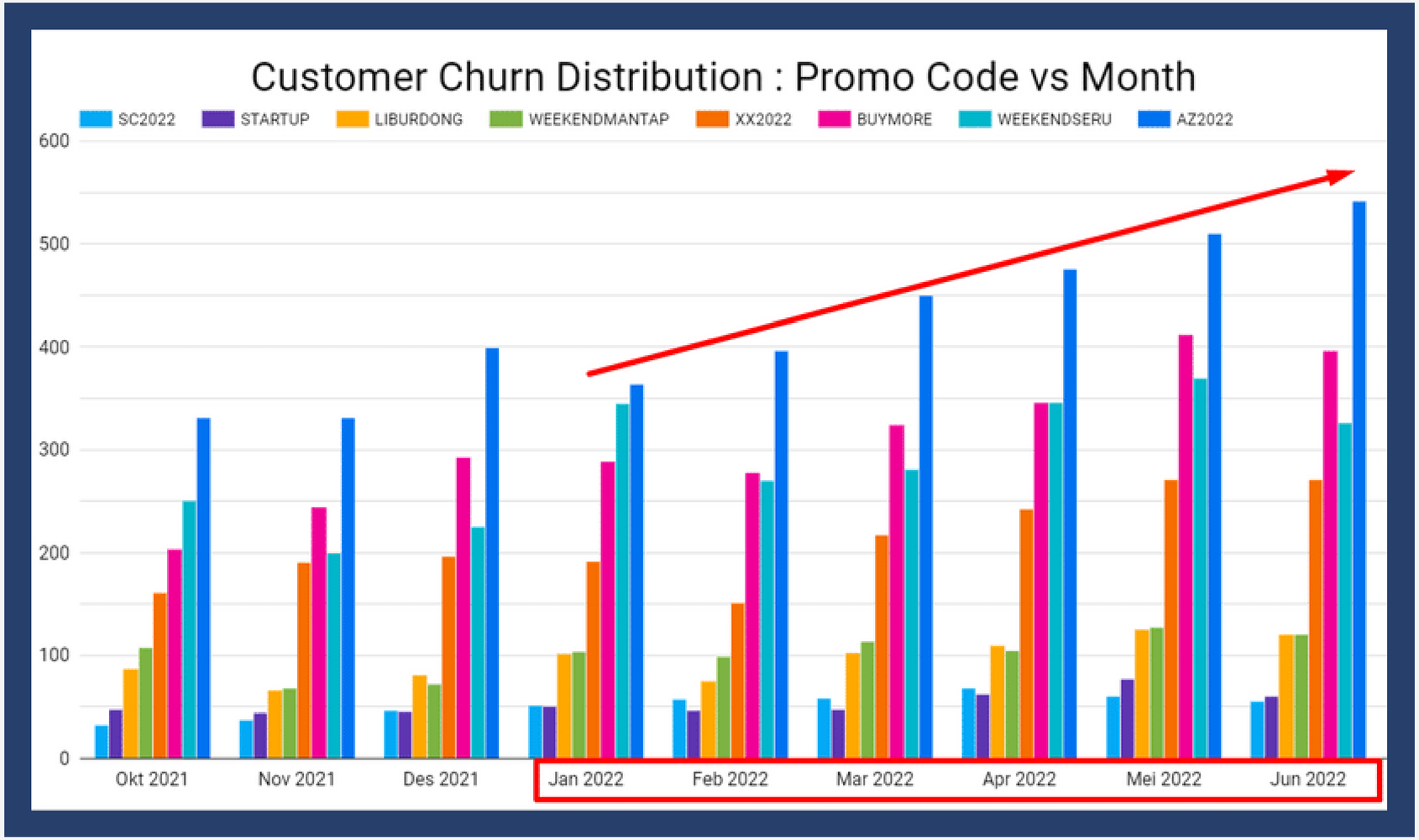
Recency:

With total users of 38879. Around 95% of customers were not returned to the application in the last 30 days and there is 5% of customer who still used the application in the last 30 days.

Age:

The range of customers' age is from 17 to 34 years old. Most customers who churn are between the age range of above 20 to 29 years old.

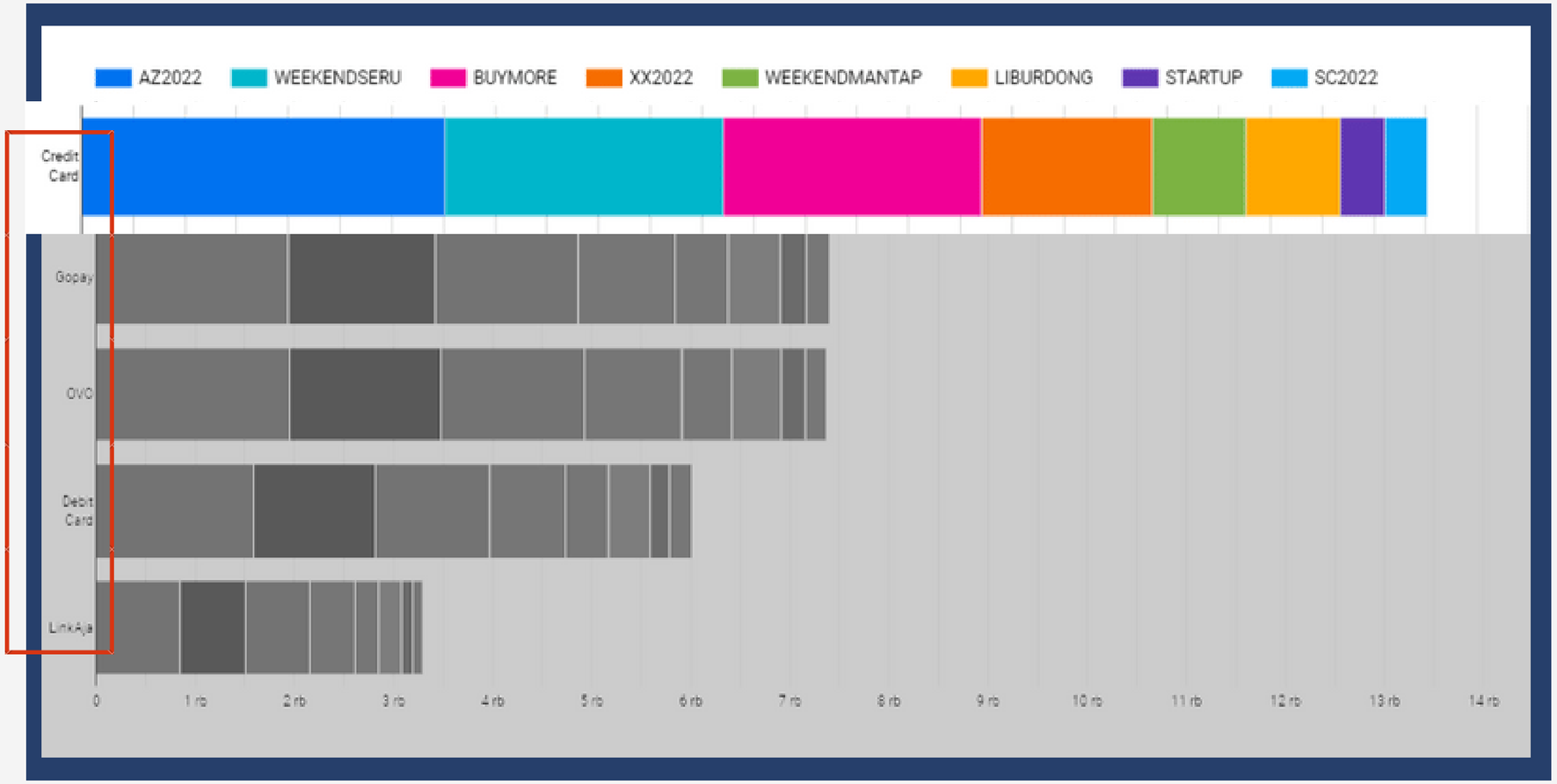
DATA VISUALIZATION



Churn occurs more frequently in early 2022. More customers who churn use the "AZ2022" promo code

DATA VISUALIZATION

Customer Churn Distribution: Payment Method vs Promo Code

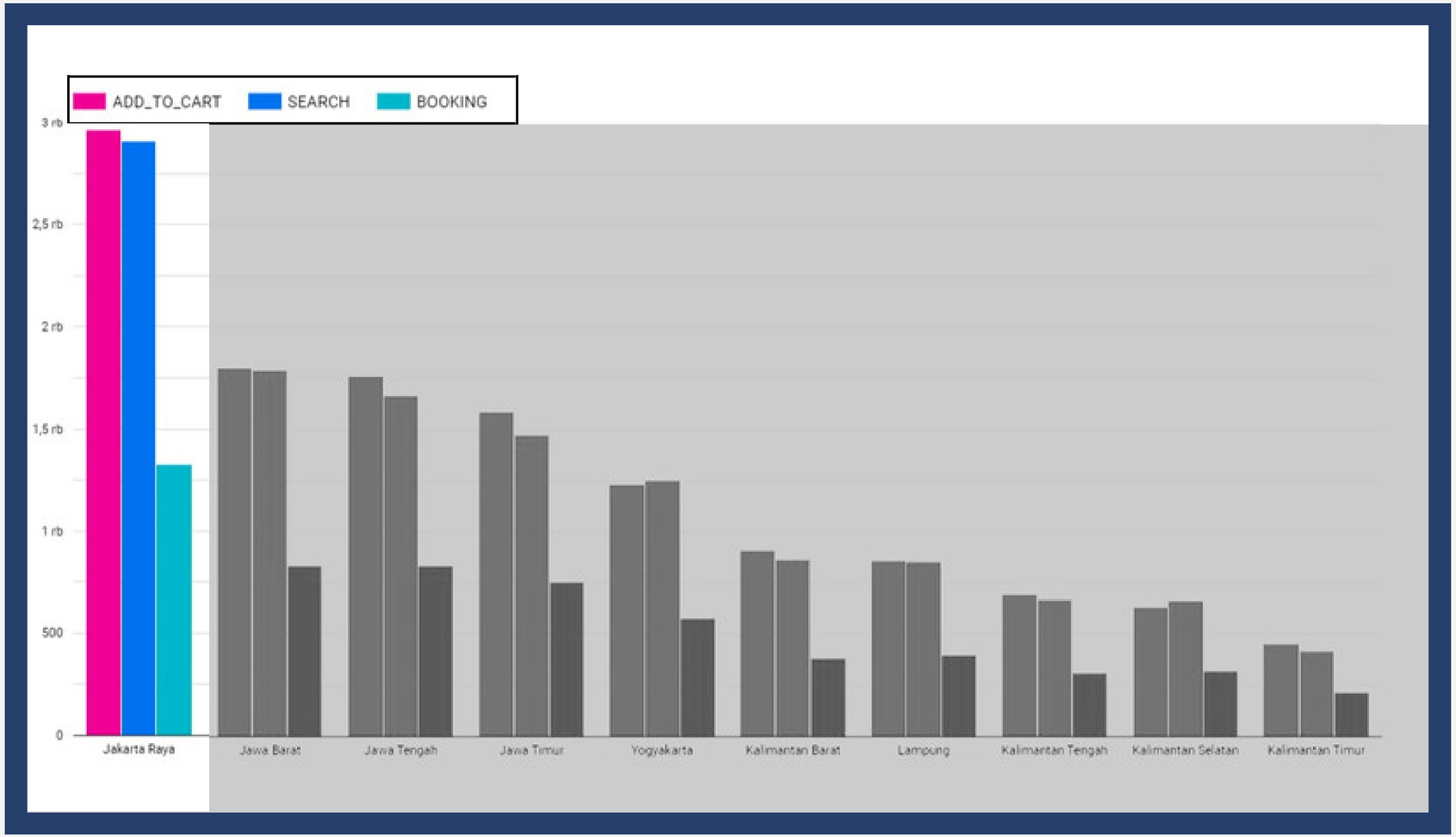


The Fashion Campus company provides 5 payment methods and the most frequently used is credit cards.

The bar chart shows that credit cards are often used together with the AZ2022 promo.

DATA VISUALIZATION

Customer Churn Distribution by Event

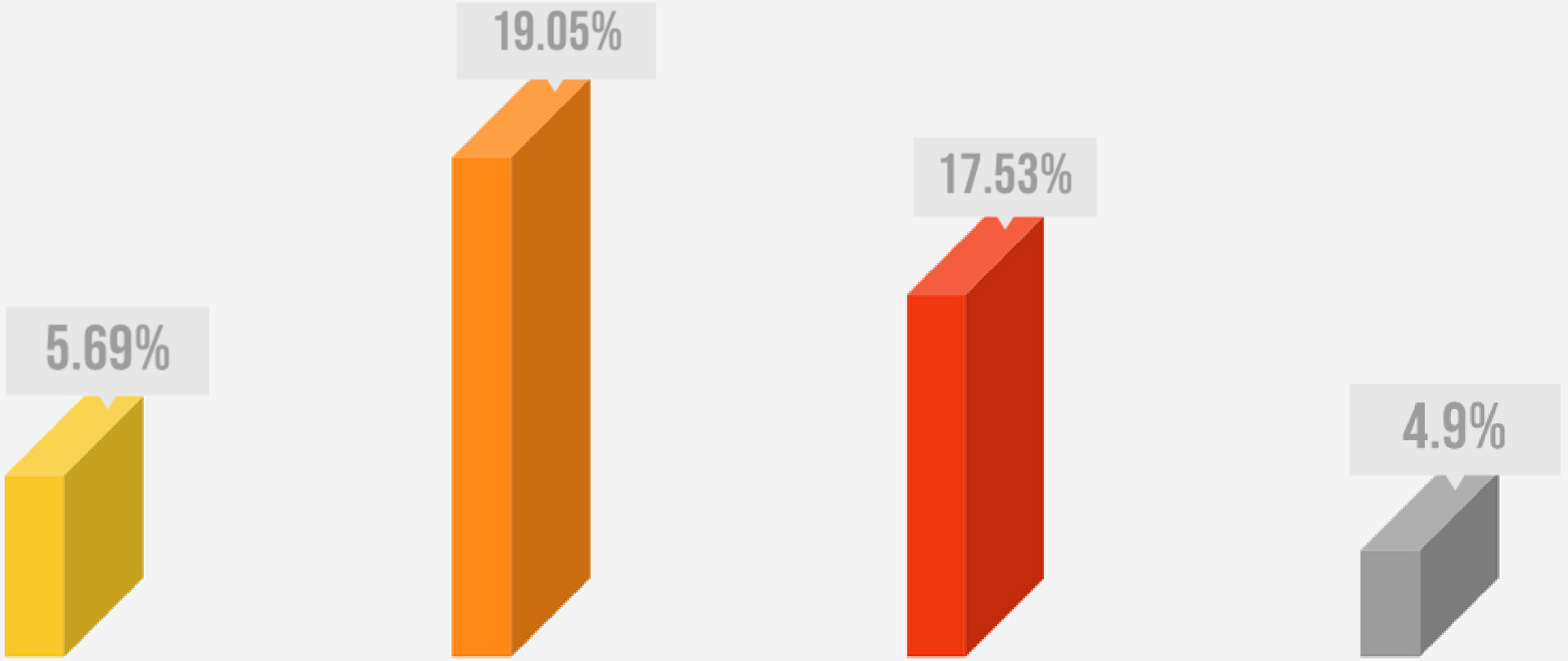


There are 3 events that customers do when makin transactions, is "add to cart" "search" and "booking"

Most customer churn comes from users who often click "add to cart" and follow by users who clicks "search".

An example is Jakarta Raya, which is the highest city to do "Add to Cart" events, but thats causing low "booking" events.

RFM ANALYSIS



CHAMPION

Visited most recently, buy often, and spend highest amount on transaction

POTENTIAL LOYALIST

A recent user, who spent a good amount

AT RISK

Spend a good amount but but long ago (not visited recently)

LOST CUSTOMER

Users who visited a long time ago, rarely visit and haven't spent much money on transactions.

DATA PREPROCESSING

In this step, we need to create a ready-to-use dataset in modelling process. The dataset is used to increasing the efficiency of resource, especially the Random Access Memory (RAM).

DATA ENCODING → **HANDLING OUTLIERS AND IMBALANCE DATA** → **FEATURE SELECTION**

ABOUT THE DATASET

0 Missing Value

customer_id	0
first_name	0
last_name	0
username	0
email	0
gender_x	0
birthdate	0
device_type	0
device_id	0
device_version	0
home_location_lat	0
home_location_long	0
home_location	0
home_country	0
first_join_date	0
created_at	0
booking_id	0
session_id	0
product_metadata	0
payment_method	0
payment_status	0
promo_amount	0
promo_code	0
shipment_fee	0
shipment_date_limit	0
shipment_location_lat	0
shipment_location_long	0
total_amount	0
event_name	0
event_time	0
event_id	0
traffic_source	0
event_metadata	0
product_id	0

0 Duplicates

```
data_final2.duplicated().sum()
```

```
0
```

Top Categorical Values

Gender : Woman

Master Category : Apparel

Device Type : Android

Product Usage : Casuals

Month : July

Payment Method : Credit Card

In this step we use merged dataset. Data encoding is divided into Label and One Hot Encoding.

For Feature enginering, we are using one hot encoding, create Age dataset, and churn dataset.

DATA ENCODING

LABEL ENCODING

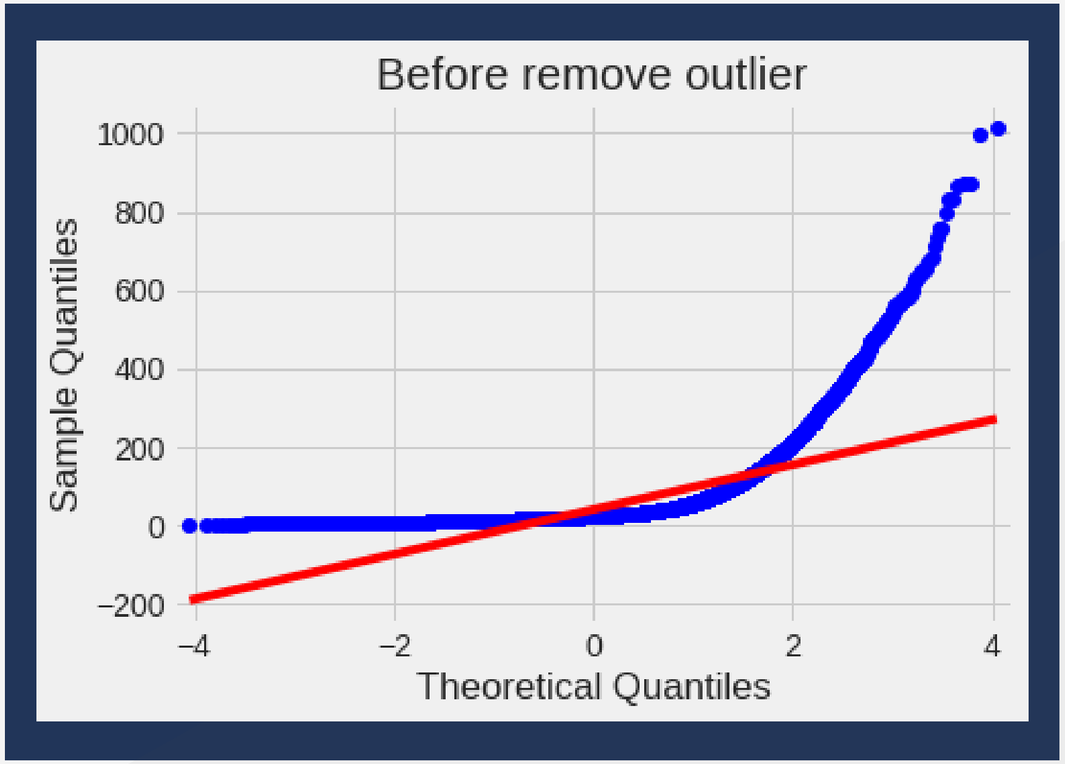
- Event name
- Created time
- Payment method
- Promo code

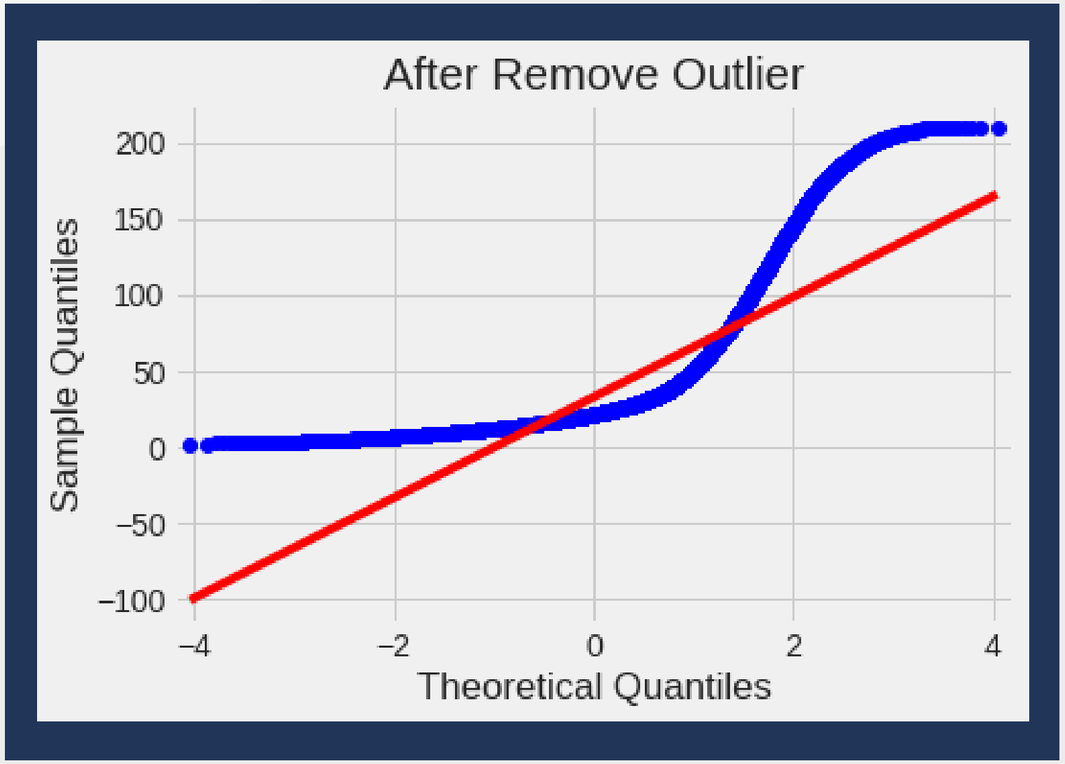
ONE HOT ENCODING

- Gender
- Payment status
- Device type

REMOVING OUTLIERS

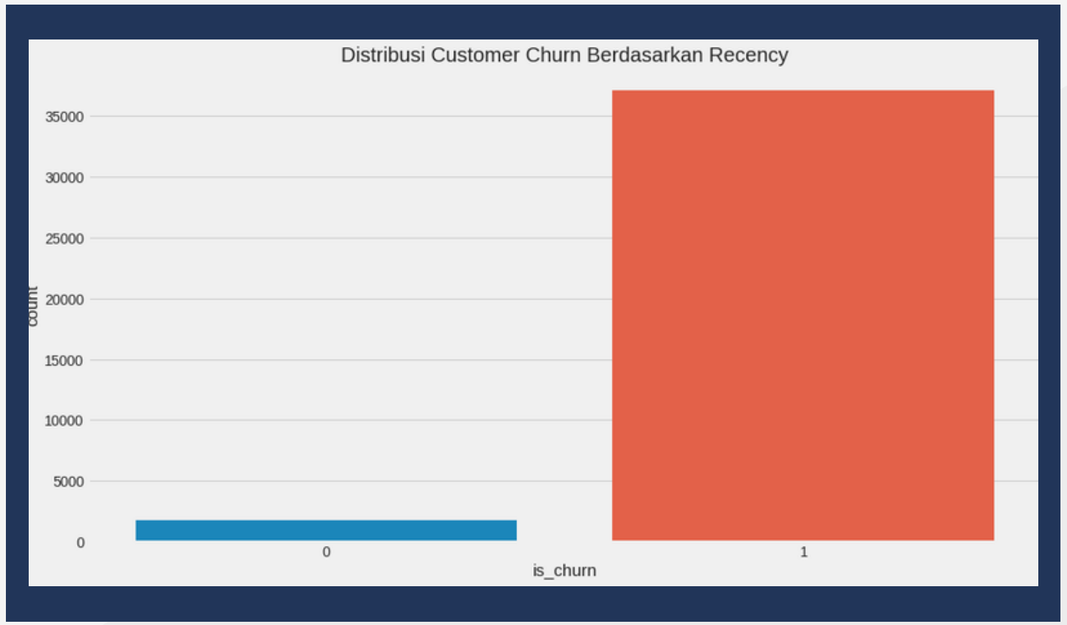
We found outliers from total amount, then we **redefining** the data using upper and lower limit.

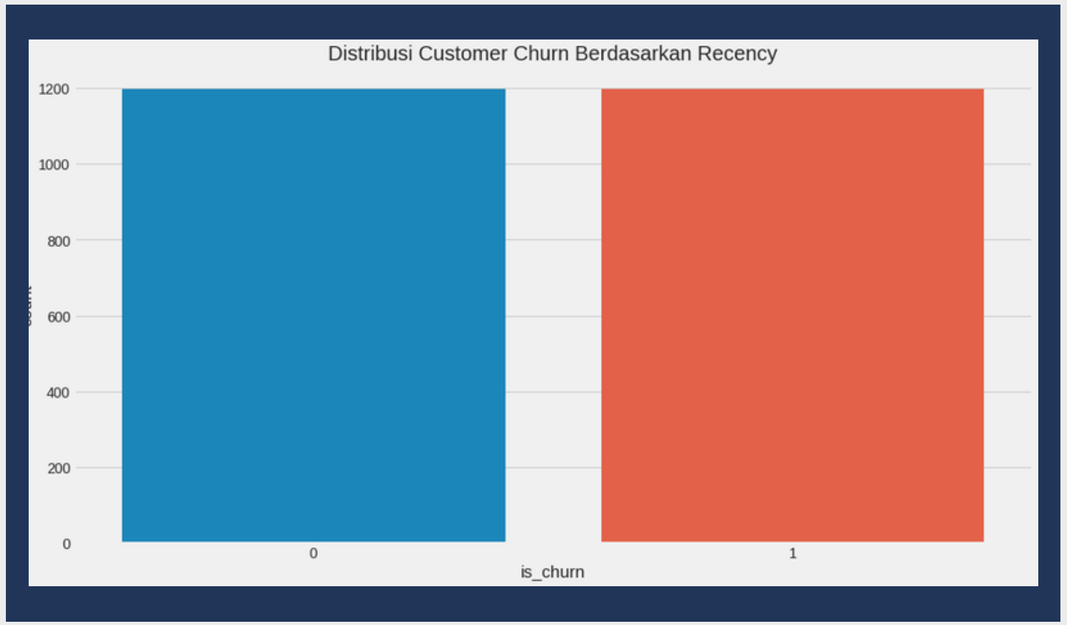




HANDLING IMBALANCE DATA

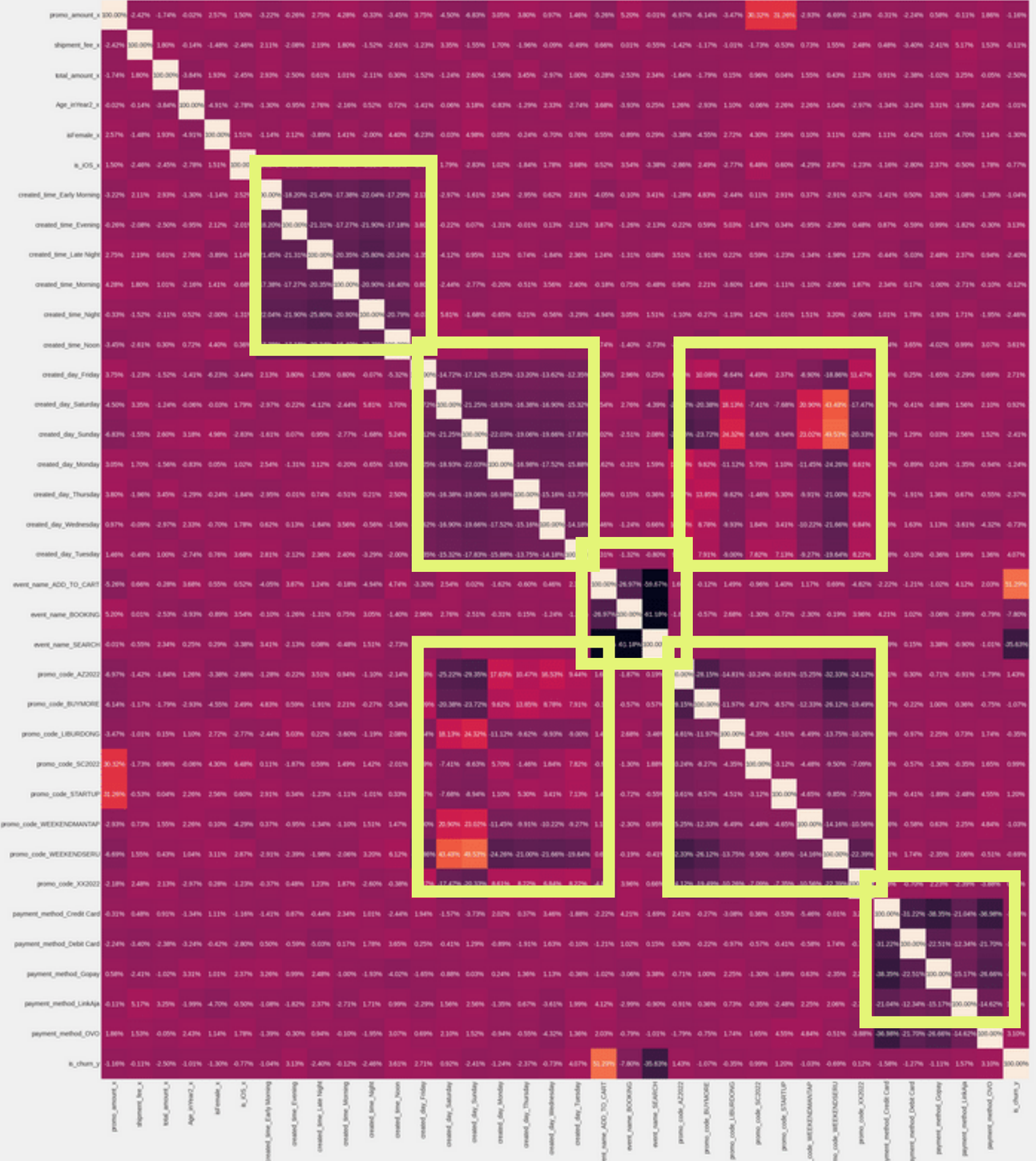
We check the balance of the data through the proportion of the is_churn variable. With the upsampled method, the data is balanced with the values of 1 and 0 being the same, each with 1198 data.





NOW, WE HAVE NEW DATASET

After the label encoding, removing outlier, and balancing data, now the dataset has 2396 rows and 89 columns. We also transforming the Label Encoded data to One Hot Encoding.

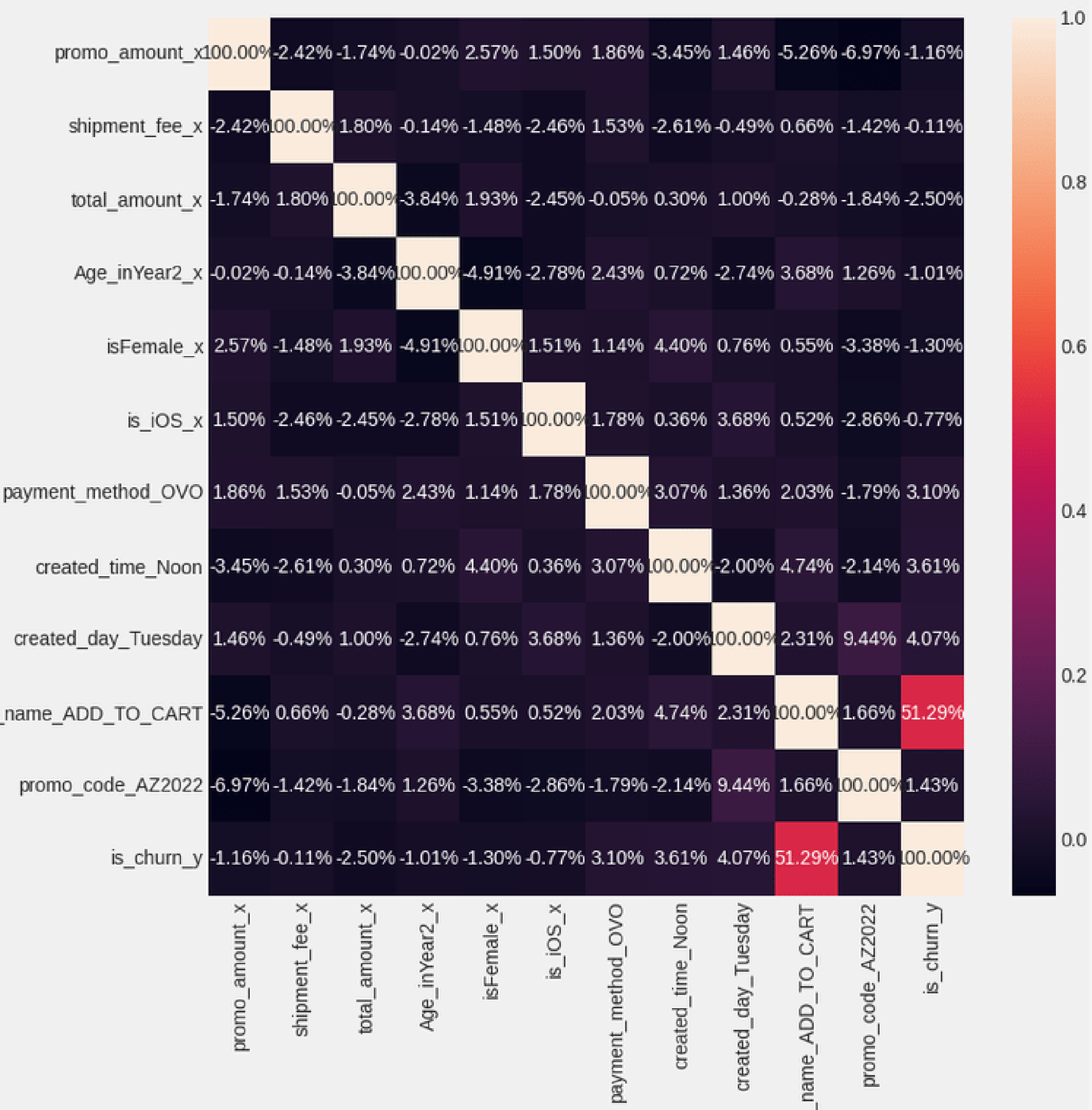


CORRELATION

We found that there was a lot of multicollinearity in the data, as indicated by the green box beside.

For example feature payment_method_Credit Card has a strong correlation with payment_method_Gopay at -38.35%, beside the correlation with is_churn just -1.11%.

To avoid bias in the modeling process, several variables that do not have multicollinearity problems were selected.



FEATURE SELECTION

Now we have features that can be used for modeling without multicollinearity and has correlation with churn. Now, we have:

INDEPENDENT VAR.

- Promo amount
- Shipment fee
- Total amount
- Age
- is_female
- is_iOS
- OVO payment method
- Created time Noon
- Created day Tuesday
- Promo code AZ2022.
- Add to Cart

DEPENDENT VAR.

- is_churn

MODELING

After knowing the dependent and independent variables, we tested several models to find out the best model in explaining the effect of the independent variable on the dependent variable.

The model we tested as the baseline models are:

- Decision Tree Classifier
- Logistic Regression
- K-Neighbors Classifier
- Gaussian Naive Bayes (GNB)
- Support Vector Classification (SVC)
- Random Forest Classifier
- Extra Trees Classifier
- XGB Classifier

BASELINE MODELS

We rank the baseline models using the accuracy in test data.
The highest accuracy is on the top of list.

	model	Accuracy training	Accuracy test	Precision	Recall	AUC	gap
7	XGBClassifier	0.741235	0.696989	0.898373	0.445694	0.696953	0.044245
5	RandomForestClassifier	1.000000	0.692816	0.771100	0.548393	0.692809	0.307184
6	ExtraTreesClassifier	1.000000	0.671954	0.709846	0.580924	0.671933	0.328046
0	DecisionTreeClassifier	1.000000	0.650249	0.643442	0.674432	0.650248	0.349751
3	GaussianNB	0.507721	0.502504	0.502300	0.800558	0.502585	0.005216
2	KNeighborsClassifier	0.680509	0.493736	0.493753	0.522531	0.493727	0.186773
1	LogisticRegression	0.504382	0.488321	0.491728	0.466726	0.488452	0.016061
4	SVC	0.509077	0.487891	0.461672	0.753996	0.488204	0.021187

WHAT NEXT?

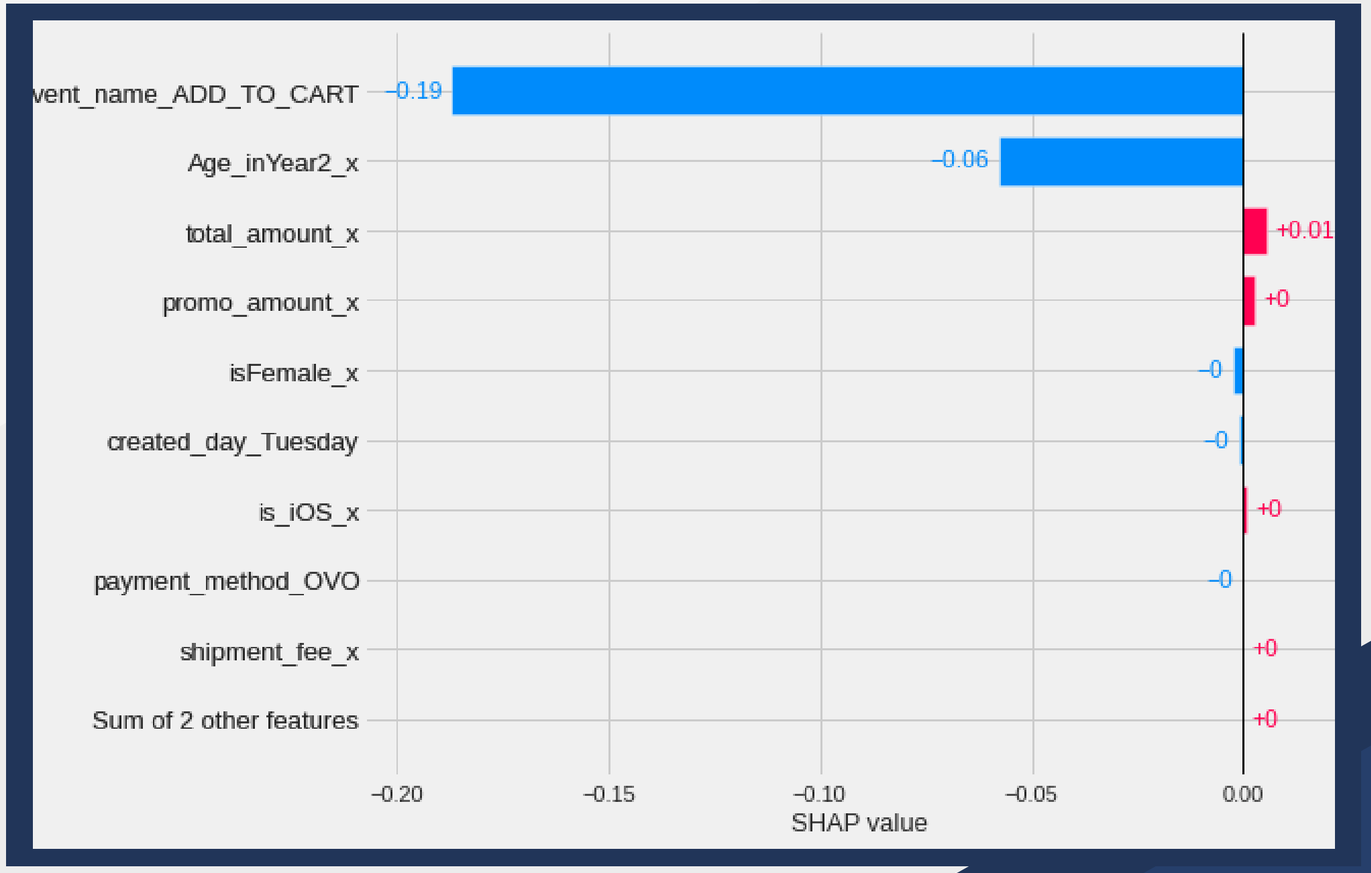
- To explain how the features contribute to the row prediction output, the SHAP Values (SHapley Additive ExPlanations) method is used.

- With this method, the transparency and interpretability of machine learning model interpretation can be improved.

XGB CLASSIFICATION

This plot shows us what are the main features affecting the prediction of a single observation, and the magnitude of the SHAP value for each feature.

We can see that Add to Cart and age can influence the prediction in negative direction, then total amount can influence the prediction in positive direction.



XGB CLASSIFICATION
(TUNING HYPERPARAMETER)

The tuning hyperparameter process is carried out using the Grid Search CV method.

The results show that in this model, only add to cart variables can affect churn predictions, in a positive direction.

Model	Accuracy (f-1 score)
Untuned Model	0.72
Tuned Model	0.72

CONCLUSION AND RECOMMENDATION

CONCLUSION

In the rise of churn rate, we need to providing solutions to increase and maintain user retention.

Churn increased every month. The Most churn customers are Customers with their age between the ages of 20-29 years old, with 30 days they not returned to using the application.

Feature has the most churned customers are payment method using credit cards, and event name "Add to Cart".

We need a ready dataset for modelling. So, we are doing the feature engineering, label encoding, handling outliers, and balancing data.

We decide to use is_churn as dependent variable and the independent variables are Promo amount, Shipment fee, Total amount, Age, is_female, is_iOS, OVO payment method, Created time Noon, Created day Tuesday, Promo code AZ2022, and Add to Cart.

From the baseline models comparison, we decided to using XGB Classifier, then after performing the hyperparameter tuning process, it can be seen that the add to cart variable has a positive effect on churn.

RECOMMENDATION

Only "Buy Now" Promos

New promo called LebaranAZ only available if customer buy using the "Buy Now" feature.

Total Amount Increased

Based on our finding in correlation map.

1 → 2 → 3 → 4 → 5

Create Feature "Buy Now"

This feature always appear on every product page.

"Add to Cart" Decreased

Based on our finding in XGB Classifier model, and EDA.

Churn Decreased

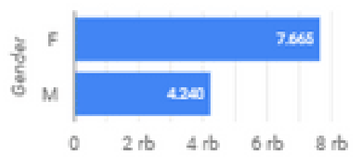
FASHION CAMPUS DASHBOARD

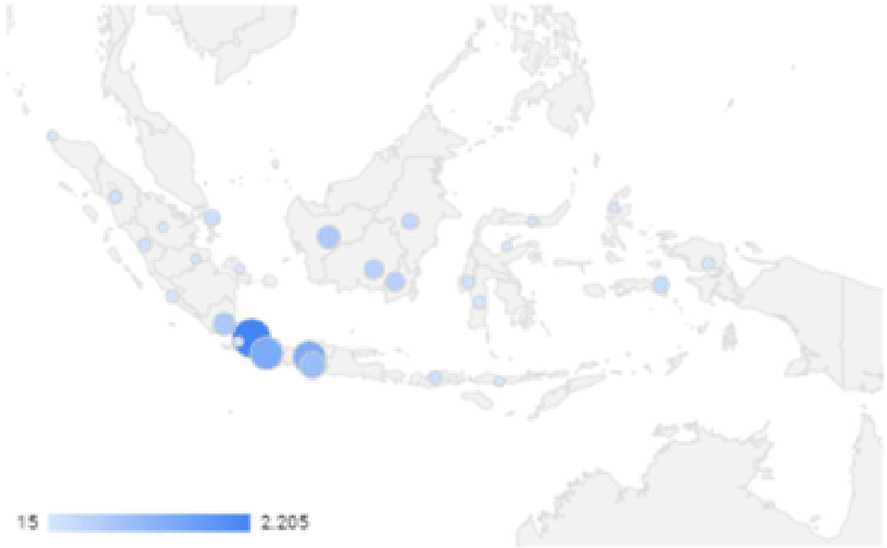
1 Jan 2022 - 18 Nov 2022

Fashion Campus that operates in the fashion industry with the "Indonesia Young Urbans" as the target market. This company provides local to international brands that are loved by youngsters.

Record Count
11,9 rb
↑ 21.4%

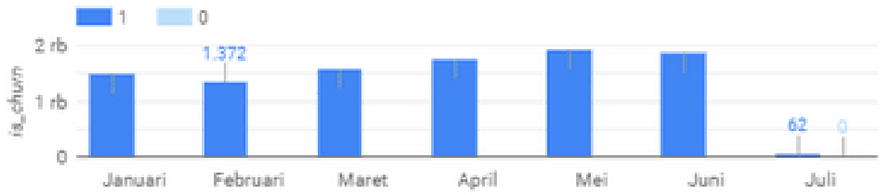
Total Transaction
Rp6,42 M
↑ 17.8%

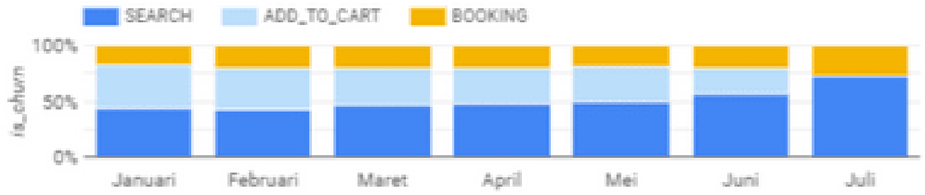


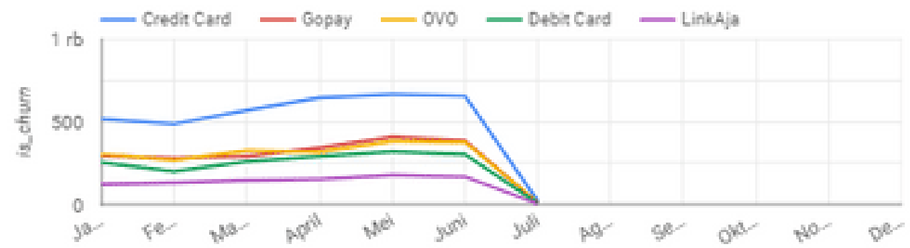


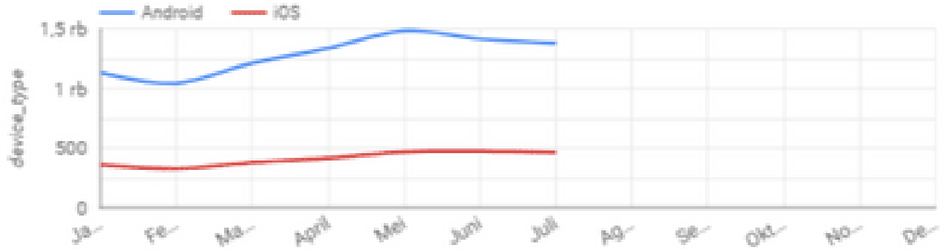
	Location	Total Customer	Total Transaction	Total Churn
1.	Jakarta Raya	2.205	Rp1.184.798.170	1.884
2.	Jawa Barat	1.357	Rp760.491.958	1.163
3.	Jawa Tengah	1.351	Rp709.788.434	1.153
4.	Jawa Timur	1.142	Rp626.616.183	965
5.	Yogyakarta	925	Rp466.186.350	768
6.	Kalimantan Barat	668	Rp347.701.950	561
7.	Lampung	659	Rp353.916.383	568
8.	Kalimantan Selatan	503	Rp272.053.551	435
9.	Kalimantan Tengah	497	Rp274.072.511	402
10.	Kalimantan Timur	359	Rp193.471.719	306

1 - 33 / 33









TASK DIVISIONS

bit.ly/PembagianTugasItaly

NOTEBOOK

bit.ly/AllNotebookItaly

DASHBOARD

bit.ly/DashboardItaly

THANK YOU