Welcome

DQLab

Data Analyst Project: Business Decision Research

Dataset Retail Toko Olahraga

Created by: Huan Wendy Ariono
Last Update: 27 November 2022

Program language

python™

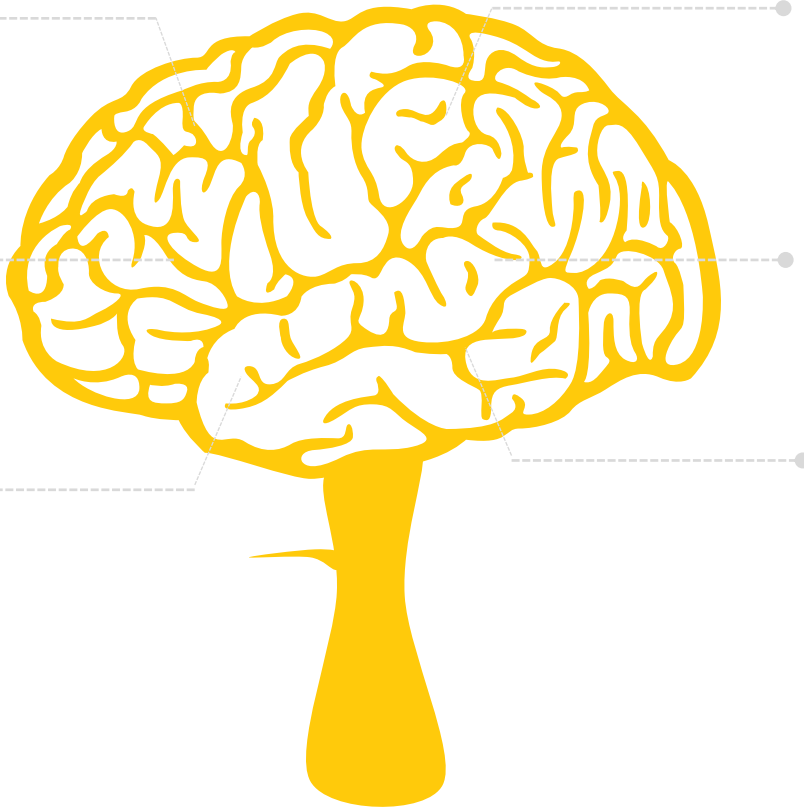
knowledge

Capaian

Dapat menerapkan bahasa pemograman Python untuk menyelesaikan persoalan untuk analisis data.

Dapat mengolah dataset skala kecil hingga besar

Dapat menerapkan exploratory data analysis (EDA) untuk persoalan bisnis



Dapat menghasilkan visualisasi data yang tepat sebagai representasi persoalan bisnis

Dapat membuat dan menghasilkan model prediktif hingga menguji tingkat akurasi untuk memilih model yang tepat

Dapat menerapkan teknik investigasi data-data yang memiliki anomali

Kunjungi medium/huans502 untuk penjelasan lebih detail

6 FASE ANALISIS DATA

Bertanya/ ask
mendefinisikan masalah yang harus dipecahkan

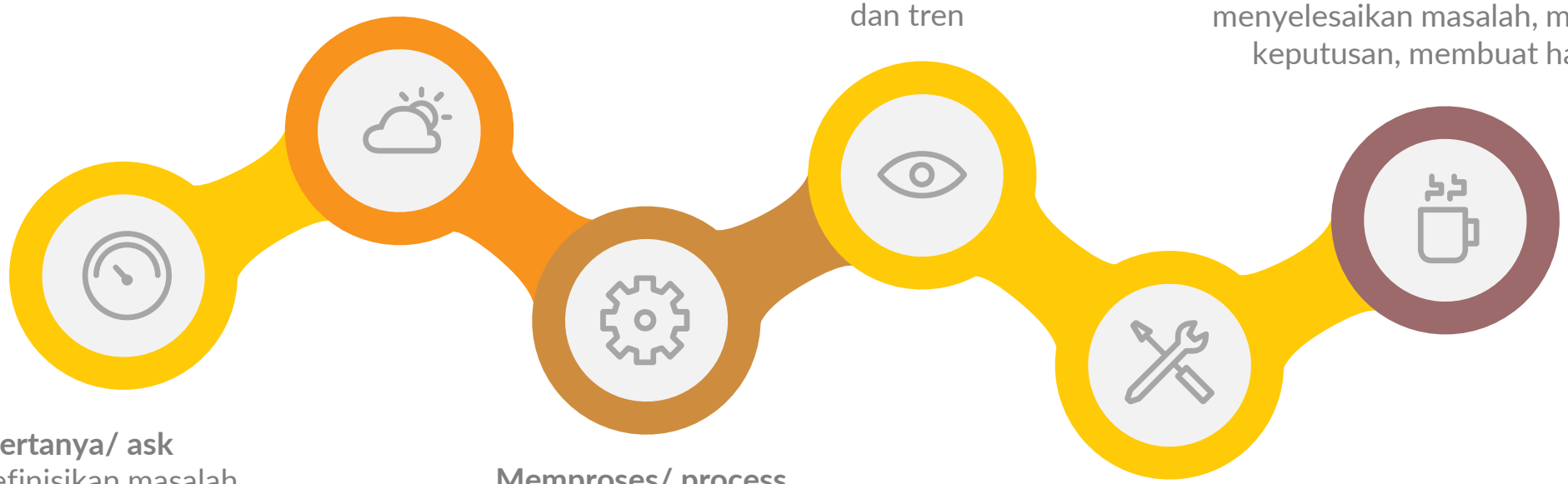
Mempersiapkan/ prepare
mengumpulkan dan menyimpan data yang akan kita gunakan untuk proses analisis

Memproses/ process
memastikan data yang lengkap dan benar

Menganalisis/ analyze
menemukan pola, hubungan, dan tren

Membagikan/share
menggunakan visulaisasi agar pihak lain mudah memahami hasil analisis

Bertindak/ act
menerapkan wawasan kita untuk menyelesaikan masalah, mengambil keputusan, membuat hal baru



Python

Library Yang Digunakan

Pandas

Digunakan untuk load, prepare, manipulate, modelling, dan analysis data

Matplotlib

digunakan untuk visualisasi data agar memahami karakteristik maupun trend di dalam data yang dimiliki

sklearn

untuk membantu melakukan processing data ataupun melakukan training data untuk kebutuhan machine learning atau data science

Study Case

Sebuah toko olahraga yang menjual berbagai kebutuhan olahraga seperti Jaket, Baju, Tas, dan Sepatu. Toko ini mulai berjualan sejak tahun 2013, sehingga sudah memiliki pelanggan tetap sejak lama, dan tetap berusaha untuk mendapatkan pelanggan baru sampai saat ini. Di awal tahun 2019, manajer toko tersebut ingin kita memecahkan masalah yang ada di tokonya, yaitu menurunnya pelanggan yang membeli kembali ke tokonya. Manajer toko mendefinisikan bahwa customer termasuk sudah bukan disebut pelanggan lagi (churn) ketika dia sudah tidak bertransaksi ke tokonya lagi sampai dengan 6 bulan terakhir dari update data terakhir yang tersedia. Manajer toko pun memberikan data transaksi dari tahun 2013 sampai dengan 2019 dalam bentuk csv (comma separated value) bernama data_retail.csv dengan jumlah baris 100.000 baris data.

Untuk coding lebih jelas silahkan kunjungi https://github.com/huwea/Project/

Bertanya/ ask

Dari study case diatas dapat kita ketahui pemangku kepentingannya adalah manajer toko. Harapan dari manajer tersebut adalah agar kita menyelidiki churn

Mempersiapkan/ prepare

Churn adalah pelanggan yang sudah tidak melakukan transaksi lagi dalam waktu tertentu, dalam study case ini rentang waktu yang digunakan adalah 6 bulan terakhir dari data transaksi yang ada. Untuk menyelidiki churn tentunya kita memerlukan data transaksi yang mana telah disediakan oleh pihak toko berupa file csv yang berisi data transaksi dari tahun 2013 sampai dengan 2019 dengan jumlah baris 100.000 baris data.

Memproses/ process

Dari study case diatas dapat kita ketahui pemangku kepentingannya adalah manajer toko. Harapan dari manajer tersebut adalah agar kita menyelidiki churn

```
---------------------------------
Cek tipe data
---------------------------------

RangeIndex: 100000 entries, 0 to 99999
Data columns (total 8 columns):
 #   Column                     Non-Null Count   Dtype
---  ------                     --------------   -----
 0   no                         100000 non-null  int64
 1   Row_Num                    100000 non-null  int64
 2   Customer_ID                100000 non-null  int64
 3   Product                    100000 non-null  object
 4   First_Transaction          100000 non-null  int64
 5   Last_Transaction           100000 non-null  int64
 6   Average_Transaction_Amount 100000 non-null  int64
 7   Count_Transaction          100000 non-null  int64
dtypes: int64(7), object(1)
memory usage: 6.1+ MB
None
```

Cek tipe data sebelum diproses

```
---------------------------------
Lima data teratas:
---------------------------------
   no  Row_Num  Customer_ID Product  First_Transaction  Last_Transaction  \
0   1        1        29531   Jaket      1466304274396     1538718482608
1   2        2        29531  Sepatu      1406077331494     1545735761270
2   3        3       141526     Tas      1493349147000     1548322802000
3   4        4       141526   Jaket      1493362372547     1547643603911
4   5        5        37545  Sepatu      1429178498531     1542891221530

   Average_Transaction_Amount  Count_Transaction
0                     1467681                 22
1                     1269337                 41
2                      310915                 30
3                      722632                 27
4                     1775036                 25
```

Cek 5 data teratas sebelum di proses

Lanjutan

Dari study case diatas dapat kita ketahui pemangku kepentingannya adalah manajer toko. Harapan dari manajer tersebut adalah agar kita menyelidiki churn

```
-----------------------------------
Cek null
-----------------------------------
False


-----------------------------------
Detail setiap jumlah null column
-----------------------------------
no                          0
Row_Num                     0
Customer_ID                 0
Product                     0
First_Transaction           0
Last_Transaction            0
Average_Transaction_Amount  0
Count_Transaction           0
dtype: int64
```

```
-----------------------------------
Info dataset:
-----------------------------------

RangeIndex: 100000 entries, 0 to 99999
Data columns (total 8 columns):
 #   Column                      Non-Null Count   Dtype
---  ------                      --------------   -----
 0   no                          100000 non-null  int64
 1   Row_Num                     100000 non-null  int64
 2   Customer_ID                 100000 non-null  int64
 3   Product                     100000 non-null  object
 4   First_Transaction           100000 non-null  datetime64[ns]
 5   Last_Transaction            100000 non-null  datetime64[ns]
 6   Average_Transaction_Amount  100000 non-null  int64
 7   Count_Transaction           100000 non-null  int64
dtypes: datetime64[ns](2), int64(5), object(1)
memory usage: 6.1+ MB
None
```

```
-----------------------------------
Lima data teratas:
-----------------------------------
   Customer_ID Product            First_Transaction  \
0        29531   Jaket 2016-06-19 02:44:34.396000000
1        29531  Sepatu 2014-07-23 01:02:11.493999872
2       141526     Tas 2017-04-28 03:12:27.000000000
3       141526   Jaket 2017-04-28 06:52:52.546999808
4        37545  Sepatu 2015-04-16 10:01:38.530999808

                Last_Transaction  Average_Transaction_Amount  \
0 2018-10-05 05:48:02.608000000                     1467681
1 2018-12-25 11:02:41.269999872                     1269337
2 2019-01-24 09:40:02.000000000                      310915
3 2019-01-16 13:00:03.911000064                      722632
4 2018-11-22 12:53:41.529999872                     1775036

   Count_Transaction is_churn
0                 22    False
1                 41    False
2                 30    False
3                 27    False
4                 25    False
```

Memastikan dataset bebas dari null value

Memastikan tipe data sesuai

Hapus column yang tidak digunakan, tambahkan column churn, lalu cek 5 data teratas

Menganalisis/ analyze & Membagikan/share

1

Mengecek Data Historical Penjualan Kebutuhan Olahraga

Mengeksplore bagaimana penjualan kebutuhan olahraga di toko berkembang dari tahun ke tahun

2

Memeriksa Pola Konsumen Dalam Pembelian Produk

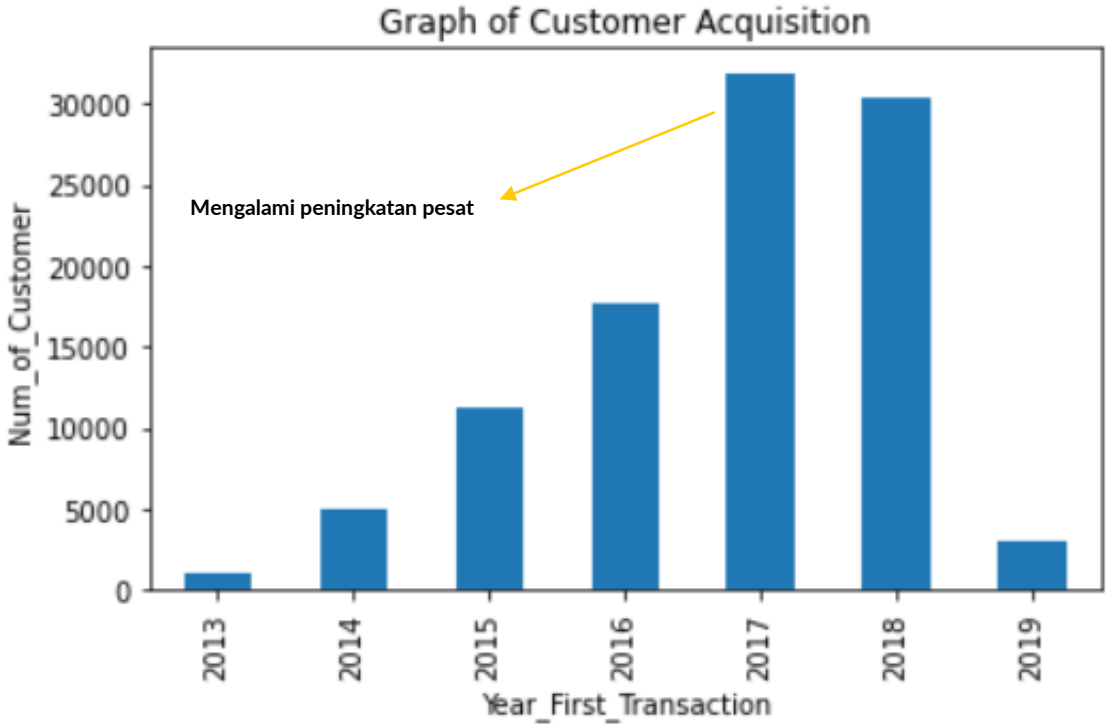
Memahami perilaku konsumen dalam membeli produk yang mana datanya dapat digunakan untuk perencanaan promo produk untuk menarik konsumen

3

Membuat Model Logistic Regression

Output berupa grafik heatmap confusion matrix, accuracy, precision, dan recall

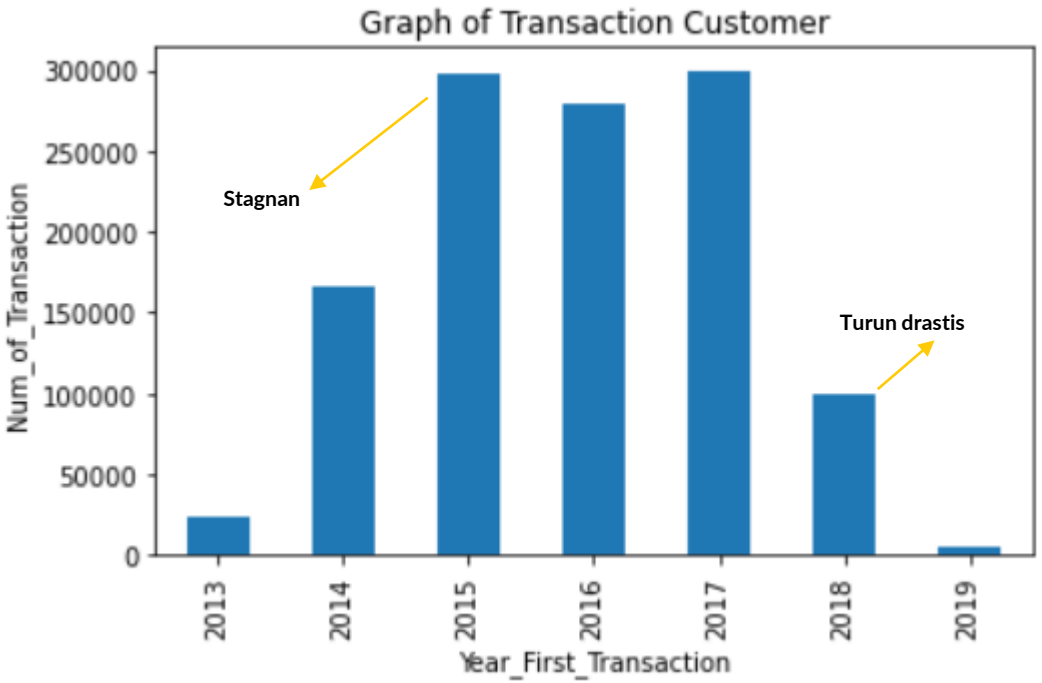
Jumlah customer yang melakukan transaksi mengalami peningkatan dari tahun 2013, transaksi meningkat pesat pada tahun 2017



Detail

- Data diambil dari data transaksi toko olahraga dalam rentang waktu tahun **2013 - 2019**
- Penghitungan jumlah customer yang melakukan transaksi mencakup produk olahraga **baju, jaket, sepatu, dan tas**
- Dari barchart disamping kita mendapatkan inshight jumlah customer yang melakukan transaksi terus meningkat dari tahun 2013 sampai 2017 setelah itu menurun sedikit di tahun 2018. Kemudian pada tahun 2019 menurun drastis tapi kita harus jeli pada tahun 2019 transaksi terakhir dari data yang tersedia adalah hanya sampai bulan februari 2019 sehingga kita tidak bisa menggunakan data pada tahun 2019 ini sebagai acuan karena kurangnya data.

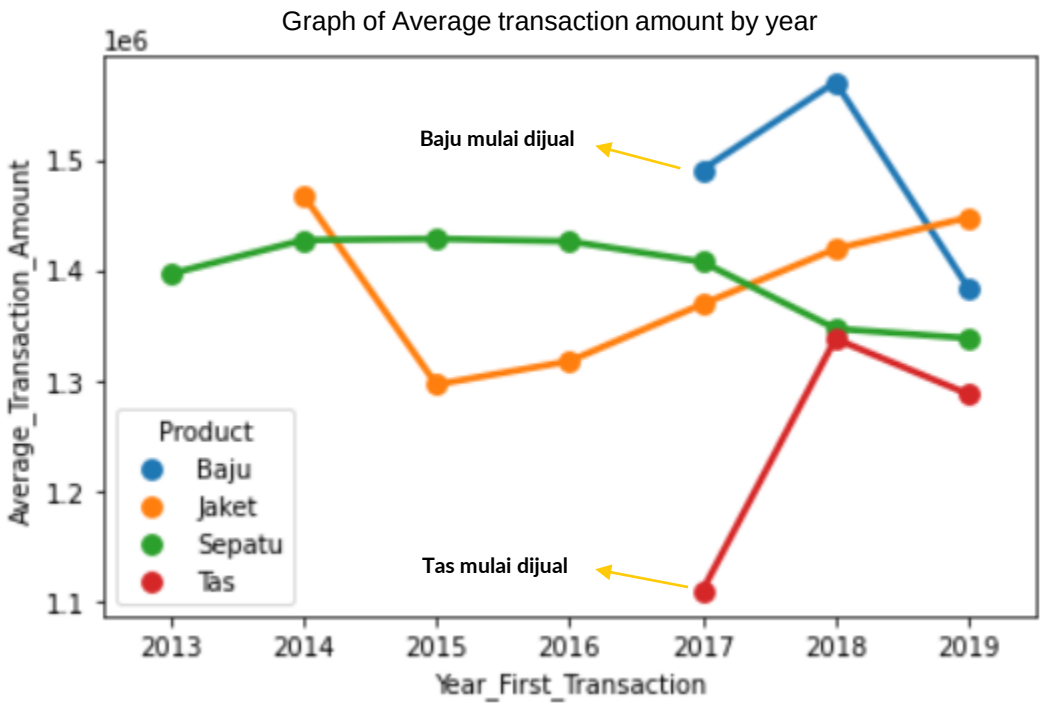
Jumlah transaksi produk yang dilakukan customer terus meningkat dari tahun 2013 hingga 2015, kemudian stagnan pada tahun 2015 – 2017, lalu mengalami penurunan drastis pada tahun 2018



Detail

- **Penghitungan jumlah total transaksi seluruh produk pertahunnya**
- **Kenaikan jumlah transaksi** menandakan setiap customer sering berbelanja
- **Penurunan jumlah transaksi** menandakan setiap customer jarang berbelanja
- Dari barchart disamping dapat kita lihat jumlah transaksi produk yang dilakukan customer terus meningkat dari tahun 2013 hingga 2015 kemudian stagnan hingga tahun 2017. Tapi sayangnya turun drastis di tahun 2018 jika kita kaitkan dengan graph Graph of Customer sebelumnya ini artinya pada tahun 2018 jumlah orang yang melakukan transaksi berkurang sedikit dari tahun 2017 tetapi transaksi total yang dilakukan perorangnya menurun secara drastis artinya customer mulai jarang belanja, seperti kasus Customer acquisition by year sebelumnya untuk data di tahun 2019 tidak bisa digunakan sebagai acuan karena kurangnya data.

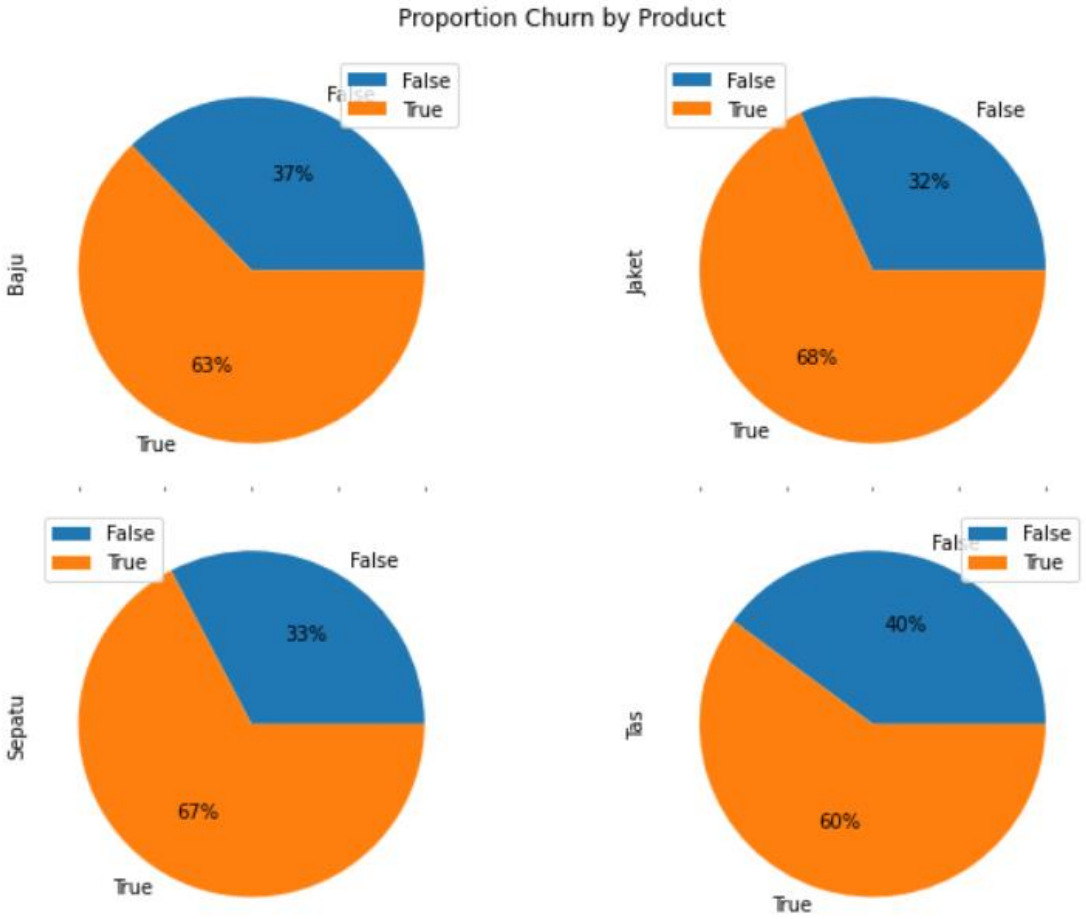
Tren dari tahun ke tahun rata-rata jumlah transaksi untuk tiap-tiap produknya



Detail

- **Average transaction amount by year** merupakan rata-rata dari total price semua transaksi produk tertentu dibagi jumlah transaksi produk tertentu
- **Data dalam satuan juta**
- Disini dapat kita ketahui alasan mengapa pada grafik Customer acquisition by year pada tahun 2017 mengalami peningkatan yang pesat hal ini dikarenakan produk baju mulai dijual pada tahun tersebut sedangkan tahun sebelumnya belum ada sehingga hal ini meningkatkan jumlah customer yang melakukan transaksi pada tahun tersebut. Sedangkan untuk graph Transaction by year pada tahun 2017 hanya berefek peningkatan sedikit pada jumlah transaksi produk yang telah dilakukan customer per tahunya. Dari grafik diatas dapat kita lihat pola yang tidak beraturan dari Graph of Average transaction amount by year untuk setiap produk.

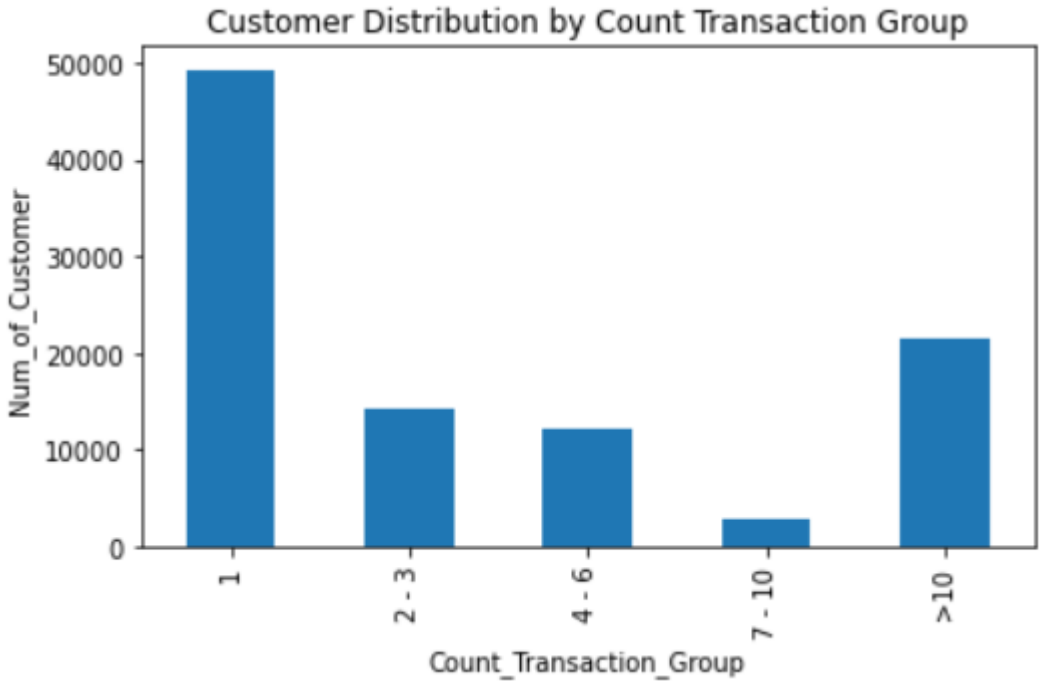
Proporsi churned customer untuk setiap produk



Detail

- **True** artinya churn
- **False** artinya tidak churn
- **Diagram pie chart** disamping dapat kita diketahui bahwa produk jaket merupakan produk dengan tingkat churn paling tinggi, walaupun begitu selisih churnnya dengan produk lain tidak begitu jauh
- Arti churn setiap produk ini adalah jumlah presentase customer yang sudah 6 bulan tidak beli produk tersebut disbanding yang masih membeli produk

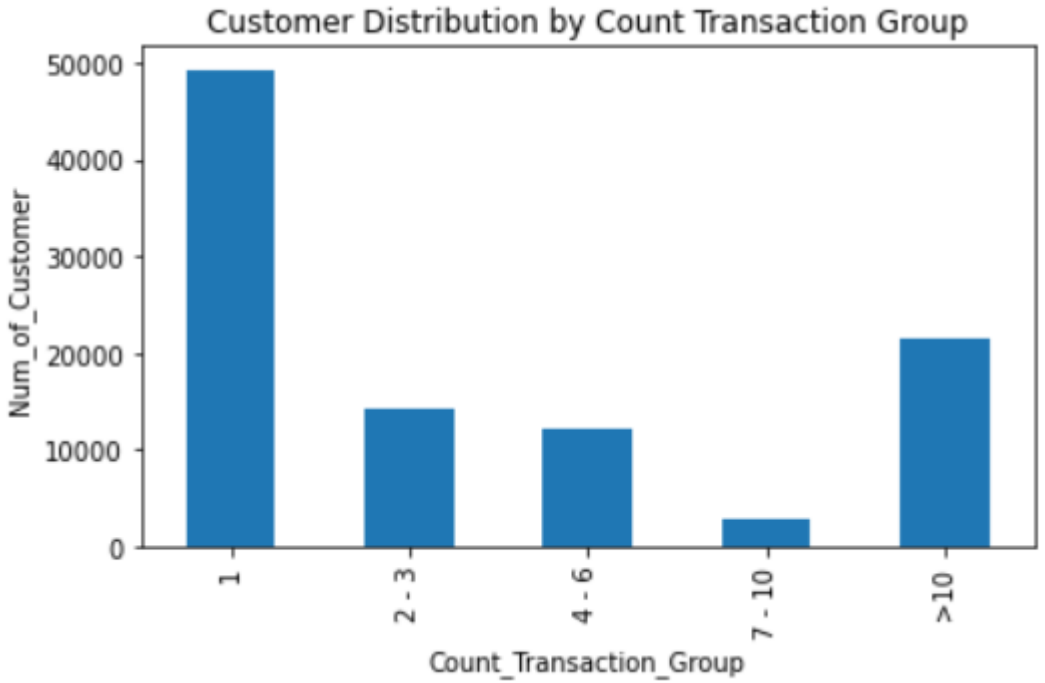
Kategorisasi jumlah transaksi



Data overview

- Dari grafik diatas dapat kita lihat customer kebanyakan hanya membeli 1 produk saja alias ngecer

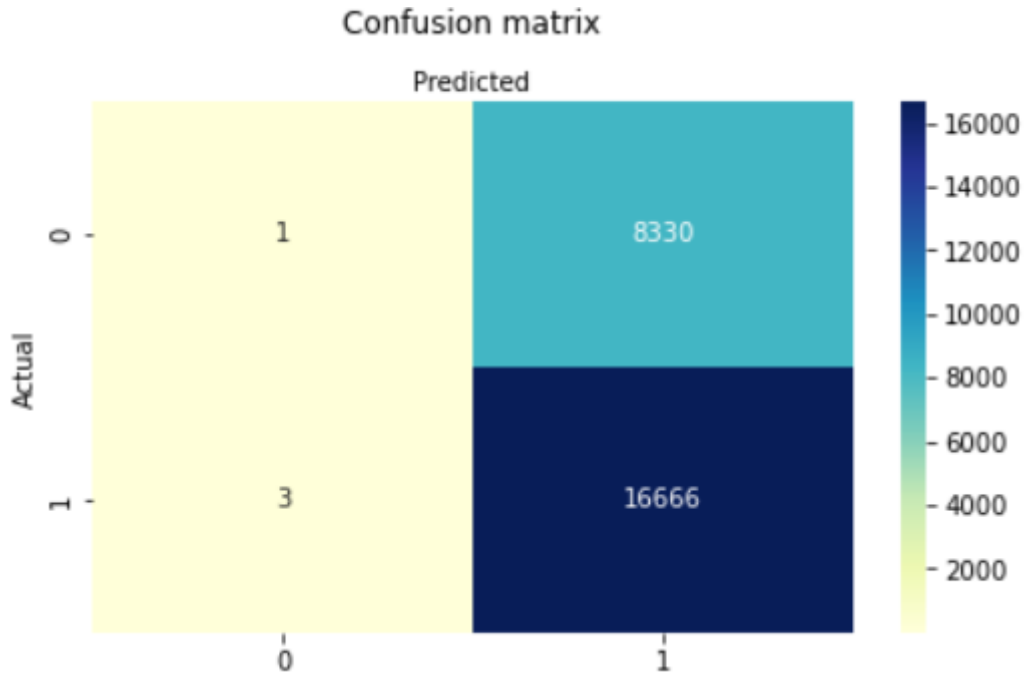
Distribusi kategorisasi average transaction amount



Data overview

- Dari grafik samping dapat kita simpulkan **Average_Transaction_Amount** para customer kebanyakan adalah sekitar 1.000.000 - 2.500.000

Model Logistic Regression menggunakan dataset data transaksi toko olahraga dalam rentang waktu tahun 2013 - 2019




```
Accuracy : 0.66668
Precision: 0.66668
Recall : 0.66668
```

Data overview

- Dari nilai accuracy, precision, dan recall diatas dapat disimpulkan nilai dari model Logistic Regression tidak terlalu tinggi

#DQLABDVIZ2OHLREE

CERTIFICATE

OF COMPLETION

This certificate is proudly presented to

Huan Wendy Ariono

Has Completed in

Data Analyst Project: Business Decision Research

Sep 7, 2022

Selesai

THANKS FOR WATCHING

Huan Wendy Ariono

Link Portofolio

Github

https://github.com/huwea

Rpubs

https://rpubs.com/Worstone57

Tableau

https://public.tableau.com/app/profile/huan.wendy.ariono

Medium

https://medium.com/@huans502

About

My name is Huan Wendy Ariono, I am Fresh Graduate of Informatics Engineering at University Muhammadiyah Surakarta. Currently I focus on data analysis. I am also active in adding new knowledge in the field of data by attending courses, workshops, reading articles and writing articles related to data in the medium.

My experience in data field is being able to use python, R, SQL, excel, googlesheet and tableau as well as other tools to analyze data and get valuable input to solve problems. I got these skills through lectures, independent projects and taking courses related to data.

LINKEDIN

https://www.linkedin.com/in/huanwendyariono/