






Lanjut Tugas PDSD

<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
 Due Date	@November 21, 2024 8:40 → 10:40
 Created time	@November 19, 2024 11:27 PM
 Status	Done
 Area	College
 Priority	➔ <u>Medium</u>

- Data csv yang perlu dicleaning :

1. geolocation_dataset :

- ☐ Menghapus duplikasi data.

2. order_items_dataset :

- ☐ Mengubah tipe data kolom shipping_limit_date menjadi datetime.

3. order_reviews_dataset :

- ☐ Memperbaiki data yang null pada kolom review_comment_title.
 - ☐ Untuk yang bintang 5 dan datanya null, ganti NaN menjadi Super recomendo.
 - ☐ Untuk yang bintang 4 dan datanya null, ganti NaN menjadi recomendo.
 - ☐ Untuk yang bintang 3 dan datanya null, ganti NaN menjadi Bom (bagus)
 - ☐ Untuk yang bintang 2 dan datanya null, ganti NaN menjadi bom, mas o produto demorou muito para chegar (bagus, tapi produknya lama sekali sampainya)
 - ☐ Untuk yang bintang 1 dan datanya null, ganti menjadi Não chegou meu produto. (Produk saya tidak sampai)
- ☐ Memperbaiki data yang null pada kolom review_comment_message.

- ☐ Untuk yang bintang 1 dan datanya null, ganti NaN menjadi Nada de chegar o meu pedido. (Pesanan saya tidak pernah sampai.)
- ☐ Untuk yang bintang 2 dan datanya null, ganti NaN menjadi Demora mais para entrega (Butuh waktu lebih lama untuk pengiriman)
- ☐ Untuk yang bintang 3 dan datanya null, ganti NaN menjadi Entrega no prazo (Pengiriman tepat waktu)
- ☐ Untuk yang bintang 4 dan datanya null, ganti NaN menjadi Atendeu minha expectativa (Memenuhi harapan saya)
- ☐ Untuk yang bintang 5 dan datanya null, ganti NaN menjadi Recebi bem antes do prazo estipulado (Saya menerimanya jauh sebelum batas waktu yang ditentukan)
- ☐ Memperbaiki tipe data pada kolom review_creation_date dan review_answer_timestamp.

4. orders_dataset :

- ☐ Memperbaiki data yang null pada kolom order_approved_at.
 - Di bagian Cleaning Data, coba pake beberapa cara :
 1. Cek selisih waktu dari kolom order_purchase_timestamp dengan kolom order_approved_at. Taro di kolom tambahan (misal kolom bantu_approved_at) hasilnya.
 2. Hitung rata-rata, median, std dari selisih waktu.
 3. Lihat outlier data, kalo outliernya besar, maka gunakan median aja untuk ganti nilai NaN.
- ☐ Memperbaiki data yang null pada kolom order_delivered_carrier_date.
 - Di bagian Cleaning Data, coba pake beberapa cara :
 1. Cek selisih waktu dari kolom order_approved_at dengan kolom order_delivered_carrier_date. Taro di kolom tambahan (misal kolom bantu_delivered_carrier) hasilnya.
 2. Hitung rata-rata, median, std dari selisih waktu.

3. Lihat outlier data, kalo outliernya besar, maka gunakan median aja untuk ganti nilai NaN.

☐ Memperbaiki data yang null pada kolom order_delivered_customer_date.

- Di bagian Cleaning Data, coba pake beberapa cara :
 1. Cek selisih waktu dari kolom order_delivered_carrier_date dengan kolom order_delivered_customer_date . Taro di kolom tambahan (misal kolom bantu_delivered_customer) hasilnya.
 2. Hitung rata-rata, median, std dari selisih waktu.
 3. Lihat outlier data, kalo outliernya besar, maka gunakan median aja untuk ganti nilai NaN.

☐ Memperbaiki tipe data pada kolom order_purchase_timestamp, order_approved_at, order_delivered_carrier_date, order_delivered_customer_date, order_estimated_delivery_date.

5. products_dataset :

- Dalam kasus ini, kita boleh menghapus kolom product_category_name, tetapi harus hitung persentasenya dulu.
- Sebelum dihapus, coba cek persentase baris yg NaN berapa. Jika persentasenya :
 1. Kalo persentasenya $< 5\%$ → Boleh dihapus.
 2. Kalo persentasenya $5\% - 15\%$ → Pertimbangkan kolom dan jumlah total data. Kalo datanya banyak, penghapusan mungkin bisa diterima.
 3. Kalo persentasenya $> 15\%$ → Hindari hapus data dalam jumlah besar. Cari solusi lain untuk mengisi nilai yg NaN, seperti **imputation**, dll.

☐ Pada proses cleaning data, hitung persentase baris yg memiliki data NaN.

☐ Jika sudah, ambil keputusan berdasarkan pertimbangan persentase yang ada.