

UNO

Laporan Final Project Bank UNO Marketing Targets

Ketua : Daniel Rowin

Anggota :

Febby Maghfirani Aziz

Amodya Subagio

Nur Rahman Shalahudin

Fakhry Abdurrohman

Daviro Yota Nagasan Wahyudi

Putri Vina Fajriyani

Asmiyeni Islamiati



UNO

Laporan Final Project
Bank UNO Marketing Targets

Business Understanding

Final Project - Stage 0



01 Problem Statement

Jumlah nasabah yang membuka deposito berjangka pada UNO Bank dari total 45.211 nasabah yang ada, hanya sekitar 5.289 nasabah atau 11,7 % nasabah saja yang membuka deposito berjangka.

Permintaan dari manajemen UNO Bank itu minimal sebesar 15% dari total nasabah UNO bank yang membuka deposito berjangka.

Diantaranya terdapat 4.369 nasabah atau 9.66% nasabah yang dihubungi melalui telepon cellular dan 390 nasabah atau 0.86% nasabah yang dihubungi melalui telepon rumah.

02 Role

Sebagai tim data scientist dalam suatu perusahaan bernama UNO Bank, kami diminta oleh manajemen UNO Bank untuk memprediksi apakah nasabah akan berlangganan deposito berjangka berdasarkan data yang tersedia guna meningkatkan performa dari perusahaan tersebut.

03 Goal

Meningkatnya jumlah nasabah yang membuka deposito berjangka menjadi sebesar 15%. Dimana ini berdasarkan permintaan manajemen UNO Bank.

04 Objective

1. Menganalisis profil nasabah berdasarkan variabel-variabel seperti usia, pekerjaan, status perkawinan, pendidikan, dan fitur lainnya sehingga nasabah tertarik untuk membuka deposito berjangka di UNO Bank.
2. Membangun model yang dapat memprediksi dengan akurat apakah seorang nasabah akan berlangganan deposito berjangka setelah kampanye pemasaran telepon dilakukan berdasarkan klasifikasi nasabahnya.

05 Business Metrics

Conversion Rate : Persentase nasabah UNO Bank yang membuka deposito berjangka

UNO

Laporan Final Project
Bank UNO Marketing Targets

Exploratory Data Analysis

Final Project - Stage 1



1. Descriptive Statistics (15 poin)

Gunakan function `info` dan `describe` pada dataset final project kalian. Tuliskan hasil observasinya, seperti:

- A. Apakah ada kolom dengan tipe data kurang sesuai, atau nama kolom dan isinya kurang sesuai?
- B. Apakah ada kolom yang memiliki nilai kosong? Jika ada, apa saja?
- C. Apakah ada kolom yang memiliki nilai summary agak aneh?
(min/mean/median/max/unique/top/freq)

* Untuk masing-masing jenis observasi, tuliskan juga jika tidak ada masalah, misal untuk A: "Semua tipe data sudah sesuai"

01. 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

```

```
df['duration'].describe()
```

```
count    45211.000000
mean     258.163080
std      257.527812
min      0.000000
25%     103.000000
50%     180.000000
75%     319.000000
max     4918.000000
Name: duration, dtype: float64
```

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

	age	balance	campaign	pdays	previous
count	45211.000000	45211.000000	45211.000000	45211.000000	45211.000000
mean	40.936210	1362.272058	2.763841	40.197828	0.580323
std	10.618762	3044.765829	3.098021	100.128746	2.303447
min	18.000000	-8019.000000	1.000000	-1.000000	0.000000
25%	33.000000	72.000000	1.000000	-1.000000	0.000000
50%	39.000000	448.000000	2.000000	-1.000000	0.000000
75%	48.000000	1428.000000	3.000000	-1.000000	0.000000
max	95.000000	102127.000000	63.000000	871.000000	275.000000

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

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

```
df.isna().sum()
```

```
age  
job  
marital  
education  
default  
balance  
housing  
loan  
contact  
day  
month  
duration  
campaign  
pdays  
previous  
poutcome  
y  
dtype: int64
```

01. Descriptive Statistics

A. Apakah ada kolom dengan tipe data kurang sesuai, atau nama kolom dan isinya kurang sesuai?

Sebaiknya dtype untuk kolom day adalah object karena menunjukkan tanggal.

B. Apakah ada kolom yang memiliki nilai kosong? Jika ada, apa saja?

Semua kolom tidak memiliki nilai kosong, karena nilai kosong pada dataset sudah dikonversikan menjadi unknown, oleh karena itu terdapat beberapa feature yang memiliki value unknown.

C. Apakah ada kolom yang memiliki nilai summary agak aneh?

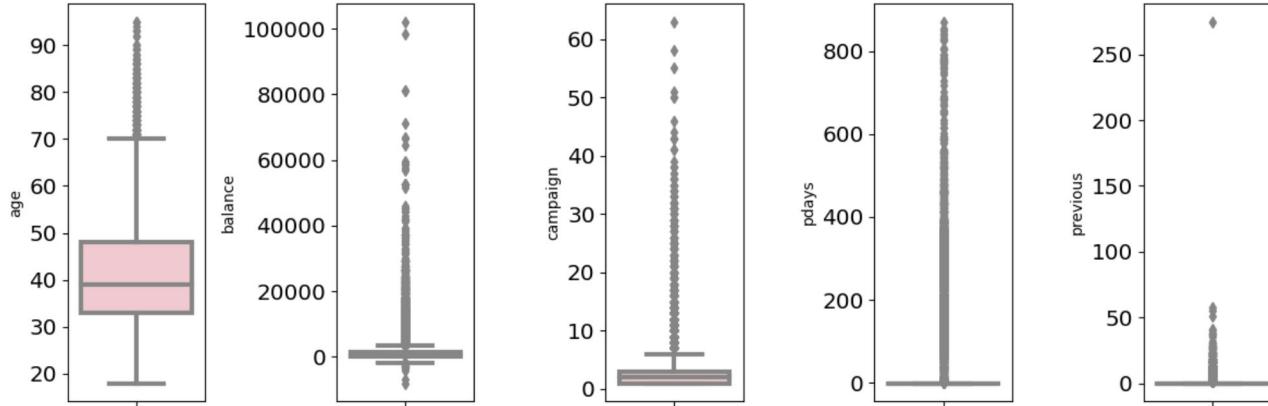
(min/mean/median/max/unique/top/freq)

- Balance terendah adalah -8019, mungkin diperlukan pemeriksaan lanjutan.***
- Terdapat contact dan outcome yang unknown***
- Nilai minimum pada kolom "duration" adalah 0***

2. Univariate Analysis (25 poin)

Gunakan visualisasi untuk melihat distribusi masing-masing kolom (feature maupun target). Tuliskan hasil observasinya, misalnya jika ada suatu kolom yang distribusinya menarik (misal skewed, bimodal, ada outlier, ada nilai yang mendominasi, kategorinya terlalu banyak, dsb). Jelaskan juga apa yang harus di-follow up saat data pre-processing.

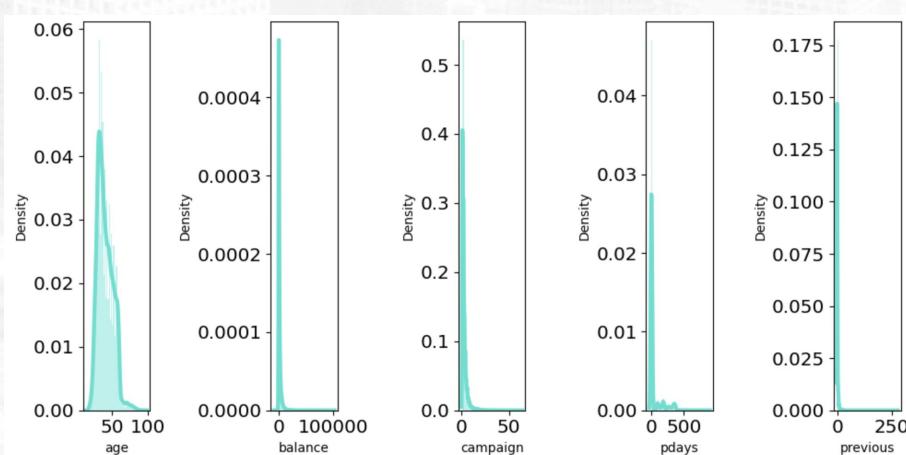
02. Univariate Analysis



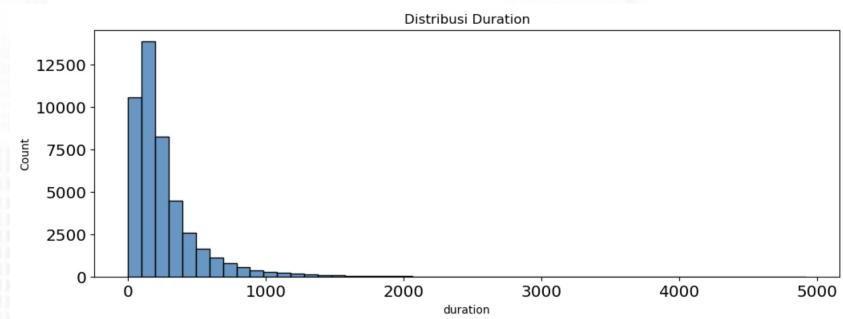
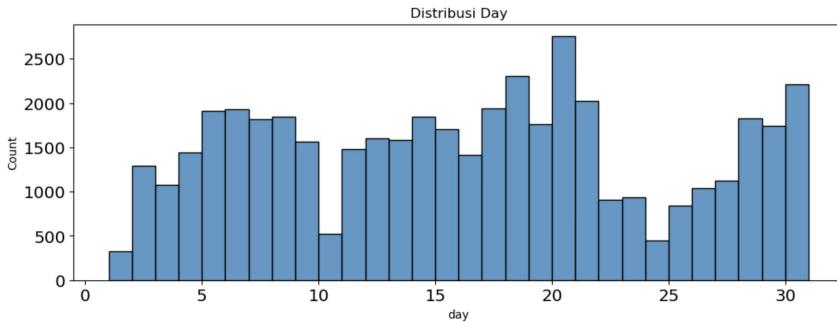
- Terdapat banyak outliers dan skewed pada feature balance, campaign, pdays, dan previous. Menangani nilai outlier pada kolom balance, campaign, dan pdays jika diperlukan.
- Feature age sudah hampir mendekati distribusi normal.

Follow-up untuk Data Pre-processing:

- Menangani nilai outlier pada feature balance, campaign, pdays dan previous jika diperlukan.
- Mungkin perlu melakukan normalisasi atau transformasi pada kolom yang memiliki skewness signifikan untuk mendapatkan distribusi yang lebih normal, terutama pada kolom balance.

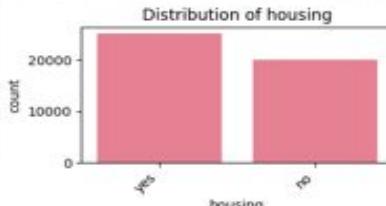
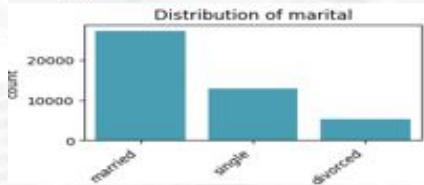
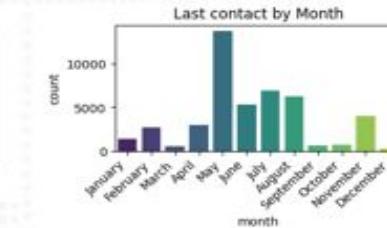
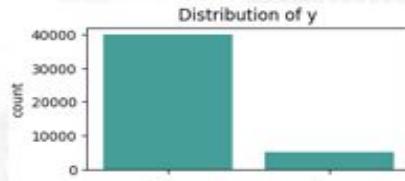
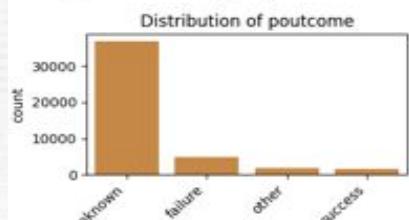
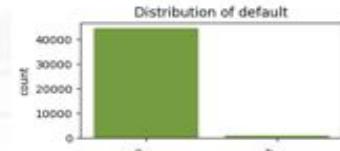
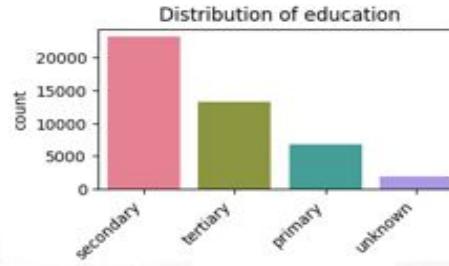
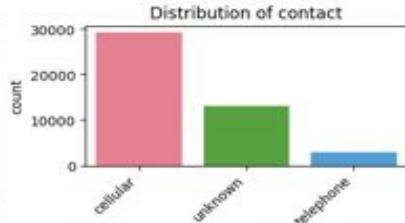
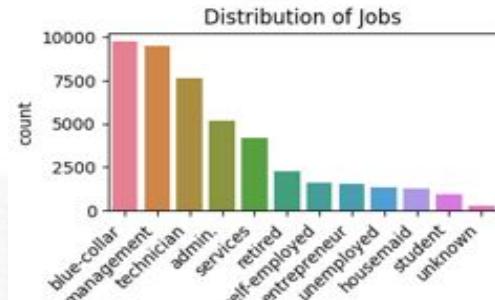


02. Univariate Analysis



- Kebanyakan data menunjukkan last contact day of the month adalah tanggal 20-21.
- Last contact duration kebanyak berlangsung kurang dari 1000 detik.

02. Univariate Analysis



- "Blue-collar", "management", "technician", "admin", dan "services" memiliki distribusi data dengan jumlah terbanyak
- Kebanyakan dari data menunjukkan status "married"
- "Secondary" education menunjukkan jumlah hasil terbanyak
- Kebanyakan dari data menunjukkan "no" default
- Kebanyakan nasabah dihubungi via cellular dan tidak memiliki loan
- Last contact month terbanyak adalah bulan May
- Untuk kolom job, pendidikan (education), dan kontak (contact) yang memiliki nilai "unknown", jika dianalogi perlu, pada tahap pre-prosesing kemungkinan bisa diganti dengan nilai modusnya.
- Kebanyakan outcome marketing campaign (poutcome) memiliki nilai "unknown".

3. Multivariate Analysis (15 poin)

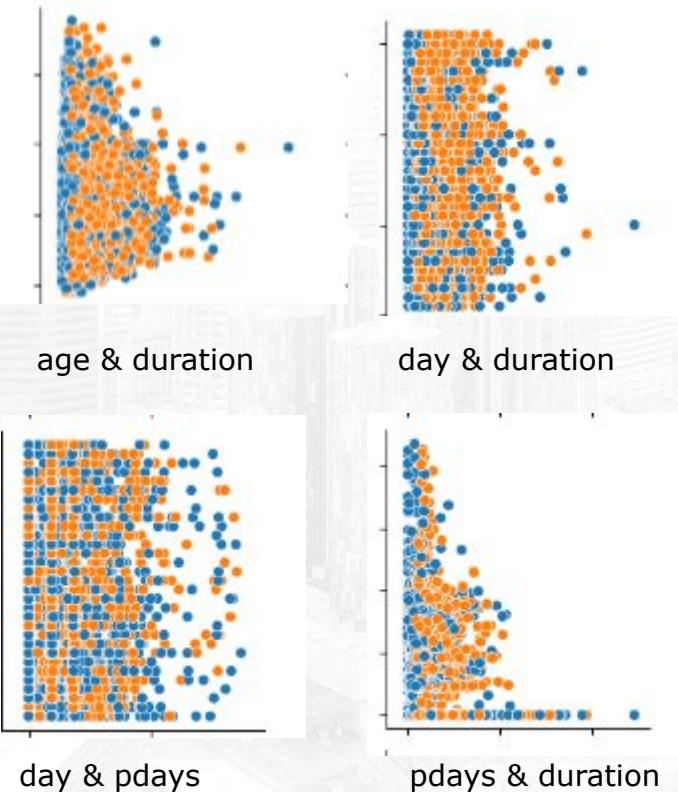
Lakukan multivariate analysis (seperti correlation heatmap dan category plots, sesuai yang diajarkan di kelas). Tuliskan hasil observasinya, seperti:

- A. Bagaimana korelasi antara masing-masing feature dan label. Kira-kira feature mana saja yang paling relevan dan harus dipertahankan?
- B. Bagaimana korelasi antar-feature, apakah ada pola yang menarik? Apa yang perlu dilakukan terhadap feature itu?

* Tuliskan juga jika memang tidak ada feature yang saling berkorelasi

3. Multivariate Analysis

A. Korelasi antar feature dan label

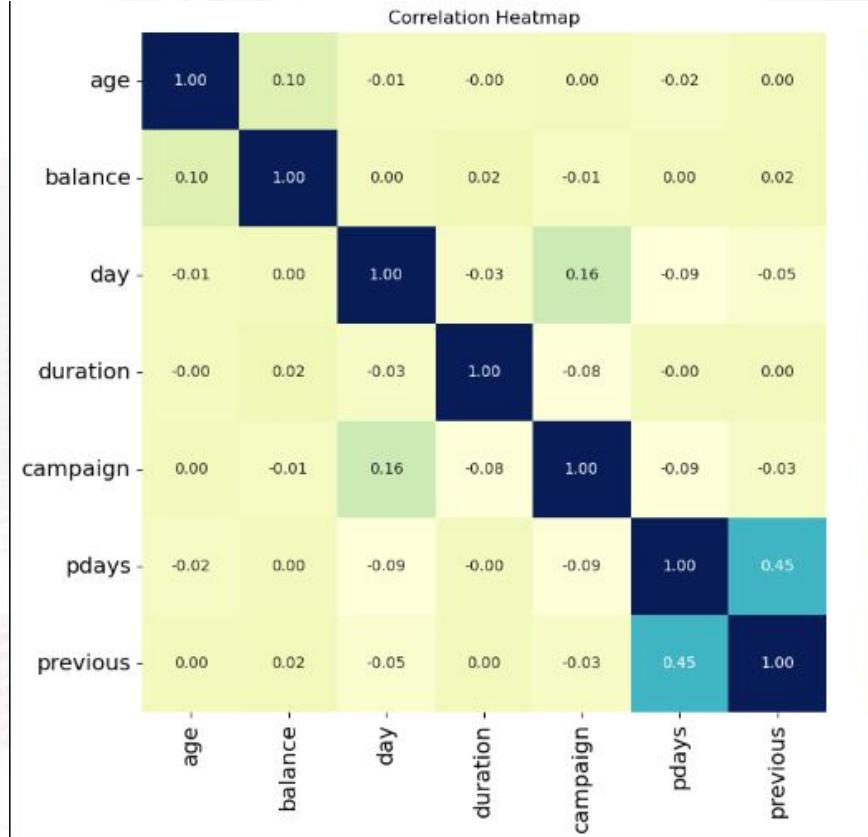


Korelasi antara feature dan label yang mengindikasikan kombinasi yang baik :

- age & duration
- day & duration
- day & pdays
- pdays & duration

Untuk korelasi antara feature dan label apabila dalam scatter plot terdapat kecenderungan terpisahnya antara plot feature dan label mengindikasikan kombinasi yang baik.

3. Multivariate Analysis



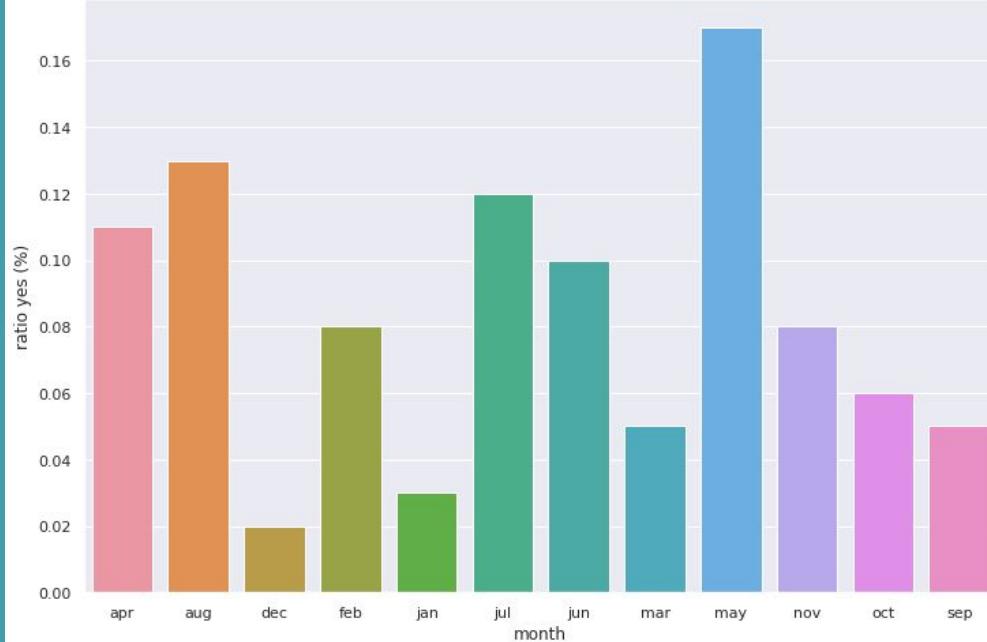
B. Korelasi antar feature :

Berdasarkan heatmap ini, dapat dilihat bahwa korelasi antar feature tidak terlalu kuat karena sebagian besar nilai korelasi tidak melebihi angka 0,7. Namun ada beberapa fitur yang memiliki nilai yang lebih tinggi dengan yang lain yaitu korelasi antara pdays & previous yang memiliki korelasi positif sebesar (0.45), day & campaign (0.16), age & balance (0.10), campaign & pdays, pdays & days dengan nilai korelasi negatif (-0,09)

Tindakan yang perlu dilakukan adalah menjaga fitur fitur yang memiliki nilai korelasi tinggi untuk dilakukan analisis lebih lanjut guna memahami hubungan hubungan antar fitur tersebut.

3. Multivariate Analysis

Hubungan Antara Fitur "month" Dengan Nasabah Yang Memilih "yes"



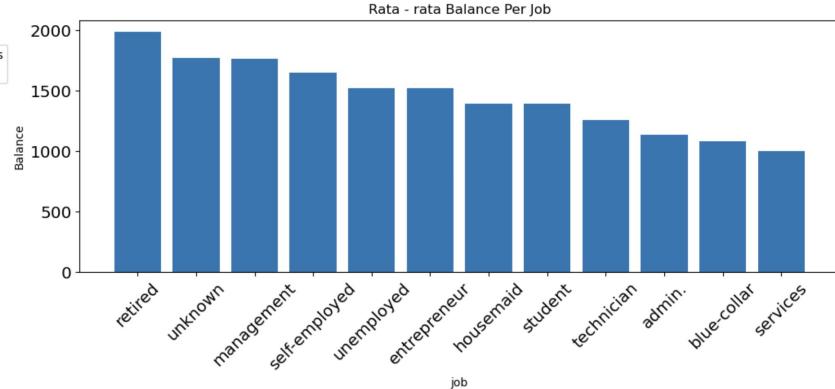
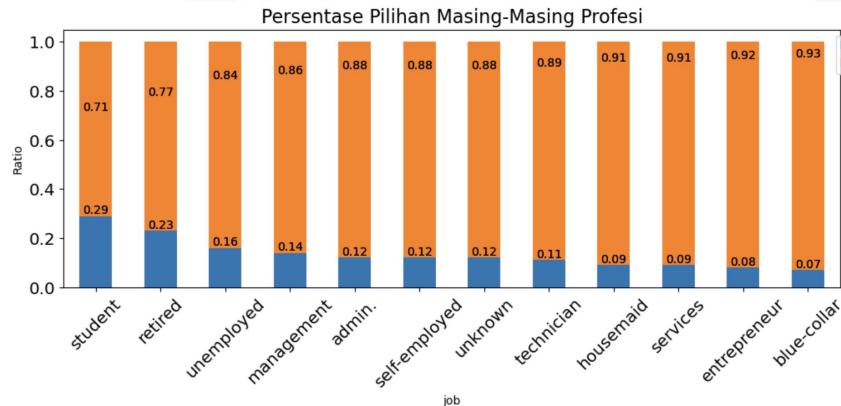
C. Ratio Yes (%)

Dapat dilihat bahwa pada bulan may persentase customer yang melakukan pendaftaran deposito sedikit lebih tinggi dibandingkan bulan-bulan lainnya. Namun jika kita melihat secara keseluruhan pada setiap bulannya jumlah customer yang mendaftar tidaklah berbeda jauh antara satu bulan ke bulan lainnya, bahkan semua bulan tidak ada yg memiliki persentase "yes" lebih besar dari 0,2%. Hal ini dapat menjadi landasan jika kita memutuskan untuk tidak menggunakan feature month terhadap model yang akan kita buat, dikarenakan feature month tidak memiliki pengaruh yang signifikan perihal seorang customer tertarik mendaftar deposito berjangka.

4. Business Insight (30 poin)

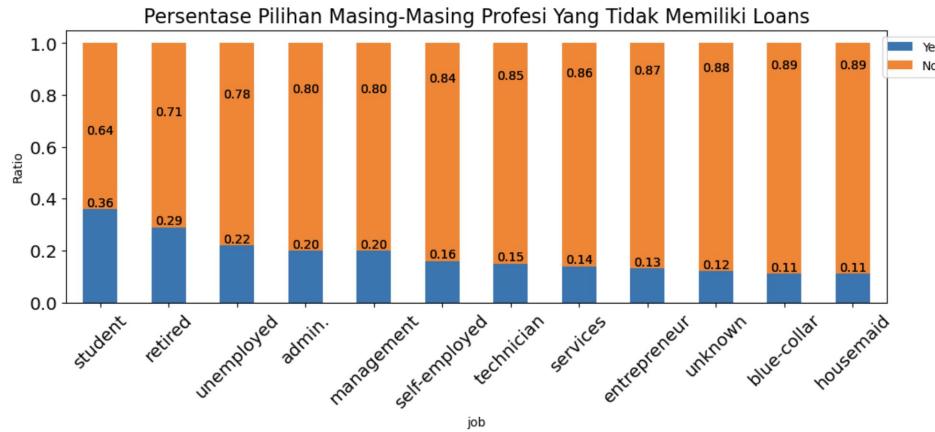
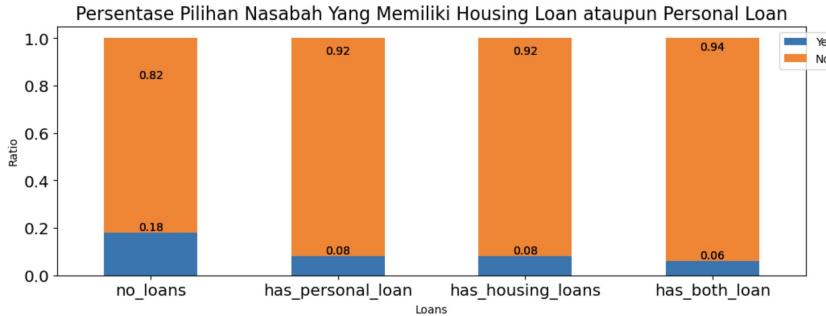
Selain EDA, lakukan juga beberapa analisis dan visualisasi untuk menemukan suatu business insight. Tuliskan minimal 3 insight, dan berdasarkan insight tersebut jelaskan rekomendasinya untuk bisnis.

04. Business Insight



- Nasabah yang membuka term deposit didominasi oleh nasabah yang berprofesi sebagai **student, retired, unemployed, dan management**.
- Nasabah yang memiliki saldo rata - rata tahunan tertinggi dimiliki oleh nasabah dengan profesi **retired, unknown, management, self-employed, unemployed, dan entrepreneur** dengan saldo rata - rata tahunan di atas 1500.

04. Business Insight



Nasabah yang membuka term deposit didominasi oleh nasabah yang **tidak memiliki housing loans maupun personal loans**. Dan jika nasabah yang sama sekali tidak memiliki loans dan dilihat berdasarkan pekerjaannya, customer yang membuka term deposit kebanyakan beprofesi sebagai **student, retired, unemployed, admin, dan management**. Direkomendasikan untuk menghubungi calon nasabah yang berprofesi sebagai **self-employed, dan entrepreneur** mengingat nasabah tersebut memiliki saldo rata - rata tahunan yang cukup tinggi.

04. Business Insight

y campaign

0 no 2.846350

1 yes 2.141047

```
import scipy.stats as st
# Hypothesis Testing using mann whitney
stat, p_value= st.mannwhitneyu(yes['campaign'],no['campaign'],alternative='two-sided')
p_value

1.9484904873905108e-71

alpha = 0.05
print('P-Value :',p_value)

if p_value >= alpha:
    print('Tidak cukup bukti jumlah campaign mampu membedakan user untuk membuka akun atau tidak')
else:
    print('cukup bukti jumlah campaign mampu membedakan user untuk membuka akun atau tidak')
```

P-Value : 1.9484904873905108e-71

cukup bukti jumlah campaign mampu membedakan user untuk membuka akun atau tidak

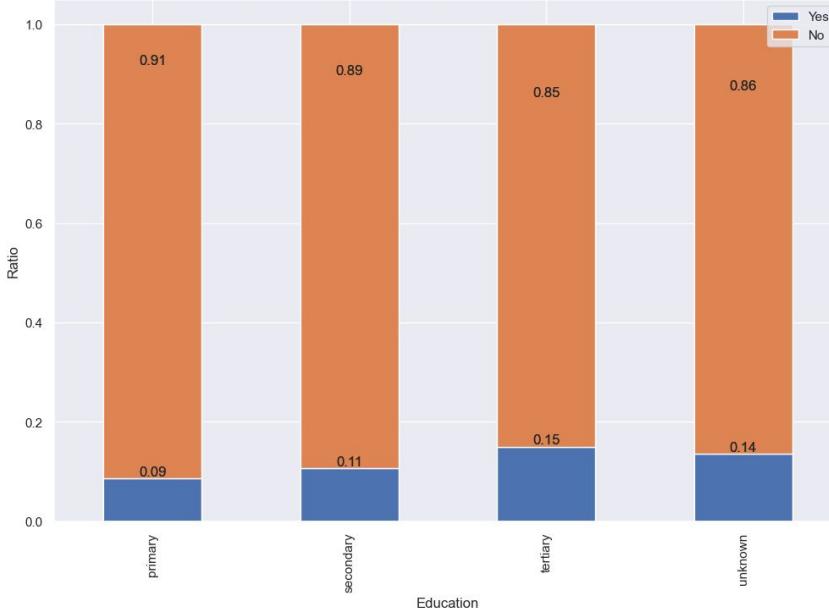
Hasil uji hipotesis menunjukkan bahwa $p\text{-value} < \alpha$, maka kita akan mengambil keputusan bahwa jumlah campaign berpengaruh terhadap nasabah untuk membuka akun atau tidak secara signifikan.

Namun berdasarkan berdasarkan rata-rata ternyata semakin banyak campaign yang diberikan ternyata user akan semakin menolak membuka akun.

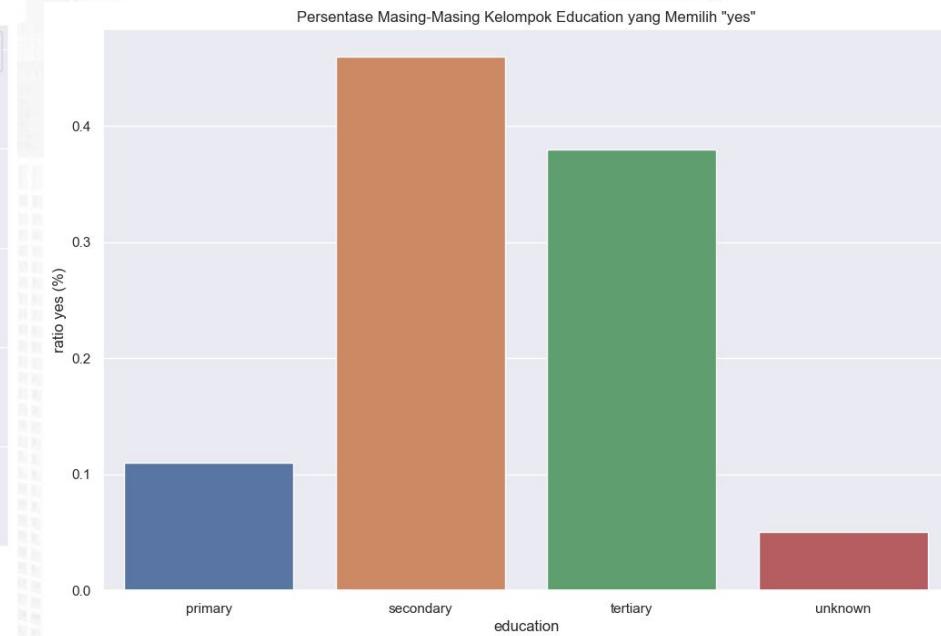
Jadi kesimpulannya, jumlah campaign berhubungan terbalik dengan user membuka akun.

04. Business Insight

Percentase Masing-Masing Kelompok Education yang Memilih "yes"



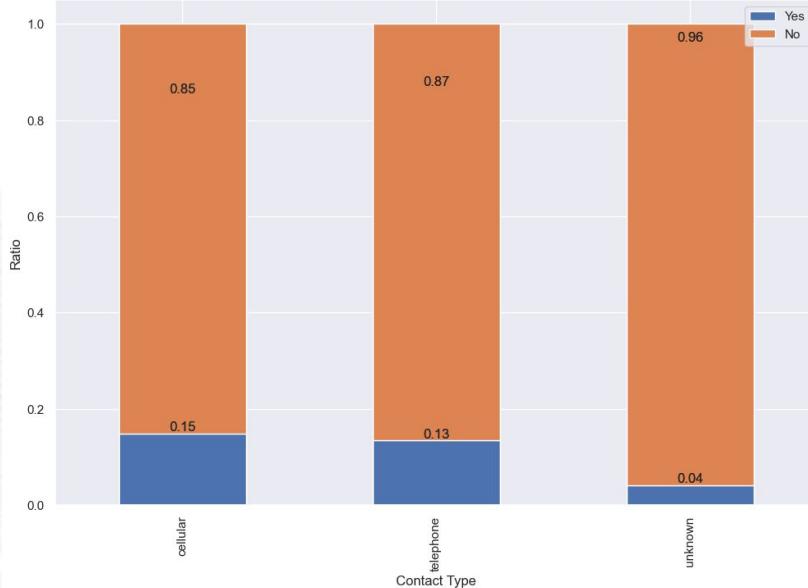
Percentase Masing-Masing Kelompok Education yang Memilih "yes"



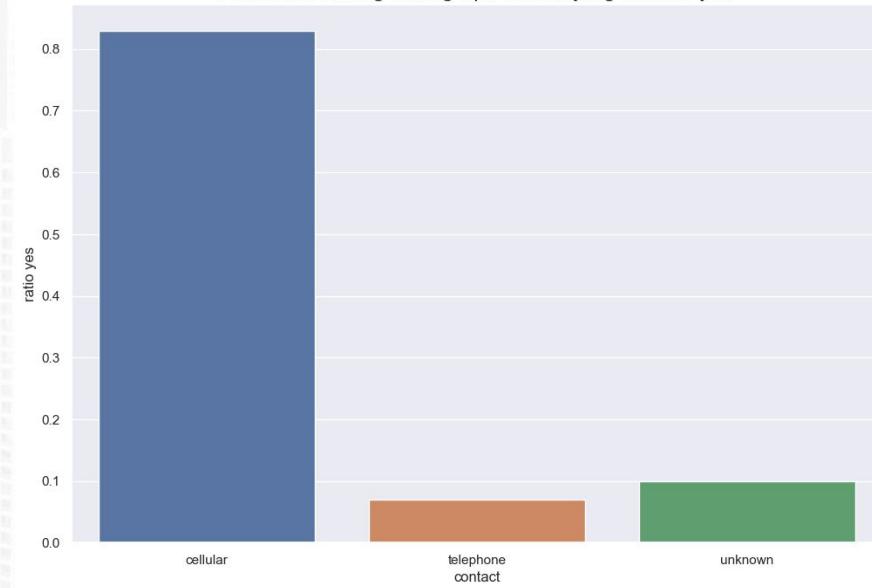
- Pada visualisasi diatas dapat dilihat bahwa kelompok education **tertiary** (lulusan S1 atau di atasnya) paling banyak memilih untuk mendaftar deposito berjangka sebanyak 15% dibanding total yang mendaftar di kelompoknya, yang diikuti oleh kelompok education **yang tidak diketahui**, lalu **secondary** (lulusan SMP dan SMA) dan terakhir **primary** (Lulusan SD ke bawah).
- Dengan membandingkan terhadap mereka yang mengambil yes saja, terlihat bahwa kelompok education **secondary** paling banyak memilih yes, diikuti dekat kelompok **tertiary**, lalu **primary** dan kelompok Education **yang tidak diketahui**.
- Dari insight tersebut, rekomendasi bisnis yang dapat diberikan adalah menargetkan campaign terhadap kelompok **tertiary**. Bisa di daerah kampus, universitas, mengajak investasi sebagai kesempatan untuk mendapatkan passive income untuk uang yang mungkin telah mereka tabung.
- Rekomendasi ke dua, bisa ditargetkan untuk mereka yang masih di jenjang sekolah, utamanya **secondary**. Bisa melakukan campaign ke sekolah-sekolah, melibatkan orang tua, untuk membuka rekening dan deposito dini.

04. Business Insight

Percentase Pilihan Masing-Masing Cara Contact



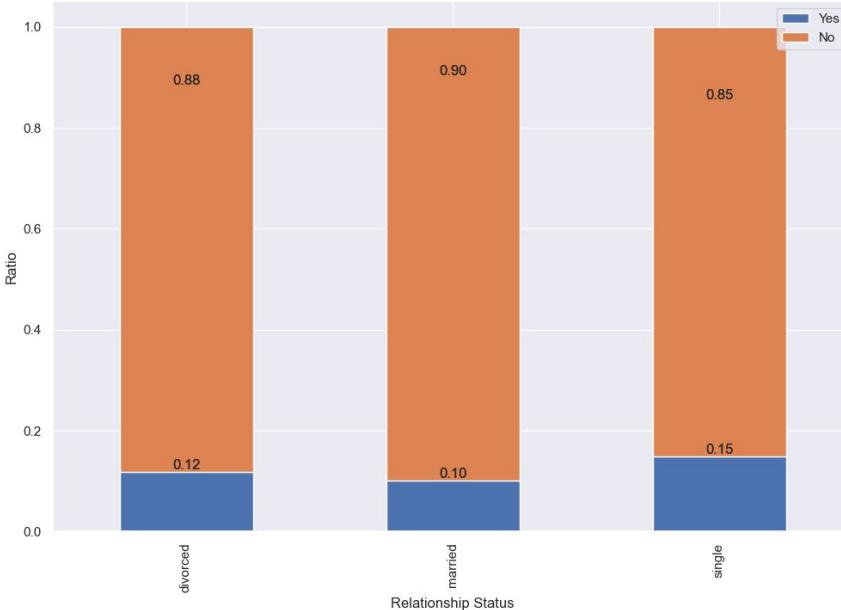
Percentase Masing-Masing Tipe Contact yang Memilih "yes"



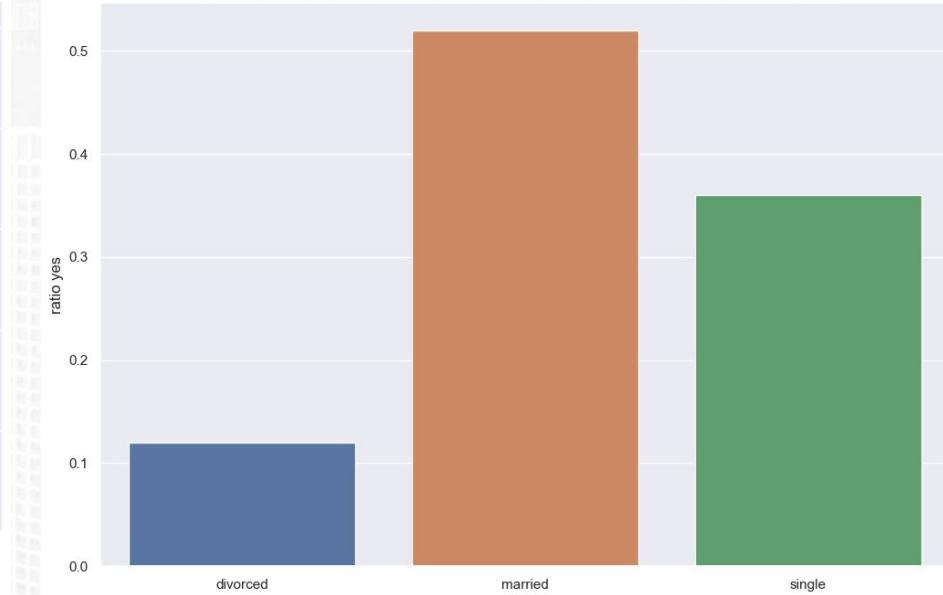
Bisa dilihat melalui visualisasi diatas bahwa tipe contact **cellular** (HP) merupakan tipe contact yang paling banyak menghasilkan customer untuk mendaftar deposito berjangka. Hal ini bisa terjadi karena di era sekarang ini orang-orang lebih banyak melakukan komunikasi melalui telepon cellular dibandingkan telepon biasa (telepon rumah), dan hampir setiap orang pasti mempunyai telepon cellular. Dari sini pihak bank **direkomendasi** mulai merubah strategi campaign untuk memprioritaskan melakukan contact melalui telepon cellular. Selain itu pihak bank sebelum melakukan panggilan secara langsung ke nomor telepon cellular, bisa juga bisa melakukan campaign melalui email atau melalui whatsapp sebagai tahapan awal dalam menawarkan deposito berjangka. Dengan demikian peluang customer yang mendaftar deposito berjangka akan meningkat.

04. Business Insight

Percentase Pilihan Masing-Masing Status Pernikahan



Percentase Masing-Masing Tipe Status Pernikahan yang Memilih "yes"



- Pada visualisasi diatas dapat dilihat bahwa **yang masih single** paling banyak memilih untuk mendaftar deposito berjangka sebanyak 15% dibanding total yang mendaftar di antara nasabah yang masih single, lalu diikuti oleh yang sudah **bercerai**, dan terakhir mereka yang sudah **menikah** dalam 10%.
- Dengan membandingkan terhadap mereka yang mengambil yes saja, terlihat bahwa **yang sudah menikah** paling banyak memilih yes, diikuti mereka yang masih **single**, dan terakhir mereka yang sudah **ceraia**.
- Rekomendasi untuk meningkatkan mereka yang mendaftar deposito berjangka pihak bank saat melakukan campaign kepada masing-masing individu dapat melakukan promosi dengan strategi pendekatan yang berbeda. Misalkan kepada kelompok "**single**" bisa melakukan promosi seperti "dengan bunga deposito dalam setahun adalah x% maka kira-kira dalam y tahun dana menikah akan dapat terkumpul". Sedangkan untuk kelompok "**married**" bisa melakukan promosi seperti "dengan bunga deposito dalam setahun adalah x% maka dana pendidikan untuk anak/dana pensiun saat masa tua akan terjamin"

5. Git (15 poin)

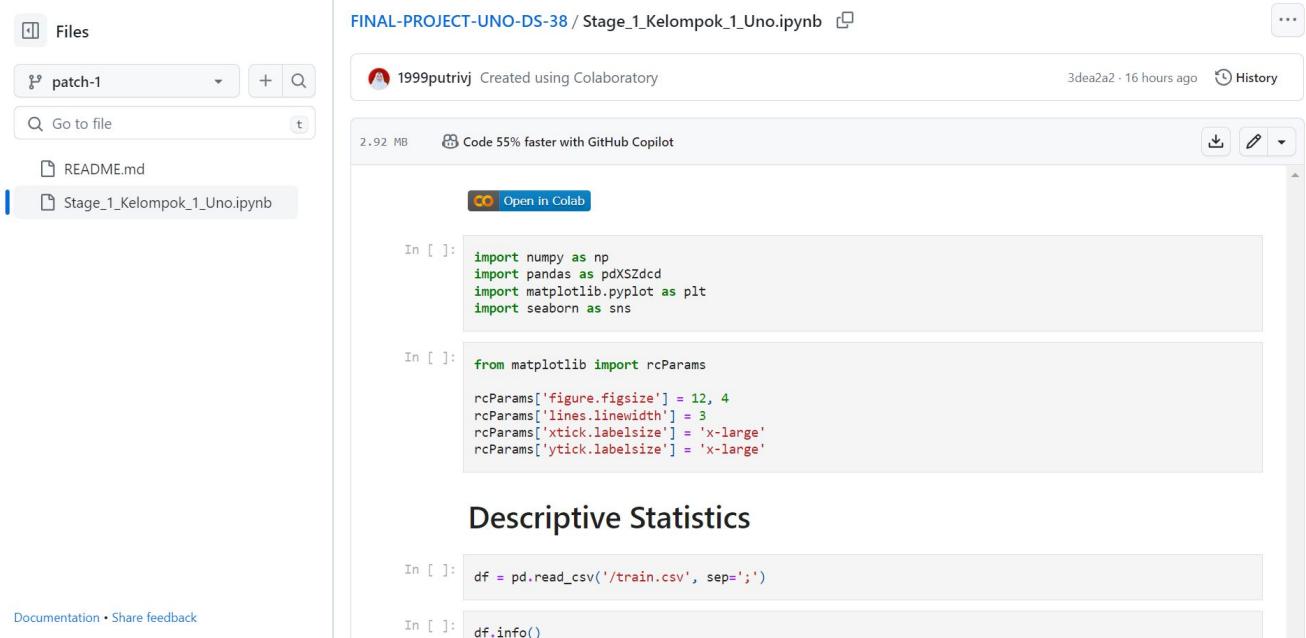
Upload project teman-teman di sebuah repository git. Berkolaborasilah di Git jika ada perubahan version dari waktu ke waktu.

- A. Buat Repository Git
- B. Upload file notebook atau file penggerjaan lainnya pada repository tersebut

Untuk file README, dapat merupakan summary insight yang telah didapatkan dari EDA.

5. Git (15 poin)

Git - Uno



The screenshot shows a GitHub repository interface. On the left, there's a sidebar with a 'Files' section containing a 'patch-1' folder, a 'README.md' file, and the 'Stage_1_Kelompok_1_Uno.ipynb' notebook, which is currently selected. The main area displays the contents of the notebook:

FINAL-PROJECT-UNO-DS-38 / Stage_1_Kelompok_1_Uno.ipynb

Created using Colaboratory · 16 hours ago · History

Code 55% faster with GitHub Copilot

[Open in Colab](#)

```
In [ ]:  
import numpy as np  
import pandas as pdXSZdc  
import matplotlib.pyplot as plt  
import seaborn as sns
```

```
In [ ]:  
from matplotlib import rcParams  
  
rcParams['figure.figsize'] = 12, 4  
rcParams['lines.linewidth'] = 3  
rcParams['xtick.labelsize'] = 'x-large'  
rcParams['ytick.labelsize'] = 'x-large'
```

Descriptive Statistics

```
In [ ]:  
df = pd.read_csv('/train.csv', sep=';')
```

```
In [ ]:  
df.info()
```

Laporan Final Project
Bank UNO Marketing Targets

Data Pre-Processing

**Final Project - Stage
2**



Estimasi Waktu Penggerjaan

 **3 - 5 jam**

Jumlah Soal

 **2 Soal**

Total Point

 **100 poin**

Teknis Pengerjaan

1. Pekerjaan dilakukan secara **berkelompok, sesuai kelompok Final Project**
2. Masing-masing anggota kelompok tetap perlu submit ke LMS (jadi bukan perwakilan)
3. File yang perlu dikumpulkan:
 - o File **jupyter notebook** (.ipynb) yang berisi source code.
 - o File **laporan homework** (.pdf) yang berisi rangkuman dari apa saja yang telah dilakukan.
4. Upload hasil pengerjaanmu melalui LMS.
 - o Masukkan semua file ke dalam **1 file** dengan format **ZIP**.
 - o Nama File:
Preprocessing - <Nama Kelompok>.zip

1. Data Cleansing (50 poin)

Lakukan pembersihan data, sesuai yang diajarkan di kelas, seperti:

- A. Handle missing values
- B. Handle duplicate data
- C. Handle outliers
- D. Feature transformation
- E. Feature encoding
- F. Handle class imbalance

Di laporan homework, tuliskan apa saja yang telah dilakukan dan metode yang digunakan.

* Tetap tuliskan jika memang ada tidak yang perlu di-handle (contoh: "Tidak perlu feature encoding karena semua feature sudah numerical" atau "Outlier tidak di-handle karena akan fokus menggunakan model yang robust terhadap outlier").

A. Handle missing values

```
df.isna().sum()
```

```
age          0
job          0
marital      0
education    0
default      0
balance      0
housing      0
loan          0
contact      0
day           0
month         0
duration     0
campaign     0
pdays         0
previous     0
poutcome     0
y             0
dtype: int64
```

```
[ ] df['poutcome'].replace({'unknown': 'never'}, inplace=True)
```

Tidak terdapat missing values pada dataset, namun untuk fitur poutcome yang memiliki value unknown dilakukan replacement menjadi never yang menunjukkan bahwa customer tidak pernah di hubungi karena memiliki pdays = -.

B. Handle duplicated data

```
[180] df.duplicated().sum()
```

```
0
```

Tidak terdapat duplicated values pada dataset

C. Handle Outliers

```
[ ] nums2 = ['age', 'campaign']
for num in nums2:
    df[num] = np.log(df[num])
```

```
▶ from scipy import stats

print("Before removing outlier: ", len(df))

for num in nums2:
    z_scores = np.abs(stats.zscore(df[num]))
    df = df[z_scores < 3]

print("After removing outlier: ", len(df))
```

```
➡ Before removing outlier: 45211
After removing outlier: 44790
```

Beberapa fitur numerik pada dataset memiliki sebaran yang right skewed (long right tailed), oleh karena itu dilakukan log transformation terlebih dahulu sebelum me-remove outliers. Log transformation hanya dilakukan pada fitur age dan campaign saja, karena jika semua fitur numerik dilakukan log transformation jumlah data setelah remove outlier akan hilang secara keseluruhan atau berjumlah 0. Setelah log transformation pada fitur age dan campaign, selanjutnya menghapus outliers menggunakan z-scores yang mana jumlah data berkurang dari 45.211 menjadi 44.790.\

D. Feature transformation

```
[ ] from imblearn.over_sampling import SMOTE
from sklearn.model_selection import train_test_split

X = df.drop(['y'], axis=1)
y = df['y']

X_train, X_test, y_train, y_test = train_test_split(X, y, random_state=42, test_size=0.2)

print(f'Number of Train Data: {y_train.shape[0]}')
print(f'Number of Test Data: {y_test.shape[0]}')

Number of Train Data: 35832
Number of Test Data: 8958
```

Sebelum dilakukan feature transformation pada, dataset dibagi menjadi train data dan test data terlebih dahulu. Yang mana 80% data merupakan train data yang berjumlah 35.835 dan test data sebanyak 20% dari keseluruhan dataset yang berjumlah 8.958

D. Feature transformation

```
▶ from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
columns_to_standardize = ['age', 'balance', 'campaign', 'pdays', 'previous']
X_train[columns_to_standardize] = scaler.fit_transform(X_train[columns_to_standardize])
X_test[columns_to_standardize] = scaler.transform(X_test[columns_to_standardize])
print("DataFrame setelah distandardisasi:")
X_train.head()
```

Selanjutnya dilakukan feature scaling pada train data dan test data dengan standarization untuk fitur numerik.

E. Feature Encoding

```
[ ] mapping_default = {  
    'no' : 0,  
    'yes' : 1,  
}  
X_train['default'] = X_train['default'].map(mapping_default)  
X_test['default'] = X_test['default'].map(mapping_default)  
  
[ ] mapping_housing = {  
    'no' : 0,  
    'yes' : 1,  
}  
X_train['housing'] = X_train['housing'].map(mapping_housing)  
X_test['housing'] = X_test['housing'].map(mapping_housing)  
  
▶ mapping_loan = {  
    'no' : 0,  
    'yes' : 1,  
}  
X_train['loan'] = X_train['loan'].map(mapping_loan)  
X_test['loan'] = X_test['loan'].map(mapping_loan)
```

```
[ ] X_train_encoded_education = pd.get_dummies(X_train['education'], prefix = 'pendidikan')  
X_test_encoded_education = pd.get_dummies(X_test['education'], prefix = 'pendidikan')  
  
[ ] X_train_encoded_kerja = pd.get_dummies(X_train['job'], prefix = 'kerja')  
X_test_encoded_kerja = pd.get_dummies(X_test['job'], prefix = 'kerja')  
  
[ ] X_train_encoded_marital = pd.get_dummies(X_train['marital'], prefix = 'status')  
X_test_encoded_marital = pd.get_dummies(X_test['marital'], prefix = 'status')  
  
[ ] X_train_encoded_contact = pd.get_dummies(X_train['contact'], prefix = 'contact')  
X_test_encoded_contact = pd.get_dummies(X_test['contact'], prefix = 'contact')  
  
[ ] X_train_encoded_poutcome = pd.get_dummies(X_train['poutcome'], prefix = 'poutcome')  
X_test_encoded_poutcome = pd.get_dummies(X_test['poutcome'], prefix = 'poutcome')
```

- Feature Encoding merupakan proses mengubah feature categorical menjadi feature numeric.
- Pada data yang bertipe ordinal dan distinct values = 2 (ya/tidak) diubah menggunakan Label Encoding, sisanya diubah menggunakan one hot encoding

F. Handle class imbalance

```
▶ y_train.value_counts()
```

```
→ no      31603  
yes     4229  
Name: y, dtype: int64
```

```
[ ] # OVERSAMPLING
```

```
from imblearn import over_sampling  
X_oversampling , y_oversampling = over_sampling.SMOTE(random_state=42).fit_resample(X_train_combined,y_train)  
print(pd.Series(y_oversampling).value_counts())
```

```
no      31603  
yes    31603  
Name: y, dtype: int64
```

- Oversampling SMOTE pada data train yang memiliki ketimpangan pada ditribusi target

2. Feature Engineering (35 poin)

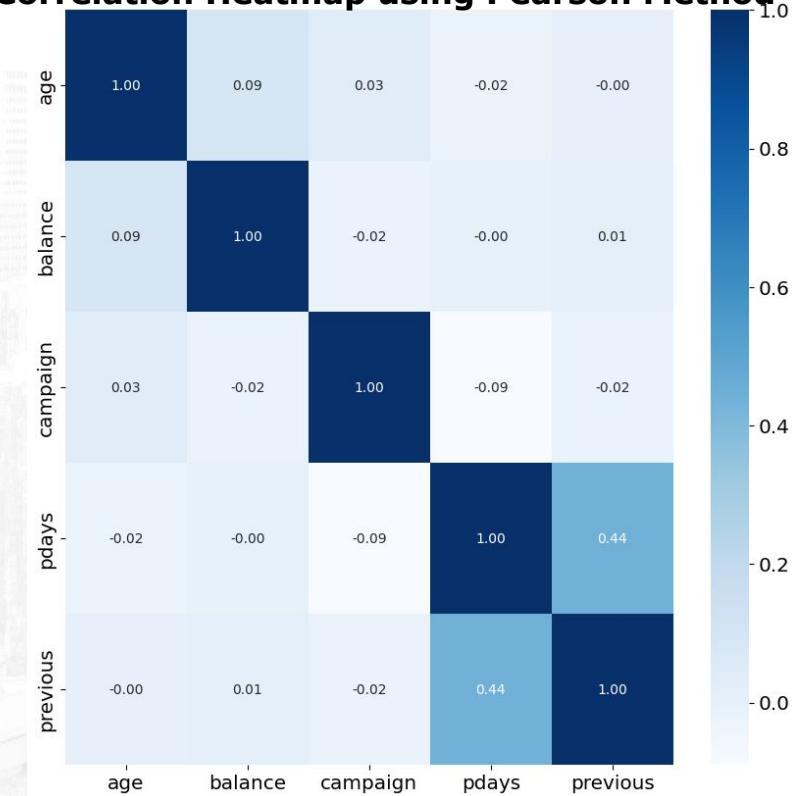
Cek feature yang ada sekarang, lalu lakukan:

- A. Feature selection (membuang feature yang kurang relevan atau redundan)
- B. Feature extraction (membuat feature baru dari feature yang sudah ada)
- C. Tuliskan minimal 4 feature tambahan (selain yang sudah tersedia di dataset) yang mungkin akan sangat membantu membuat performansi model semakin bagus (ini hanya ide saja, untuk menguji kreativitas teman-teman, tidak perlu benar-benar dicari datanya dan tidak perlu diimplementasikan)

* Untuk 2A & 2B, tetap tuliskan jika memang tidak bisa dilakukan (contoh: "Semua feature digunakan untuk modelling (tidak ada yang dihapus), karena semua feature relevan")

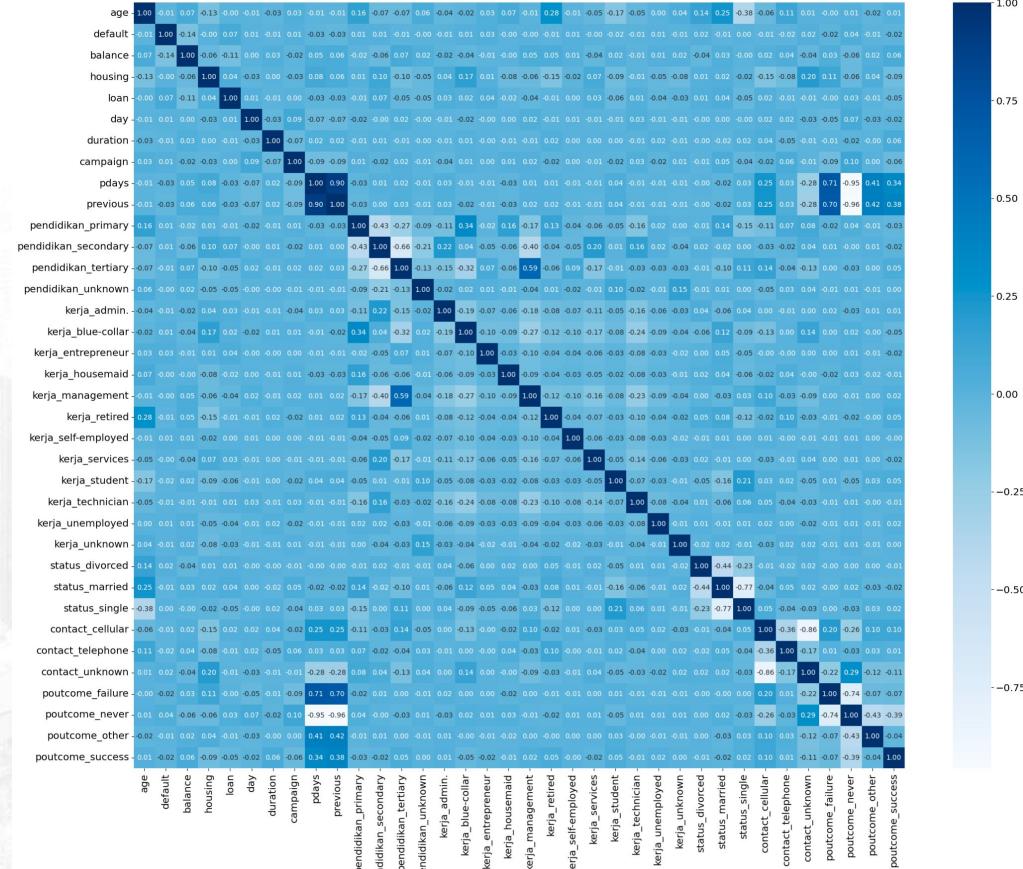
Feature Engineering

Numerical-Numerical Correlation Heatmap using Pearson Method

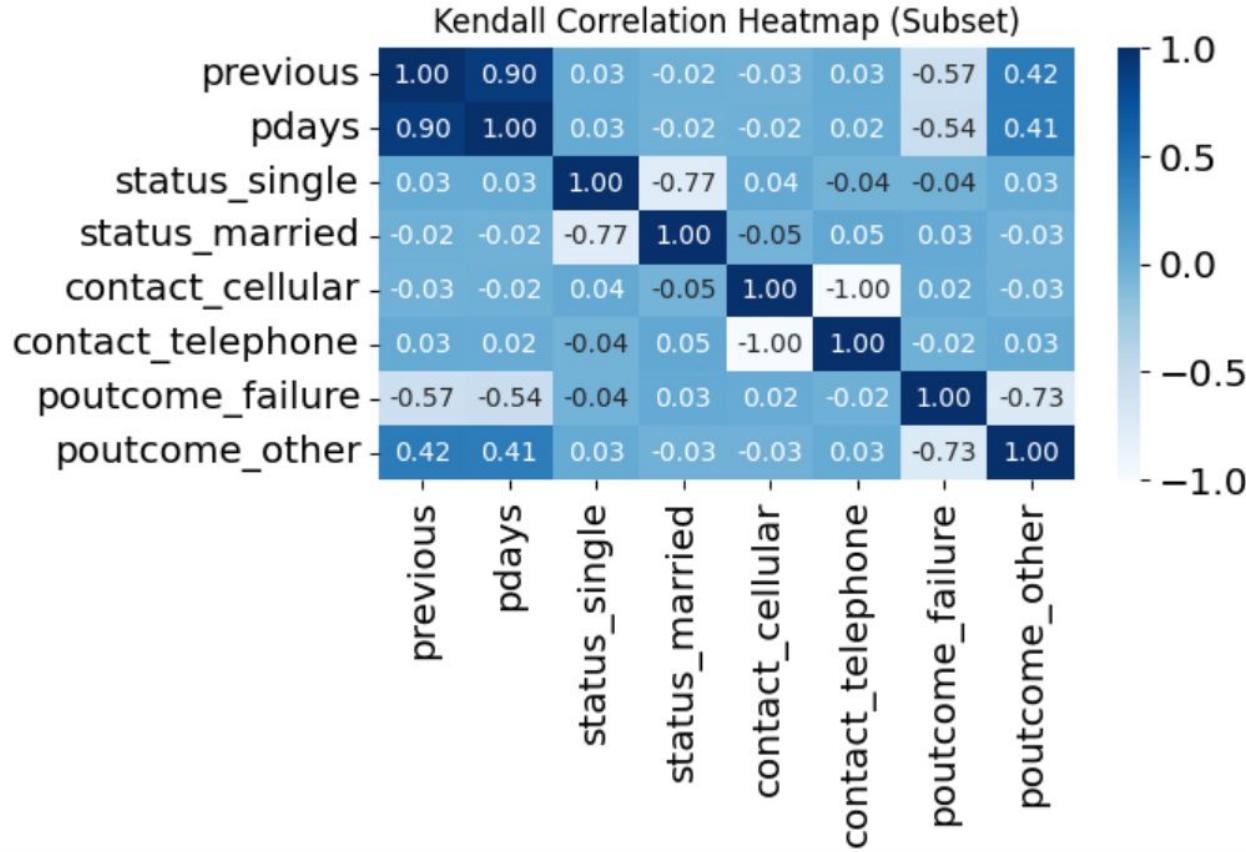


Feature Engineering

Categorical-Numerical and Categorical-Categorical Correlation Heatmap using Kendall Method



Feature Engineering



Feature Engineering

Delete Unnecessary Columns

Sebab tidak ada feature tanggal yang lengkap, hanya akan didrop '**day**' sebab korelasinya terhadap **y** kecil, serta '**month**' sebab dari apa yang telah disimpulkan pada EDA Multivariate Analysis. Ditambah, day dan month tanpa time series year dan tanpa adanya info kapan campaign dilaksanakan, menjadi sulit untuk menarik kesimpulan dari hal tersebut.

Delete Redundant Data

Feature '**previous**' dengan '**pdays**', '**status_single**' dengan '**status_married**', '**contact_cellular**' dengan '**contact_unknown**', '**poutcome_failure**' dengan '**poutcome_never**', memiliki korelasi di atas 0.7 yang menyebabkan mereka redundant untuk dijadikan feature bersama, sehingga akan digunakan salah satu saja.

Feature Engineering

Feature Extraction

Dengan adanya 32 Feature, dirasa telah cukup feature yang dibutuhkan, dengan feature ini memiliki relevansi tersendiri terhadap 'target' yang ingin dicapai. Sehingga kami **tidak akan mengeluarkan feature baru.**

Ide Feature *for future references*

1. Jumlah anak/tanggungan
2. Memiliki produk deposito berjangka pada bank lain (y/n)
3. Sudah berapa lama menjadi nasabah bank tersebut
4. Memiliki produk investasi lain selain deposito berjangka (y/n)
5. Durasi setiap campaign (dengan informasi waktu yang lebih memadai)

3. Git (15 poin)

Upload project teman-teman di sebuah repository git. Berkolaborasilah di Git jika ada perubahan version dari waktu ke waktu.

- A. Buat Repository Git
- B. Upload file notebook atau file penggerjaan lainnya pada repository tersebut

Untuk file README, dapat merupakan summary dari proses data preproses yang telah dilakukan. Boleh menggunakan repositori yang sama atau membuat baru.

Git

README

B. Handle Duplicate Data

Tidak terdapat duplicated values pada dataset

C. Handle Outliers

Beberapa fitur numerik pada dataset memiliki sebaran yang right skewed (long right tailed), oleh karena itu dilakukan log transformation terlebih dahulu sebelum me-remove outliers. Log transformation hanya dilakukan pada fitur age dan campaign saja, karena jika semua fitur numerik dilakukan log transformation jumlah data setelah remove outlier akan hilang secara keseluruhan atau berjumlah 0. Setelah log transformation pada fitur age dan campaign, selanjutnya menghapus outliers menggunakan z-scores yang mana jumlah data berkurang dari 45.211 menjadi 44.790.

D. Feature Transformation

- Sebelum dilakukan feature transformation, dataset dibagi menjadi train data dan test data terlebih dahulu. Yang mana 80% data merupakan train data yang berjumlah 35.835 dan test data sebanyak 20% dari keseluruhan dataset yang berjumlah 8.958
- Selanjutnya dilakukan feature scaling pada train data dan test data dengan standarization untuk fitur numerik.

README

Feature Engineering

Delete Unnecessary Columns

Sebab tidak ada feature tanggal yang lengkap, hanya akan didrop 'day' sebab korelasinya terhadap y kecil, serta 'month' sebab dari apa yang telah disimpulkan pada EDA Multivariate Analysis. Ditambah, day dan month tanpa time series year dan tanpa adanya info kapan campaign dilaksanakan, menjadi sulit untuk menarik kesimpulan dari hal tersebut.

Delete Redundant Data

Feature 'previous' dengan 'pdays', 'status_single' dengan 'status_married', 'contact_cellular' dengan 'contact_unknown', 'poutcome_failure' dengan 'poutcome_never', memiliki korelasi di atas 0.7 yang menyebabkan mereka redundant untuk dijadikan feature bersama, sehingga akan digunakan salah satu saja.

Feature Extraction

Dengan adanya 32 Feature, dirasa telah cukup feature yang dibutuhkan, dengan feature ini memiliki relevansi tersendiri terhadap 'target' yang ingin dicapai. Sehingga kami tidak akan mengeluarkan feature baru.

[Link Github Final Project Stage 2 - Data Pre-Processing](#)

TERIMA KASIH

[Link Google Collab Stage 1](#)

[Link Google Collab Stage 2](#)