Rakamin Academy

peerData

Exploratory Data Analysis

Laporan Homework

Dataset

Deskripsi :
Deposito berjangka adalah investasi tunai yang disimpan di bank. Kampanye pemasaran melalui telepon menjadi salah satu cara efektif untuk menjangkau orang. Namun, mereka membutuhkan investasi besar karena pusat panggilan besar disewa untuk melaksanakan kampanye. Oleh karena itu, sangat penting untuk mengidentifikasi nasabah yang kemungkinan besar akan berkonversi terlebih dahulu sehingga mereka dapat ditargetkan secara khusus melalui panggilan.

Dataset :
Memprediksi pelanggan yang berpotensi untuk men-deposito uangnya (berlangganan) atau tidak (tidak berlangganan)

Data :
Setiap satu baris data mewakili satu nasabah bank, satu kolom berisi data dari nasabah tersebut

Dataset

Kolom	Deskripsi (Mewakili)
Age	Usia nasabah bank
Job	Pekerjaan nasabah bank
Marital	Status pernikahan nasabah bank
Education	Pendidikan terakhir nasabah bank
Default	Apakah nasabah mempunyai saldo di rekening ?
Balance	Saldo tahunan nasabah bank (Euro)
Housing	Apakah nasabah memiliki pinjaman rumah ?
Loan	Apakah nasabah memiliki pinjaman pribadi ?
Contact	Jenis komunikasi kontak

Dataset

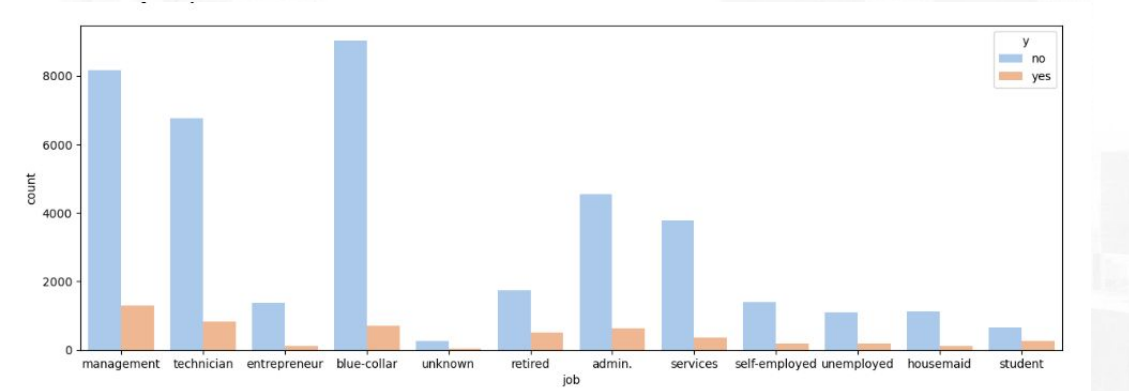
Kolom	Deskripsi (Mewakili)
Day	Hari kontak terakhir dalam sebulan
Month	Bulan kontak terakhir dalam setahun
Duration	Durasi kontak terakhir (detik)
Campaign	Jumlah kontak yang dilakukan selama kampanye untuk nasabah bank
Pdays	Jumlah hari yang berlalu setelah klien terakhir dihubungi dari kampanye sebelumnya (numerik, -1 berarti klien tidak dihubungi sebelumnya)
Previous	Jumlah kontak yang dilakukan sebelum kampanye untuk nasabah bank
Poutcome	Hasil dari kampanye pemasaran sebelumnya
Y	Apakah klien sudah pernah melakukan deposito di bank ?

STAGE 1 - *EDA, INSIGHT* & Visualisasi

- Eksplorasi Data
- *Exploratory Data Analysis*
- Insight Bisnis dan Visualisasi

EKSPLORASI DATA

JOB

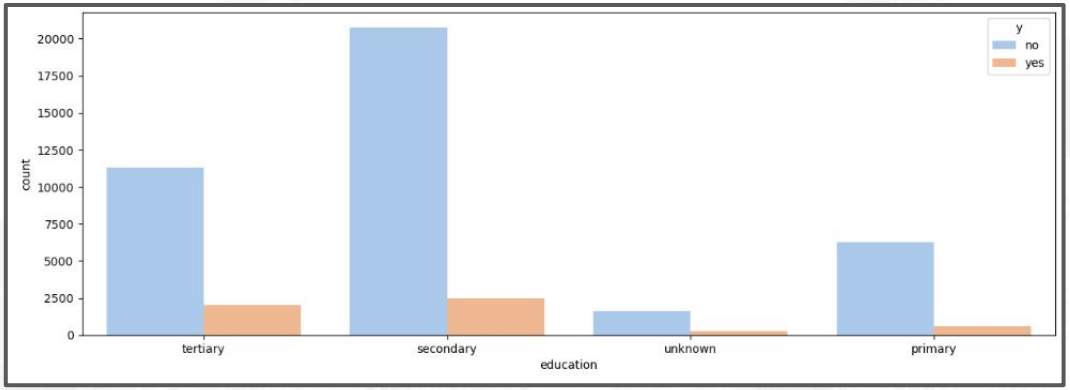


	index	Job	Do not deposit	Deposit	Percent who deposit
0	8	student	669	269	28.678038
1	5	retired	1748	516	22.791519
2	10	unemployed	1101	202	15.502686
3	4	management	8157	1301	13.755551
4	0	admin.	4540	631	12.202669
5	6	self-employed	1392	187	11.842939
6	11	unknown	254	34	11.805556
7	9	technician	6757	840	11.056996
8	7	services	3785	369	8.883004
9	3	housemaid	1131	109	8.790323
10	2	entrepreneur	1364	123	8.271688
11	1	blue-collar	9024	708	7.274969

Mayoritas jumlah nasabah yang tertarik untuk menyimpan di deposit secara jumlah adalah nasabah yang bekerja di **manajemen** dengan jumlah 1301 nasabah. Disusul dengan teknisi 840 dan blue-collar 708 (pekerjaan di bidang manufaktur, mining dsb). Hal ini mungkin disebabkan karena pekerja manajemen memiliki gaji yang cenderung tetap dan cukup besar untuk bisa diinvestasikan.

Namun, secara **conversion rate,** nasabah yang berkesibukan sebagai **pelajar** memiliki persentase ketertarikan **tertinggi**. Hal ini mungkin karena berlangganan deposito menjadi target yang ingin dilakukan saat pertama kali membuka rekening di bank (mungkin berpikir untuk investasi sejak dini). Ini juga berarti peluang untuk nasabah pelajar untuk convert adalah yang tertinggi. Untuk retired juga persentasenya kedua tertinggi (di atas 20%), hal ini mungkin nasabah menyimpan tabungan pensiunnya pada deposito berjangka.

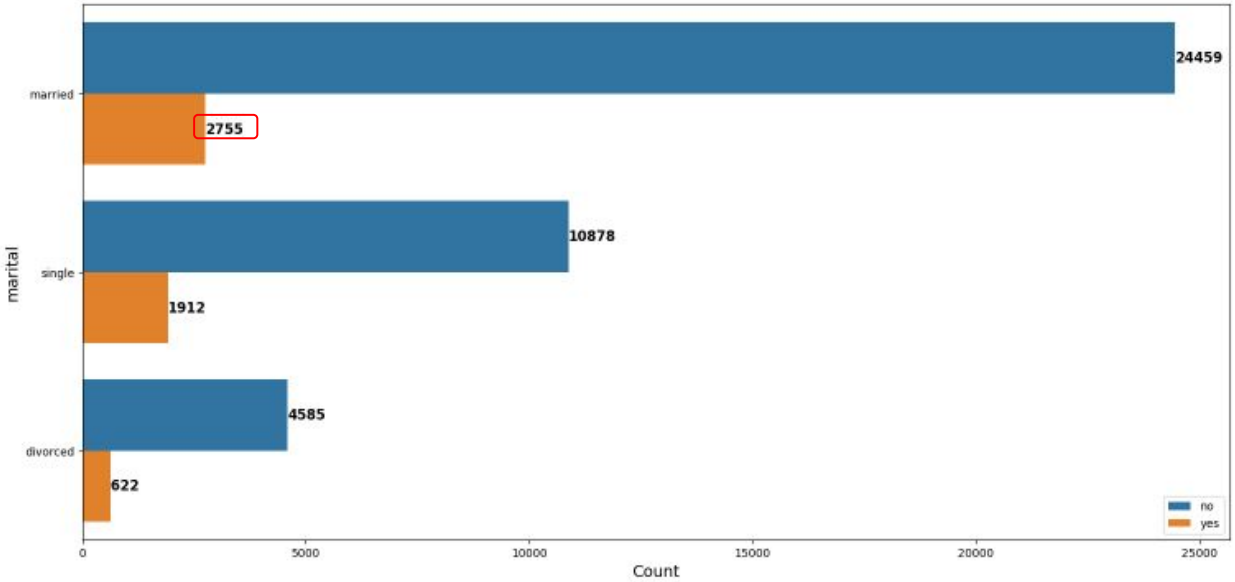
EDUCATION



	index	Education	Do not deposit	Deposit	Percent who deposit
0	2	tertiary	11305	1996	15.006390
1	3	unknown	1605	252	13.570275
2	1	secondary	20752	2450	10.559435
3	0	primary	6260	591	8.626478

Berdasarkan **jumlah yang berlangganan**, pendidikan **secondary** memiliki jumlah nasabah deposit terbanyak. Jika dilihat dari conversion ratenya, maka dapat disimpulkan bahwa **semakin tinggi pendidikan** nasabah, peluang **convertnya** juga semakin tinggi.

Pendidikan yang tinggi juga dapat **mencerminkan stabilitas keuangan** dan minat yang lebih besar dalam produk atau layanan yang ditawarkan, yang pada akhirnya dapat **meningkatkan kemungkinan** nasabah untuk **melakukan tindakan konversi.** Hal tersebut mungkin karena **semakin tinggi pendidikan formal**, **semakin tinggi pula iterasi keuangan seseorang.**

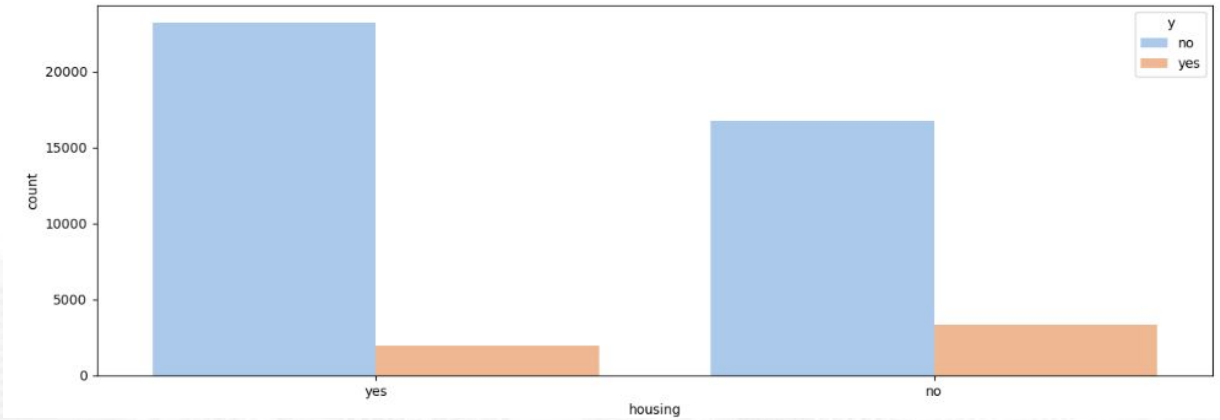


MARITAL STATUS

Marital Status	Percent who Deposit
Married	0.10
Single	0.14
Divorced	0.11

Jika dilihat dari marital status, nasabah dengan status **'menikah'** memiliki **jumlah** ketertarikan lebih tinggi dalam berlangganan deposito. Hal ini mungkin dikarenakan nasabah yang telah menikah membutuhkan akun deposito untuk menyimpan dana tabungan berdasarkan target yang telah mereka rencanakan. Namun, untuk **persentase/conversion rate**, status '**single'** memiliki persentase paling tinggi meskipun tidak begitu jauh dibanding yang lain, mungkin karena banyak dari nasabah bank memiliki spare uang lebih untuk bisa di investasikan. Namun, perbedaan conversion rate **tidak terlihat signifikan.**

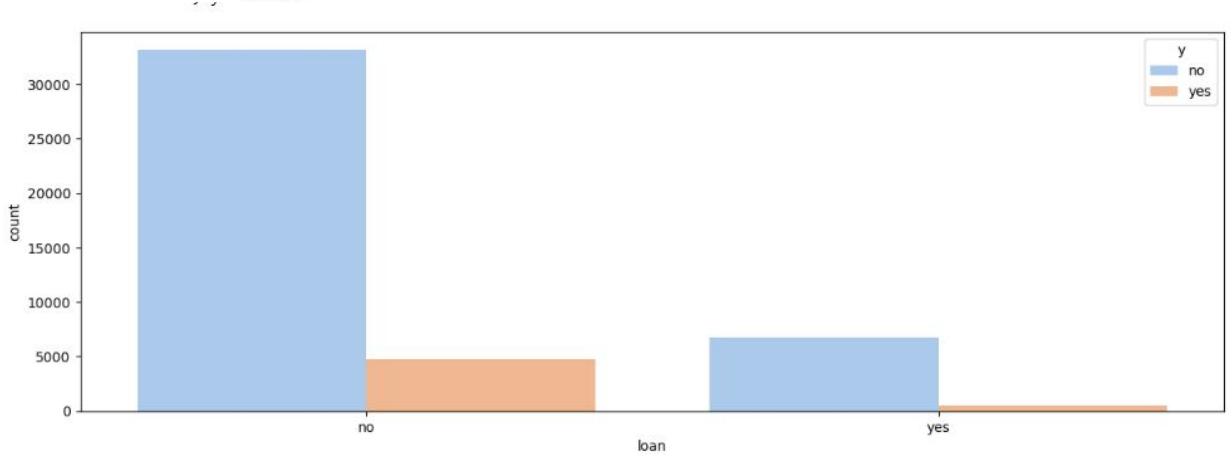
HOUSING



	index	Housing Loan	Do not deposit	Deposit	Percent who deposit
0	0	no	16727	3354	16.702355
1	1	yes	23195	1935	7.699960

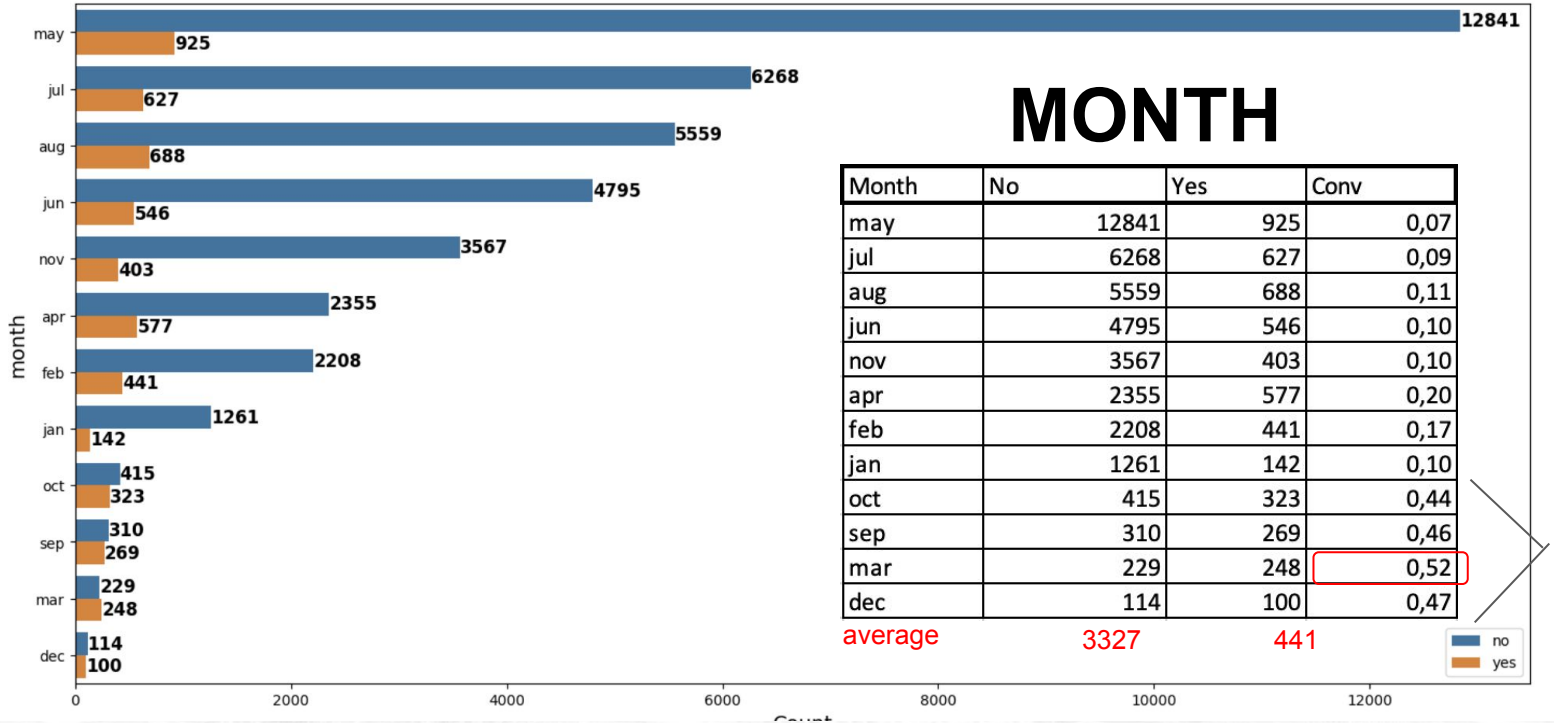
Nasabah yang **belum pernah mengajukan KPR** cenderung **berlangganan** produk deposit. Hal ini dapat disimpulkan dari **jumlah dan persentase** convertnya. Ini karena nasabah yang tidak mengajukan KPR cenderung memiliki spare dana untuk berlangganan deposito berjangka. Persentase konversi nasabah yang tidak memiliki KPR juga memiliki nilai **lebih dari 2 kali lipat** dari yang punya KPR.

LOAN



	index	Loan	Do not deposit	Deposit	Percent who deposit
0	0	no	33162	4805	12.655727
1	1	yes	6760	484	6.681391

Senada dengan faktor KPR, nasabah yang **belum pernah mengajukan pinjaman** juga lebih tinggi dari segi **jumlah** dan **persentase convertnya**. Hal ini juga karena spare dana lebih banyak jika tidak memiliki pinjaman lain. Perbedaan persentase yang tidak punya pinjaman juga hampir **2 kali lipat** nasabah yang memiliki pinjaman.



MONTH

Month	No	Yes	Conv
may	12841	925	0,07
jul	6268	627	0,09
aug	5559	688	0,11
jun	4795	546	0,10
nov	3567	403	0,10
apr	2355	577	0,20
feb	2208	441	0,17
jan	1261	142	0,10
oct	415	323	0,44
sep	310	269	0,46
mar	229	248	0,52
dec	114	100	0,47
average	3327	441	

average:
267 no
235 yes
0.47 conv

- Rata-rata nasabah yang dikontak setiap bulan memiliki tingkat konversi yang rendah atau tidak tertarik dengan deposito
- Nasabah yang **tidak tertarik** paling banyak saat dikontak pada bulan Mei
- **Conversion rate tertinggi** ada pada bulan **Maret**, untuk bulan Desember, September dan Oktober juga cenderung tinggi.
- Secara rata-rata setiap bulan, ada **441** nasabah yang berlangganan dan **3327** menolak berlangganan.
- Secara **rata-rata 4 bulan** dengan CR tertinggi, sebanyak **235 berlangganan** dan **267 menolak**, dan rata-rata CR nya sebesar **0,47**.

EXPLORATORY DATA ANALYSIS

Descriptive Statistics

Sekilas tentang dataset :
Terdapat 17 kolom (16 kolom fitur dan 1 kolom target)

```
df = pd.read_csv('banking_dataset_train.csv', delimiter = ";")
df.sample(5)
```

	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	y
23547	32	management	married	tertiary	no	0	no	no	cellular	28	aug	15	13	-1	0	unknown	no
16662	40	blue-collar	married	secondary	no	3131	yes	no	cellular	24	jul	401	1	-1	0	unknown	no
11145	48	management	single	tertiary	no	0	no	no	unknown	18	jun	96	3	-1	0	unknown	no
24101	31	admin.	married	secondary	no	352	no	no	telephone	28	oct	60	1	-1	0	unknown	no
2632	52	admin.	divorced	secondary	no	26	yes	no	unknown	13	may	215	1	-1	0	unknown	no

Descriptive Statistics

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 45211 entries, 0 to 45210
Data columns (total 17 columns):
 #   Column      Non-Null Count  Dtype 
---  ------      --------------  ----- 
 0   age         45211 non-null  int64 
 1   job         45211 non-null  object
 2   marital     45211 non-null  object
 3   education   45211 non-null  object
 4   default     45211 non-null  object
 5   balance     45211 non-null  int64 
 6   housing     45211 non-null  object
 7   loan        45211 non-null  object
 8   contact     45211 non-null  object
 9   day         45211 non-null  int64 
 10  month       45211 non-null  object
 11  duration    45211 non-null  int64 
 12  campaign    45211 non-null  int64 
 13  pdays       45211 non-null  int64 
 14  previous    45211 non-null  int64 
 15  poutcome    45211 non-null  object
 16  y           45211 non-null  object
dtypes: int64(7), object(10)
memory usage: 5.9+ MB
```

Terdapat 2 jenis tipe data, yaitu **int64** dan **object**. Semua tipe data sudah sesuai, terdapat 7 data numerik dengan tipe data "int64" dan sisanya kategorikal dengan tipe data "object". Semua tipe data **sudah sesuai** dengan kolom **fitur.**

Kolom **categorical** dengan tipe data **object**, dan kolom **numerical** dengan tipe data **int64**, karena kolom **numerical mempunyai nilai yang bulat.**

Descriptive Statistics

```
df.isna().sum()/len(df)*100
#dalam bentuk percentage
```

```
age          0.0
job          0.0
marital      0.0
education    0.0
default      0.0
balance      0.0
housing      0.0
loan         0.0
contact      0.0
day          0.0
month        0.0
duration     0.0
campaign     0.0
pdays        0.0
previous     0.0
poutcome     0.0
y            0.0
dtype: float64
```

Berdasarkan df.isna( ) diketahui bahwa **tidak ada data yang kosong** pada dataset yang digunakan, sehingga nilai persentase missing value terhadap keseluruhan data adalah 0

Descriptive Statistics

```
df[nums].describe()
```

	age	balance	day	duration	campaign	pdays	previous
count	45211.000000	45211.000000	45211.000000	45211.000000	45211.000000	45211.000000	45211.000000
mean	40.936210	1362.272058	15.806419	258.163080	2.763841	40.197828	0.580323
std	10.618762	3044.765829	8.322476	257.527812	3.098021	100.128746	2.303441
min	18.000000	-8019.000000	1.000000	0.000000	1.000000	-1.000000	0.000000
25%	33.000000	72.000000	8.000000	103.000000	1.000000	-1.000000	0.000000
50%	39.000000	448.000000	16.000000	180.000000	2.000000	-1.000000	0.000000
75%	48.000000	1428.000000	21.000000	319.000000	3.000000	-1.000000	0.000000
max	95.000000	102127.000000	31.000000	4918.000000	63.000000	871.000000	275.000000

- Berdasarkan hasil perhitungan statistika, terdapat **perbedaan angka antara mean dan median** pada **kolom balance, duration, pdays, campaign, previous.** Nilai mean **lebih besar** dari median-nya, mengindikasikan bahwa **grafik distribusi frekuensi menceng kanan atau kemencengan positif.**
- Dalam hal nilai minimum, kolom 'balance' memiliki nilai negatif (-8019), yang mungkin tidak sesuai untuk saldo rekening bank.
- Kolom 'duration' memiliki nilai minimum 0, yang mungkin tidak sesuai untuk durasi panggilan. Hal ini mungkin menunjukkan panggilan yang terlewat atau masalah lain.
- Kolom pdays memiliki nilai minimum -1, nilai tersebut merupakan representasi nasabah yang belum pernah dihubungi di campaign sebelumnya.

Descriptive Statistics

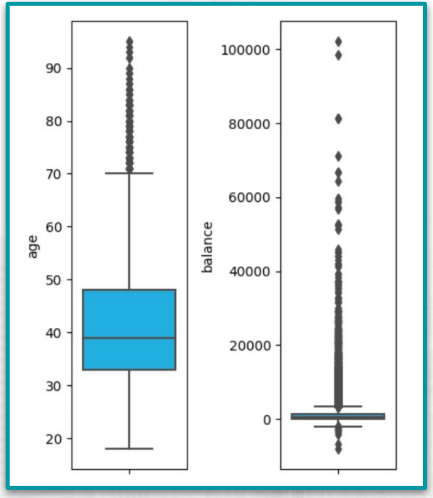
```
df[cats].describe()
```

	job	marital	education	default	housing	loan	contact	month	poutcome	y
count	45211	45211	45211	45211	45211	45211	45211	45211	45211	45211
unique	12	3	4	2	2	2	3	12	4	2
top	blue-collar	married	secondary	no	yes	no	cellular	may	unknown	no
freq	9732	27214	23202	44396	25130	37967	29285	13766	36959	39922

- Berdasarkan observasi kolom-kolom categorical, kebanyakan dari nasabah bank adalah orang yang **memiliki pekerjaan *"blue-collar"*** atau para pekerja kasar, yang **sudah menikah dengan pendidikan menengah.**
- Para nasabah tersebut sebagian besar **memiliki pinjaman rumah,** yang dapat **dihubungi lewat telepon seluler.** Namun, sebagian besar dari para nasabah tersebut **tidak mendepositkan uang mereka pada bank sebelumnya.**
- Dalam hal unique data, **tidak ada data yang terlalu beragam**.
- Dalam hal frequency, variabel 'default' memiliki jumlah nilai "no" yang terlalu banyak, hal ini juga terjadi pada variabel 'month', 'poutcome' dan 'y' yang cukup ada ketimpangan data.

Univariate Analysis

Boxplot - Numerical Columns



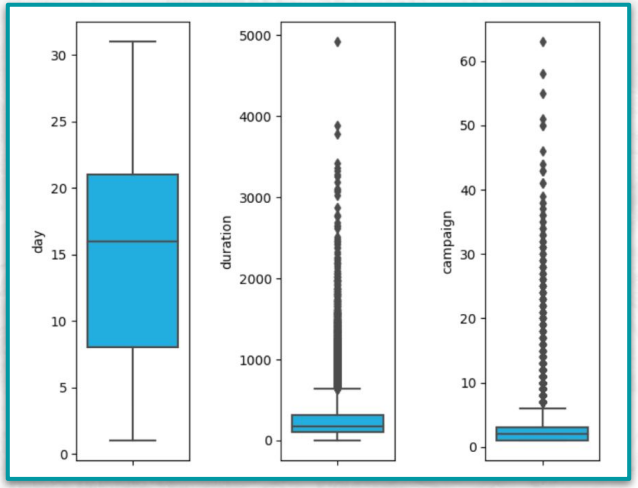
"age" :
- Rentang usia responden antara 18-95 tahun.
- Median usia responden adalah sekitar 39 tahun.
- 50% dari keseluruhan data terpusat pada usia antara 33 hingga 48 tahun.
- Distribusi usia cenderung normal tanpa adanya outlier.

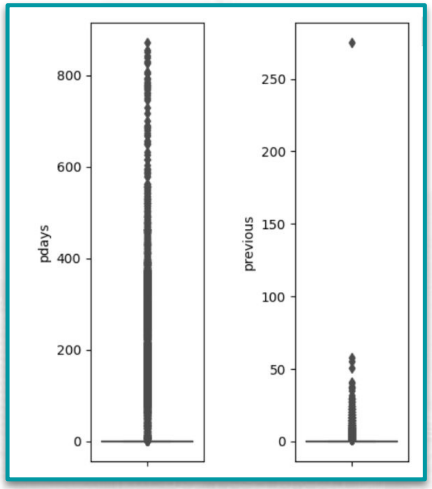
"balance" :
- Saldo rekening responden memiliki variasi yang sangat besar, dengan rentang dari -8019 hingga 102127.
- Sebagian besar responden memiliki saldo di bawah 1428.
- Terdapat outlier ekstrim pada sisi atas distribusi, menunjukkan adanya responden dengan saldo rekening yang sangat tinggi.

"day":
- Mayoritas hari terakhir responden dikontak antara hari ke 8 hingga ke 21. Secara umum, distribusi data cenderung normal tanpa terlihat outliers.

"duration":
- Durasi panggilan memiliki variasi yang cukup besar, dengan rentang dari 0 hingga 4900 detik.
- Median durasi panggilan adalah sekitar 180 detik.
- Adanya outlier di bagian atas boxplot yang disertai dengan whisker bagian atas yang lebih panjang, menunjukkan bahwa distribusi data menjulur ke arah kanan (positive skewness).

"campaign":
- Keseluruhan data terpusat pada jumlah campaign antara 0 hingga 8.
- Terdapat banyak outlier pada sisi atas distribusi, menunjukkan adanya beberapa responden dengan jumlah kontak yang sangat tinggi.



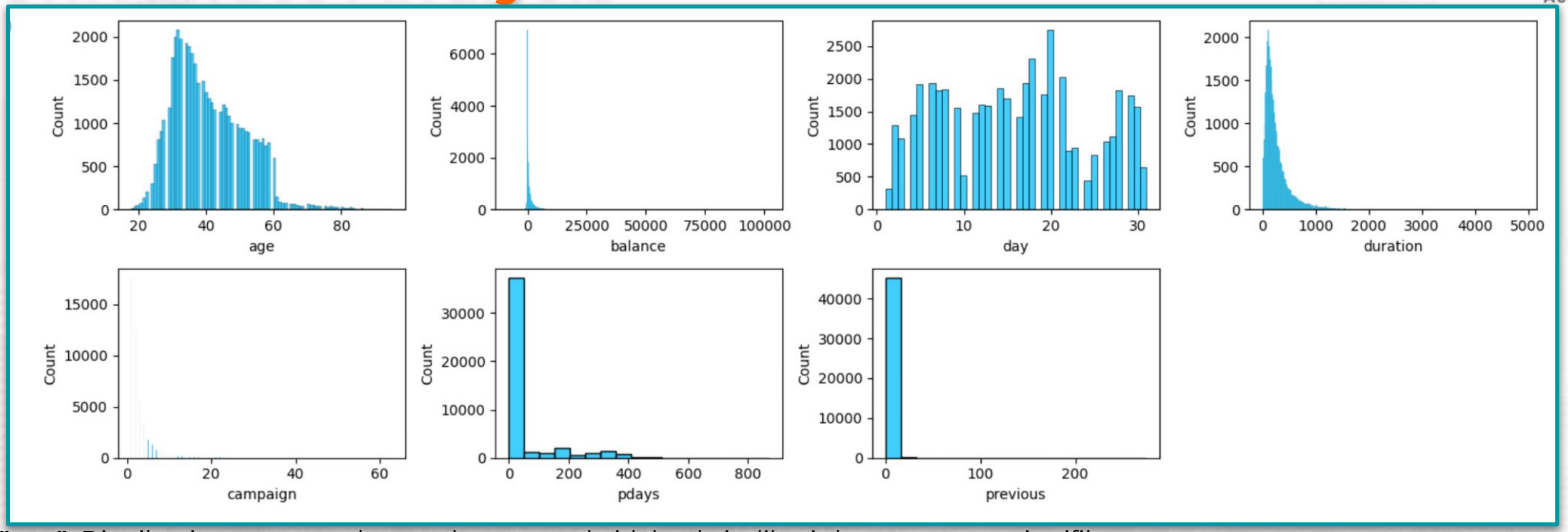


“pdays":
- Median jumlah hari adalah -1, yang mungkin menunjukkan bahwa mayoritas responden belum pernah dikontak sebelumnya.
- Terdapat banyak nilai ekstrem dan outlier pada sisi atas distribusi

"previous":
- Mayoritas responden tidak memiliki kontak sebelumnya sebelum kampanye saat ini, seperti yang ditunjukkan oleh nilai median 0 .
- Terdapat beberapa outlier dan nilai ekstrem yang sangat tinggi(275) pada sisi atas distribusi.

Univariate Analysis

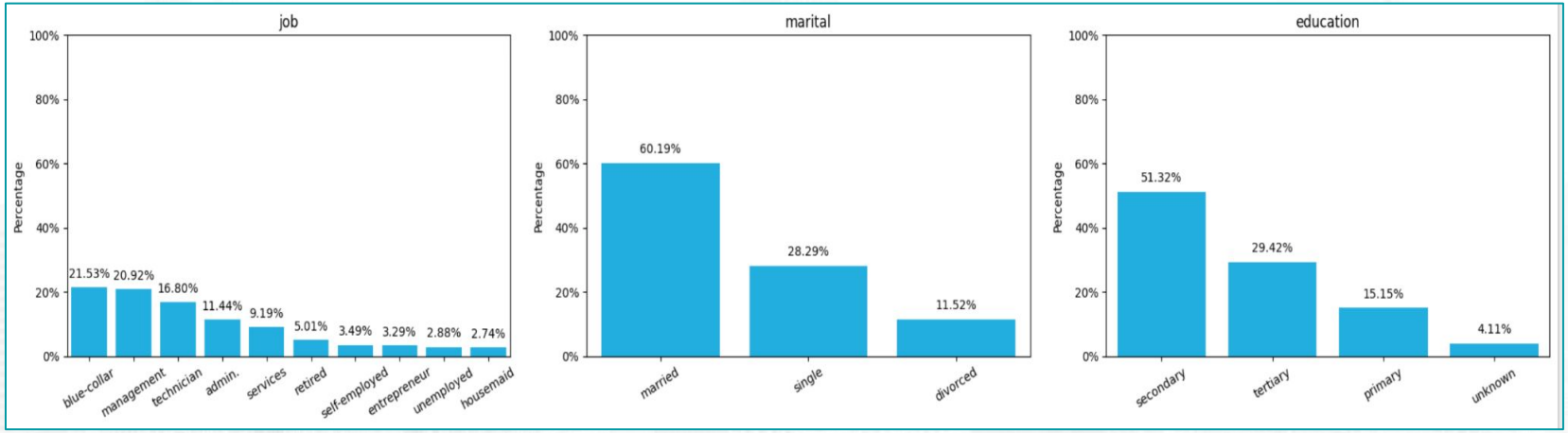
Histplot - Numerical Columns



- **"age":** Distribusi umur tampaknya cukup normal, tidak ada indikasi skewness yang signifikan.
- **"balance":** Distribusi saldo tampaknya sangat skew ke kanan (positively skewed). Perlu dilakukan penghapusan outlier yang berada di luar kisaran nilai yang masuk akal dengan menentukan batasan atas dan bawah atau melakukan transformasi data (log transformation)
- **"day":** Distribusi kolom ini tidak menunjukkan karakteristik yang mencolok.
- **"duration":** Distribusi durasi panggilan juga sangat skew ke kanan. Perlu dilakukan penanganan outlier dengan transformasi data (misalnya log transform) atau penggunaan teknik penggantian outlier (misalnya menggunakan batas atas atau bawah yang relevan).
- **"campaign":** Distribusi jumlah panggilan kampanye cenderung positively skewed, dengan sebagian besar nasabah menerima panggilan dalam jumlah yang sedikit. Terdapat nilai maksimum yang jauh lebih tinggi dari nilai-nilai lainnya, menunjukkan adanya beberapa nasabah yang menerima panggilan kampanye dalam jumlah yang sangat banyak.
- **"pdays":** Distribusi nilai pdays sangat skew ke kanan, dengan sebagian besar nilai berada pada -1 (non-called). Saat melakukan pra-pemrosesan data, nilai -1 dapat diganti dengan nilai yang lebih bermakna seperti NaN untuk menandai klien yang tidak pernah dihubungi sebelumnya.
- **"previous":** Distribusi jumlah kontak sebelum kampanye saat ini juga sangat skew ke kanan. Perlu dilakukan penghapusan

Univariate Analysis

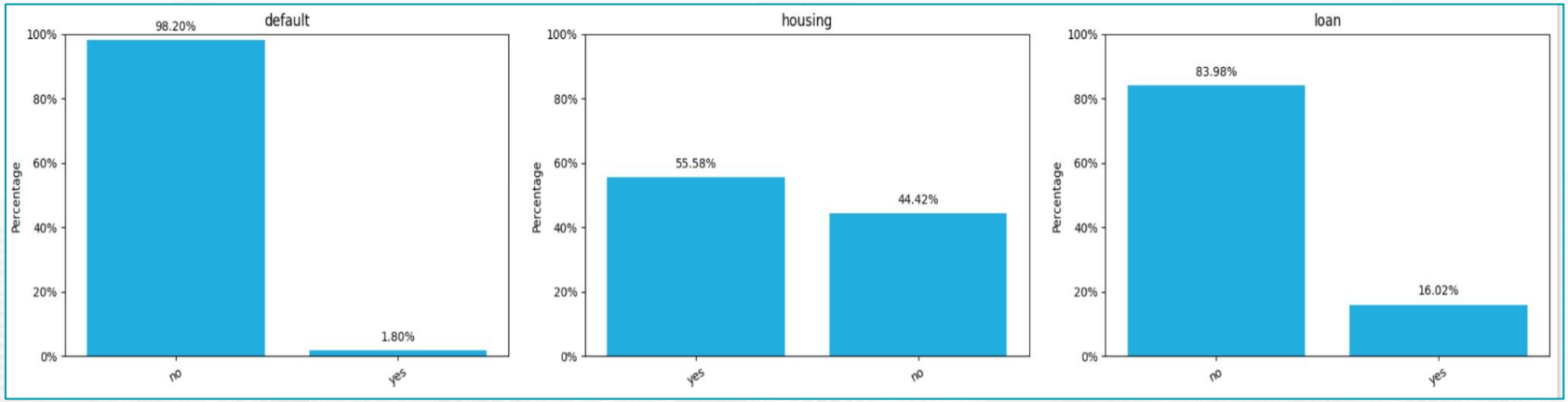
Barplot - Categorical Columns



- Variabel "job":
  - **Blue-collar , management,** dan **technician** adalah tiga pekerjaan paling **umum** dalam dataset ini.
  - **Housemaid , unemployed , entrepreneur ,**dan **self-employed** adalah pekerjaan yang paling **jarang** ditemui dalam dataset ini.
- Variabel "marital":
  - Mayoritas responden (60%) dalam dataset ini adalah yang sudah **menikah**.
- Variabel "education":
  - Sekitar **setengah** dari responden memiliki pendidikan tingkat menengah (**secondary**)

Univariate Analysis

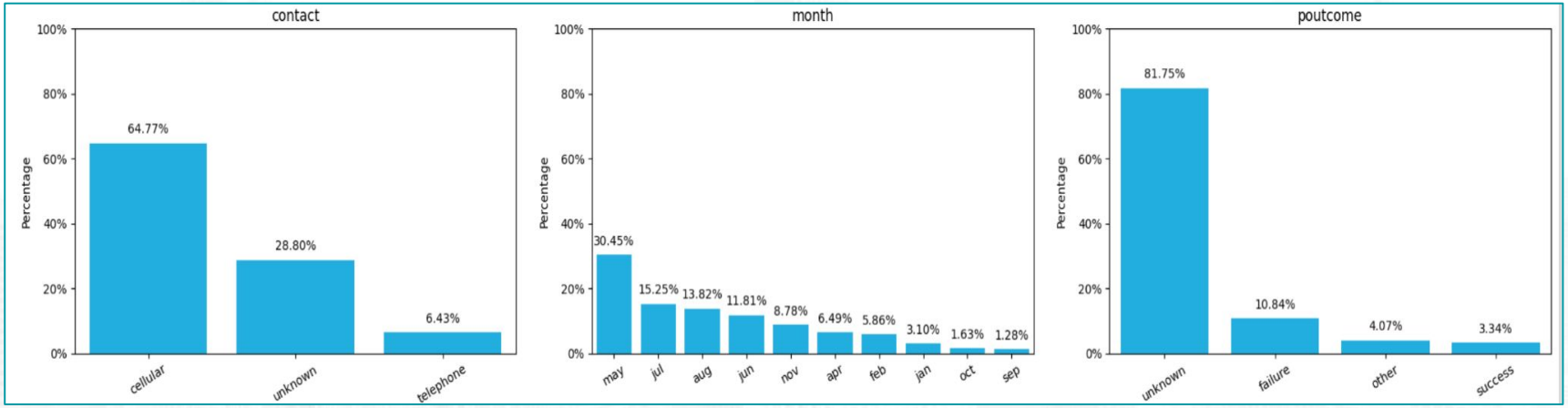
Barplot - Categorical Columns



- Variabel "default":
  - Mayoritas responden (98%) **tidak memiliki masalah default** pada pinjaman atau kredit.
- Variabel "housing":
  - Lebih dari setengah responden (55%) memiliki **kepemilikan rumah (housing).**
- Variabel "loan":
  - Mayoritas responden (83%) **tidak memiliki pinjaman.**

Univariate Analysis

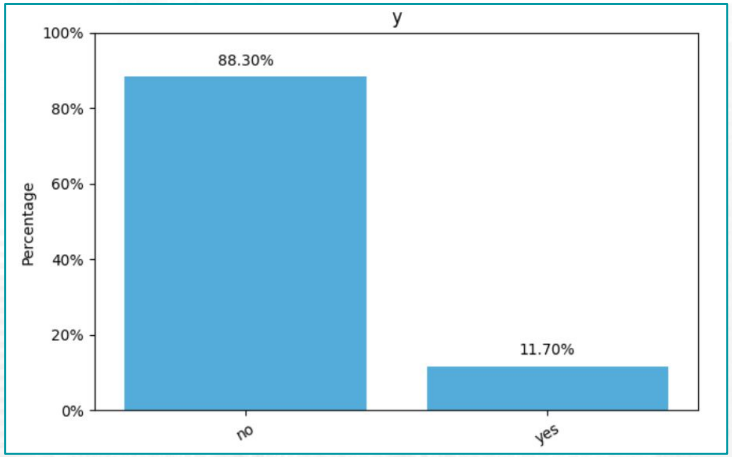
Barplot - Categorical Columns

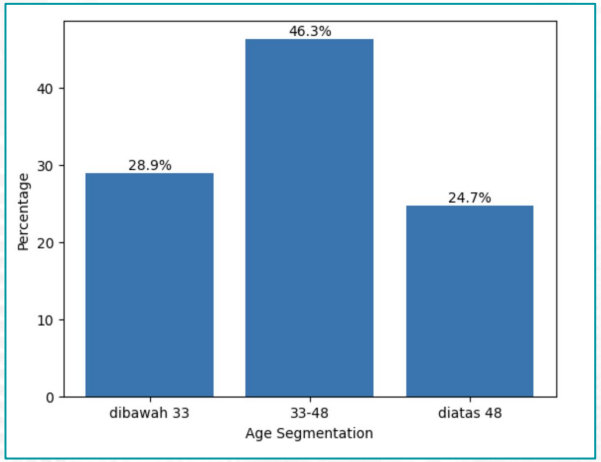


- Variabel "contact":
  - Sebagian besar komunikasi (64%) dilakukan melalui **telepon seluler (cellular)**
- Variabel "month":
  - Bulan terbanyak dalam dataset ini adalah **Mei (30%) ,**diikuti oleh bulan **juli ,Agustus** dan **Juni**
- Variabel "poutcome":
  - Sebagian besar responden(80%) memiliki hasil pemasaran sebelumnya yang tidak diketahui **(unknown).**

Univariate Analysis

Barplot - Categorical Columns





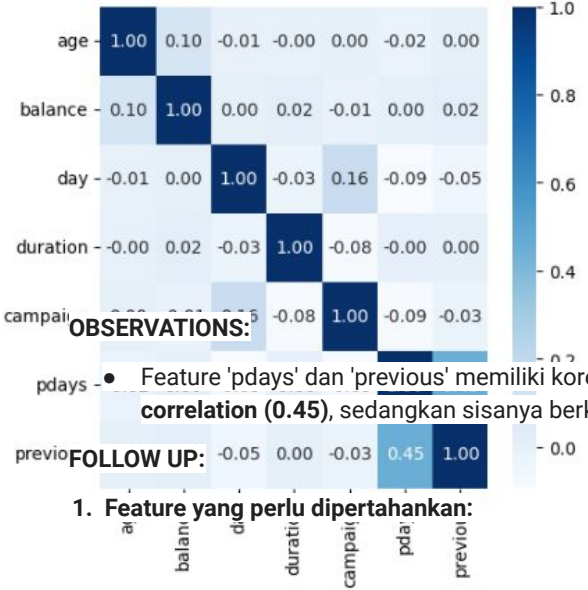
- Variabel "y":
  - Mayoritas (88%) responden **tidak berlangganan produk** atau layanan yang ditawarkan.
- Variabel "age segmentation":
  - Mayoritas (46%) responden berada di rentang **umur 33-48 tahun**

Multivariate Analysis

Heatmap - Numerical Columns

Rakamin Academy

	age	balance	day	duration	campaign	pdays	previous
age	1.000000	0.097783	-0.009120	-0.004648	0.004760	-0.023758	0.001288
balance	0.097783	1.000000	0.004503	0.021560	-0.014578	0.003435	0.016674
day	-0.009120	0.004503	1.000000	-0.030206	0.162490	-0.093044	-0.051710
duration	-0.004648	0.021560	-0.030206	1.000000	-0.084570	-0.001565	0.001203
campaign	0.004760	-0.014578	0.162490	-0.084570	1.000000	-0.088628	-0.032855
pdays	-0.023758	0.003435	-0.093044	-0.001565	-0.088628	1.000000	0.454820
previous	0.001288	0.016674	-0.051710	0.001203	-0.032855	0.454820	1.000000



OBSERVATIONS:

- Feature 'pdays' dan 'previous' memiliki korelasi yang masuk dalam kategori **moderate correlation (0.45)**, sedangkan sisanya berkorelasi lemah atau sangat lemah.

FOLLOW UP:

1. **Feature yang perlu dipertahankan:**
   - Previous : Memuat informasi mengenai jumlah kontak yang dihubungi pada campaign sebelumnya.
2. Feature yang perlu d

OBSERVATIONS:

- Feature 'pdays' dan 'previous' memiliki korelasi yang masuk
  correlation (0.45), sedangkan sisanya berkorelasi lemah at

FOLLOW UP:

1. **Feature yang perlu dipertahankan:**
   - Feature **'pdays'** dipertimbangkan untuk **tidak dimasukkan** d feature **'previous' dipertahankan**. Hal ini karena dari sisi per memiliki banyak **outliers** yang lebih ekstrim, memiliki nilai **n** relevan untuk merepresentasikan nasabah yang belum pern memiliki **standar deviasi** yang sangat besar ketimbang 'prev
   - Feature **'duration'** juga **dipertimbangkan untuk tidak dimasu** dikarenakan banyaknya nasabah yang belum pernah dihubu sebelumnya sehingga **banyak nilai 0**, namun ini butuh **peny**

3. Business Recommendation

- Berdasarkan analisis hubungan antara nasabah yang berlangganan dengan pinjaman dan KPR, bahwa marketer diharapkan bisa langsung **menghubungi** "**hanya**" nasabah yang **tidak memiliki KPR** dan **tidak memiliki Pinjaman**, karena conversion rate nya cukup **signifikan** lebih tinggi ketimbang yang memiliki keduanya.
- Mengingat tingkat conversion rate **pelajar** yang **tertinggi**, bank dapat mengembangkan program atau produk yang ditujukan secara khusus kepada mereka. Misalnya, program tabungan pendidikan atau deposito berjangka khusus untuk pelajar dengan manfaat dan suku bunga yang menarik. Dalam melakukan pemasaran, bank juga dapat bekerja sama dengan institusi pendidikan atau mengadakan kegiatan yang relevan dengan kebutuhan dan minat pelajar.
- Marketer bisa mencoba memaksimalkan untuk menghubungi nasabah pada bulan **Maret, September, Oktober** dan **Desember**, karena memiliki **conversion rate** yang secara **signifikan** lebih tinggi ketimbang bulan lainnya.
- Dari ke 4 bulan yang direkomendasikan di atas, dengan jumlah rata-rata **conversion rate 47%,** dibandingkan dengan total percobaan sebanyak **42.511** sepanjang tahun, maka marketer disarankan menghubungi **maksimal** hanya sebanyak **21.356 nasabah** (47% dari total percobaan) saja dan fokus di 4 bulan tersebut.
- Marketer bisa mencoba memaksimalkan untuk menghubungi dan menawarkan layanan eksklusif, seperti penawaran suku bunga yang lebih baik bagi deposito jangka panjang nasabah dengan status pendidikan yang tinggi (**diatas primary education**) untuk lebih memaksimalkan conversion rate nasabah.