Nama Kelompok: peerData
Stage: Stage 1
Mentor: Mas Ridho Aryo Pratama

Pembagian tugas di stage ini:

1. Muhammad Iqbal : Markdown insights, business recommendation, laporan project pdf, notulensi
2. A Nahda La Roiba : Visualization (uni-multivariate), readme and git repo
3. Ilham Maulana : Markdown insights, business recommendation
4. Firstandy Edgar Dhafa : Readme and git repo
5. Clara Natalie S : Visualization (uni-multivariate)
6. R. Rani Indah Salamah : Visualization (uni-multivariate), laporan project pdf
7. Eka Apriyani : Visualization (uni-multivariate), notulensi
8. Sekar Ayu Larasati : Info dan description table, Google Colaboratory, laporan project pdf

Poin pembahasan:

1. Presenting progress and role
2. Descriptive data analysis
3. Univariate visualization
4. Multivariate visualization
5. Insights
6. Business recommendations
7. Coding simplification

Hasil Diskusi:

1. Orang yang bertanggung jawab di bagian nya harus bisa menjelaskan harus bisa menjelaskan apa yang dikerjakan kepada teman lainnya dalam satu tim agar arah pemikiran sejalan
2. Untuk kolom heatmap tetap kolom yang numerik, tidak perlu encode categorical ke numerical untuk mencari heatmap yang categorical karena data categorical ada yang nominal dan ordinal. Sehingga kurang tepat untuk kolom categorical dianalisa menggunakan heatmap untuk melihat correlation, tetapi lebih tepat menggunakan probability
3. Kolom categorical walaupun terlihat seperti ordinal (contoh kasus tingkat Pendidikan) tetapi di real life nya kita perlakukan seperti nominal. Sehingga kolom categorical di encode lebih prefer menggunakan OHE (One Hot Encoding)
4. Customer yang tidak memiliki pinjaman maupun KPR memiliki probabilitas untuk deposit lebih tinggi dibandingkan dengan yang memiliki pinjaman maupun KPR

5. Feature yang memiliki std tinggi kemungkinan outlier
6. Heatmap correlation menjadi indikasi awal apakah data yang dianalisa memiliki multicollinearity (adanya hubungan antar variable bebas). Jika ada indikasi multicollinearity maka bisa ambil salah satu
7. Walaupun dalam pairplot seperti tidak ada correlation tetapi jika ada batas yang memisahkan pada setiap targetnya maka akan memudahkan ML
8. ML pada classification memiliki decision boundary (semakin terpisah datanya maka semakin mudah ML dalam mempelajari data)

Tindak lanjut:

1. Didalam coding usahakan buat function agar tidak repeat code
2. Business recommendation (tidak semua insight dibuat recommendation) dan masih ingin dianalisa kembali
3. Untuk bulan-bulan yang memiliki conversion rate tinggi, maka untuk approach customer tersebut di bulan sebelumnya (customer perlu waktu untuk memutuskan deposit atau tidak)
4. Untuk memisahkan kolom categorical dan numerical maka menggunakan df.select_dtypes(include='object').columns atau df.select_dtypes(exclude='object').columns

5. Untuk melihat feature yang null menggunakan presentase agar dapat mengetahui tingkat persentase feature null yang tinggi dan rendah
6. Untuk univariate bisa ambil salah satu atau keduanya (visualisasi maupun angka) tetapi lebih prefer menggunakan visualisasi dan dalam bentuk persentase
7. KDE dan violin plot univariate tidak perlu ditampilkan (lebih prefer menggunakan histplot)
8. Subplot univariate diganti ke percentage bukan frequensi
9. Value_counts versi percentage dibuat visualisasi (tidak perlu didetailkan angkanya lebih ke persentase saja)

10. Ambil df.pairplot() yang ada targetnya untuk memudahkan dalam menganalisa
11. Fokuskan analisa ke target
12. Kolom job atau kolom categorical lainnya yang frequensinya dikit bisa dibuat grouping saja dalam bentuk others sehingga sama frequensinya dengan value lainnya, karena kalau terlalu detail maka dapat terjadi overfitting dan tidak baik untuk model ML
13. Tanyakan penjelasan langkah-langkah untuk meningkatkan performa model ke tutor