Laporan Praktikum Minggu 2: Data Wrangling & EDA

Nama **:** Salwa Hafizhah Yusuf

NIM **:** 2411070055

URL W&B Project **:** https://wandb.ai/salwahafizhahyus-stikomelrahma/titanic-eda-2026?nw=nwusersalwahafizhahyus

1. Pendahuluan

Praktikum Data Wrangling dan Interactive EDA pada minggu ke-2 ini bertujuan sebagai tahap awal dalam pengolahan data sebelum digunakan untuk machine learning. Mahasiswa belajar melakukan data loading dari cloud (URL GitHub), membersihkan data dengan teknik imputasi untuk menangani missing values, mengeksplorasi data secara interaktif menggunakan W&B Tables, serta melakukan feature engineering agar data mentah menjadi fitur yang siap digunakan untuk pelatihan model.

Data Wrangling adalah proses membersihkan, mengubah, dan menyiapkan data agar lebih terstruktur dan siap dianalisis. Dataset Titanic digunakan sebagai standar pembelajaran karena memiliki kombinasi data numerik dan kategorikal, terdapat missing values untuk latihan data cleaning, serta cocok untuk tugas klasifikasi (memprediksi penumpang yang selamat atau tidak), sehingga sangat ideal untuk latihan dasar data science.

2. Analisis Data Mentah (Inspeksi)

- **Kolom A** : Age (177 missing values)
- **Kolom B** : Cabin (687 missing values)
- **Kolom C** : Embarked (2 missing values)

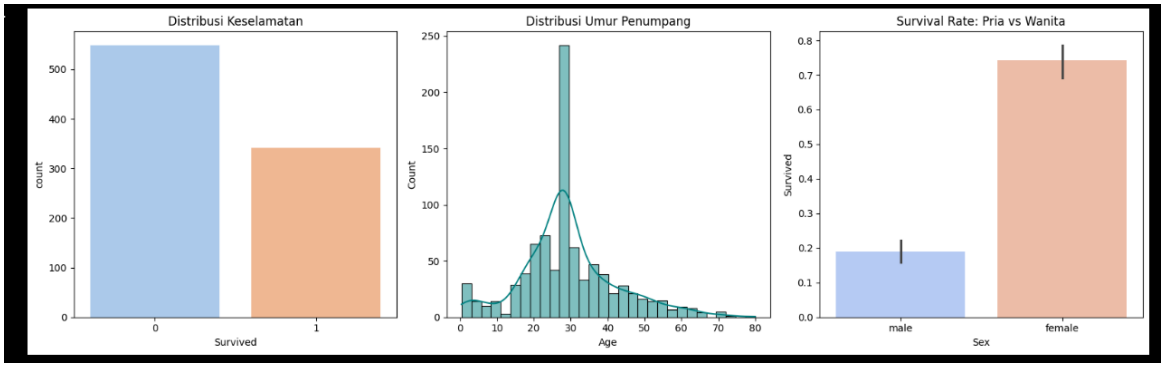
3. Strategi Pembersih Data

- Median digunakan karena data **Age** cenderung memiliki distribusi yang tidak simetris (skewed) dan bisa mengandung outlier (misalnya usia sangat tua atau sangat muda). Median lebih stabil terhadap outlier dibandingkan mean, sehingga menghasilkan imputasi yang lebih representatif dan tidak bias.
- Kolom **Embarked** adalah data kategorikal (S, C, Q). Untuk data kategorikal, nilai yang paling tepat digunakan sebagai imputasi adalah **modus** (nilai yang paling sering muncul), karena mean atau median tidak dapat diterapkan pada data kategori.
- Kolom **Cabin** memiliki jumlah missing values yang sangat besar (687 dari 891 data). Jika diimputasi, hasilnya akan kurang akurat dan berpotensi menimbulkan bias karena sebagian besar datanya kosong. Oleh karena itu, secara teknis lebih baik kolom ini dihapus agar tidak merusak kualitas model.

4. Visualisasi & Temuan Utama

- **Survival Rate by Gender :** Perempuan memiliki tingkat keselamatan lebih tinggi dibandingkan laki-laki. Hal ini karena dalam konteks historis tragedi Titanic berlaku prinsip **“women and children first”**, sehingga wanita diprioritaskan saat evakuasi.

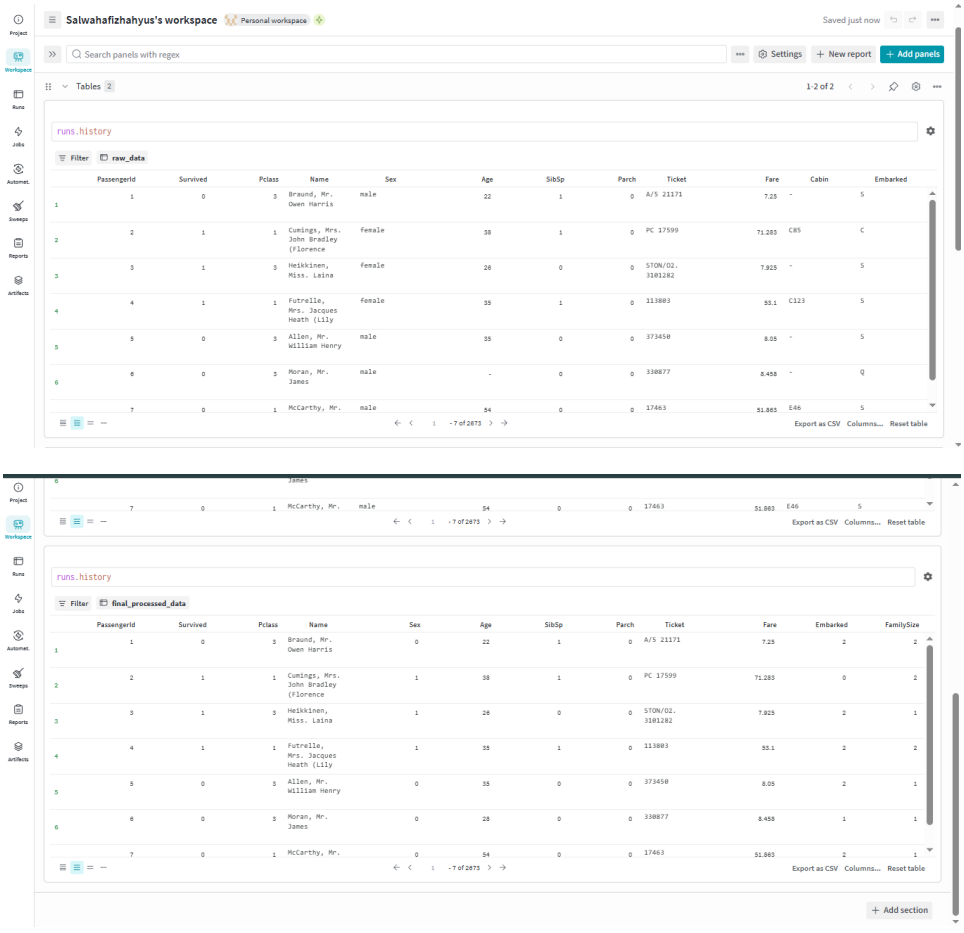
- **Distribusi Umur** **:** Mayoritas penumpang berusia sekitar **20–40 tahun**, dengan jumlah usia dewasa lebih banyak dibandingkan anak-anak dan lansia. Distribusi umur sedikit condong ke kanan karena ada beberapa penumpang usia lanjut, tetapi jumlahnya tidak banyak.



5. Feature Engineering

- Rumus FamilySize (LaTeX) : FamilySize=SibSp+Parch+1
- Analisis : Korelasi **FamilySize–Survived = 0.02** → sangat kecil (mendekati 0). Artinya, ukuran keluarga **hampir tidak berpengaruh signifikan** terhadap peluang selamat (secara linear).
- Encoding : Data teks seperti *Sex* harus diubah jadi angka (0/1) karena model Machine Learning hanya bisa memproses **data numerik**, bukan teks.

6. Integrasi Weights & Biases (W&B)



7. Kesimpulan dan Refleksi

EDA sangat penting sebelum tahap modeling karena membantu memahami karakteristik data, menemukan pola, mendeteksi outlier, serta mengetahui adanya missing values atau ketidakseimbangan data. Dengan EDA, kita dapat menyiapkan data dengan lebih baik sehingga model machine learning menjadi lebih akurat dan tidak menghasilkan kesimpulan yang bias.