



EXPLORATORY DATA ANALYSIS



ANGGOTA TEAM



- M RIZQI FADHILAH
- M ARVIN FADRIANSYAH
- MELLIZA NASTASIA IZAZI
- THUFAEL BINTANG ALFATTAH
- ZULFIKAR FAUZI
- ANNISA SULISTYANINGSIH
- NIKEN MUSTIKAWENI
- GALIH REFA



DAFTAR PEMBAHASAN

DESCRIPTIVE STATISTICS

UNIVARIATE ANALYSIS

MULTIVARIATE ANALYSIS

BUSINESS INSIGHT





DESCRIPTIVE STATISTICS



Descriptive Statistics

Data Information

```
<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    bool   
 5   balance     45211 non-null    int64  
 6   housing     45211 non-null    bool   
 7   loan         45211 non-null    bool   
 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    object  
 13  pdays       45211 non-null    int64  
 14  previous    45211 non-null    int64  
 15  poutcome   45211 non-null    object  
 16  y           45211 non-null    bool   
dtypes: bool(4), int64(6), object(7)
memory usage: 4.7+ MB
```

Berikut adalah perubahan tipe data yang dilakukan beserta penjelasannya:

1. default: Tipe data diubah dari object menjadi boolean.
 - Alasan: Kolom ini hanya berisi dua opsi nilai: "yes" atau "no", sehingga lebih tepat disimpan sebagai tipe boolean untuk mempermudah proses analisis dan pengolahan data.
2. housing: Tipe data diubah dari object menjadi boolean.
 - Alasan: Sama seperti kolom default, kolom ini juga hanya memiliki dua pilihan nilai: "yes" atau "no". Dengan menggunakan tipe boolean, representasi dan analisis data menjadi lebih efisien.
3. loan: Tipe data diubah dari object menjadi boolean.
 - Alasan: Karena kolom ini hanya terdiri dari dua nilai: "yes" atau "no", mengonversinya ke tipe boolean akan mempermudah pengolahan data.
4. y: Tipe data diubah dari object menjadi boolean.
 - Alasan: Kolom ini merupakan label target yang hanya memiliki dua nilai: "yes" atau "no". Mengubahnya menjadi boolean akan membantu dalam proses klasifikasi.



Descriptive Statistics

Handle Missing Values

Untuk menangani masalah missing values, meskipun data tampak lengkap, terkadang terdapat anomali di mana data kosong sebenarnya diisi dengan nilai "unknown". Hal ini bisa diverifikasi melalui beberapa sumber berikut:

1. Artikel Ilmiah:

- S. Moro, P. Cortez, dan P. Rita. Pendekatan Berbasis Data untuk Memprediksi Keberhasilan Telemarketing Bank. *Decision Support Systems*, Elsevier, 62:22-31, Juni 2014.
- S. Moro, R. Laureano, dan P. Cortez. Penggunaan Data Mining dalam Pemasaran Langsung Bank: Penerapan Metodologi CRISP-DM. Dalam P. Novais et al. (Eds.), *Proceedings of the European Simulation and Modelling Conference - ESM'2011*, hal. 117-121, Guimarães, Portugal, Oktober 2011. EUROSIS.

2. Dataset asli: Bank Marketing Dataset

Mengatasi anomali ini penting untuk memastikan data yang digunakan dalam analisis dan model machine learning lebih akurat serta dapat diandalkan.



Descriptive Statistics

Handle Missing Values

age	0
job	288
marital	0
education	1857
default	0
balance	0
housing	0
loan	0
contact	13020
day	0
month	0
duration	0
campaign	0
pdays	0
previous	0
poutcome	36959
y	0
dtype:	int64

Berdasarkan data yang ada, kolom-kolom berikut memiliki nilai "unknown":

1. job: terdapat 288 nilai "unknown"
2. education: terdapat 1857 nilai "unknown"
3. contact: terdapat 13020 nilai "unknown"
4. poutcome: terdapat 36959 nilai "unknown"

Hal ini menunjukkan bahwa meskipun data terlihat lengkap, nilai "unknown" ini adalah anomali yang perlu diatasi agar hasil analisis dan model machine learning lebih akurat dan dapat diandalkan.



Descriptive Statistics

Handle Missing Values

Penanganan yang dilakukan :

1. Menghapus nilai "unknown" pada kolom job dan education
 - Kita akan menghapus baris yang berisi nilai "unknown" di kolom-kolom ini untuk menjaga kualitas data dan mengurangi data yang tidak perlu. Jumlah nilai "unknown" di sini sedikit, jadi tidak terlalu berdampak pada data keseluruhan. Tapi, kita akan cek lagi untuk memastikan cara ini yang paling tepat.
2. Mengganti nilai "unknown" pada kolom contact dengan nilai modus
 - Nilai "unknown" di kolom ini akan diganti dengan nilai yang paling sering muncul, yaitu "cellular" atau "telephone", karena menurut info di Kaggle, kontak hanya dilakukan lewat dua cara itu.
3. Mengganti nilai "unknown" pada kolom poutcome dengan nilai "nonexistent"
 - Nilai "unknown" di kolom ini akan diganti dengan "nonexistent" karena biasanya itu berarti pelanggan belum pernah dihubungi sebelumnya. Ini sesuai dengan data yang sebenarnya dan sumber yang terpercaya.



Descriptive Statistics

Handle Duplicated Data

```
⌚ # Check for duplicated rows  
duplicates = df.duplicated().sum()  
print(f'Duplicated rows: {duplicates}')  
  
⌚ Duplicated rows: 0
```

Dataset ini tidak memiliki baris yang berulang, jadi kita bisa langsung ke langkah berikutnya

INFORMATION INOVATORS



Descriptive Statistics

Negative Values

	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	y
5666	34	blue-collar	married	secondary	False	-436	True	False	cellular	26	may	317	1	-1	0	nonexistent	False
35159	35	blue-collar	married	secondary	False	-701	True	True	cellular	7	may	531	3	-1	0	nonexistent	False
22662	33	technician	single	secondary	False	-32	False	False	cellular	25	aug	196	12	-1	0	nonexistent	False

Penjelasan:

1. Kolom balance:

- Penanganan: Nilai negatif di kolom balance diubah menjadi positif menggunakan fungsi `abs()`. Ini dilakukan karena nilai negatif mungkin merupakan kesalahan dalam memasukkan data, dan mengubahnya menjadi positif akan membuat analisis lebih mudah. Proses ini akan dicek lagi untuk memastikan keputusan yang diambil sudah tepat.

2. Kolom pdays:

- Penjelasan: Nilai -1 di kolom pdays menunjukkan bahwa pelanggan belum pernah dihubungi sebelumnya, sesuai dengan data asli dan referensi ilmiah.
- Penanganan: Nilai -1 diganti dengan 999 supaya sesuai dengan data asli yang telah ditemukan.



INFLVATORS
INFORMATION INOVATORS



Descriptive Statistics

Zero Values

```
# Menghitung jumlah nilai 0 pada kolom numerik
zero_counts = (df.select_dtypes(include=['number']) == 0).sum()

# Menampilkan hasil
print(f'Jumlah nilai 0 di setiap kolom numerik:\n{zero_counts}')

# Jumlah nilai 0 di setiap kolom numerik:
age          0
balance     3366
day          0
duration     3
pdays         0
previous    35281
dtype: int64

# Menambahkan 1 pada semua nilai di kolom 'balance'
df['balance'] = df['balance'] + 1

# Cek kembali
zero_counts = (df.select_dtypes(include=['number']) == 0).sum()

# Menampilkan hasil
print(f'Jumlah nilai 0 di setiap kolom numerik:\n{zero_counts}')

# Jumlah nilai 0 di setiap kolom numerik:
age          0
balance     0
day          0
duration     3
pdays         0
previous    35281
dtype: int64
```

Penjelasan:

- Kolom balance: Menambahkan angka 1 pada semua nilai untuk memastikan tidak ada nilai 0, sehingga distribusi nilainya lebih adil.
- Kolom previous & duration: Catatan bahwa kolom ini akan dianalisis lebih mendalam pada tahap selanjutnya.



Data Describe Numerical

```
[ ] # Memisahkan kolom numerik (nums) dan kategorikal (cats)
nums = df.select_dtypes(include=['float64', 'int64', 'int8'])
cats = df.select_dtypes(include=['object', 'bool'])

# Describe numerical data
nums.describe()
```

→

	age	balance	day	duration	pdays	pr
count	43193.000000	43193.000000	43193.000000	43193.000000	43193.000000	43193.000000
mean	40.764082	1408.555877	15.809414	258.323409	857.226240	0.5
std	10.512640	3017.708989	8.305970	258.162006	303.431026	2.1
min	18.000000	1.000000	1.000000	0.000000	1.000000	0.0
25%	33.000000	137.000000	8.000000	103.000000	999.000000	0.0
50%	39.000000	482.000000	16.000000	180.000000	999.000000	0.0
75%	48.000000	1424.000000	21.000000	318.000000	999.000000	0.0
max	95.000000	102128.000000	31.000000	4918.000000	999.000000	275.0

Penjelasan :

- Penyebaran umur cukup luas, dengan nilai umur 25% dari data berada di sekitar 33 tahun dan 75% berada di sekitar 48 tahun. Ini menunjukkan bahwa sebagian besar pelanggan berusia antara 33 hingga 48 tahun.
- Distribusi hari tidak menunjukkan adanya pola yang sangat mencolok, tetapi angka rata-rata mendekati tengah bulan menunjukkan distribusi yang relatif merata.
- Durasi pada kuartil ke-25 adalah 103 detik dan kuartil ke-75 adalah 318 detik. Durasi rata-rata menunjukkan bahwa banyak kontak memiliki durasi yang lebih lama daripada median.
- 99% data memiliki nilai pdays sebesar 999, menunjukkan bahwa banyak pelanggan belum pernah dihubungi sebelumnya. Nilai 999 tampaknya menjadi indikator khusus untuk "belum dihubungi".



Data Describe Categorical

```
# Describe categorical data
cats.describe()
```

	job	marital	education	default	housing	loan	contact	month	campaign	poutcome	y
count	43193	43193	43193	43193	43193	43193	43193	43193	43193	43193	43193
unique	11	3	3	2	2	2	2	12	47	4	2
top	blue-collar	married	secondary	False	True	False	cellular	may	1	nonexistent	False
freq	9278	25946	23131	42411	24292	36086	40499	13192	16742	35286	38172

Penjelasan :

- Sebagian besar pelanggan memiliki pekerjaan di sektor blue-collar, menunjukkan bahwa kategori ini adalah yang paling umum dalam dataset ini.
- Pelanggan yang sudah menikah adalah yang paling umum. Ini menunjukkan bahwa status perkawinan married lebih dominan dibandingkan single atau divorced.
- Pendidikan tingkat secondary adalah yang paling umum, menunjukkan bahwa sebagian besar pelanggan memiliki pendidikan menengah.
- Mayoritas pelanggan tidak memiliki kredit macet (False), yang menunjukkan bahwa pelanggan cenderung memiliki catatan kredit yang baik.
- Sebagian besar pelanggan memiliki pinjaman rumah (True), yang menunjukkan bahwa pinjaman rumah adalah fitur umum di antara pelanggan.

- Banyak pelanggan tidak memiliki pinjaman pribadi (False), menunjukkan bahwa pinjaman pribadi kurang umum dibandingkan pinjaman rumah.
- Kontak melalui cellular adalah yang paling umum, menunjukkan bahwa sebagian besar interaksi dilakukan melalui telepon seluler.
- Bulan May adalah yang paling umum untuk interaksi, menunjukkan bahwa kampanye telemarketing cenderung lebih sering terjadi pada bulan tersebut.
- Jumlah kontak per kampanye bervariasi, dengan nilai 1 sebagai yang paling sering terjadi. Ini mungkin menunjukkan bahwa banyak pelanggan dihubungi hanya sekali dalam kampanye.
- Banyak pelanggan memiliki hasil kampanye yang tidak ada sebelumnya (nonexistent), menunjukkan bahwa sebagian besar interaksi tidak memiliki hasil kampanye yang terdokumentasi.
- Sebagian besar pelanggan tidak berlangganan produk (False), menunjukkan bahwa keputusan akhir untuk berlangganan adalah hasil yang kurang umum dibandingkan tidak berlangganan.



INNOVATORS
INFORMATION INOVATORS



Descriptive Statistics

Kesimpulan

- Data Kategorikal (cats) cocok untuk analisis frekuensi, visualisasi seperti bar chart, dan proses encoding untuk model machine learning.
- Data Numerik (nums) memungkinkan kita untuk menganalisis pola distribusi, mendekripsi outlier, dan melakukan transformasi seperti normalisasi atau scaling untuk model machine learning.

Pemisahan antara data numerik dan kategorikal ini penting dalam proses analisis dan preprocessing, karena setiap jenis data memerlukan teknik analisis dan persiapan yang berbeda.



INFOLVATORS
INFORMATION INOVATORS

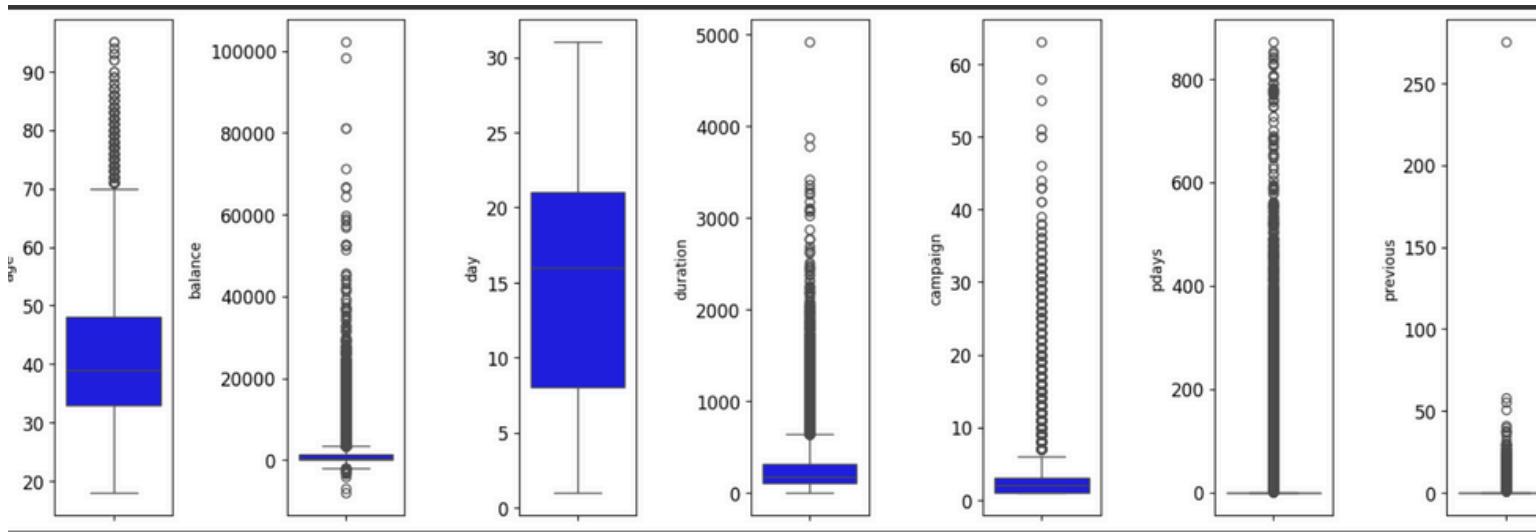




UNIVARIATE ANALYSIS



Box Plot



AGE: RENTANG USIA BERKISAR ANTARA SEKITAR 20 HINGGA 90 TAHUN, DENGAN MAYORITAS USIA TERLETAK ANTARA SEKITAR 30 DAN 50 TAHUN. ADA BEBERAPA OUTLIER DI SISI ATAS, MENUNJUKKAN ADANYA INDIVIDU YANG BERUSIA DI ATAS 70 TAHUN.

BALANCE: VARIABEL INI MENUNJUKKAN DISTRIBUSI SALDO YANG SANGAT MIRING (SKEWED), DENGAN BANYAK OUTLIER DI SISI ATAS, MENUNJUKKAN ADANYA INDIVIDU DENGAN SALDO YANG JAUH LEBIH TINGGI DIBANDINGKAN MAYORITAS. NILAI MEDIAN SALDO SANGAT RENDAH.

DAY: DISTRIBUSI "HARI" RELATIF LEBIH SERAGAM, DENGAN RENTANG HARI BERKISAR ANTARA 1 HINGGA 30. TIDAK ADA OUTLIER YANG SIGNIFIKAN.

DURATION: BOX PLOT UNTUK DURASI MENUNJUKKAN BAHWA SEBAGIAN BESAR DURASI PERCAKAPAN SANGAT RENDAH, DENGAN BEBERAPA OUTLIER SIGNIFIKAN YANG MENUNJUKKAN ADANYA PERCAKAPAN YANG SANGAT PANJANG.

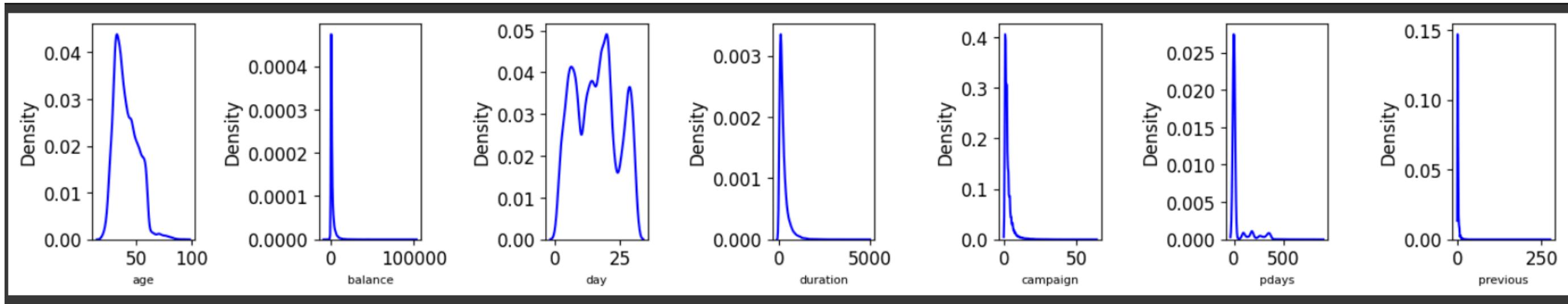
CAMPAIN: VARIABEL INI MENUNJUKKAN BAHWA SEBAGIAN BESAR INDIVIDU HANYA DIHUBUNGI BEBERAPA KALI SELAMA KAMPANYE, DENGAN BANYAK OUTLIER DI SISI ATAS, MENUNJUKKAN BEBERAPA INDIVIDU YANG DIHUBUNGI LEBIH DARI 8 KALI.

PDAYS: "PDAYS" MEMILIKI BANYAK OUTLIER DI SISI ATAS, MENUNJUKKAN BAHWA SEBAGIAN BESAR INDIVIDU DIHUBUNGI DALAM WAKTU DEKAT ATAU TIDAK DIHUBUNGI SAMA SEKALI, DENGAN BEBERAPA INDIVIDU YANG DIHUBUNGI DALAM WAKTU YANG LEBIH LAMA.

PREVIOUS: BOX PLOT INI MENUNJUKKAN BAHWA SEBAGIAN BESAR INDIVIDU TIDAK DIHUBUNGI SEBELUMNYA, DENGAN BEBERAPA OUTLIER YANG MENUNJUKKAN INDIVIDU YANG DIHUBUNGI HINGGA LEBIH DARI 50 KALI.



KDE Plot



AGE: DISTRIBUSI UMUR CONDONG KE KANAN, YANG BERARTI SEBAGIAN BESAR INDIVIDU TERKONSENTRASI PADA USIA YANG LEBIH MUDA, DENGAN SEDIKIT INDIVIDU YANG BERUSIA LEBIH TUA.

BALANCE: DISTRIBUSI SALDO SANGAT CONDONG KE KANAN. SEBAGIAN BESAR INDIVIDU MEMILIKI SALDO YANG LEBIH RENDAH, DENGAN BEBERAPA MEMILIKI SALDO YANG JAUH LEBIH TINGGI.

DAY: DISTRIBUSI VARIABEL "HARI," YANG MUNGKIN MEWAKILI HARI DALAM BULAN, MENUNJUKKAN BEBERAPA PUNCAK, YANG MENGINDIKASIKAN BAHWA HARI-HARI TERTENTU LEBIH SERING MUNCUL DALAM DATA.

DURATION: DISTRIBUSI DURASI JUGA SANGAT CONDONG KE KANAN, MENUNJUKKAN BAHWA SEBAGIAN BESAR DURASI BERSIFAT PENDEK, DENGAN SEDIKIT KASUS DURASI YANG LEBIH LAMA. DISTRIBUSI DURASI JUGA SANGAT CONDONG KE KANAN, MENUNJUKKAN BAHWA SEBAGIAN BESAR DURASI BERSIFAT PENDEK, DENGAN SEDIKIT KASUS DURASI YANG LEBIH LAMA.

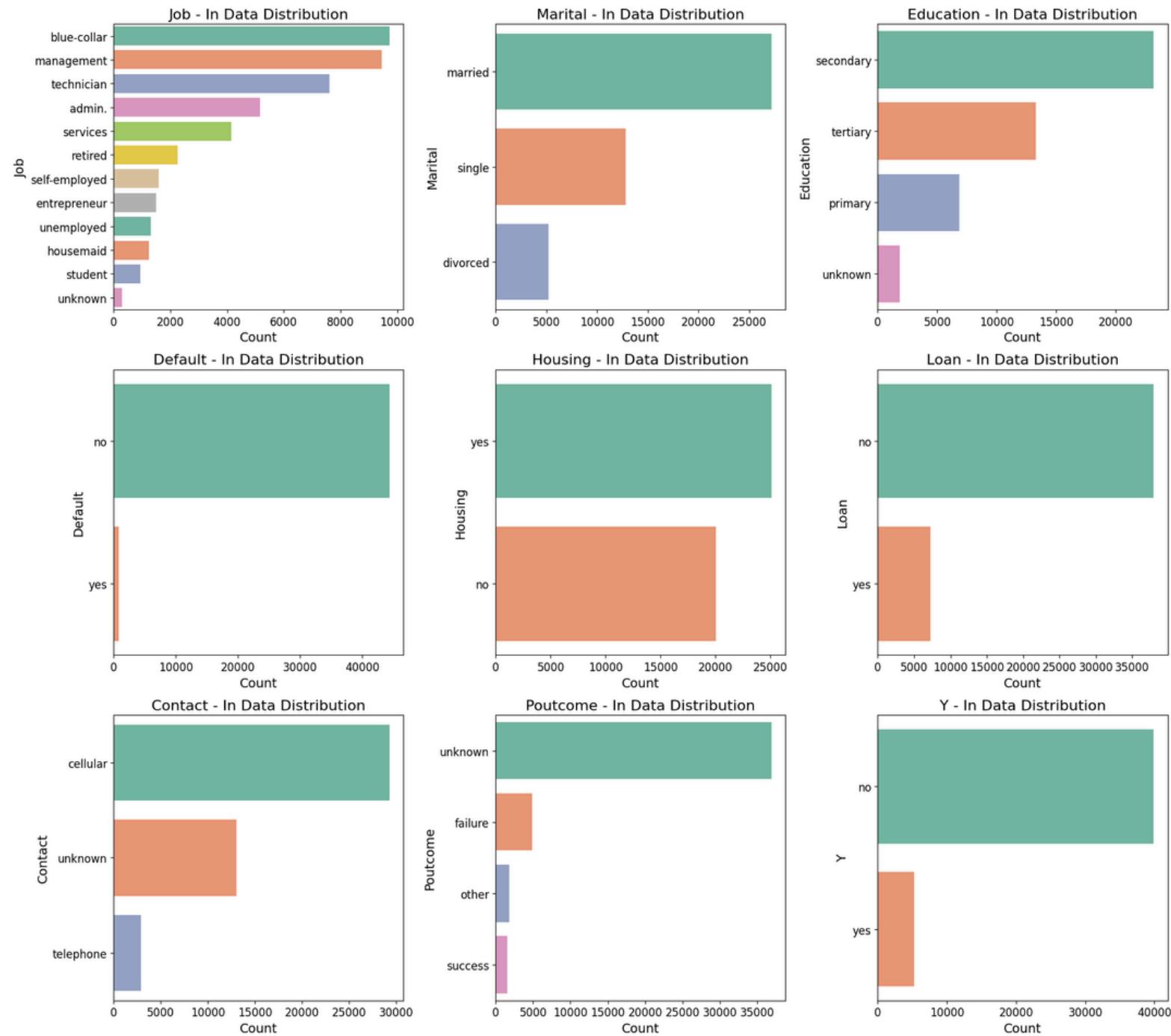
CAMPAIN: VARIABEL INI KEMUNGKINAN MEWAKILI JUMLAH KONTAK YANG DILAKUKAN SELAMA KAMPANYE, DAN GRAFIKNYA JUGA MENUNJUKKAN CONDONG KE KANAN, MENGINDIKASIKAN BAHWA SEBAGIAN BESAR INDIVIDU HANYA DIHUBUNGI BEBERAPA KALI, DENGAN BEBERAPA OUTLIER DI MANA INDIVIDU DIHUBUNGI LEBIH SERING.

PDAYS: DISTRIBUSI "PDAYS" MENUNJUKKAN BAHWA SEBAGIAN BESAR NILAINYA TERKONSENTRASI DI SEKITAR 0, DENGAN SEDIKIT EKOR YANG MEMANJANG KE NILAI YANG LEBIH TINGGI. INI BISA MENGINDIKASIKAN BAHWA SEBAGIAN BESAR INDIVIDU TIDAK DIHUBUNGI SEBELUMNYA ATAU DIHUBUNGI SUDAH CUKUP LAMA.

PREVIOUS: VARIABEL "SEBELUMNYA," YANG MUNGKIN MEWAKILI JUMLAH KONTAK SEBELUMNYA, JUGA SANGAT CONDONG KE KANAN, MENUNJUKKAN BAHWA SEBAGIAN BESAR INDIVIDU HANYA DIHUBUNGI BEBERAPA KALI SEBELUMNYA



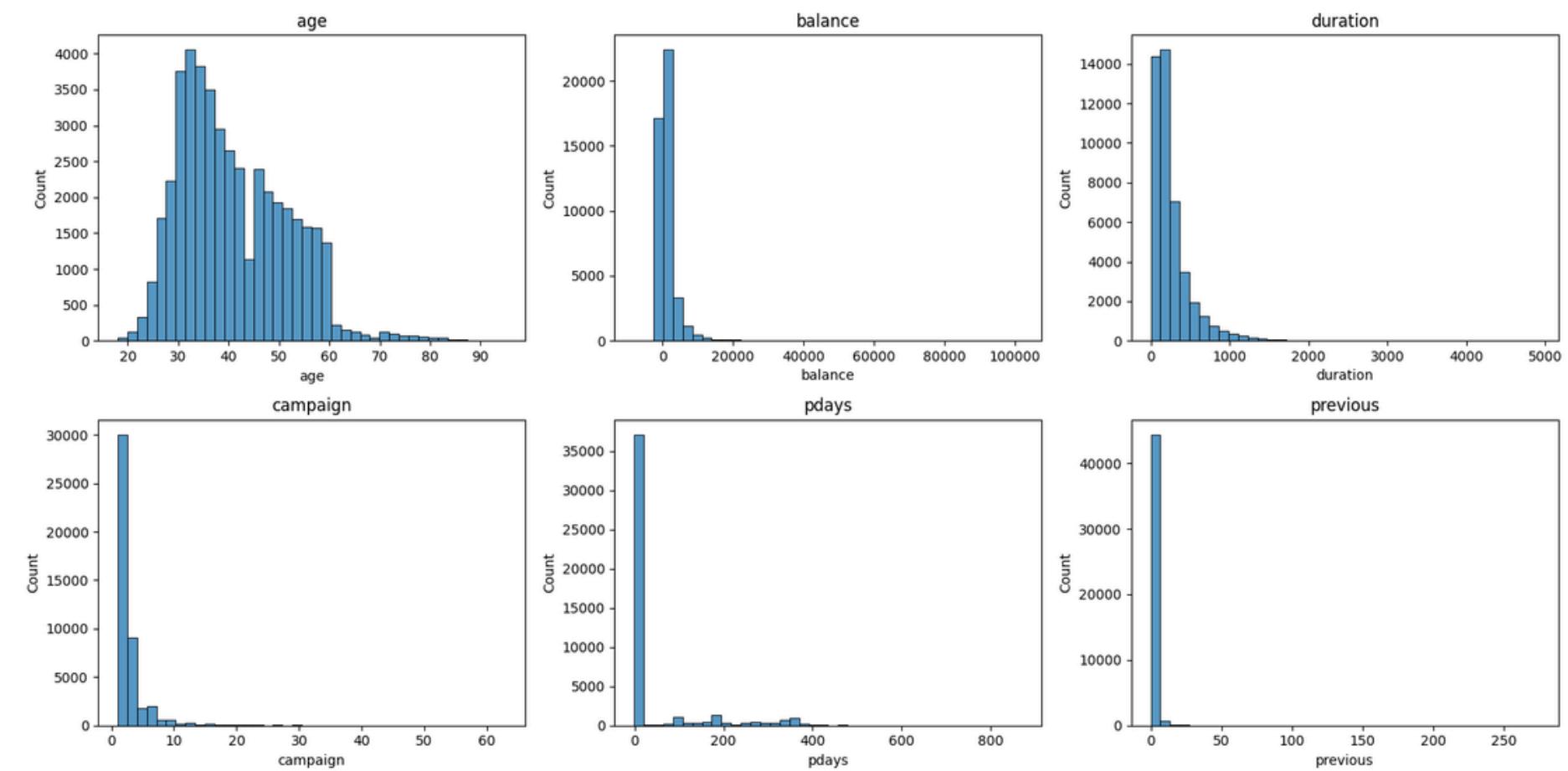
Bar Plot



- **JOB - IN DATA DISTRIBUTION (PEKERJAAN):** GRAFIK INI MENUNJUKKAN DISTRIBUSI PEKERJAAN DI ANTARA INDIVIDU DALAM DATASET. PEKERJAAN PALING UMUM ADALAH "BLUE-COLLAR" (PEKERJA KERAH BIRU), DIIKUTI OLEH "MANAGEMENT" DAN "TECHNICIAN". PEKERJAAN SEPERTI "STUDENT" DAN "UNKNOWN" MEMILIKI FREKUENSI YANG PALING RENDAH.
- **MARITAL - IN DATA DISTRIBUTION (STATUS PERKAWINAN):** GRAFIK INI MENUNJUKKAN BAHWA MAYORITAS INDIVIDU DALAM DATASET INI BERSTATUS "MARRIED" (MENIKAH), DENGAN JUMLAH YANG LEBIH SEDIKIT BERSTATUS "SINGLE" (LAJANG) DAN "DIVORCED" (BERCERAI).
- **EDUCATION - IN DATA DISTRIBUTION (PENDIDIKAN):** SEBAGIAN BESAR INDIVIDU MEMILIKI TINGKAT PENDIDIKAN "SECONDARY" (MENENGAH), DIIKUTI OLEH "TERTIALY" (TINGGI) DAN "PRIMARY" (DASAR). ADA SEBAGIAN KECIL YANG MEMILIKI STATUS PENDIDIKAN "UNKNOWN" (TIDAK DIKETAHUI).
- **DEFAULT - IN DATA DISTRIBUTION (KREDIT MACET):** SEBAGIAN BESAR INDIVIDU TIDAK MEMILIKI STATUS KREDIT MACET ("NO"), DENGAN HANYA SEDIKIT YANG MEMILIKI KREDIT MACET ("YES").
- **HOUSING - IN DATA DISTRIBUTION (KEPEMILIKAN RUMAH):** SEBAGIAN BESAR INDIVIDU MEMILIKI RUMAH ("YES"), SEMENTARA SISANYA TIDAK MEMILIKI RUMAH ("NO").
- **LOAN - IN DATA DISTRIBUTION (PINJAMAN):** KEBANYAKAN INDIVIDU TIDAK MEMILIKI PINJAMAN ("NO"), DENGAN SEBAGIAN KECIL MEMILIKI PINJAMAN ("YES").
- **CONTACT - IN DATA DISTRIBUTION (KONTAK):** SEBAGIAN BESAR KONTAK DILAKUKAN MELALUI "CELLULAR" (TELEPON SELULER), DIIKUTI OLEH "UNKNOWN" (TIDAK DIKETAHUI), DAN PALING SEDIKIT MELALUI "TELEPHONE" (TELEPON RUMAH).
- **POUTCOME - IN DATA DISTRIBUTION (HASIL SEBELUMNYA):** GRAFIK INI MENUNJUKKAN HASIL DARI KAMPANYE PEMASARAN SEBELUMNYA. SEBAGIAN BESAR HASIL TIDAK DIKETAHUI ("UNKNOWN"), DENGAN SEBAGIAN KECIL MENGALAMI "FAILURE" (GAGAL) DAN "SUCCESS" (SUKSES).
- **Y - IN DATA DISTRIBUTION (RESPONS TARGET):** GRAFIK INI MENGAMBARKAN RESPON TARGET, YAITU APAKAH INDIVIDU TERSEBUT SETUJU UNTUK BERLANGGANAN PRODUK YANG DITAWARKAN ATAU TIDAK. SEBAGIAN BESAR MENJAWAB "NO" (TIDAK SETUJU), SEMENTARA SISANYA MENJAWAB "YES" (SETUJU).



Bar Plot

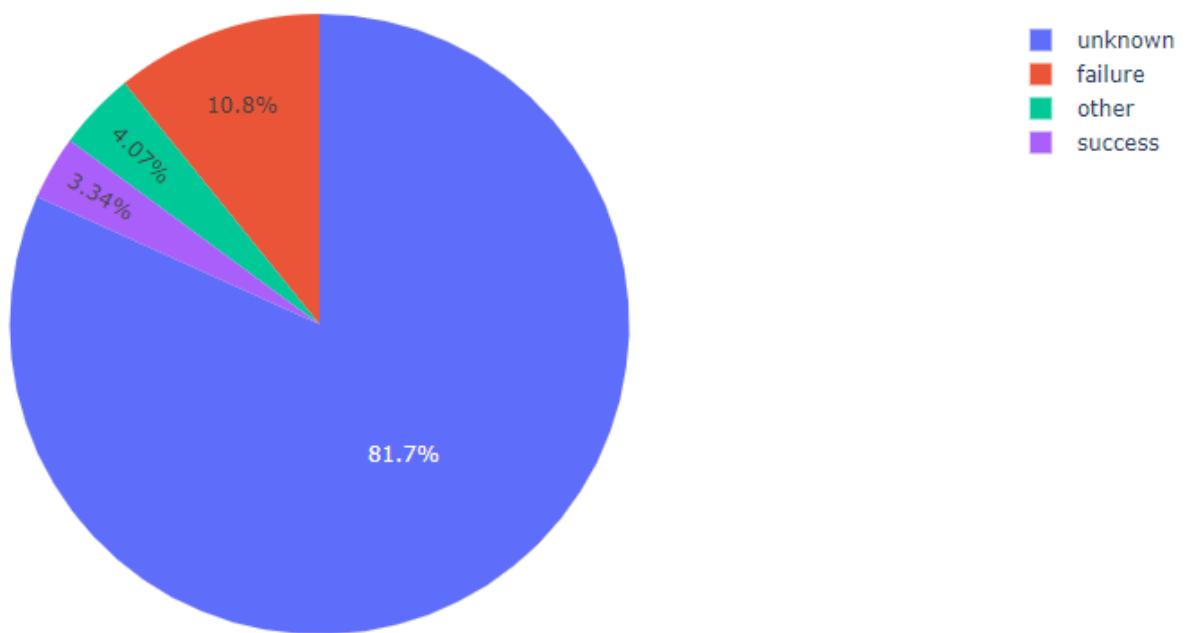


- **AGE (USIA):** HISTOGRAM UNTUK AGE MENUNJUKKAN BAHWA SEBAGIAN BESAR INDIVIDU DALAM DATASET BERUSIA ANTARA 20 HINGGA 60 TAHUN, DENGAN PUNCAK SEKITAR USIA 30-40 TAHUN. DISTRIBUSINYA CENDERUNG MIRING KE KANAN, YANG BERARTI ADA LEBIH SEDIKIT INDIVIDU YANG LEBIH TUA (DI ATAS 60 TAHUN).
- **BALANCE (SALDO):** HISTOGRAM BALANCE SANGAT MIRING KE KANAN, MENUNJUKKAN BAHWA SEBAGIAN BESAR INDIVIDU MEMILIKI SALDO BANK YANG RENDAH, DENGAN SEDIKIT YANG MEMILIKI SALDO SANGAT TINGGI. MAYORITAS DATA TERKUMPUL DI SEKITAR NILAI SALDO RENDAH, DAN SANGAT SEDIKIT YANG MEMILIKI SALDO DI ATAS 20.000.
- **DURATION (DURASI):** HISTOGRAM DURATION JUGA MIRING KE KANAN. SEBAGIAN BESAR DURASI RENDAH, MENUNJUKKAN BAHWA MAYORITAS INTERAKSI ATAU PANGGILAN (JIKA INI ADALAH DATA TELEMARKETING) BERLANGSUNG SINGKAT, DENGAN PENURUNAN YANG TAJAM SEIRING MENINGKATNYA DURASI. ADA SEDIKIT KASUS DENGAN DURASI YANG SANGAT LAMA.
- **CAMPAIN (KAMPANYE):** HISTOGRAM CAMPAIGN MENUNJUKKAN BERAPA KALI SESEORANG DIHUBUNGI SELAMA KAMPANYE. SEBAGIAN BESAR ORANG DIHUBUNGI HANYA SEKALI ATAU DUA KALI, DENGAN SANGAT SEDIKIT INDIVIDU YANG DIHUBUNGI LEBIH DARI 10 KALI.
- **PDAYS:** HISTOGRAM PDAYS (KEMUNGKINAN MEWAKILI JUMLAH HARI SEJAK KLIEN TERAKHIR KALI DIHUBUNGI DARI KAMPANYE SEBELUMNYA) SANGAT MIRING, DENGAN SEBAGIAN BESAR NILAI BERADA DI 0, YANG MENUNJUKKAN BAHWA BANYAK INDIVIDU TIDAK PERNAH DIHUBUNGI SEBELUMNYA ATAU SUDAH LAMA TIDAK DIHUBUNGI. SANGAT SEDIKIT KASUS DENGAN NILAI PDAYS LEBIH DARI 0.
- **PREVIOUS (SEBELUMNYA):** HISTOGRAM PREVIOUS MENUNJUKKAN BERAPA KALI SESEORANG DIHUBUNGI DALAM KAMPANYE SEBELUMNYA. HISTOGRAM INI MENUNJUKKAN BAHWA SEBAGIAN BESAR INDIVIDU TIDAK PERNAH DIHUBUNGI SEBELUMNYA (PREVIOUS BERNILAI 0). SANGAT SEDIKIT INDIVIDU YANG DIHUBUNGI BEBERAPA KALI DI MASA LALU, YANG SESUAI DENGAN DISTRIBUSI YANG MIRING KE KANAN.



Pie Chart

Distribution of Poutcome



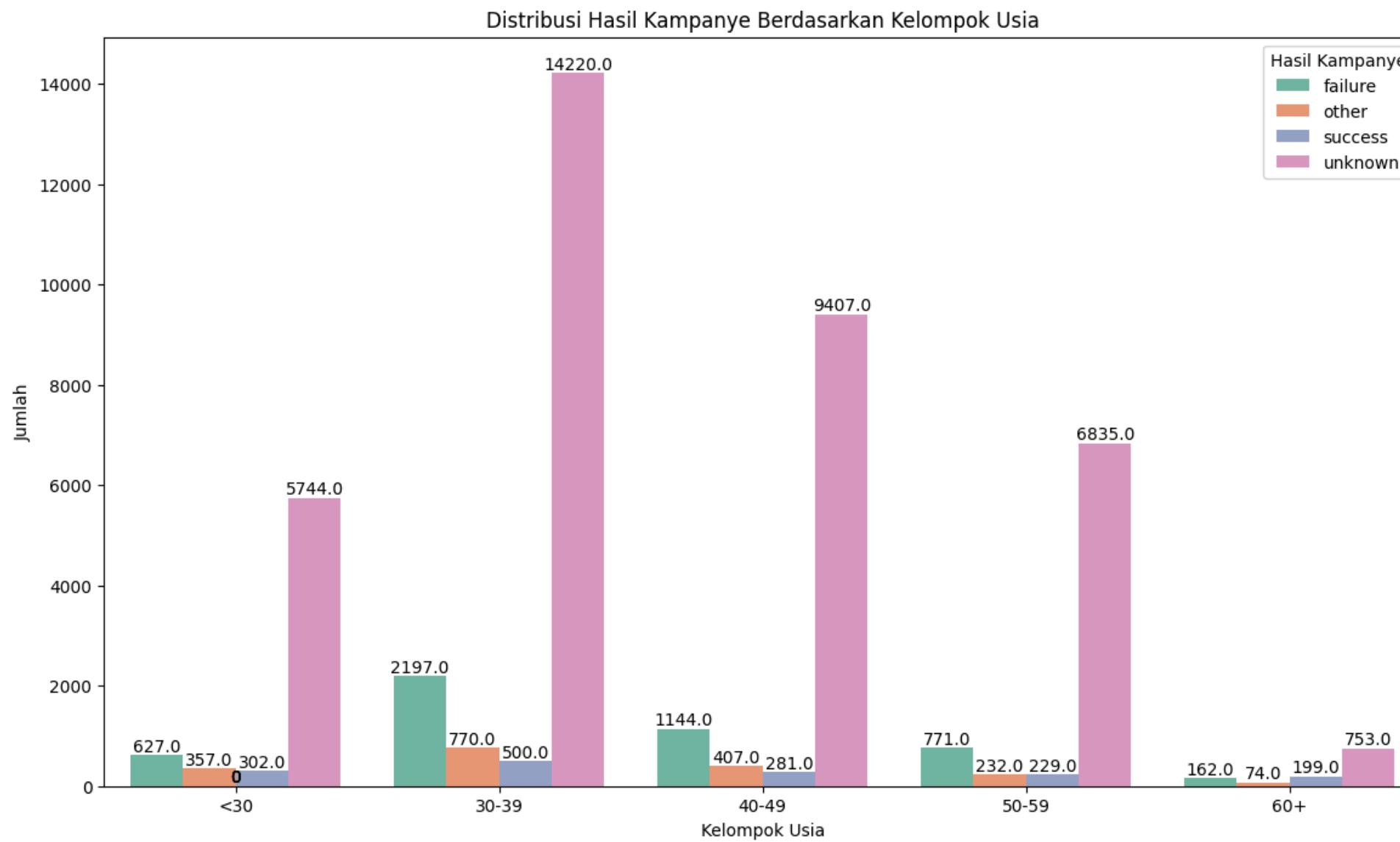
- DISTRIBUSI NILAI PADA KOLOM "POUTCOME":
 - 1. UNKNOWN (81.7%): SEBAGIAN BESAR DATA PADA KOLOM INI MEMILIKI NILAI "UNKNOWN," YANG MENUNJUKKAN BAHWA HASIL DARI KAMPANYE SEBELUMNYA TIDAK DIKETAHUI ATAU TIDAK TERCATAT. INI ADALAH KATEGORI TERBESAR DALAM DATASET.
 - 2. FAILURE (10.8%): SEBAGIAN KECIL DATA MEMILIKI NILAI "FAILURE," YANG MENANDAKAN BAHWA KAMPANYE SEBELUMNYA TIDAK BERHASIL.
 - 3. OTHER (4.07%): ADA JUGA SEBAGIAN KECIL DATA YANG TERMASUK DALAM KATEGORI "OTHER," YANG MENCAKUP HASIL-HASIL YANG TIDAK TERMASUK DALAM KATEGORI UTAMA SEPERTI "SUCCESS" ATAU "FAILURE."
 - 4. SUCCESS (3.34%): JUMLAH TERKECIL DARI DATA INI MEMILIKI NILAI "SUCCESS," YANG MENUNJUKKAN BAHWA HANYA SEBAGIAN KECIL KAMPANYE SEBELUMNYA YANG BERHASIL.
- DISTRIBUSI INI MENUNJUKKAN BAHWA KEBANYAKAN DATA TERKAIT HASIL KAMPANYE SEBELUMNYA TIDAK DIKETAHUI, DENGAN HANYA SEBAGIAN KECIL YANG BERHASIL ATAU GAGAL. INI DAPAT MEMBERIKAN WAWASAN PENTING SAAT MENGANALISIS EFEKTIVITAS KAMPANYE ATAU MEMBUAT KEPUTUSAN BERDASARKAN DATA YANG ADA.



INFLVATORS
INFORMATION INOVATORS



Distribusi Poutcome berdasarkan group umur



Penjelasan Distribusi Hasil Kampanye Berdasarkan Kelompok Usia:

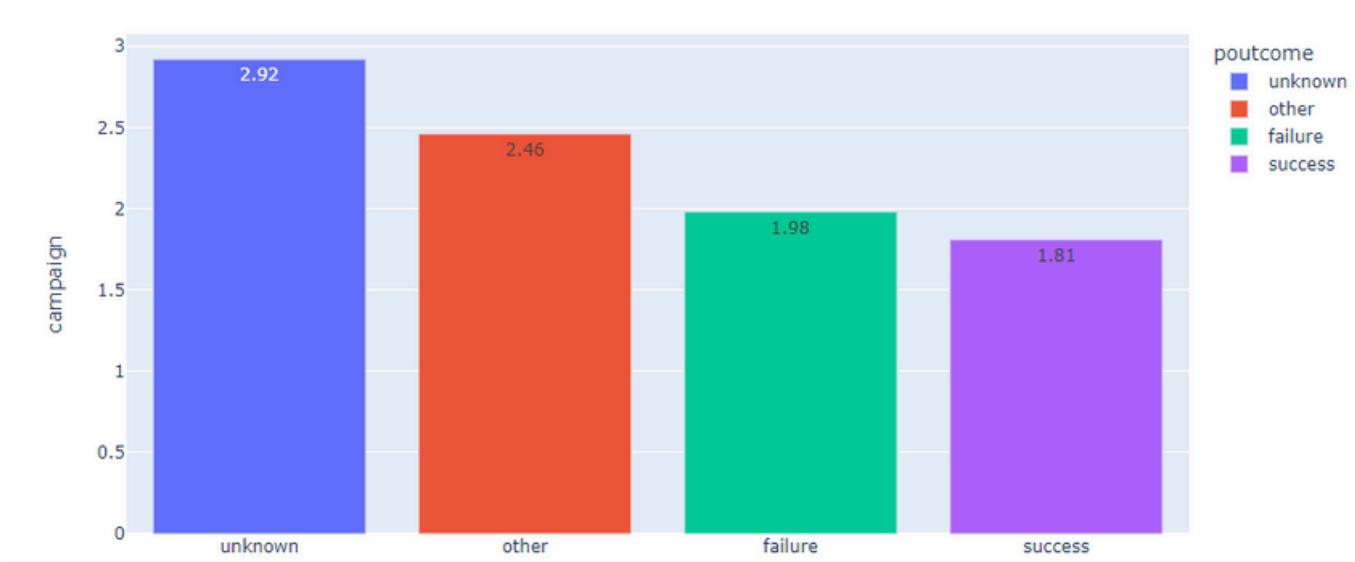
- <30 (Di bawah 30 tahun):
 - Sebagian besar hasil kampanye dalam kelompok usia ini tidak diketahui ("unknown"), dengan jumlah sekitar 5744 kasus. Ada juga beberapa kasus "failure" (gagal) dan "other" (lain-lain), tapi jumlahnya jauh lebih sedikit. Hasil "success" (berhasil) sangat sedikit.
- 30-39 tahun:
 - Kelompok usia ini memiliki jumlah kasus "unknown" terbesar, yaitu sekitar 14220. Kampanye yang berakhir dengan "failure" dan "other" juga ada, tetapi jauh lebih sedikit. Hasil "success" (berhasil) juga sangat jarang.
- 40-49 tahun:
 - Lagi-lagi, hasil "unknown" mendominasi dengan sekitar 9407 kasus. Ada beberapa kasus "failure" dan "other," tetapi jumlahnya tidak sebanyak "unknown". Kampanye yang berhasil (success) juga ada, tapi sangat sedikit.
- 50-59 tahun:
 - Di kelompok usia ini, hasil "unknown" juga yang paling banyak, sekitar 6835 kasus. Kasus "failure" dan "other" ada, namun sedikit. Sama seperti kelompok usia lain, jumlah kampanye yang berhasil (success) sangat kecil.
- 60+ (Di atas 60 tahun):
 - Kelompok usia ini memiliki jumlah kasus yang paling sedikit dibandingkan kelompok usia lainnya. Sebagian besar hasilnya "unknown," dengan sedikit kasus "failure," "other," dan hampir tidak ada yang "success".

Dari grafik ini, kita bisa melihat bahwa sebagian besar hasil kampanye, terutama pada kelompok usia yang lebih muda, tidak diketahui ("unknown"). Sementara itu, kampanye yang benar-benar berhasil (success) sangat jarang terjadi di semua kelompok usia.

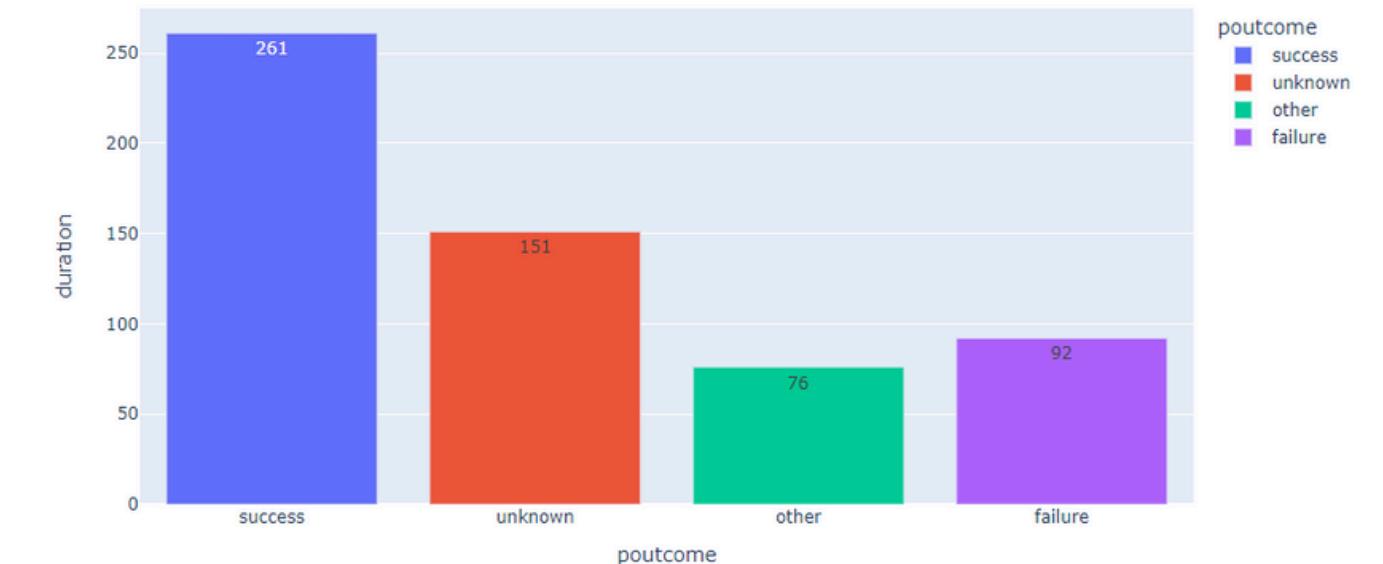


Distribusi Poutcome berdasarkan group umur

Rata-rata Jumlah Banyaknya Campaign Telepon untuk Berbagai Hasil Kampanye



Rata-rata Durasi Panggilan Telepon untuk Setiap Hasil Kampanye



- Rata-rata Jumlah Panggilan Telepon untuk Berbagai Hasil Kampanye:
- Unknown (2.92): Kampanye dengan hasil "unknown" memiliki rata-rata panggilan telepon tertinggi, yaitu sekitar 2.92 kali panggilan. Ini mungkin menunjukkan bahwa meskipun banyak usaha dilakukan, hasil kampanye tidak tercatat atau tidak diketahui.
- Other (2.46): Hasil kampanye yang masuk kategori "other" memiliki rata-rata sekitar 2.46 panggilan. Ini menunjukkan usaha yang cukup signifikan dilakukan, tetapi hasilnya tidak masuk dalam kategori utama seperti "success" atau "failure."
- Failure (1.98): Kampanye yang berakhir dengan "failure" rata-rata melibatkan 1.98 panggilan telepon. Ini menunjukkan usaha yang lebih rendah dibandingkan dengan kategori "unknown" dan "other".
- Success (1.81): Kampanye yang berhasil ("success") rata-rata melibatkan jumlah panggilan telepon paling sedikit, yaitu sekitar 1.81 kali. Ini menunjukkan bahwa kampanye yang sukses cenderung membutuhkan lebih sedikit usaha dalam hal jumlah panggilan telepon.
- Grafik ini menggambarkan hubungan antara jumlah panggilan telepon dan hasil kampanye, di mana lebih banyak panggilan telepon tidak selalu berbanding lurus dengan keberhasilan kampanye. Kampanye yang sukses justru melibatkan panggilan yang lebih sedikit.

- Rata - rata durasi panggilan telepon untuk setiap hasil kampanye
- Success (261) Kampanye dengan hasil sukses memiliki rata rata durasi panggilan selama 261 detik, yang ini menunjukkan bahwa lamanya durasi telepon pada suatu kampanye memiliki rate kesuksesan
- Unknown (151) Kampanye dengan hasil unknown memiliki rata - rata durasi panggilan 151 yang hal ini menunjukkan bahwa durasi dengan rata rata tersebut belum terdata atau tidak tercatat hasil kampanye nya.
- Other (76) kampanye dengan hasil other memiliki rata rata durasi panggilan 76 detik dimana hasil ini cukup berusaha namun hasil yang didapat tidak termasuk dalam kategori gagal atau sukses
- Failure (92) kampanye dengan hasil other memiliki rata rata durasi rendah yaitu 96 dimana hal ini menunjukkan bahwa kampanye dengan durasi yang sedikit berkemungkinan atau cenderung gagal dalam hasil kampanye.



Distribusi Poutcome berdasarkan group umur



- Berikut merupakan gambaran dari grafik pie chart
- "distribusi data konsumen yang berlangganan deposito jangka panjang. Dimana hanya 11.7% saja yang yes atau berlangganan jangka panjang dengan pelanggan sebanyak 5289 dari total 45211 data pelanggan yang menunjukan hasil tersebut merupakan hasil yang cukup rendah".
- Dari grafik tersebut didapatkan hasil dengan succes Poutcome tertinggi ada di kategori usia 30-39 namun jika dibandingkan dengan failure yang terjadi pada kelompok umur tersebut, maka kelompok umur 60+ memiliki succes rate yang jauh lebih tinggi mengingat perbandingan failure dengan succes tidak terlalu jauh atau doimnan seperti pada distribusi kelompok umur 30 - 39.





MULTIVARIATE ANALYSIS



Multivariate Analysis

	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

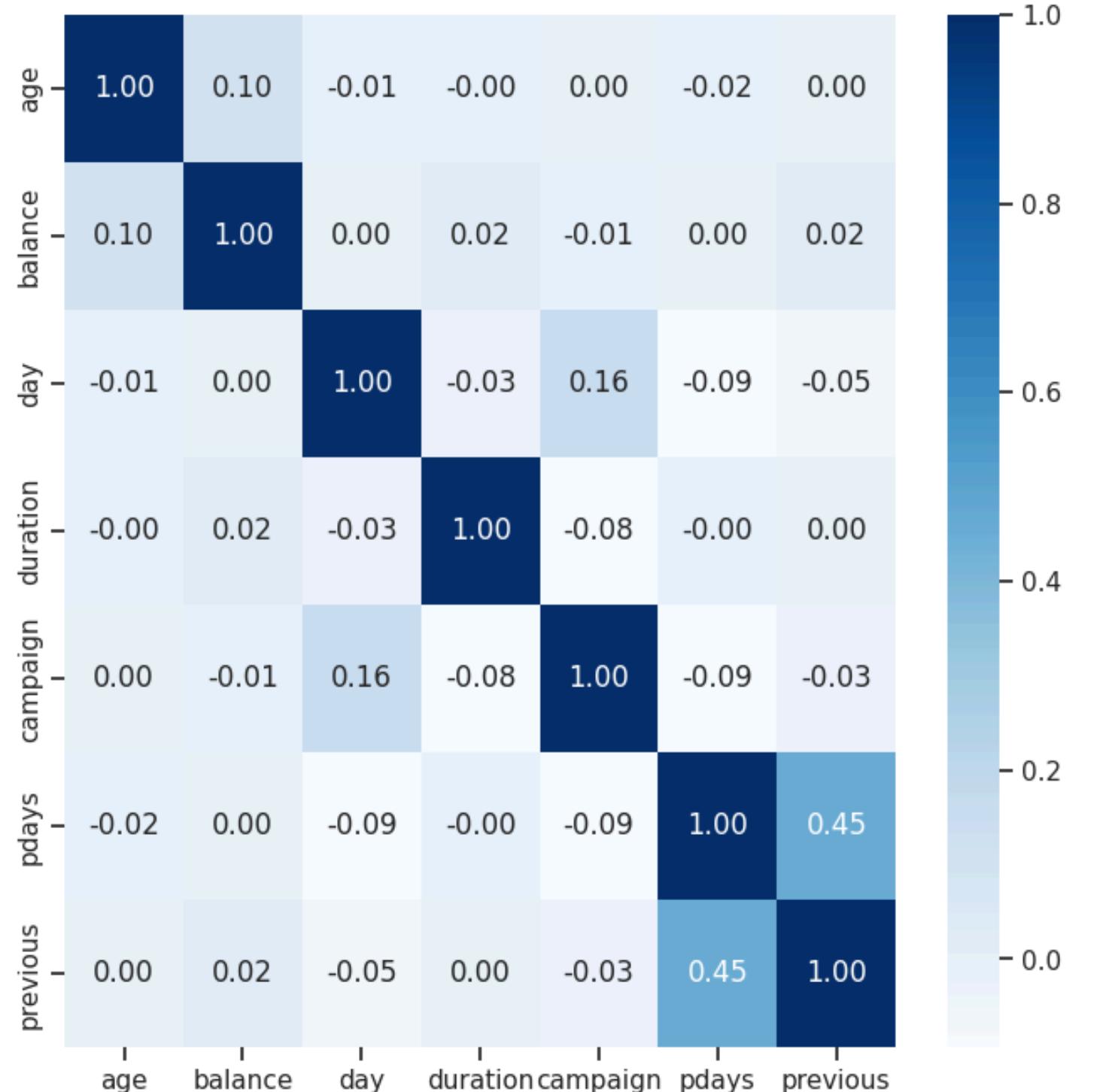
Ini merupakan table dari Matrix pada Corelation pada
Dataset Bank Marketing Target



INFOLVATORS
INFORMATION INOVATORS



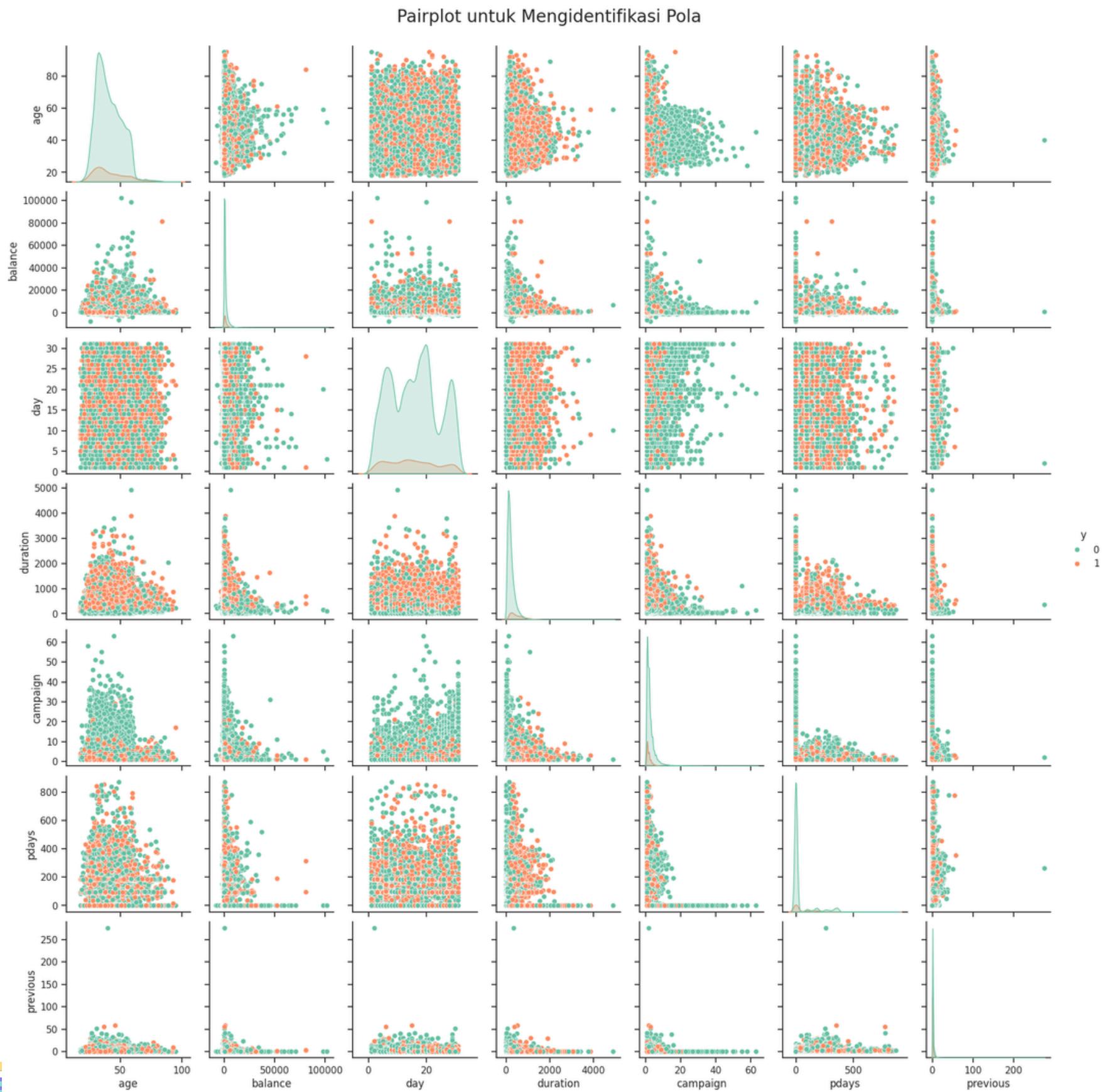
Multivariate Analysis



Dari corelation heatmap disamping didapatkanlah bahwa

- pdays dan previous:** 2 Features ini memiliki nilai korelasi yang positif dan cukup tinggi dibanding dengan features lainnya. acdan merupakan korelasi positif. atau bisa disebut **Strong potensial features**.
- day,pdays dan campaign:** Mmeiliki korelasi atau hubungan paling rendah atau korelasi negatif paling tinggi.
- Campaing :** Taret analisis ini adalah campiagn dimana campaing memiliki korelasi yang cukup baik dengan days. dan bisa menjadi **decent potensial features**.





Multivariate Analysis

Dari Grafik yang disamping dapat dilihat bahwa :

1. Data didominasi oleh warna biru dengan nilai 0 atau bisa disebut nilai no pada kolom y. yang berarti pelanggan berdeposito jangka panjang lebih sedikit dibanding yang tidak berdeposito jangka panjang





BUSINESS INSIGHT



Business Insight

Berikut insight bussines yang didapat dan Rekomendasi yang dapat diberikan:

Insight :

1.Pada Visualiasi Data yang berjudul " Rata - rata jumlah Campaign terhadap hasil poutcome "

Hasil : Dari hasil yang sudah ditampilkan didapat bahwa rata rata campaign dengan lebih dari 1 atau banyak nya 2-3 tidak menjamin succes rate, justru succes rate dari Poutcome didapatkan dari Campaign dengan hanya sekali atau dibawah 2 kali Campaign.

2.Pada Visualisasi Data yang berjudul " Rata - rata jumlah durasi panggilan telpon terhadap hasil Poutcome"

Hasil : Diketahui berdasarkan grafik durasi telpon dengan rata rata durasi 261 detik itu merupakan succes rate dibanding dengan durasi telpon dibawah 261 detik.

3.Pada Visualisasi Data yang berjudul "Distribusi Data Konsumen yang Berlangganan pada Deposito Jangka Panjang"

Hasil : Diketahui pelanggan yang berlangganan deposito jangka panjang hanya 11.7% dari 45211 data pelanggan, yang berarti ada 5289 pelanggan yang berlangganan deposito jangka panjang.

4.Pada Visualisasi Data yang berjudul "Distribusi pelanggan yang berlangganan sesuai kelompok usia"

Hasil : Didapatkan hasil dengan succes Poutcome tertinggi ada di kategori usia 30-39 namun jika dibandingkan dengan failure yang terjadi pada kelompok umur tersebut, maka kelompok umur 60+ memiliki succes rate yang jauh lebih tinggi mengingat perbandingan failure dengan succes tidak terlalu jauh atau dominan seperti pada distribusi kelompok umur 30 - 39.



INNOVATORS
INFORMATION INOVATORS



Business Insight

Rekomendasi :

- **Korelasi Insight 1 dan 2 :** " Didapatkan bahwa untuk meningkatkan rate keberhasilan atau success rate kita bisa melakukan campaign yang sedikit atau hanya dengan sekali . namun dengan durasi yang panjang untuk meningkatkan success rate deposito jangka panjang".
- **Korelasi Insight 3 dan 4 :** " Didapatkan bahwa untuk meningkatkan rata keberhasil , kita bisa melakukan pengelompokan umur untuk identifikasi mana yang lebih memungkin succes rate, berdasarkan hasil grafik menunjukan rate yang lebih tinggi pada pelanggan di umur 60+ oleh karena itu perlu memberikan perhatian khusus terhadap distribusi kelompok dengan umur tersebut.



GIT

Menggunakan Github sebagai media untuk melakukan kolaborasi antar anggota tim , untuk mempermudah pengerajan Google Collabs



Dengan langkah langkah

1. Membuat Repository untuk menyimpan File yang telah dibuat
2. Melakukan Uploading Pengerajan yang telah dikerjakan
3. Membuat File Readme yang berisi summary atau rangkuman dari yang sudah dikerjakan.

Berikut link Github : https://github.com/aizenciel/EDA_Infolvators





SEKIAN TERIMAKASIH

