

# **[Team 12 DS 3] - Final Presentation**

[Data Science]

Kelas	TIM	NAMA
DS-3	12	- Muhamad Shidqi
		- Mohammad Rifqi Nur Faroza

# **Chapter 1. Covid-19 New Cases in Indonesia**

### Studi kasusss??

Terdapat data yang sangat banyak tentang kasus covid-19 yang terjadi di Indonesia. Sebagai junior data scientist tim teknologi kesehatan, disini kita akan memberikan beberapa informasi dan insight terkait data covid-19 di Indonesia.

# Sub Topic

Kasus Covid 19 di Indonesia.csv

Buka dengan

Bagikan

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
Date	Location ISO Code	Location	New Cases	New Deaths	New Recovered	New Active Cases	Total Cases	Total Deaths	Total Recovered	Total Active Cases	Location Level	City or Regency	Province	Country	Continent	Island	Time Zone	Special Status	Total Regencies
3/1/2020	ID-JK	DKI Jakarta	2	0	0	0	2	39	20	75	-66	Province	DKI Jakarta	Indonesia	Asia	Jawa	UTC+07:00	Daerah Khusus Ibu K	1
3/2/2020	ID-JK	DKI Jakarta	2	0	0	0	2	41	20	75	-64	Province	DKI Jakarta	Indonesia	Asia	Jawa	UTC+07:00	Daerah Khusus Ibu K	1
3/2/2020	ID-JK	Indonesia	2	0	0	0	2	2	0	0	2	Country		Indonesia	Asia	Jawa	UTC+07:00		416
3/2/2020	ID-RI	Riau	1	0	0	0	1	0	1	0	0	Province	Riau	Indonesia	Asia	Sumatera	UTC+07:00		10
3/3/2020	ID-JK	DKI Jakarta	2	0	0	0	2	43	20	75	-62	Province	DKI Jakarta	Indonesia	Asia	Jawa	UTC+07:00	Daerah Khusus Ibu K	1
3/3/2020	ID-JK	Indonesia	0	0	0	0	0	2	0	0	2	Country		Indonesia	Asia	Jawa	UTC+07:00		416
3/3/2020	ID-JB	Jawa Barat	1	1	0	0	0	1	1	60	-60	Province	Jawa Barat	Indonesia	Asia	Jawa	UTC+07:00		18
3/3/2020	ID-RI	Riau	0	0	0	0	0	1	0	1	0	Province	Riau	Indonesia	Asia	Sumatera	UTC+07:00		10
3/4/2020	ID-JK	DKI Jakarta	2	0	0	0	2	45	20	75	-60	Province	DKI Jakarta	Indonesia	Asia	Jawa	UTC+07:00	Daerah Khusus Ibu K	1
3/4/2020	ID-JK	Indonesia	0	0	0	0	0	2	0	0	2	Country		Indonesia	Asia	Jawa	UTC+07:00		416
3/4/2020	ID-JB	Jawa Barat	1	0	0	0	1	2	1	60	-59	Province	Jawa Barat	Indonesia	Asia	Jawa	UTC+07:00		18
3/4/2020	ID-RI	Riau	0	0	0	0	1	0	1	0	0	Province	Riau	Indonesia	Asia	Sumatera	UTC+07:00		10
3/5/2020	ID-JK	DKI Jakarta	0	1	0	0	-1	45	21	75	-61	Province	DKI Jakarta	Indonesia	Asia	Jawa	UTC+07:00	Daerah Khusus Ibu K	1
3/5/2020	ID-JK	Indonesia	0	0	0	0	0	2	0	0	2	Country		Indonesia	Asia	Jawa	UTC+07:00		416
3/5/2020	ID-JB	Jawa Barat	1	0	0	0	1	3	1	60	-58	Province	Jawa Barat	Indonesia	Asia	Jawa	UTC+07:00		18
3/5/2020	ID-RI	Riau	0	0	0	0	0	1	0	1	0	Province	Riau	Indonesia	Asia	Sumatera	UTC+07:00		10
3/6/2020	ID-BT	Banten	1	0	1	0	1	5	111	-115	Province	Banten	Indonesia	Asia	Jawa	UTC+07:00		4	
3/6/2020	ID-JK	DKI Jakarta	0	0	0	0	45	21	75	-61	Province	DKI Jakarta	Indonesia	Asia	Jawa	UTC+07:00	Daerah Khusus Ibu K	1	
3/6/2020	ID-JK	Indonesia	2	0	0	0	2	4	0	0	4	Country		Indonesia	Asia	Jawa	UTC+07:00		416
3/6/2020	ID-JB	Jawa Barat	1	0	0	0	1	4	1	60	-57	Province	Jawa Barat	Indonesia	Asia	Jawa	UTC+07:00		18
3/6/2020	ID-RI	Riau	0	0	0	0	0	1	0	1	0	Province	Riau	Indonesia	Asia	Sumatera	UTC+07:00		10
3/7/2020	ID-BT	Banten	0	0	0	0	0	1	5	111	-115	Province	Banten	Indonesia	Asia	Jawa	UTC+07:00		4
3/7/2020	ID-JK	DKI Jakarta	0	2	0	0	-2	45	23	75	-63	Province	DKI Jakarta	Indonesia	Asia	Jawa	UTC+07:00	Daerah Khusus Ibu K	1
3/7/2020	ID-JK	Indonesia	0	0	0	0	0	4	0	0	4	Country		Indonesia	Asia	Jawa	UTC+07:00		416
3/7/2020	ID-JB	Jawa Barat	0	0	0	0	0	4	1	60	-57	Province	Jawa Barat	Indonesia	Asia	Jawa	UTC+07:00		18
3/7/2020	ID-RI	Riau	0	0	0	0	0	1	0	1	0	Province	Riau	Indonesia	Asia	Sumatera	UTC+07:00		10
3/8/2020	ID-BT	Banten	1	0	3	-2	2	5	114	-117	Province	Banten	Indonesia	Asia	Jawa	UTC+07:00		4	
3/8/2020	ID-JK	DKI Jakarta	0	0	0	0	45	23	75	-53	Province	DKI Jakarta	Indonesia	Asia	Jawa	UTC+07:00	Daerah Khusus Ibu K	1	
3/8/2020	ID-JK	Indonesia	2	0	0	0	2	6	0	0	6	Country		Indonesia	Asia	Jawa	UTC+07:00		416
3/8/2020	ID-JB	Jawa Barat	0	0	0	0	0	4	1	60	-57	Province	Jawa Barat	Indonesia	Asia	Jawa	UTC+07:00		18
3/8/2020	ID-JT	Jawa Tengah	1	1	0	0	0	1	10	11	-20	Province	Jawa Tengah	Indonesia	Asia	Jawa	UTC+07:00		29
3/8/2020	ID-RI	Riau	0	0	0	0	0	1	0	1	0	Province	Riau	Indonesia	Asia	Sumatera	UTC+07:00		10
3/9/2020	ID-BT	Banten	0	0	0	0	0	2	5	114	-117	Province	Banten	Indonesia	Asia	Jawa	UTC+07:00		4
3/9/2020	ID-JK	DKI Jakarta	0	1	0	0	-1	45	24	75	-64	Province	DKI Jakarta	Indonesia	Asia	Jawa	UTC+07:00	Daerah Khusus Ibu K	1
3/9/2020	ID-JK	Indonesia	13	0	0	0	13	0	19	0	19	Country		Indonesia	Asia	Jawa	UTC+07:00		416
3/9/2020	ID-JB	Jawa Barat	0	0	0	0	0	4	1	60	-57	Province	Jawa Barat	Indonesia	Asia	Jawa	UTC+07:00		18
3/9/2020	ID-JT	Jawa Tengah	0	0	0	0	0	1	10	11	-20	Province	Jawa Tengah	Indonesia	Asia	Jawa	UTC+07:00		29
3/9/2020	ID-RI	Riau	0	0	0	0	0	1	0	1	0	Province	Riau	Indonesia	Asia	Sumatera	UTC+07:00		10
3/10/2020	ID-BT	Banten	0	0	8	-8	2	5	122	-125	Province	Banten	Indonesia	Asia	Jawa	UTC+07:00		4	
3/10/2020	ID-JK	DKI Jakarta	0	0	0	0	0	45	24	75	-64	Province	DKI Jakarta	Indonesia	Asia	Jawa	UTC+07:00	Daerah Khusus Ibu K	1
3/10/2020	ID-JK	Indonesia	8	0	2	6	27	0	2	25	Country		Indonesia	Asia	Jawa	UTC+07:00		416	
3/10/2020	ID-JB	Jawa Barat	0	0	0	0	0	4	1	60	-57	Province	Jawa Barat	Indonesia	Asia	Jawa	UTC+07:00		18
3/10/2020	ID-JT	Jawa Tengah	0	0	0	0	0	1	10	11	-20	Province	Jawa Tengah	Indonesia	Asia	Jawa	UTC+07:00		29
3/10/2020	ID-RI	Riau	0	0	0	0	0	1	0	1	0	Province	Riau	Indonesia	Asia	Sumatera	UTC+07:00		10
3/10/2020	ID-SG	Sulawesi Tenggara	1	0	0	0	1	1	0	0	1	Province	Sulawesi Tenggara	Indonesia	Asia	Sulawesi	UTC+08:00		15
3/11/2020	ID-BA	Bali	1	1	0	0	0	1	1	3	-3	Province	Bali	Indonesia	Asia	Nusa Tenggara	UTC+08:00		8
3/11/2020	ID-BT	Banten	0	0	8	-8	2	5	130	-133	Province	Banten	Indonesia	Asia	Jawa	UTC+07:00		4	
3/11/2020	ID-JK	DKI Jakarta	0	0	0	0	0	45	24	75	-64	Province	DKI Jakarta	Indonesia	Asia	Jawa	UTC+07:00	Daerah Khusus Ibu K	1
3/11/2020	ID-JK	Indonesia	7	1	0	0	6	34	1	2	31	Country		Indonesia	Asia	Jawa	UTC+07:00		416
3/11/2020	ID-JB	Jawa Barat	0	1	0	0	-1	4	2	60	-58	Province	Jawa Barat	Indonesia	Asia	Jawa	UTC+07:00		18
3/11/2020	ID-JT	Jawa Tengah	2	1	1	0	3	11	12	-20	Province	Jawa Tengah	Indonesia	Asia	Jawa	UTC+07:00		29	
3/11/2020	ID-RI	Riau	0	0	0	0	0	1	0	1	0	Province	Riau	Indonesia	Asia	Sumatera	UTC+07:00		10

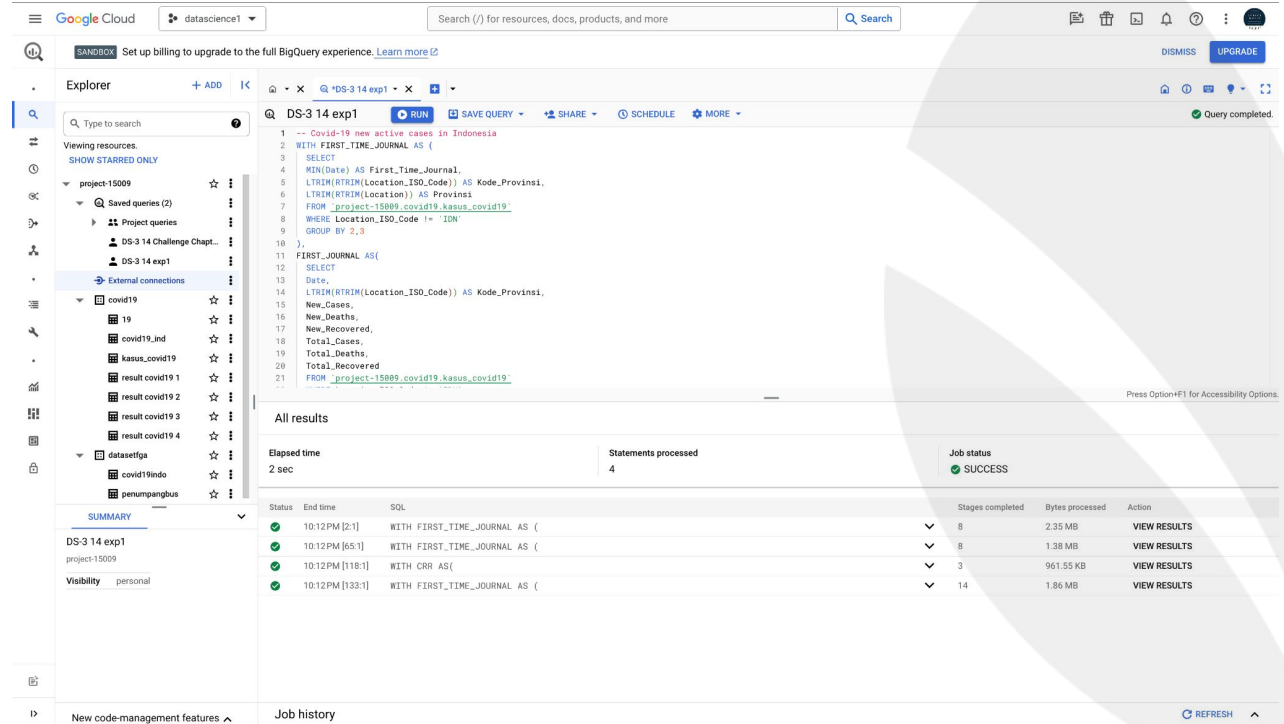
Ada apa sih?

Ada apa aja sih di dalam data covid-19 di Indonesia??

## Sub Topic

### Step 1

Pertama tentunya kita akan import data ke Big Query



The screenshot displays the Google Cloud BigQuery console. The top navigation bar includes the Google Cloud logo, a project selector (data science1), a search bar, and utility icons. Below the navigation bar, the Explorer panel on the left shows a tree view of resources, including a project named 'project-15009' with various datasets and queries. The main editor area shows a SQL query titled 'DS-3 14 exp1' with the following content:

```
1 -- Covid-19 new active cases in Indonesia
2 WITH FIRST_TIME_JOURNAL AS (
3   SELECT
4     MIN(Date) AS First_Time_Journal,
5     LTRIM(RTRIM(Location_ISO_Code)) AS Kode_Provinsi,
6     LTRIM(RTRIM(Location)) AS Provinsi
7   FROM `project-15009.covid19.kasus_covid19`
8   WHERE Location_ISO_Code != 'IDN'
9   GROUP BY 2,3
10 ),
11 FIRST_JOURNAL AS(
12   SELECT
13     Date,
14     LTRIM(RTRIM(Location_ISO_Code)) AS Kode_Provinsi,
15     New_Cases,
16     New_Deaths,
17     New_Recovered,
18     Total_Cases,
19     Total_Deaths,
20     Total_Recovered
21   FROM `project-15009.covid19.kasus_covid19`
22 )
```

The query has been executed successfully, as indicated by the 'Query completed.' status. Below the query editor, the 'All results' section shows the execution details:

Elapsed time	Statements processed	Job status
2 sec	4	SUCCESS

The 'All results' section also includes a table with the following columns: Status, End time, SQL, Stages completed, Bytes processed, and Action. The table contains four rows of results, each with a green checkmark in the Status column and a 'VIEW RESULTS' link in the Action column.

Status	End time	SQL	Stages completed	Bytes processed	Action
✓	10:12 PM [2:1]	WITH FIRST_TIME_JOURNAL AS (	8	2.35 MB	VIEW RESULTS
✓	10:12 PM [6:5:1]	WITH FIRST_TIME_JOURNAL AS (	8	1.38 MB	VIEW RESULTS
✓	10:12 PM [118:1]	WITH CRR AS(	3	961.55 KB	VIEW RESULTS
✓	10:12 PM [133:1]	WITH FIRST_TIME_JOURNAL AS (	14	1.86 MB	VIEW RESULTS

At the bottom of the console, there is a 'Job history' section with a 'REFRESH' button.

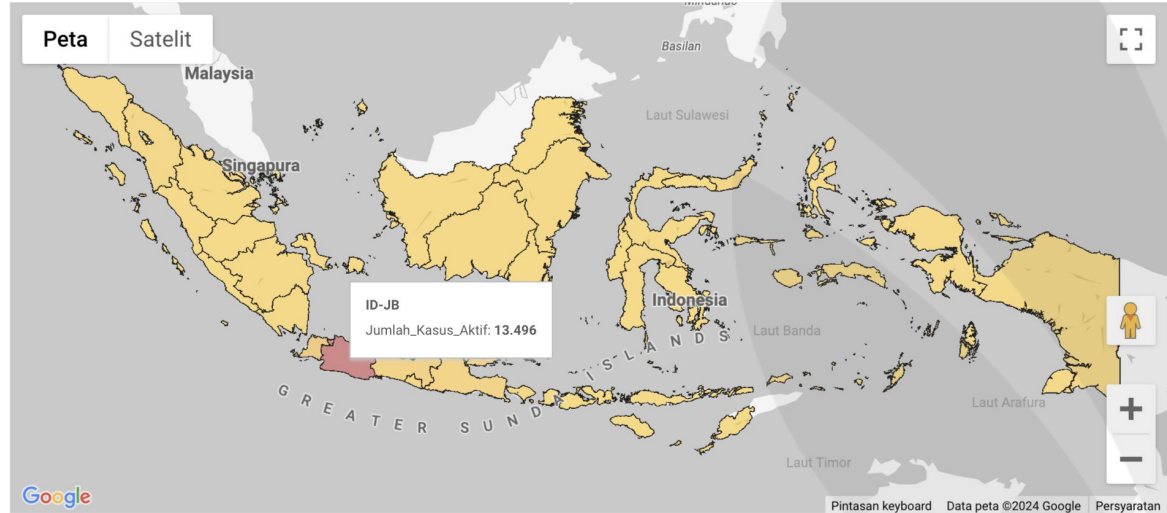
### Step 2

Setelah data dataset di import ke BigQuery kita dapat melakukan beberapa query

```
1  -- Covid-19 new active cases in Indonesia
2  WITH FIRST_TIME_JOURNAL AS (
3    SELECT
4      MIN(Date) AS First_Time_Journal,
5      LTRIM(RTRIM(Location_ISO_Code)) AS Kode_Provinsi,
6      LTRIM(RTRIM(Location)) AS Provinsi
7    FROM `project-15009.covid19.kasus_covid19`
8    WHERE Location_ISO_Code != 'IDN'
9    GROUP BY 2,3
10 ),
11 FIRST_JOURNAL AS(
12   SELECT
13     Date,
14     LTRIM(RTRIM(Location_ISO_Code)) AS Kode_Provinsi,
15     New_Cases,
16     New_Deaths,
17     New_Recovered,
18     Total_Cases,
19     Total_Deaths,
20     Total_Recovered
21   FROM `project-15009.covid19.kasus_covid19`
22   WHERE Location_ISO_Code != 'IDN'
23 ),
24 OLD_CASES AS(
25   SELECT
26     a.Kode_Provinsi AS Kode_Provinsi,
27     a.Provinsi,
28     (b.Total_Cases - b.New_Cases) AS Kasus_Aktif_Awal,
29     (b.Total_Deaths - b.New_Deaths) AS Kematian_Awal,
30     (b.Total_Recovered - b.New_Recovered) AS Sembuh_Awal,
31   FROM FIRST_TIME_JOURNAL AS a
32   LEFT JOIN FIRST_JOURNAL AS b
33   ON a.First_Time_Journal = b.Date AND a.Kode_Provinsi = b.Kode_Provinsi
34 ),
35 NEW_CASES AS(
36   SELECT
37     LTRIM(RTRIM(Location_ISO_Code)) AS Kode_Provinsi,
38     LTRIM(RTRIM(Location)) AS Provinsi,
39     SUM(New_Cases) AS Kasus_Baru,
40     SUM(New_Deaths) AS Kematian_Baru,
41     SUM(New_Recovered) AS Sembuh_Baru,
42   FROM `project-15009.covid19.kasus_covid19`
43   WHERE Location_ISO_Code != 'IDN'
```

## Hasil query

Rank	Kode_Provinsi	Provinsi	Jumlah_Kasus_Aktif
1	ID-JB	Jawa Barat	13496
2	ID-JK	DKI Jakarta	10959
3	ID-BT	Banten	2558
4	ID-JT	Jawa Tengah	1423
5	ID-JI	Jawa Timur	1147
6	ID-YO	Daerah Istimewa Yogyakarta	669
7	ID-SU	Sumatera Utara	664
8	ID-SA	Sulawesi Utara	565
9	ID-BA	Bali	474
10	ID-SS	Sumatera Selatan	313
11	ID-KI	Kalimantan Timur	272
12	ID-PA	Papua	237
13	ID-LA	Lampung	226
14	ID-RI	Riau	224
15	ID-KT	Kalimantan Tengah	220
16	ID-SB	Sumatera Barat	204
17	ID-KS	Kalimantan Selatan	153
18	ID-SN	Sulawesi Selatan	153
19	ID-PB	Papua Barat	137
20	ID-KR	Kepulauan Riau	130





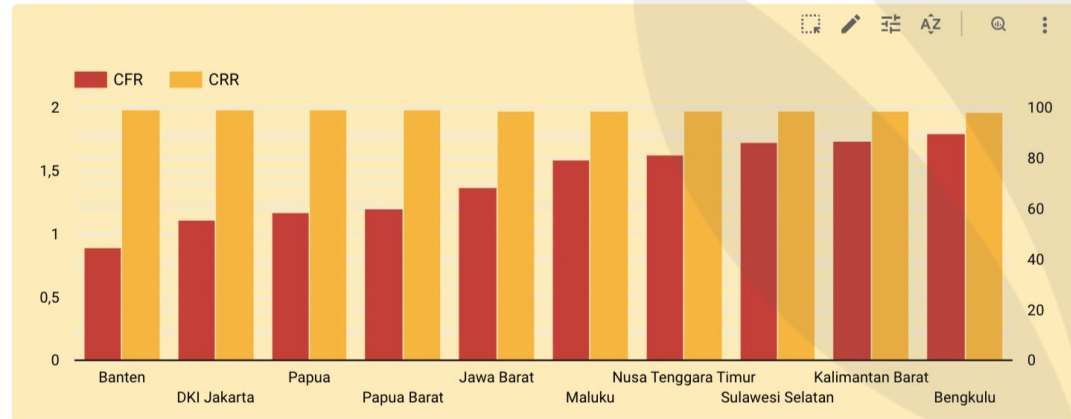
## Sub Topic

## Hasil query

CRR  
97

CFR  
3

Rank	Kode_Provinsi	Provinsi	Case Fatality Rate	Case Recovery Rate
1	ID-BT	Banten	0,89%	99,11%
2	ID-JK	DKI Jakarta	1,11%	98,89%
3	ID-PA	Papua	1,17%	98,83%
4	ID-PB	Papua Barat	1,20%	98,80%
5	ID-JB	Jawa Barat	1,37%	98,63%
6	ID-MA	Maluku	1,58%	98,42%
7	ID-NT	Nusa Tenggara Timur	1,62%	98,38%
8	ID-SN	Sulawesi Selatan	1,72%	98,28%
9	ID-KB	Kalimantan Barat	1,73%	98,27%
10	ID-BE	Bengkulu	1,79%	98,21%
11	ID-KU	Kalimantan Utara	1,89%	98,11%
12	ID-SU	Sumatera Utara	2,08%	97,92%
13	ID-SG	Sulawesi Tenggara	2,22%	97,78%
14	ID-SB	Sumatera Barat	2,27%	97,73%
15	ID-MU	Maluku Utara	2,29%	97,71%
16	ID-JA	Jambi	2,30%	97,70%
17	ID-SA	Sulawesi Utara	2,32%	97,68%
18	ID-BB	Kepulauan Bangka Belitung	2,45%	97,55%
19	ID-NB	Nusa Tenggara Barat	2,49%	97,51%
20	ID-SR	Sulawesi Barat	2,53%	97,47%



### Kesimpulan

Dalam data tersebut terlihat bahwa, penyebaran kasus covid-19 di Indonesia paling banyak terletak di provinsi pada pulau jawa karena sering dijadikan destinasi wisata dan bisnis secara lokal dan internasional. Karena alasan tersebut persebaran covid-19 semakin cepat. Dan untuk tingkat kesembuhan penanganan pasien paling banyak di daerah provinsi besar yang ada di pulau jawa, mungkin dikarenakan akan kesadaran kesehatan dan fasilitas medis yang lebih baik diantara provinsi-provinsi yang lain.

## **Chapter 2. Churn Customer**

### Business problem??

Dalam beberapa bulan terakhir perusahaan StarTelco mengalami penurunan pendapatan yang diakibatkan banyaknya persaingan antar provider perusahaan telekomunikasi untuk bersaing mendapatkan pelanggan.

Untuk menentukan strategi bisnis kedepannya, StarTelco berharap mendapatkan insight dan prediksi dari team data science dan team marketing agar perusahaan dapat mempertahankan pelanggan dan menaikan keuntungan seperti sebelumnya.

### Solution

Untuk menjawab permasalahan tersebut, team data science akan melakukan evaluasi dan mempelajari data perusahaan guna untuk memprediksi dan menentukan strategi bisnis berikutnya. Team data science akan menggunakan model XGBoosts, karena kecepatan, efisiensi, dan kemampuan menghasilkan prediksi yang akurasi akurat.

### About data

Dataset ini berisi 4250 sampel. Setiap sampel berisi 19 fitur dan 1 variabel boolean "churn" yang menunjukkan kelas sampel. 19 fitur masukan dan 1 variabel target adalah:

- **"state"**, string. 2-letter code of the US state of customer residence
- **"account\_length"**, numerical. Number of months the customer has been with the current telco provider
- **"area\_code"**, string="area\_code\_AAA" where AAA = 3 digit area code.
- **"international\_plan"**, (yes/no). The customer has international plan.
- **"voice\_mail\_plan"**, (yes/no). The customer has voice mail plan.
- **"number\_vmail\_messages"**, numerical. Number of voice-mail messages.
- **"total\_day\_minutes"**, numerical. Total minutes of day calls.
- **"total\_day\_calls"**, numerical. Total minutes of day calls.
- **"total\_day\_charge"**, numerical. Total charge of day calls.
- **"total\_eve\_minutes"**-, numerical. Total minutes of evening calls.
- **"total\_eve\_calls"**-, numerical. Total number of evening calls.
- **"total\_eve\_charge"**-, numerical. Total charge of evening calls.
- **"total\_night\_minutes"**-, numerical. Total minutes of night calls.
- **"total\_night\_calls"**-, numerical. Total number of night calls.
- **"total\_night\_charge"**-, numerical. Total charge of night calls.
- **"total\_intl\_minutes"**-, numerical. Total minutes of international calls.
- **"total\_intl\_calls"**-, numerical. Total number of international calls.
- **"total\_intl\_charge"**-, numerical. Total charge of international calls
- **"number\_customer\_service\_calls"**-, numerical. Number of calls to customer service
- **"churn"**-, (yes/no). Customer churn - target variable.

## Explotartory Data Analysis

Pada EDA ini gambar di samping menunjukkan pembersihan dan memperbaiki data guna untuk menunjang hasil terbaik saat training model.

```
1 #Delete area_code sentence from area_code column
2 train['area_code'] = train['area_code'].str.extract('(\d+)')
3
4 #Correct innappropriate data type from train dataset base on Attribute Information
5 train['international_plan'] = np.where(train['international_plan'] == 'no', 0, 1).astype('int64')
6 train['voice_mail_plan'] = np.where(train['voice_mail_plan'] == 'no', 0, 1).astype('int64')
7 train['churn'] = np.where(train['churn'] == 'no', 0, 1).astype('int64')
8
9 #Recheck The Data Summary
10 data_summary(train)
```

	Features	Data Types	Missing Values	Number of Unique Values	Unique Value
0	state	object	0	51	[OH, NJ, OK, MA, MO, LA, WV, IN, RI, IA, MT, N...
1	account_length	int64	0	215	No unique value
2	area_code	object	0	3	[415, 408, 510]
3	international_plan	int64	0	2	No unique value
4	voice_mail_plan	int64	0	2	No unique value
5	number_vmail_messages	int64	0	46	No unique value
6	total_day_minutes	float64	0	1843	No unique value
7	total_day_calls	int64	0	120	No unique value
8	total_day_charge	float64	0	1843	No unique value
9	total_eve_minutes	float64	0	1773	No unique value
10	total_eve_calls	int64	0	123	No unique value
11	total_eve_charge	float64	0	1572	No unique value
12	total_night_minutes	float64	0	1757	No unique value
13	total_night_calls	int64	0	128	No unique value
14	total_night_charge	float64	0	992	No unique value
15	total_intl_minutes	float64	0	168	No unique value
16	total_intl_calls	int64	0	21	No unique value
17	total_intl_charge	float64	0	168	No unique value
18	number_customer_service_calls	int64	0	10	No unique value
19	churn	int64	0	2	No unique value

### Kenapa sih???

Kenapa sih kita ambil model

XGBoosts????

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
xgboost	Extreme Gradient Boosting	0.9570	0.9188	0.7422	0.9407	0.8280	0.8039	0.8123	0.1250
lightgbm	Light Gradient Boosting Machine	0.9546	0.9140	0.7232	0.9433	0.8156	0.7905	0.8013	0.2300
rf	Random Forest Classifier	0.9533	0.9197	0.7016	0.9540	0.8063	0.7806	0.7938	0.0800
gbc	Gradient Boosting Classifier	0.9516	0.9177	0.7231	0.9174	0.8068	0.7796	0.7880	0.1440
et	Extra Trees Classifier	0.9301	0.9142	0.5301	0.9517	0.6790	0.6436	0.6793	0.0600
dt	Decision Tree Classifier	0.9186	0.8318	0.7110	0.7145	0.7109	0.6637	0.6648	0.0180
knn	K Neighbors Classifier	0.8884	0.6866	0.3151	0.7487	0.4417	0.3912	0.4375	0.5690
ada	Ada Boost Classifier	0.8773	0.8548	0.3772	0.6128	0.4647	0.3998	0.4160	0.0490
nb	Naive Bayes	0.8719	0.8414	0.5512	0.5491	0.5483	0.4739	0.4750	0.0160
lr	Logistic Regression	0.8693	0.8302	0.2242	0.6039	0.3251	0.2691	0.3108	0.6600
lda	Linear Discriminant Analysis	0.8649	0.8256	0.2959	0.5417	0.3806	0.3122	0.3313	0.0160
ridge	Ridge Classifier	0.8645	0.0000	0.1073	0.6148	0.1813	0.1457	0.2144	0.0170
dummy	Dummy Classifier	0.8592	0.5000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0150
svm	SVM - Linear Kernel	0.7649	0.0000	0.2660	0.3129	0.1586	0.0966	0.1280	0.0180
qda	Quadratic Discriminant Analysis	0.5742	0.4841	0.4002	0.1942	0.1740	0.0057	0.0185	0.0170



Pada gambar disamping menunjukkan variabel X dan Y untuk kebutuhan training model

```
1 #Define X and Y
2 X = train.drop('churn', axis=1)
3 Y = train['churn']
```

```
1 outliers_check(X)
```

	Feature	Outliers (%)
0	account_length	0.47
1	international_plan	9.32
2	voice_mail_plan	0.00
3	number_vmail_messages	2.02
4	total_day_minutes	0.59
5	total_day_calls	0.66
6	total_day_charge	0.61
7	total_eve_minutes	0.80
8	total_eve_calls	0.56
9	total_eve_charge	0.80
10	total_night_minutes	0.87
11	total_night_calls	0.78
12	total_night_charge	0.87
13	total_intl_minutes	1.46
14	total_intl_calls	2.35
15	total_intl_charge	1.46
16	number_customer_service_calls	7.88

## Model development process

Dalam upaya kami memprediksi churn pelanggan bagi penyedia telekomunikasi, kami memerlukan cara untuk mengukur seberapa baik kinerja model kami.

Ada beberapa metrik yang tersedia untuk mengevaluasi model klasifikasi, masing-masing memiliki tujuan berbeda. Salah satu metrik yang umum adalah akurasi, yang mengukur proporsi instance yang diklasifikasikan dengan benar di semua instance. Namun, akurasi saja mungkin bukan pilihan terbaik untuk model prediksi churn kami.



## Model performance evaluation (confusion matrix)

Dalam eksplorasi model pengklasifikasi untuk memprediksi churn dalam layanan telekomunikasi, kami telah mengevaluasi berbagai opsi. Diantaranya, XGBoost menonjol sebagai pilihan optimal. XGBoost, kependekan dari Extreme Gradient Boosting, telah menunjukkan kinerja unggul dalam hal recall dan nilai F1 dibandingkan pengklasifikasi lain yang tersedia. Kekuatannya terletak pada kemampuannya menangani kumpulan data kompleks secara efektif dan mempelajari pola rumit, yang sangat penting untuk memprediksi churn pelanggan secara akurat.



### Classification Report for Training Data

	precision	recall	f1-score	support
0	0.97	0.98	0.97	1096
1	0.84	0.82	0.83	179
accuracy			0.95	1275
macro avg	0.91	0.90	0.90	1275
weighted avg	0.95	0.95	0.95	1275

### Confusion Matrix for Training Data

```
[[1069  27]
 [  33 146]]
```

### Kesimpulan

Dalam pemodelan ini XGBoosts menunjukkan hasil dan performa yang sangat bagus diantara model yang lain. Selain itu, interpretasi XGBoost adalah aset berharga dalam konteks prediksi churn.

Kemampuannya untuk memberikan wawasan tentang fitur mana yang paling berpengaruh dalam menentukan churn memungkinkan penyedia telekomunikasi mendapatkan pemahaman yang lebih mendalam tentang perilaku pelanggan dan menyesuaikan strategi retensi.

Secara keseluruhan, XGBoost muncul sebagai model ideal untuk prediksi churn, menawarkan kombinasi kuat antara kinerja tinggi, keserbagunaan, dan kemampuan interpretasi untuk secara efektif mengatasi tantangan retensi pelanggan di industri telekomunikasi.

# **Report Pembagian Pengerjaan Tugas**

Nama	Tasklist/Deliverable
Muhamad Shidqi	Paling banyak dalam pembuatan query dan program selama pelatihan ini dan juga terlibat dalam pembuatan visualisasi.
Mohammad Rifqi Nur Faroza	Paling banyak dalam pembuatan visualisasi data dan juga turut andil dalam pembuatan program code.

# Thank You