Kelompok: Car-a-thon
Stage: 1
Mentor: Stephanie
Pukul/ Tanggal: 21.00 WIB/ 13 Mei 2022

Pembagian tugas di stage ini:

Nama: Bagus Ganjar Lugina
Tugas: Descriptive Statistics, Univariate Analysis, Business Insight, Notulen Mentoring, Source Code, Laporan Project.

Nama: M. Harun Arrasyid
Tugas: Descriptive Statistics, Multivariate Analysis, Business Insight, Notulen Mentoring, Source Code, Materi PPT.

Nama: Bernadetha Stella
Tugas: Descriptive Statistics, Multivariate Analysis, Business Insight, Notulen Mentoring, Source Code, Laporan Project.

Nama: Samuella Magdalena E
Tugas: Descriptive Statistics, Multivariate Analysis, Business Insight, Notulen Mentoring, Source Code, Materi PPT, Laporan Project.

Nama: Raihan Kurniasugianto
Tugas: Descriptive Statistics, Latar Belakang Masalah, Univariate Analysis, Business Insight, Notulen Mentoring, Source Code, Laporan Project.

Nama: M Rifqi Sarosa
Tugas: Descriptive Statistics, Univariate Analysis, Business Insight, Notulen Mentoring, Source Code, Materi PPT.

Kelompok: Car-a-thon
Stage: 1
Mentor: Stephanie
Pukul/ Tanggal: 21.00 WIB/ 13 Mei 2022

Poin pembahasan:

1. Latar belakang
   - Cari sumber eksternal untuk pendukung latar belakang, misalnya melihat kesalahan penilaian mobil berapa persen dan sefatal apa (apakah harga tinggi dijual rendah, berapa yang ga jadi beli karena harga terlalu tinggi, dsb.),
   - *Appraisal time* tidak terlalu butuh *research*, tapi bisa dicari informasi tentang penilaian harga mobil biasanya memakan waktu berapa lama,
   - Untuk latar belakang sebaiknya di-*support* dengan sumber eksternal terlebih dahulu, tetapi jika tidak ada bisa menggunakan asumsi.
2. Descriptive Statistics:
   - Untuk *present* akhir perlu dibagi 2 bagian: *overview* dan kebersihan data (*numerical* dan *categorical*, berapa baris dan kolom, kebersihan data, *missing data* dan persentasenya serta tindak lanjut yang perlu dilakukan),
   - Lebih ditekankan kembali untuk penggalian insightnya khususnya yang menjadi *highlight* saja ketika di materi PPT.
3. Univariate Analysis
   - Data numerical bisa menggunakan *box plot*, *violin plot*, *histogram* (umumnya *box plot* atau *histogram* lebih mudah dipahami awam), dikarenakan ketika presentasi akhir nanti ada beberapa stakeholder *non-technical* (tim bisnis) sehingga perlu ditampilkan distribusi dengan grafik yang mudah dipahami,
   - Melihat distribusi (dengan *histogram*), melihat *outliers* (dengan *boxplot*) dan apa yang mau dilakukan (*keep*, *cut* batas tertinggi dan terendah, dll).
4. Multivariate Analysis
   - *Heatmap* untuk melihat korelasi linear. Untuk variabel numerik yang memiliki korelasi sangat tinggi (>0.95), kemungkinan redundan. Jika diambil sebagai fitur, maka dapat menyebabkan akurasi menurun → tinjau fitur mmr. Batas suatu fitur dapat dianggap redundant adalah sebesar >0.95. Korelasi negatif maupun positif yang sangat tinggi dapat dianggap redundan,
   - Korelasi kecil dari heatmap belum tentu tidak berkorelasi dan harus drop feature, untuk melihat redundan maka selanjutnya perlu dianalisis dengan korelasi linear juga,
   - Data *categorical* biasanya menggunakan KDE Plot untuk melihat distribusi. Cek persentase *low*, *medium*, *high* untuk setiap kategorinya,
   - Dapat menggunakan *count plot* per kategori (untuk kategori yang banyak dapat dikelompokkan dulu). Bisa dibuat *threshold* untuk data categorical kurang lebih 10,
   - Untuk data *categorical* perlu melihat korelasi dari kde, tidak bisa secara angka dengan heatmap. Dengan menggunakan kde, korelasi untuk data *categorical* dapat dilihat dari semakin terpisahnya masing-masing kategori (semakin terpisah-pisah maka semakin berpengaruh).

Kelompok: Car-a-thon
Stage: 1
Mentor: Stephanie
Pukul/ Tanggal: 21.00 WIB/ 13 Mei 2022

Hasil Diskusi:

- Sales revenue bisa diukur dengan menghitung rata-rata kenaikan harga mobil setelah menggunakan model *machine learning* yang dibuat dibantu dengan sumber eksternal (Ekspektasi dari tim rakamin dalam melihat *business metrics* harus memiliki korelasi antara model yang dibuat dengan *metrics*),
- Untuk KDE plot yang memiliki beberapa puncak dapat didefinisikan bahwa kategori tersebut multimodal sehingga mendekati distribusi normal, contohnya pada kolom *condition*,
- Untuk data *categorical* dikelompokkan kembali mengacu pada banyaknya data dan hubungan antar fitur. Contoh: ambil topnya dan kelompokkan sisanya menjadi others. Intinya melihat hubungan dan tindak lanjut yang perlu dilakukan,
- Untuk melihat hubungan antar fitur, *selling price* bisa dibagi high dan low, kemudian melihat persentase tiap merk mobil di masing-masing kategori selling price. Dapat juga dengan melihat rata-rata harga setiap merk, kemudian dikelompokkan masing-masing merk yang memiliki range harga yang sama,
- Fitur *seller* dapat dikelompokkan dengan mengambil yang beberapa yang tertinggi saja (*top seller*) dan sisanya dibuat sebagai " others", sedangkan fitur selling date bisa diambil bulan atau tahunnya saja.

Tindak Lanjut:

- Mencari sumber data eksternal atau membuat suatu asumsi untuk memberikan statement yang kuat terhadap business metrics **Sales Revenue** dan **Appraisal Time**.
- Menambahkan analisis yang menampilkan overview data (jumlah row dan kolom dan persentase missing values).
- Handle lowercase dan uppercase yang ada pada setiap kolom dengan tipe data string, untuk menyederhanakan data sebelum melakukan **Univariate Analysis**.
- Ubah Univariate Analysis dengan menggunakan **boxplot**/**histogram**.
- Menganalisis dan melihat korelasi fitur kategorikal dengan menggunakan KDE Plot
- Melakukan pengelompokkan kolom yang memiliki kategori banyak berdasarkan **selling price (target)** untuk mempermudah dalam melihat distribusi. Contoh: Mengelompokkan top 10 make (merk) mobil terjual berdasarkan segmentasi harga low, med, high.