

Final Presentation

[Data Science Digital Talent Scholarship 2024]

[CHALLENGE I]
Persebaran Virus Covid 19 di Indonesia

SQL QUERY

[Studi Kasus Persebaran Covid 19 Di Indonesia]

SQL QUERY 1

1. Jumlah total kasus Covid-19 aktif yang baru di setiap provinsi lalu diurutkan berdasarkan jumlah kasus yang paling besar

Row	Province ▼	Total_New_Active_Ca
1	Jawa Barat	13496
2	DKI Jakarta	10922
3	Banten	2558
4	Jawa Tengah	1423
5	Jawa Timur	1136
6	Daerah Istimewa Yogyakarta	669
7	Sumatera Utara	664
8	Sulawesi Utara	565
9	Bali	474

```
SELECT Province
      ,SUM(New_Active_Cases) AS Total_New_Active_Cases
FROM `kasus_covid.covid19`
GROUP BY 1
ORDER BY 2 DESC
```

2. Mengambil 2 (dua) location iso code yang memiliki total kematian karena Covid-19 paling sedikit

```
SELECT Location_ISO_Code  
       ,SUM(Total_Deaths) AS Total_Deaths  
FROM `kasus_covid.covid19`  
GROUP BY 1  
ORDER BY 2 ASC  
LIMIT 2
```

Row	Location_ISO_Code	Total_Deaths
1	ID-MA	147196
2	ID-MU	167511

3. Data tentang tanggal-tanggal ketika rate kasus recovered di Indonesia paling tinggi beserta jumlahnya

Row	Date	Highest_Recovery_Rate
1	2020-03-26	28.0
2	2020-03-28	13.0
3	2020-04-01	5.8571
4	2020-04-08	2.405
5	2020-04-09	2.25
6	2020-04-10	2.1396
7	2020-04-11	2.1238
8	2020-04-12	1.916699999999...
9	2020-04-13	1.88
10	2020-04-14	1.88
11	2020-04-16	1.3824
12	2020-04-17	1.3429
13	2020-04-18	1.302

```
SELECT Date
       ,MAX(Case_Recovered_Rate) AS Highest_Recovery_Rate
FROM kasus_covid.covid19
WHERE Country = 'Indonesia'
GROUP BY 1
```

4. Total case fatality rate dan case recovered rate dari masing-masing location iso code yang diurutkan dari data yang paling rendah

SELECT

Location_ISO_Code,

SUM(Case_Fatality_Rate) AS

Total_Case_Fatality_Rate,

SUM(Case_Recovered_Rate) AS

Total_Case_Recovered_Rate

FROM

`data_covid19.kasus_covid`

GROUP BY

Location_ISO_Code

ORDER BY

Total_Case_Fatality_Rate ASC,

Total_Case_Recovered_Rate ASC;

Row	Location_ISO_Code	Total_Case_Fatality	Total_Case_Recover
1	ID-KU	14.285000000000...	733.7265999999...
2	ID-NT	15.934500000000...	700.8207999999...
3	ID-PA	16.895300000000...	608.2326000000...
4	ID-JA	17.326799999999...	760.5292000000...
5	ID-SG	19.668699999999...	741.6644000000...
6	ID-KB	20.560999999999...	771.5737999999...
7	ID-SR	21.755600000000...	732.8722999999...
8	ID-SN	22.457400000000...	775.2974000000...
9	ID-SB	24.010300000000...	754.2531
10	ID-PB	24.334100000000...	757.1986999999...

5. Data tentang tanggal-tanggal saat total kasus Covid-19 mulai menyentuh angka 30.000-an

```
SELECT Date  
FROM kasus_covid.covid19  
WHERE Total_Cases >= 30000
```

Row	Date
1	2020-06-06
2	2020-06-07
3	2020-06-08
4	2020-06-09
5	2020-06-10
6	2020-06-11
7	2020-06-12
8	2020-06-13
9	2020-06-14
10	2020-06-15
11	2020-06-16
12	2020-06-17

SQL QUERY 6

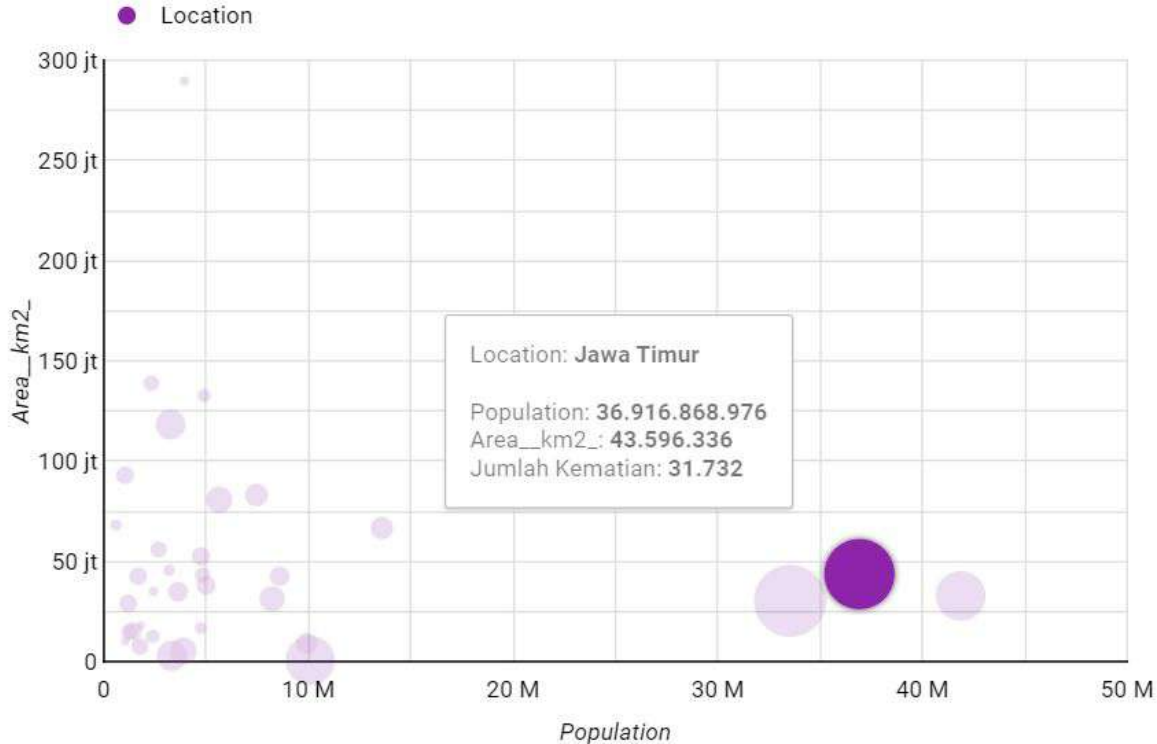
6. Jumlah data yang tercatat ketika kasus Covid-19 lebih dari atau sama dengan 30.000

```
SELECT COUNT(*)  
FROM kasus_covid.covid19  
WHERE Total_Cases >= 30000
```

Row	f0_
1	14399

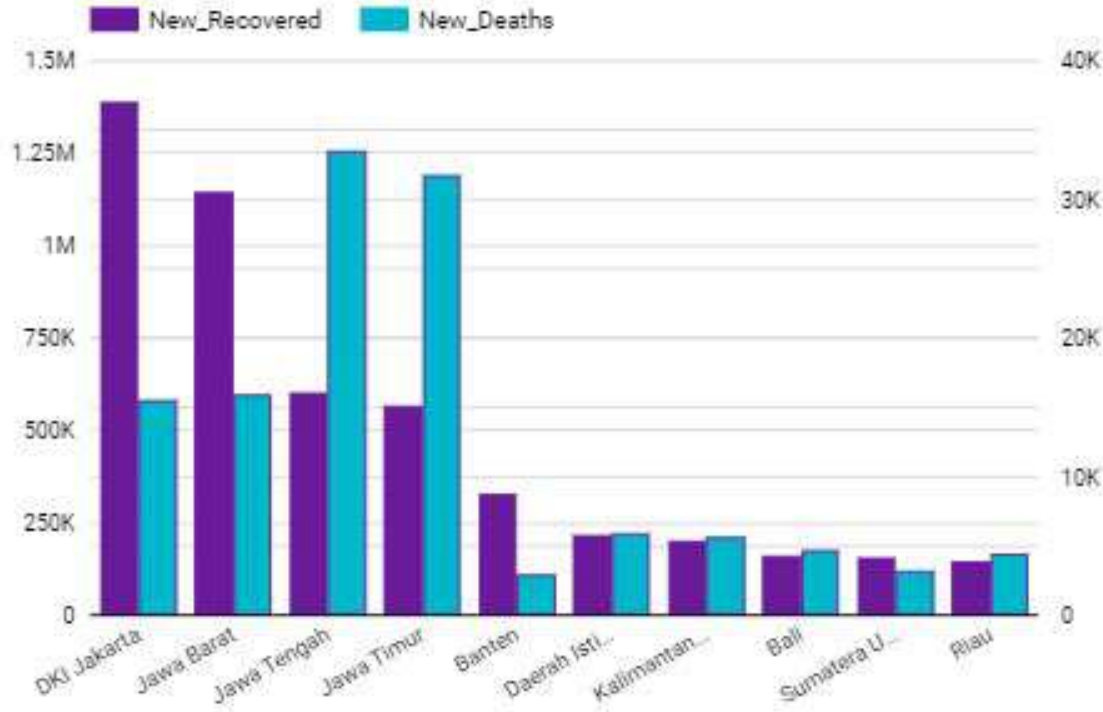
DASHBOARD

[Studi Kasus Persebaran Covid 19 Di Indonesia]



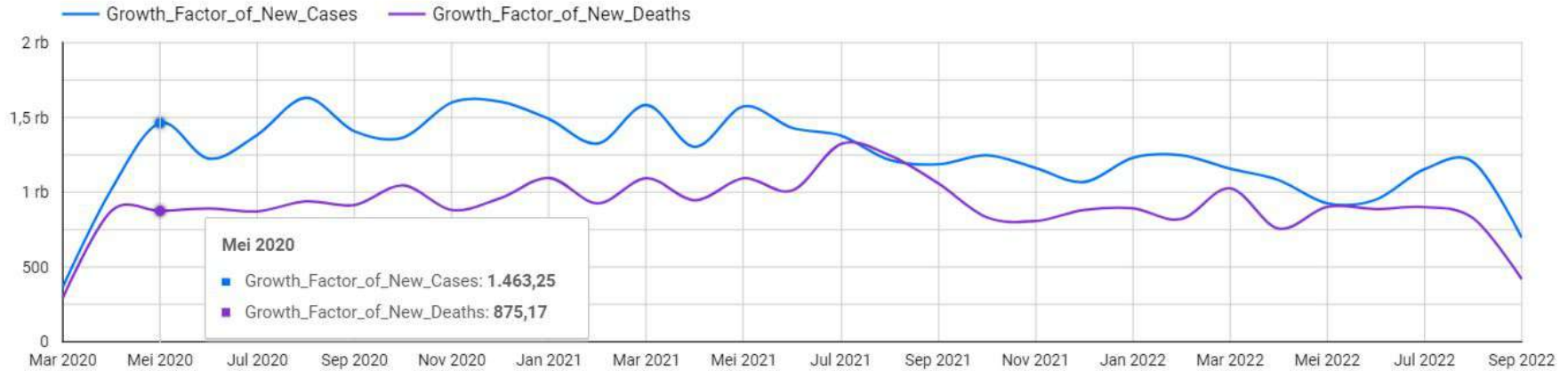
Analisis Persebaran Populasi dan Jumlah Kematian

Variabel **Population** dan **Area_km2_** dapat memberikan pemahaman tentang **kepadatan populasi dan ukuran wilayah** yang dapat mempengaruhi penyebaran virus.



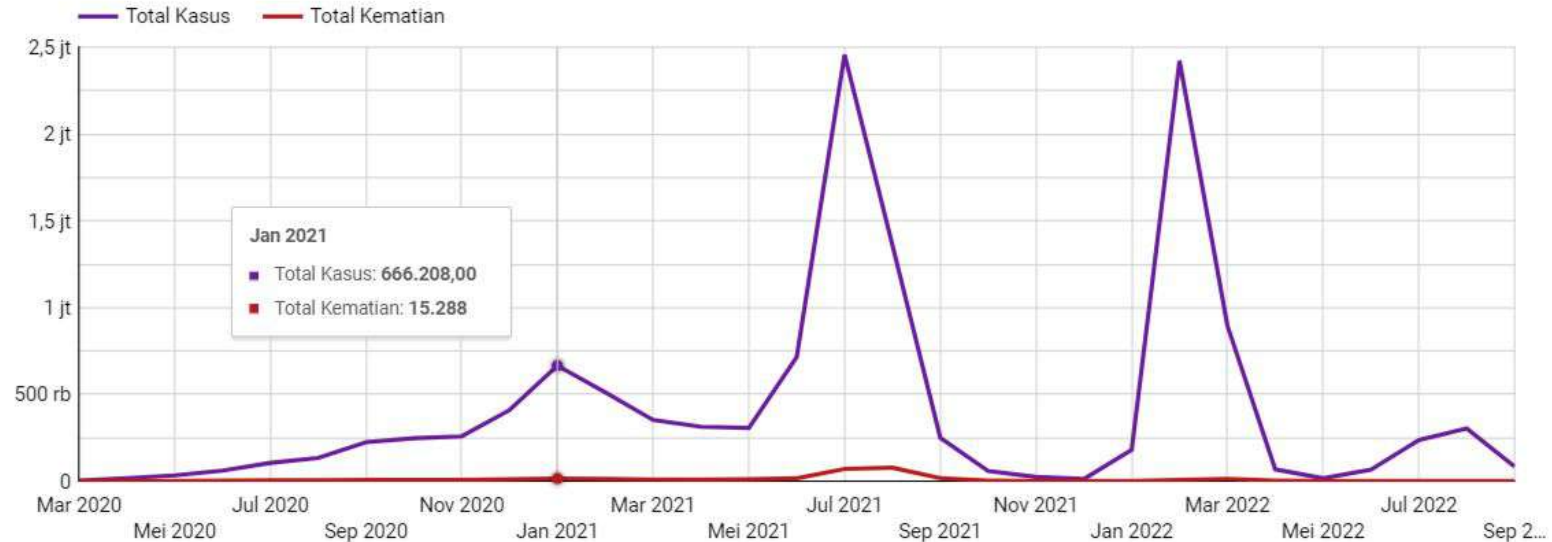
Perbandingan Tingkat Kematian dan Kesembuhan

New_Recovered, **New_Deaths**, dan **Province** memberikan pemahaman tentang **tingkat kematian dan kesembuhan** kasus COVID-19 dari berbagai provinsi di Indonesia



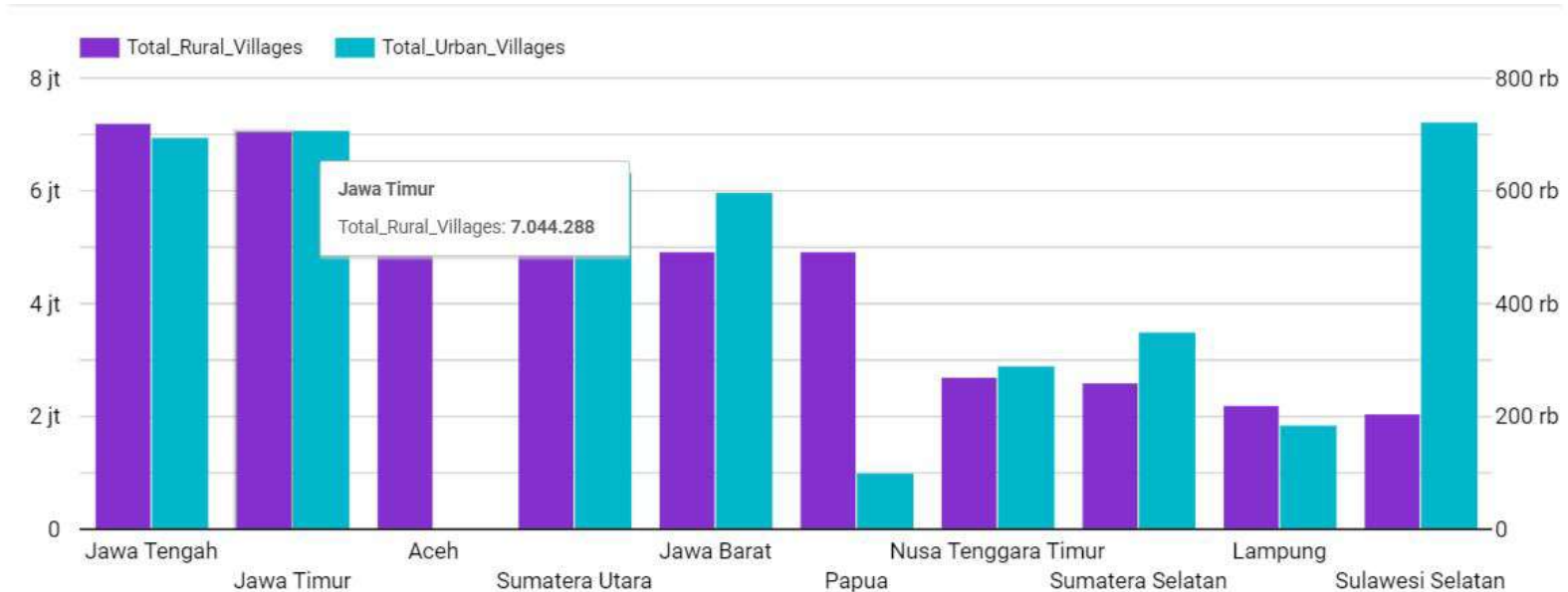
Analisis Penyebaran Kasus Baru dan Kematian

Growth_Factor_of_New_Cases dan **Growth_Factor_of_New_Deaths** memberikan gambaran tentang **kecepatan** penyebaran kasus baru dan kematian harian.



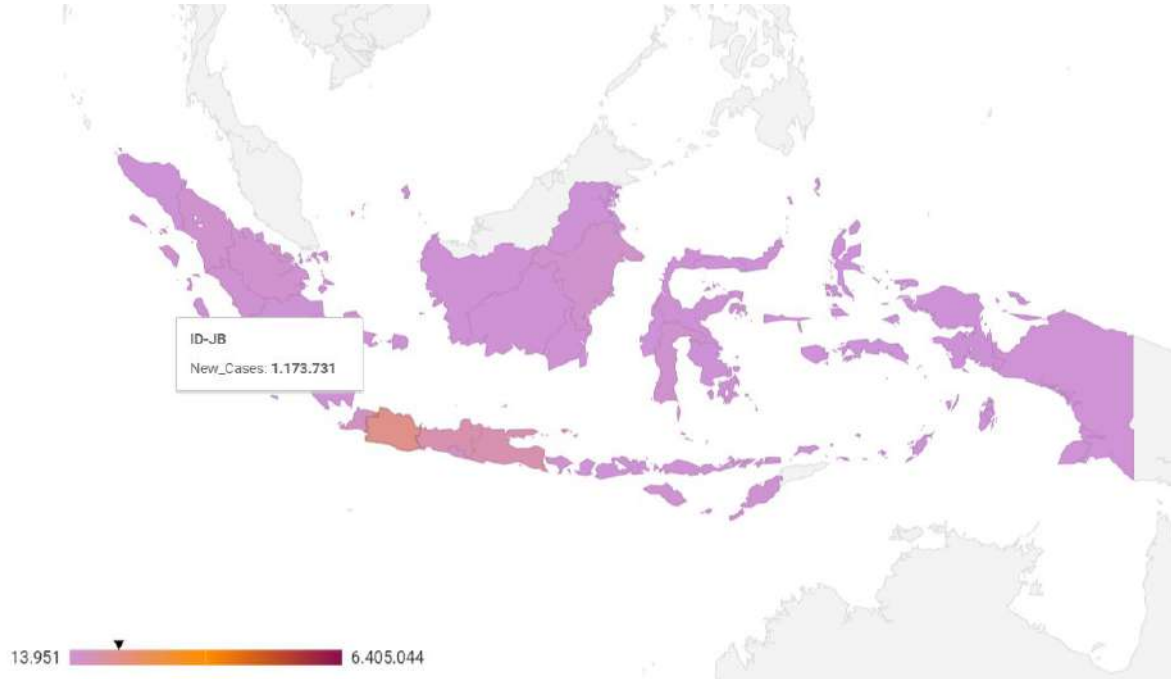
Analisis Perbandingan Total Kasus dengan Tingkat Kematian

New_Cases dan **New_Deaths** memberikan pemahaman tentang **perbandingan total kasus yang terjadi dengan tingkat kematian** karena Covid-19 **di setiap bulan**.



Analisis Perbandingan Total Urbanisasi dan Ruralisasi

Data tentang **Total_Urban Villages** dan **Total_Rural_Villages** dapat memberikan gambaran tentang **seberapa urbanisasi dan ruralisasi suatu wilayah**, yang dapat mempengaruhi penyebaran virus.



Analisis Perbandingan Demografi Kasus Covid

Dengan data **Location_ISO_code** dan **New_cases** kita dapat membandingkan **kasus COVID-19** di berbagai provinsi.

[CHALLENGE II]

Churn Classification

Project Overview

Customer churn didefinisikan sebagai ketika pelanggan atau pelanggan berhenti melakukan bisnis dengan perusahaan atau layanan.

Karena sebagian besar perusahaan memiliki banyak pelanggan, sulit untuk mempertahankan pelanggan individual. Biayanya akan lebih besar daripada pendapatan tambahannya. Namun, perusahaan dapat berkonsentrasi pada retensi pelanggan hanya pada klien yang "berisiko tinggi" jika mereka tahu pelanggan mana yang kemungkinan besar akan meninggalkan perusahaan. Memperluas cakupannya dan mendapatkan lebih banyak pelanggan adalah tujuan utamanya. Pelanggan adalah kunci sukses di pasar ini.

Karena mempertahankan pelanggan yang sudah ada jauh lebih murah daripada mendapatkan pelanggan baru, perpindahan pelanggan adalah metrik penting. **Untuk mengurangi churn pelanggan, perusahaan telekomunikasi perlu memprediksi pelanggan mana yang berisiko tinggi mengalami churn.**

Business Understanding

Problem Statements

1. Perkembangan industri telekomunikasi memperketat persaingan antar provider
2. Perusahaan harus dapat mengetahui pelanggan yang akan churn dan faktor yang mempengaruhinya

Goals

1. Melakukan prediksi agar bisa memetakan strategi bisnis untuk mempertahankan pelanggan.

Solutions Statements

1. Membuat model machine learning yang dapat memprediksi pelanggan yang berpotensi churn
2. Mengetahui pola/parameter pelanggan churn

Data Understanding

Dataset Info

Dataset yang digunakan dalam proyek ini merupakan data yang berisi informasi mengenai customer churn. Berikut adalah informasi pada dataset:

- Dataset memiliki format CSV
- Dataset memiliki 4250 sampel dengan 20 fitur yang dapat digunakan
- Dataset tidak memiliki missing value (null value) dan duplikat data

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4250 entries, 0 to 4249
Data columns (total 20 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   state                                4250 non-null   object
1   account_length                       4250 non-null   int64
2   area_code                            4250 non-null   object
3   international_plan                   4250 non-null   object
4   voice_mail_plan                      4250 non-null   object
5   number_vmail_messages               4250 non-null   int64
6   total_day_minutes                   4250 non-null   float64
7   total_day_calls                     4250 non-null   int64
8   total_day_charge                    4250 non-null   float64
9   total_eve_minutes                   4250 non-null   float64
10  total_eve_calls                     4250 non-null   int64
11  total_eve_charge                     4250 non-null   float64
12  total_night_minutes                 4250 non-null   float64
13  total_night_calls                   4250 non-null   int64
14  total_night_charge                  4250 non-null   float64
15  total_intl_minutes                  4250 non-null   float64
16  total_intl_calls                     4250 non-null   int64
17  total_intl_charge                    4250 non-null   float64
18  number_customer_service_calls       4250 non-null   int64
19  churn                               4250 non-null   object
dtypes: float64(8), int64(7), object(5)
memory usage: 664.2+ KB
```

Dataset Info

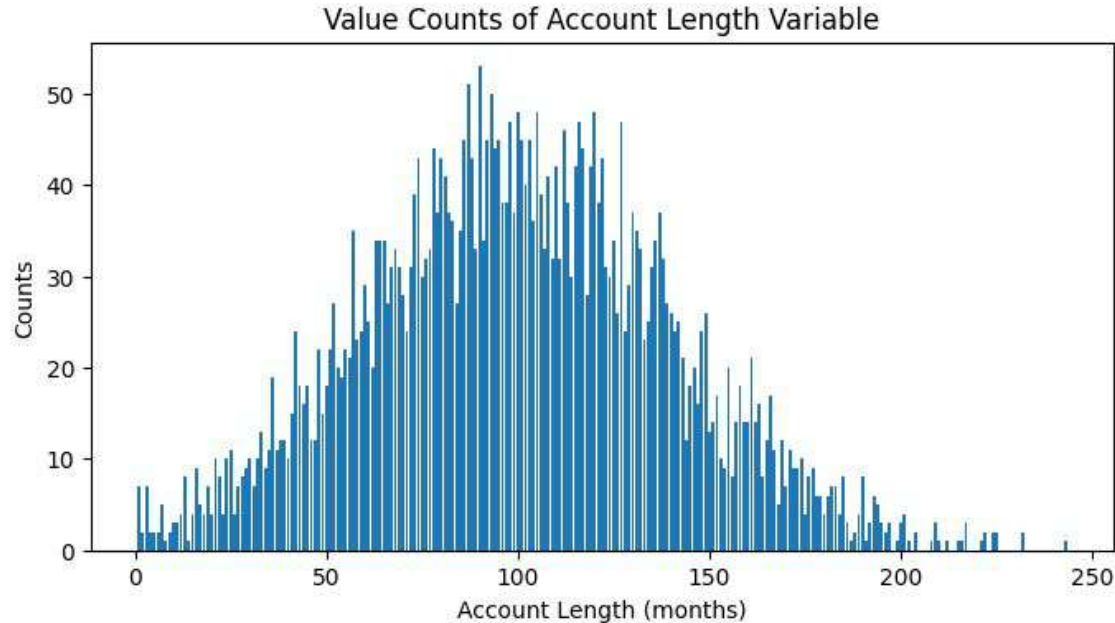
- Dataset terdiri dari 15 fitur numerikal dan 5 fitur kategorikal
- 15 fitur numerikal diantaranya sebagai berikut:

```
['account length', 'number vmail messages', 'total day minutes', 'total day calls',  
 'total day charge', 'total eve minutes', 'total eve calls', 'total eve charge',  
 'total night minutes', 'total night calls', 'total night charge', 'total intl minutes',  
 'total intl calls', 'total intl charge', 'number customer service calls']
```

- 5 fitur kategorikal diantaranya sebagai berikut:

```
['state', 'area code', 'international plan', 'voice mail plan', 'churn']
```

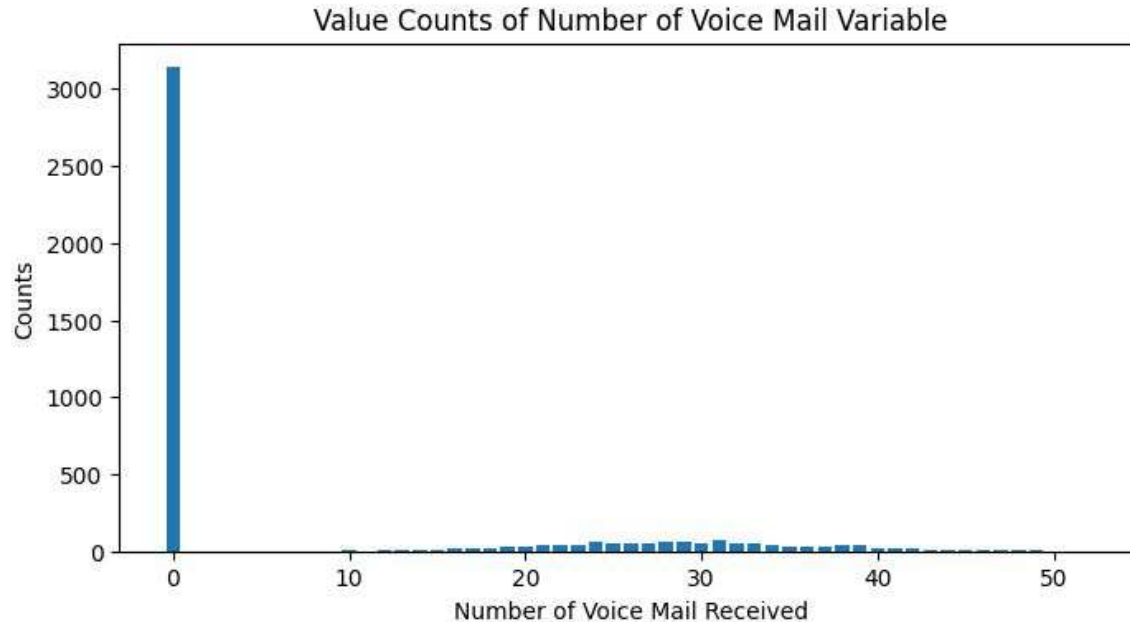

Exploratory Data Analyst (EDA)



Account Length


Grafik ini menggambarkan **banyak orang yang berlangganan** selama jangka waktu tertentu:

Bar plot tersebut menunjukkan bahwa **mayoritas customers ternyata berlangganan ke layanan ISP ini cukup lama (~100 bulan, 8 tahun)**.




Number of Voice Email

Untuk variabel ini, **bar plot** menunjukkan **tingginya frekuensi** orang yang tidak menerima pesan suara (voicemail) sama sekali. Ternyata, **mayoritas orang tidak menerima voicemail karena tidak mendaftarkan diri ke program voicemail.**



	<code>total_day_minutes</code>	<code>total_day_calls</code>	<code>total_day_charge</code>
<code>total_day_minutes</code>	1.000000	0.000747	1.000000
<code>total_day_calls</code>	0.000747	1.000000	0.000751
<code>total_day_charge</code>	1.000000	0.000751	1.000000



	<code>total_night_minutes</code>	<code>total_night_calls</code>	<code>total_night_charge</code>
<code>total_night_minutes</code>	1.000000	0.023815	0.999999
<code>total_night_calls</code>	0.023815	1.000000	0.023798
<code>total_night_charge</code>	0.999999	0.023798	1.000000

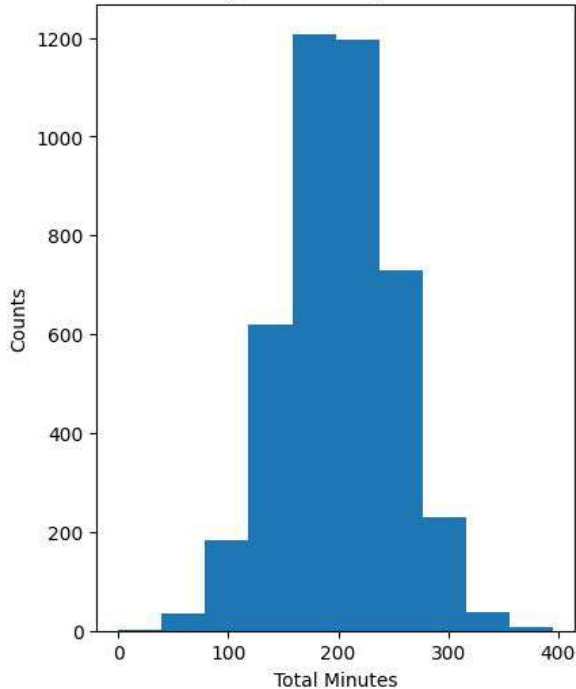
	<code>total_eve_minutes</code>	<code>total_eve_calls</code>	<code>total_eve_charge</code>
<code>total_eve_minutes</code>	1.000000	0.003101	1.000000
<code>total_eve_calls</code>	0.003101	1.000000	0.00312
<code>total_eve_charge</code>	1.000000	0.003120	1.000000

Day, Night and Eve Variables

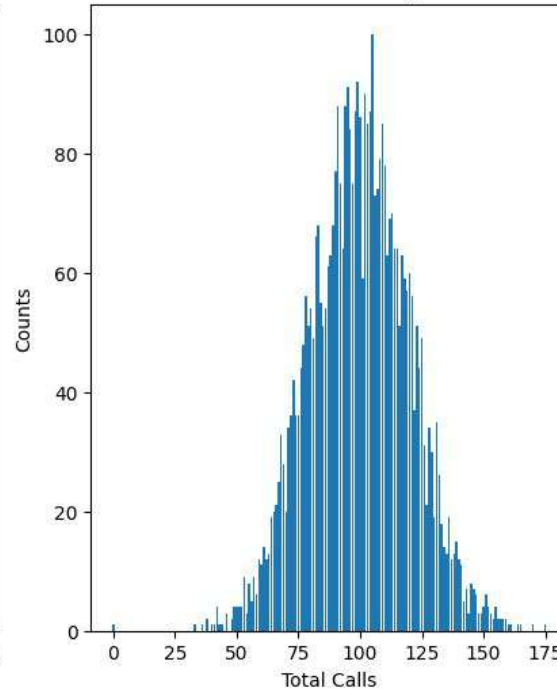
Dalam dataset ini, penggunaan telepon customer pada periode waktu tertentu dicatat dalam 3 variabel, yaitu **total calls** (jumlah panggilan), **total minutes** (lamanya panggilan), dan **total charge** (banyak tagihan).

Ternyata, apabila dicari koefisien korelasi Pearson antara ketiganya, akan didapatkan korelasi sempurna antara total minutes dan total charge. Dengan demikian, bisa dipilih salah satu variabel saja sebagai input features pada model machine learning. **Di sini, baik variable day, night, maupun eve, dipilih total_minuets dan total_calls, sedangkan total charge akan di drop.**

Histogram Total Night Minutes



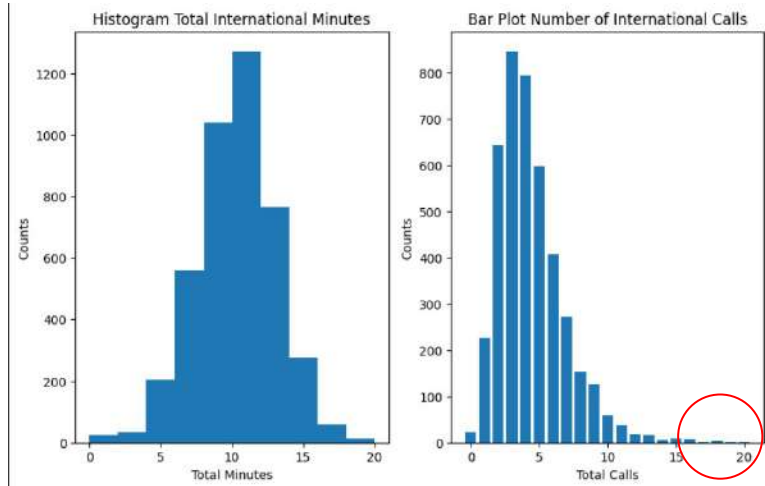
Bar Plot Number of Night Calls



Day, Night and Eve Variables*

Grafik di samping menggambarkan **jumlah dan durasi panggilan pada malam hari**, dimana terlihat ada **probable outlier** untuk variabel jumlah panggilan, yaitu pengguna dengan jumlah panggilan 0 saat malam dan intensitas normal di periode lain. Fenomena tersebut mungkin saja terjadi karena pengguna terkait benar-benar tidak melakukan aktivitas saat malam hari, sehingga data points tersebut **bukan pasti error dan tidak perlu dihapus**.

Hal ini juga terjadi untuk variabel **total_day_calls** dan **total_eve_calls**, sehingga juga tidak ada data yang dihapus untuk variabel tersebut

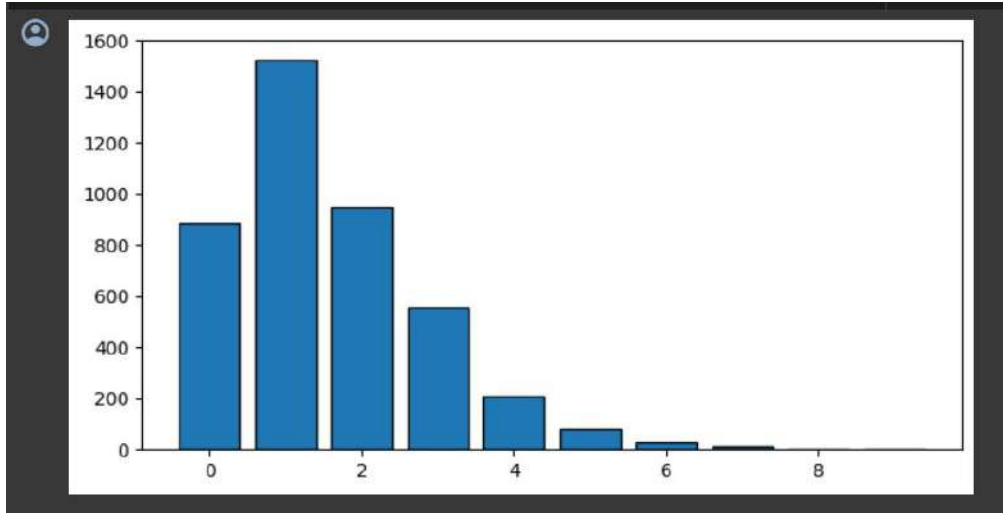


International Variables

Pada dataset ini, ada 4 variabel yang berhubungan dengan panggilan internasional yang diduga colinear satu sama lain. Ternyata, koefisien korelasi menunjukkan tingginya hubungan antara **total_intl_minutes** dan **total_intl_charge** saja, sehingga **variabel charge** akan di-drop dan yang lain di-keep.

	international_plan	total_intl_minutes	total_intl_calls	total_intl_charge
international_plan	1.000000	0.023815	0.006956	0.023799
total_intl_minutes	0.023815	1.000000	0.019328	0.999993
total_intl_calls	0.006956	0.019328	1.000000	0.019414
total_intl_charge	0.023799	0.999993	0.019414	1.000000

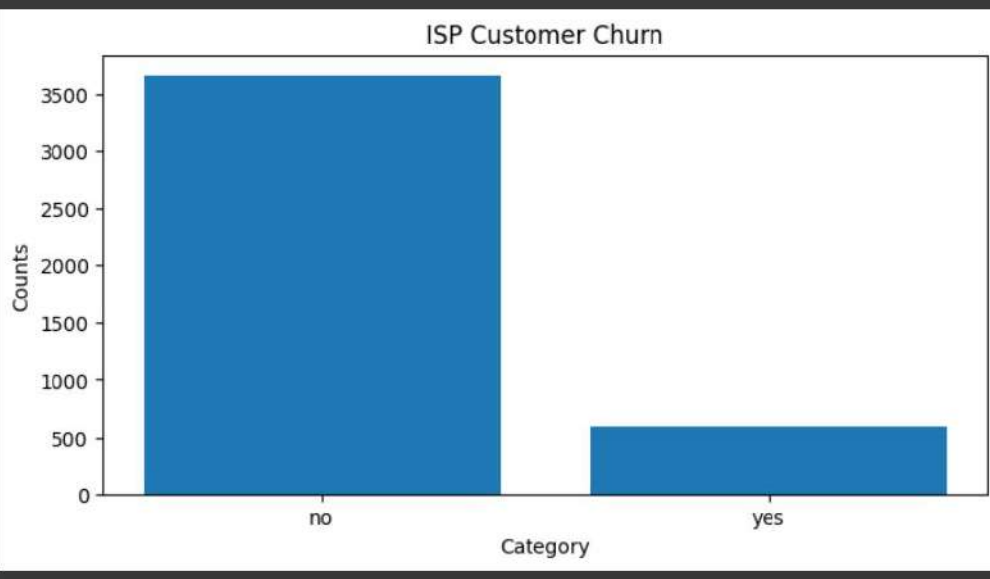
Di sisi lain, **central tendency total_intl_calls** relatif rendah dibanding jenis panggilan lain, sehingga **16 kali panggilan** sudah dianggap probable outlier. Karena masih masuk akal, maka data ini akan **tetap dipertahankan dalam dataset**.



Number Customer Service Call

Pada variabel **number_customer_service_calls**, karena central tendency variabelnya rendah, upper outer fence-nya pun cukup rendah, sehingga **6 kali panggilan** pun sudah dianggap tidak masuk akal dan dikategorikan sebagai probable outlier.

Dengan demikian, tidak dilakukan penghapusan data lagi di variabel ini.

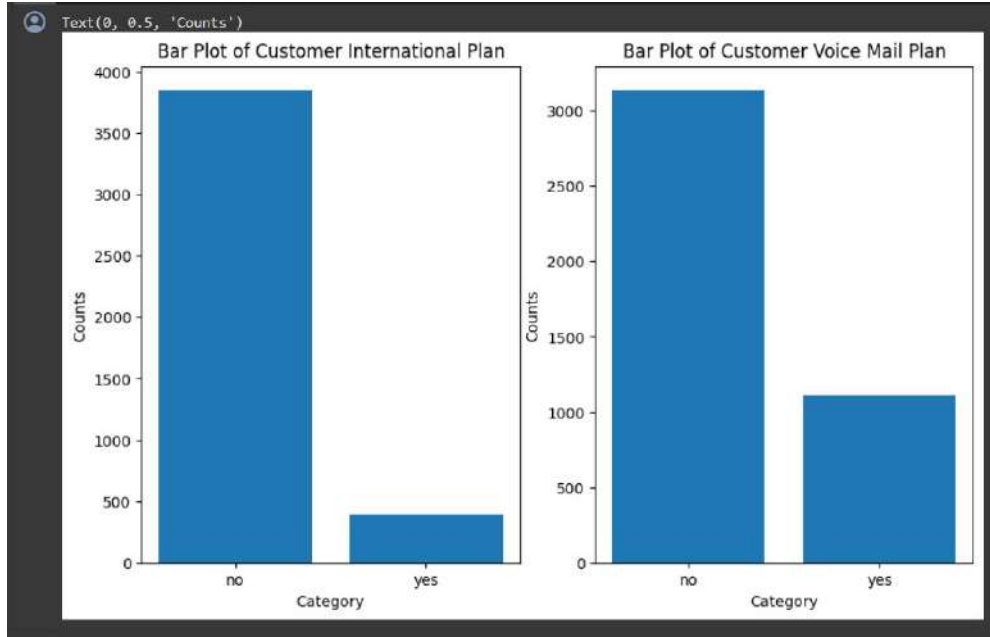


Churn

Grafik ini menggambarkan frekuensi pelanggan yang lanjut (**churn='no'**) dan berhenti (**churn='yes'**) berlangganan.

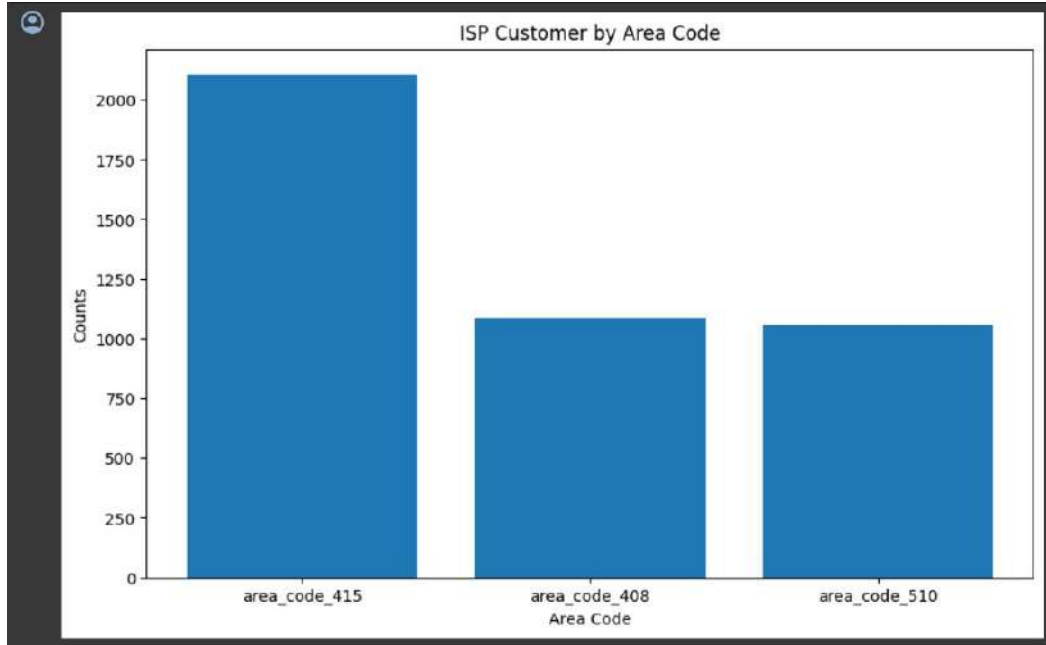
Pada bar plot, terlihat bahwa jumlah customer yang melanjutkan berlangganan jauh lebih banyak daripada yang berhenti. Idealnya, dilakukan teknik seperti *Synthetic Minority Oversampling Technique* (**SMOTE**) untuk mengatasi imbalanced classification, namun pada challenge ini dataset training akan diterima seadanya dengan asumsi dataset representatif terhadap kondisi populasi.

Label kategori biner ini juga perlu di-encode pada feature engineering dengan mengganti 'no' menjadi 0 dan 'yes' menjadi 1.



International Plan dan Voice Mail Plan

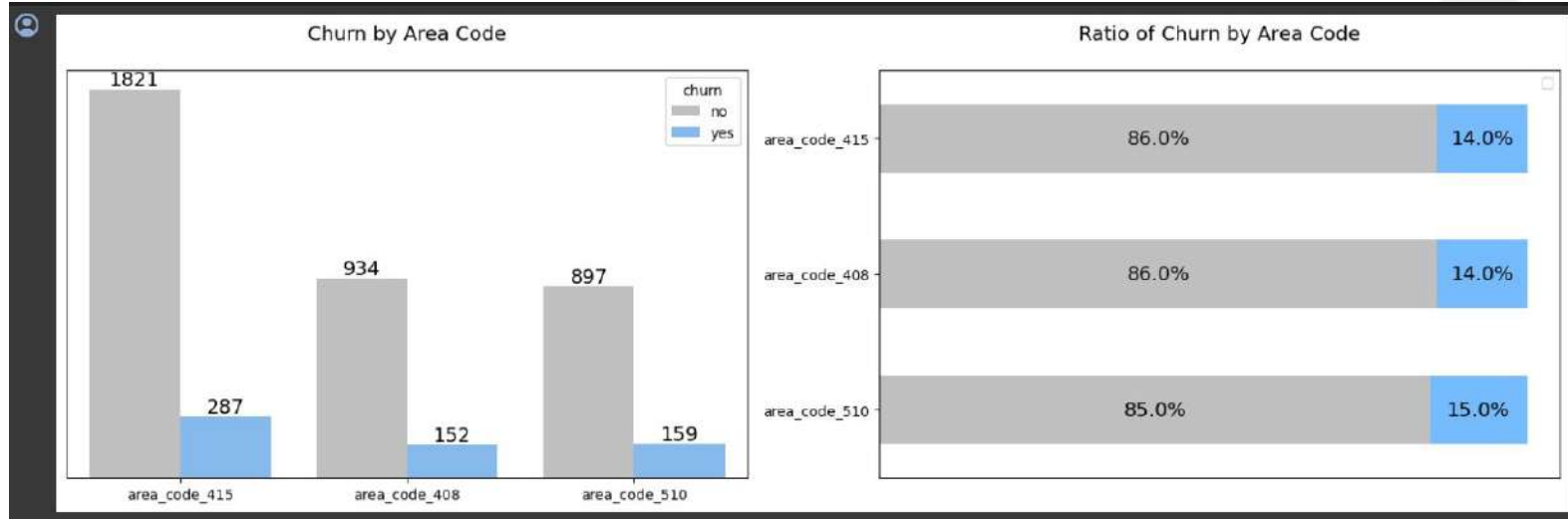
Kedua variabel ini menunjukkan status langganan customer terhadap layanan **international plan** dan **voicemail plan**. Pengguna yang tidak memiliki layanan voicemail tidak bisa menerima pesan suara sama sekali (sehingga **number_vmail_messages=0**), sedangkan orang yang tidak memiliki international plan masih bisa melakukan panggilan internasional namun dengan rate yang lebih tinggi.



Area Code*

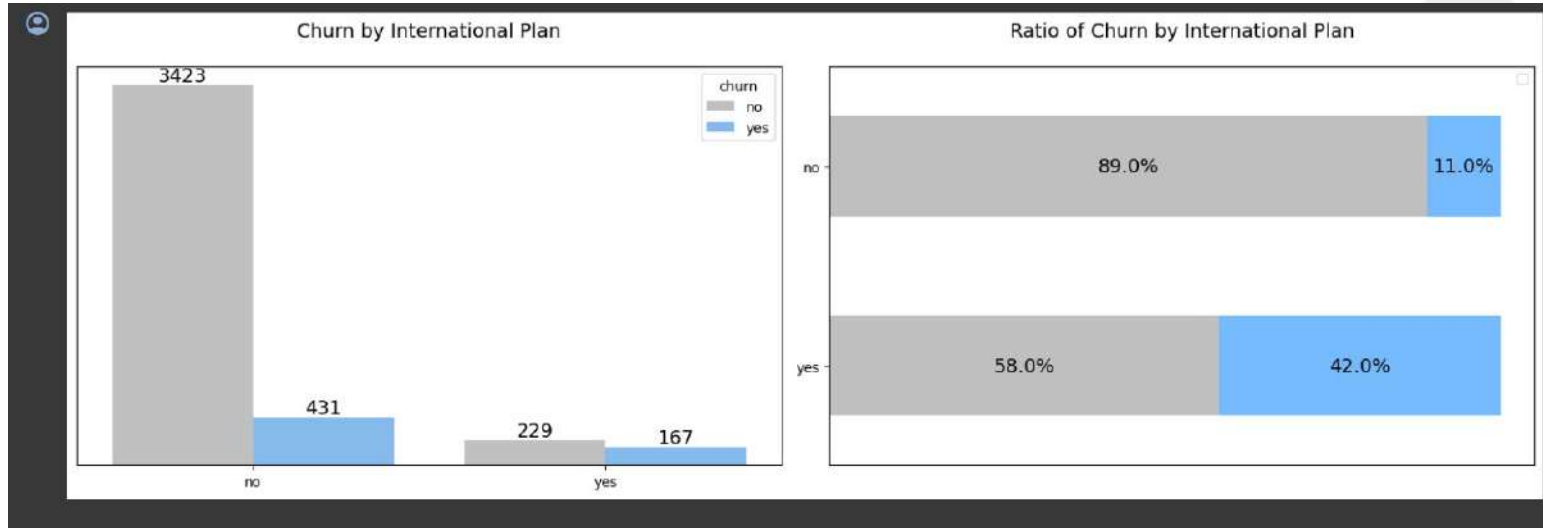
Di dataset ini, hanya ada 3 kategori dalam variabel kode area. Hal ini cukup mencurigakan karena area code di USA umumnya dibagi berdasarkan wilayah geografisnya, sehingga karena ada 51 negara bagian maka pasti ada >51 kode area.

Ternyata, untuk data dengan kode area 415 yang seharusnya ada di California (CA), tercatat ada 51 unique values pada variabel state. Ini menunjukkan variabel **area_code** kemungkinan besar adalah error yang harus di drop pada feature engineering.



Area Code

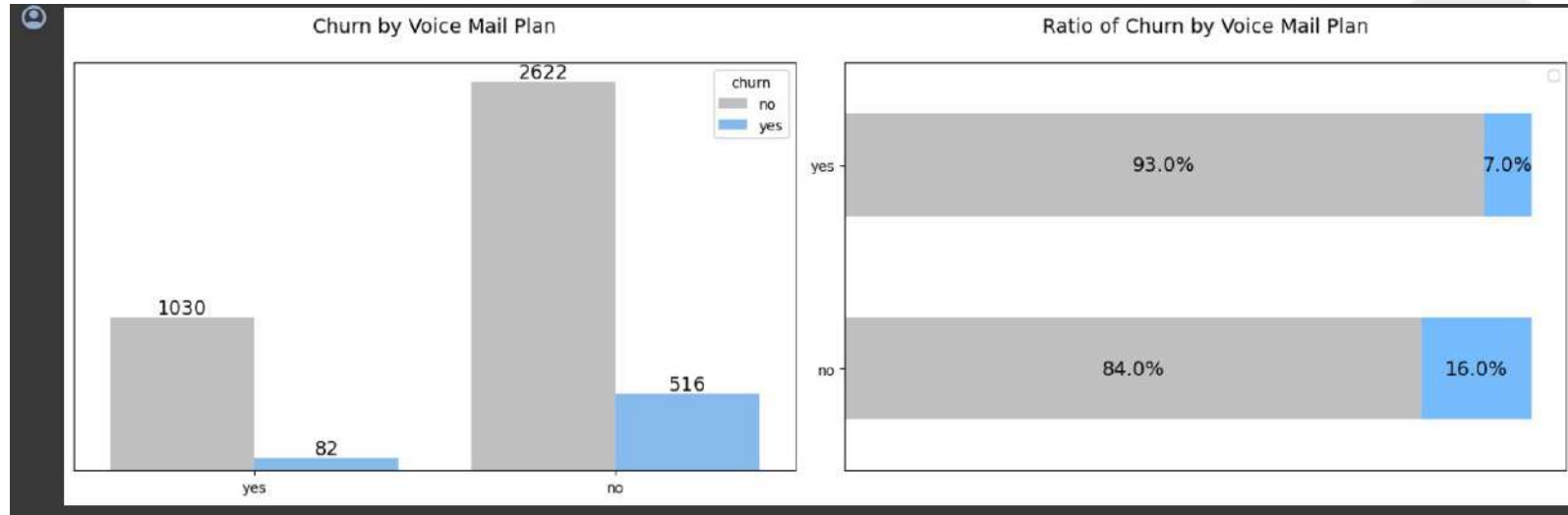
Area 415 memiliki jumlah (count) customer paling tinggi. Secara rasio, pada setiap area memiliki persentase customer yang churn tidak jauh berbeda yaitu 14 - 15 %



International Plan

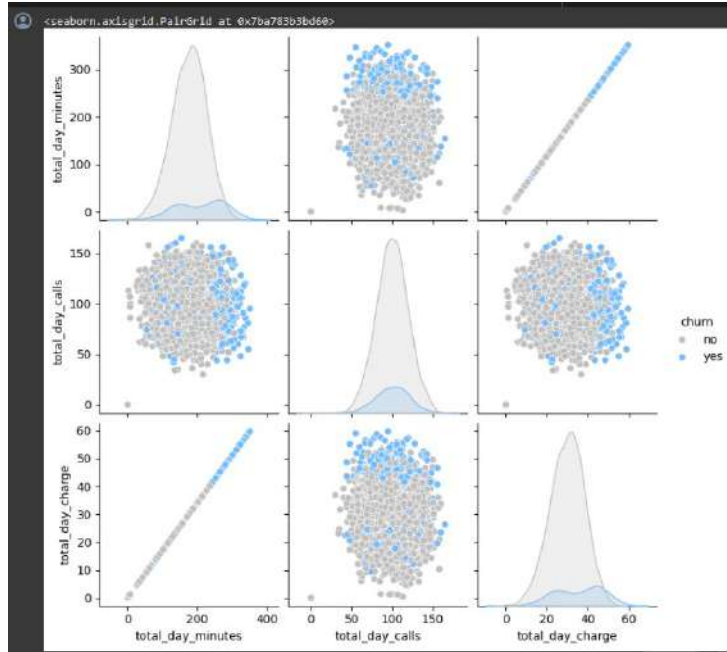
Angka pelanggan yang memiliki International Plan cukup rendah namun memiliki tingkat churn yang tinggi

- Churn pada pelanggan yang memiliki Internatiol plan ini kemungkinan dapat terjadi karena biaya roaming yang tinggi atau masalah kualitas jaringan provider



Voice Mail

Mayoritas pelanggan tidak memiliki voice mail plan, dan pelanggan yang tidak memiliki voice mail plan lebih berpotensi untuk churn. Kemungkinan pelanggan churn karena jika pelanggan tidak memiliki voice mail plan, mereka tidak akan dapat menerima pesan suara dari orang-orang yang mencoba menghubungi mereka ketika mereka tidak dapat menjawab telepon. Ini dapat menyebabkan ketidaknyamanan bagi pelanggan.



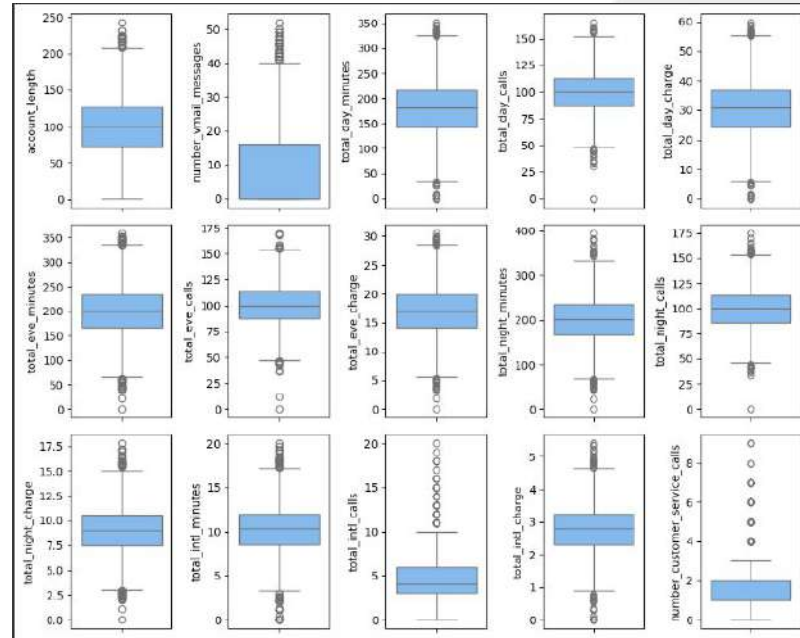
Total Charge, Calls, Minute

Total harga yang dibebankan pada panggilan di pagi hari memiliki pengaruh yang cukup besar terhadap tingkat churn pelanggan. Terlihat bahwa mayoritas pelanggan churn yang memiliki durasi menit panggilan lebih lama, mendapatkan harga panggilan yang lebih besar. Hal tersebut kemungkinan dapat disebabkan karena pelanggan tidak puas terhadap harga untuk telepon dengan durasi yang lama (mungkin terlalu mahal untuk pagi hari).

Data Preprocessing

Handling Outliers

Outliers masih **dapat ditoleransi** atau **bukan kesalahan input** data atau **nilai ekstrem**, sehingga tidak dilakukan penanganan pada outliers



Features Encoding

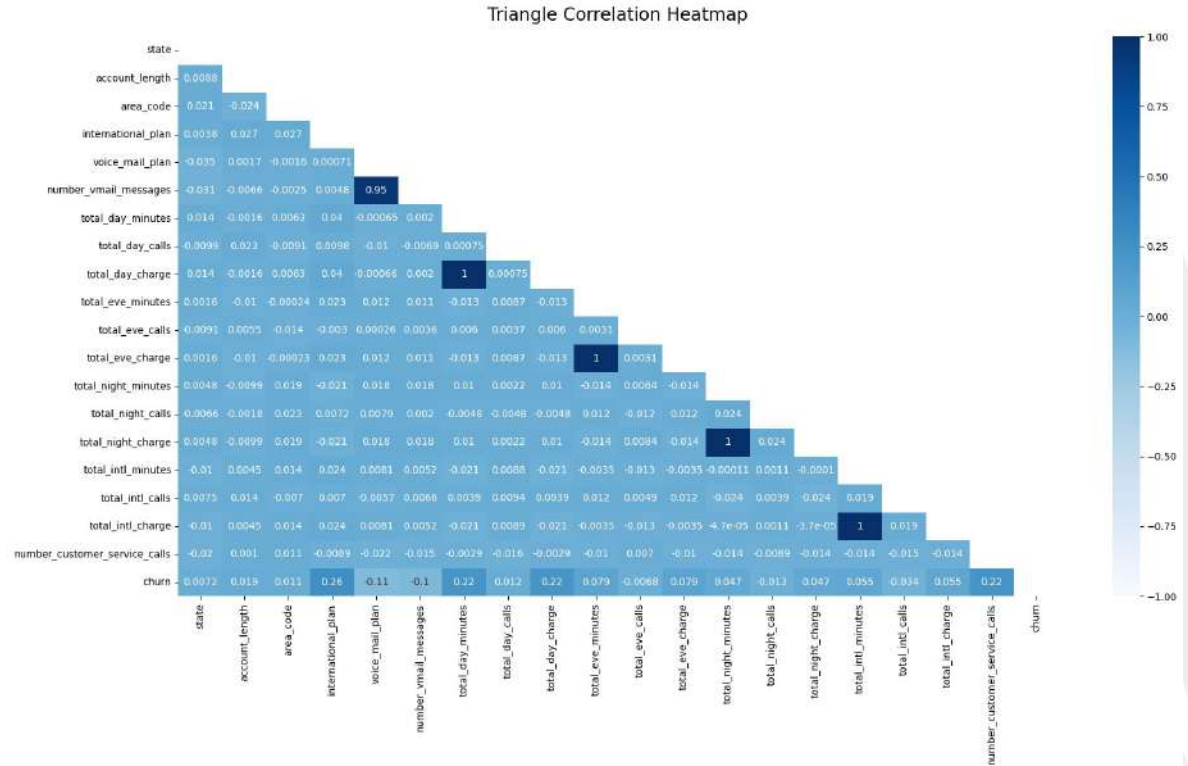
Dilakukan Label Encoding pada fitur '**international_plan**', '**voice_mail_plan**', '**churn**' (Yes – 1, No – 0), dan **area_code**

```
le = LabelEncoder()
# -----
for col in df_en.columns:
    if df_en[col].dtype == 'O':
        df_en[col] = le.fit_transform(df_en[col])
```

Features Selection

Memilih fitur yang akan digunakan pada model. Melakukan drop pada beberapa fitur diantaranya:

'number_vmail_messages',
'total_day_minutes',
'total_eve_minutes',
'total_night_minutes',
'total_intl_minutes', 'state'



Modeling Dan Evaluation

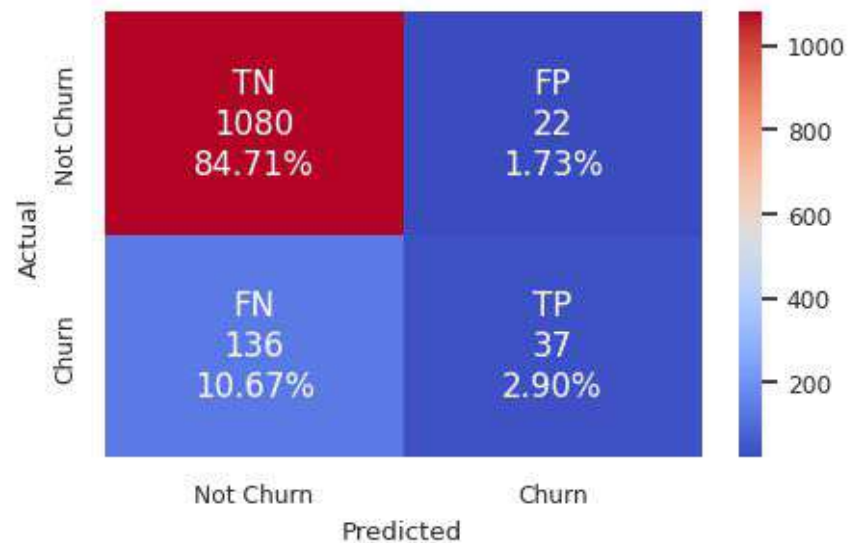
Model Evaluation

Random Forest adalah model terbaik daripada model lain yang dibandingkan, karena menggunakan algoritma pembelajaran supervisi (***Supervised Learning***) dan memiliki **nilai AUC tinggi** untuk memaksimalkan prediksi.

	Models	Recall	AUC	AUC Train
1	Random Forest Classifiefier	0.89	0.93	0.94
0	Logistic Regression	0.78	0.84	0.82

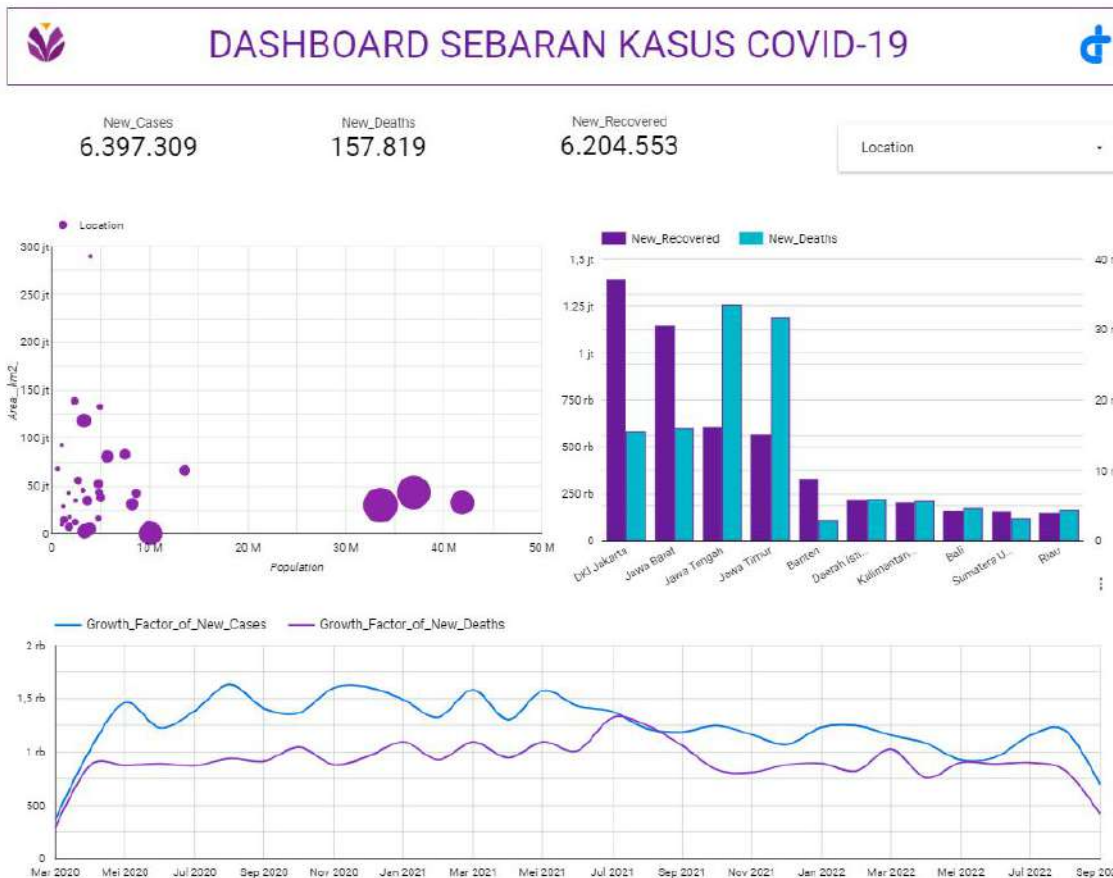
Confusion Matrix

Recall berkonsentrasi pada seberapa banyak **pelanggan churn yang dapat diidentifikasi**. Model dapat menentukan jumlah pelanggan yang mungkin **meninggalkan layanan (TP+FP)**. Hal ini dilakukan untuk membantu bisnis menghindari masalah dan mempertahankan pelanggan. Selain itu, **kesalahan prediksi customer non-churn (FN)** memiliki nilai paling rendah, menurut hasil confusion matrix.



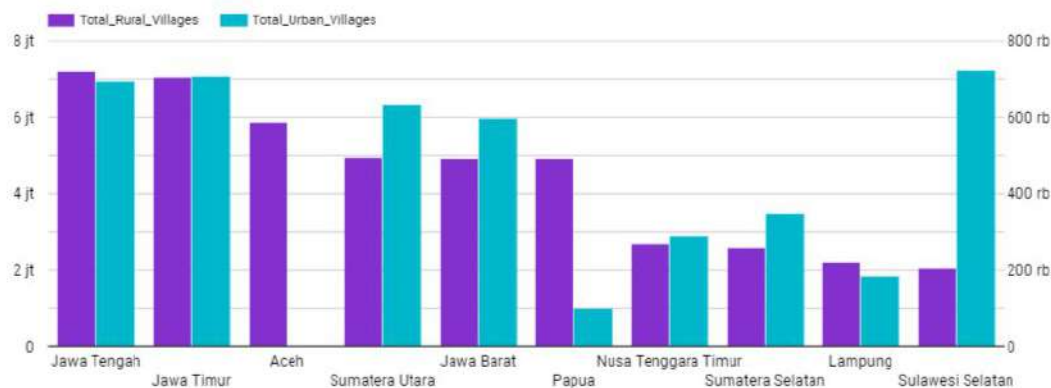
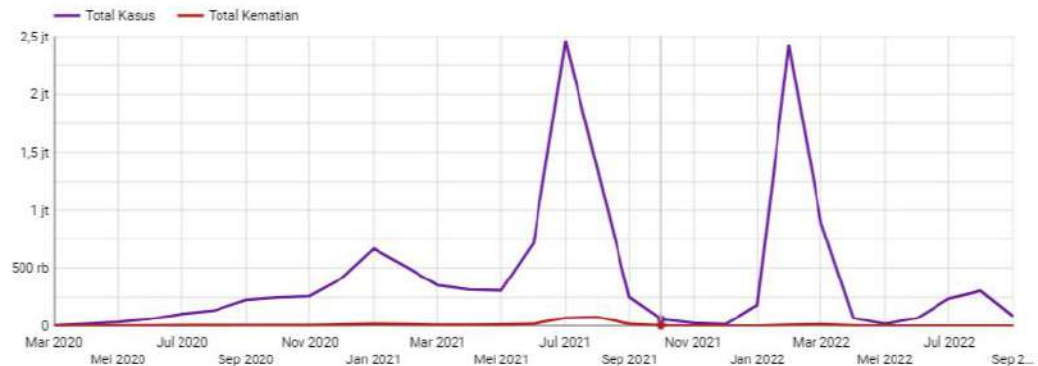
Appendix

Appendix	Link
Gdocs (SQL Query)	https://docs.google.com/document/d/1P-YbAGm5-UcLisZN4aETeSvEdKxXDxzh_jwCvpqJHGI/edit?usp=sharing
Dashboard	https://lookerstudio.google.com/reporting/1f8f8fe5-ee21-4524-9c62-cd3f382f462a
Churn Classification	https://colab.research.google.com/drive/1wut7sC3xb3568nI9z6REyS-G_2o03OE1?usp=sharing





DASHBOARD SEBARAN KASUS COVID-19





DASHBOARD SEBARAN KASUS COVID-19



Thank You