

## **BAB III**

### **PELAKSANAAN KERJA**

#### **3.1 Kedudukan dan Koordinasi**

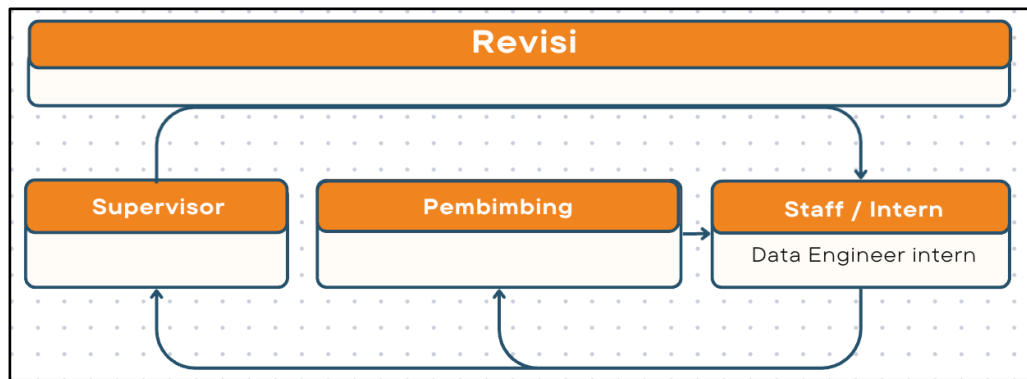
Selama melaksanakan kegiatan magang di Kawan Study, dengan menempati posisi sebagai Data Engineer Intern di bawah Divisi IT, yang berada di bawah koordinasi langsung dengan Chief Technology Officer platform Kawan Study. Sebagai intern, peran disini adalah mendukung pengelolaan data, pembuatan pipeline data, serta pengembangan sistem analisis internal yang berkaitan dengan Business Intelligence dan sistem rekomendasi akademik. Struktur kedudukan penulis dalam organisasi ditunjukkan pada bagan berikut:

##### **3.1.1 Kedudukan**

Selama proses magang sebagai Data Engineer Intern di bawah Divisi IT Kawan Study. Posisi ini berada dalam struktur organisasi di bawah koordinasi langsung Chief Technology Officer. Dalam menjalankan tugas, Koordinasi dengan pembimbing lapangan yang ditunjuk untuk memastikan setiap pekerjaan sesuai dengan standar dan kebutuhan perusahaan.

##### **3.1.2 Koordinasi**

Alur koordinasi pekerjaan dimulai dari Supervisor yang memberikan arahan kepada pembimbing lapangan. Pembimbing kemudian menugaskan pekerjaan kepada tersebut kepada intern. Hasil pekerjaan akan disampaikan kembali kepada pembimbing untuk diperiksa dan direview. Jika diperlukan revisi, pembimbing akan memberikan feedback untuk perbaikan. Setelah hasil pekerjaan disetujui, Laporan akan diberikan kepada pembimbing untuk memeriksa Alur koordinasi ini ditunjukkan pada gambar 3.1 Bagan Alur Koordinasi



Gambar 3. 1 Bagan Alur Koordinasi

### 3.2 Tugas yang Dilakukan

Selama menjalani program magang di Platform Kawan Study, posisi dan divisi ditempatkan pada bagian Data & Business Intelligence dengan fokus utama pada pengelolaan data peserta, pembersihan dataset, integrasi data keperluan analisis, serta pengembangan dashboard untuk mendukung proses pengambilan keputusan berbasis data. Lingkup pekerjaan meliputi analisis data, transformasi dataset, validasi data lintas sumber, hingga penyusunan visualisasi untuk memetakan pola dan kebutuhan pengguna platform.

Tugas dan tanggung jawab dilaksanakan secara bertahap dengan supervisi dari mentor divisi. Seluruh aktivitas diarahkan untuk memastikan kualitas data tetap terjaga, proses analitik berjalan efisien, dan output dapat digunakan secara langsung oleh stakeholder internal. Berikut merupakan rangkaian kegiatan yang telah dilakukan selama masa magang.

Tabel 3. 1 Kegiatan selama proses magang pada platform Kawan Study

No	Kegiatan	Tanggal Mulai	Tanggal Selesai
Project 1 – Beasiswa Diaspora Merdeka (Dashboard Pendaftar & Psychometric)			
1	Orientasi magang & pengenalan scope proyek	01-08-2025	04-08-2025
2	Penyusunan, scope proyek dan hasil akhir	05-08-2025	22-08-2025
3	Monitoring, validasi data dan dokumen pendaftar, dan rekap progres	25-08-2025	05-09-2025
4	Data understanding & dokumentasi variabel dataset pendaftar	08-09-2025	19-09-2025
5	Data Preparation, Cleaning, validasi, seleksi kolom relevan, dan finalisasi dataset pendaftar	22-09-2025	30-09-2025

No	Kegiatan	Tanggal Mulai	Tanggal Selesai
7	Implementasi Dashboard Overview Project 1 dan validasi KPI & insight	10-10-2025	13-10-2025
8	Pembuatan, penyempurnaan, dan finalisasi Psychometric Dashboard	14-10-2025	23-10-2025
9	Final review, dokumentasi akhir, backup, Homepage dan handover Project 1	24-10-2025	31-10-2025
Project 2 – Dashboard Tujuan Menetap WNI di Luar Negeri			
10	Data understanding & Preparation dataset WNI	03-11-2025	06-11-2025
11	Pengembangan dashboard Project	07-11-2025	19-11-2025
12	Dokumentasi final Project 2, QA KPI, arsip, dan wrap-up magang	20-11-2025	22-11-2025

Seluruh kegiatan magang dilaksanakan secara bertahap, dimulai dari pengenalan lingkungan kerja hingga pelaksanaan berbagai proyek yang mendukung operasional berbasis data. Aktivitas-aktivitas tersebut mencakup proses data cleaning, data integration, analisis dataset peserta, pembuatan dashboard interaktif, hingga validasi kualitas data sebelum digunakan pada proses pelaporan maupun kebutuhan unit lain.

Setiap proyek dirancang untuk menjawab kebutuhan bisnis internal melalui pendekatan data-driven decision making, dimana hasil pekerjaan diuji, dievaluasi, dan dikembangkan secara berkelanjutan melalui diskusi bersama mentor dan tim. Evaluasi dilakukan untuk memastikan bahwa model, visualisasi, dan pipeline yang dibangun telah sesuai standar, bebas dari kesalahan teknis, serta mampu memberikan insight yang relevan.

Adapun rincian tools yang digunakan selama pelaksanaan magang akan dijelaskan secara terstruktur pada bagian berikut, lengkap dengan visualisasi logo serta deskripsi fungsional yang relevan dalam konteks implementasi proyek di Kawan Study. Berikut merupakan tools yang dipakai untuk mendukung kegiatan magang;

#### a) Jupyter

Python merupakan salah satu bahasa pemrograman tingkat tinggi yang dirancang dengan sintaks yang sederhana dan mudah dipahami, sehingga banyak digunakan dalam bidang data science, machine learning, dan pengembangan aplikasi berbasis data [35]. Dengan kemampuan untuk memisahkan application

layer dan coordination layer, Jupyter Notebook memberikan fleksibilitas dalam pengembangan pipeline data sekaligus menjaga transparansi proses kerja. Jupyter Notebook juga dimanfaatkan untuk melakukan data cleaning, validasi, transformasi data, serta feature engineering secara terstruktur dan terdokumentasi sebelum data digunakan pada tahap visualisasi.



*Gambar 3. 2 Logo Jupyter Notebook*

Sumber: [36]

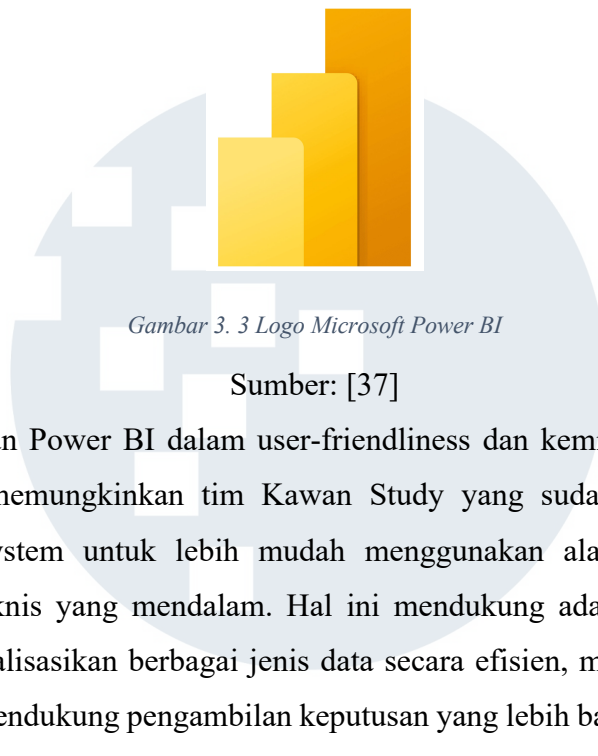
Dalam penerapannya proyek Kawan Study, penggunaan Jupyter Notebook mendukung reproducibility, traceability, dan kolaborasi tim. Setiap Langkah mulai dari data cleaning, exploratory data analysis, feature engineering, hingga membuat feature tambahan pada dataset visualisasi dapat terdokumentasi dengan jelas dalam satu alur, memungkinkan proses validasi dan review yang lebih efektif. Integrasi antara kode, output, dan penjelasan teoretis menjadikan Jupyter Notebook sebagai alat yang ideal untuk membangun sistem analitik yang sistematis, terstruktur, dan mudah dipahami.

#### **b) Power BI**

Power BI adalah alat visualisasi data yang memungkinkan pembuatan tabel dinamis dan grafik interaktif untuk menganalisis serta memfilter data dalam jumlah besar, mempercepat pengambilan keputusan. Dengan antarmuka grafis interaktif, pengguna dapat mengeksplorasi data menggunakan fitur seperti drill-down dan berbagai presentasi seperti peta, grafik, atau animasi [33]. Di Kawan Study, Power BI digunakan untuk memvisualisasikan data pengguna platform, menganalisis partisipasi, dan memberikan wawasan yang dapat langsung diterapkan dalam pengembangan konten dan peningkatan pengalaman pengguna.

Melalui pembuatan dashboard interaktif yang memvisualisasikan metrik kunci, Power BI memungkinkan tim Kawan Study untuk mengakses dan

menganalisis data dengan fleksibilitas tinggi, mendukung pemahaman yang lebih baik terhadap pola keikutsertaan dan kinerja peserta. Dengan integrasi data dari berbagai sumber seperti SQL Server dan Data Mining, Power BI memfasilitasi analisis yang mendalam terhadap perkembangan peserta serta kinerja konten [33].



*Gambar 3. 3 Logo Microsoft Power BI*

Sumber: [37]

Keunggulan Power BI dalam user-friendliness dan kemiripan rumus DAX dengan Excel memungkinkan tim Kawan Study yang sudah familiar dengan Microsoft ecosystem untuk lebih mudah menggunakan alat ini, tanpa perlu keterampilan teknis yang mendalam. Hal ini mendukung adaptabilitas platform dalam memvisualisasikan berbagai jenis data secara efisien, mempercepat proses analitik untuk mendukung pengambilan keputusan yang lebih baik di Kawan Study [37].

### **3.3 Uraian Pelaksanaan Kerja**

#### **3.3.1 Project 1 – Pendaftaran dan Visualisasi Psikotes dan Psikometrik Peserta Beasiswa Diaspora Merdeka 2025**

Selama kegiatan magang di Kawan Study, data engineer berperan dalam proses pengolahan dan analisis data pendaftaran peserta program Beasiswa Diaspora Merdeka. Aktivitas kerja dilakukan secara bertahap dan terstruktur, dimulai dari tahap pengumpulan data hingga penyusunan dashboard Business Intelligence. Fokus utama pekerjaan adalah melakukan monitoring data collection, data cleaning, transformasi data, serta mendukung proses analisis dan visualisasi data peserta. Semua pekerjaan tersebut dilakukan di bawah supervisi langsung dari pembimbing lapangan, dengan menggunakan kombinasi platform seperti Jupyter Notebook, dan Power BI.

Pada tahap awal, penerimaan dan peninjauan file CSV hasil rekap pendaftaran peserta yang telah dikumpulkan dan diserahkan oleh tim internal Kawan Study. Dataset tersebut berisi informasi identitas peserta, jalur pendaftaran, data akademik, serta status pendidikan. Kemudian melakukan proses validasi terhadap struktur dan isi dataset, termasuk pemeriksaan konsistensi email, tanggal lahir, usia, dan status pendidikan (aktif atau sudah tamat). Proses ini juga melibatkan pengecekan logika isian berdasarkan jalur beasiswa yang dipilih peserta, yaitu Jalur A (Berprestasi) dan Jalur B (Kurang Mampu). Selama tahap ini, data engineer berperan dalam monitoring serta memberikan umpan balik kepada tim internal mengenai keakuratan dan kelengkapan data.

#### **3.3.1.1 Orientasi Magang & Pengenalan Scope Proyek**

Pada tahap awal magang, kegiatan difokuskan pada orientasi dan pengenalan lingkungan kerja di Kawan Study. Peserta magang diperkenalkan dengan tim internal, budaya perusahaan, serta sistem pendukung yang digunakan. Pembahasan utama mencakup pemahaman mendalam terhadap program Beasiswa Diaspora Merdeka yang merupakan kolaborasi dengan Kementerian untuk memberikan peluang beasiswa bagi anak muda Indonesia yang ingin melanjutkan pendidikan dan bekerja di Jepang. Scope proyek diperkenalkan secara detail, yaitu pembangunan dashboard untuk memantau pendaftar serta memberikan insight melalui visualisasi data interaktif. Koordinasi awal dengan supervisor dilakukan untuk meninjau data yang tersedia dan menetapkan ekspektasi hasil akhir proyek.



*Gambar 3. 4 Poster Diaspora Merdeka - Kawan Study*

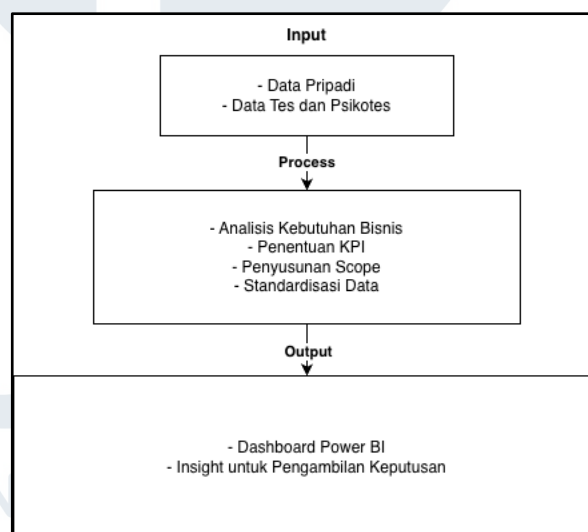
Gambar 3.4 menampilkan poster program Beasiswa Diaspora Merdeka yang digunakan sebagai konteks awal untuk memahami tujuan program dan ruang lingkup data yang akan diolah. Tujuan utama dari pengembangan ini adalah untuk membangun dashboard visualisasi data yang akan mempermudah proses pemantauan pendaftar, serta memberikan insight yang lebih jelas melalui tampilan data visualisasi yang interaktif [33]. Dengan adanya dashboard dari Power BI ini, tim Kawan Study dapat lebih mudah menganalisis data seperti hasil tes, psikotes, dan status pendaftaran, sehingga proses



seleksi dan pemberian beasiswa dapat dilakukan secara lebih efisien [21].

### 3.3.1.2 Penyusunan, scope proyek dan hasil akhir

Tahap ini mencakup penyusunan scope proyek secara rinci, termasuk identifikasi kebutuhan bisnis, target data, serta hasil akhir yang diharapkan. Program beasiswa telah mencapai lebih dari 1400 pendaftar, namun proses evaluasi data masih kurang efektif akibat volume data yang besar meliputi hasil tes, psikotes, dan data pribadi peserta. Gambar 3.5 dibawah ini menunjukkan diagram alur pelaksanaan Project 1, yang menggambarkan tahapan kerja mulai dari penentuan scope proyek, proses pengolahan dan validasi data, hingga pengembangan dashboard Business Intelligence.



Gambar 3. 5 Diagram Alur - Project

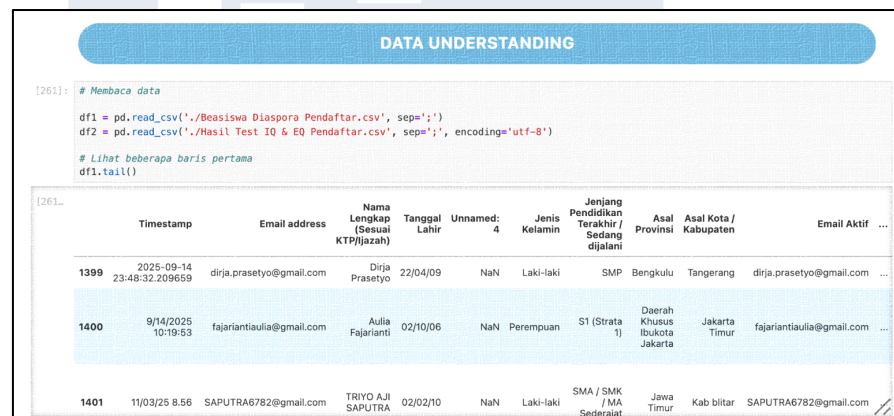
Tujuan utama proyek ditetapkan untuk membangun dashboard visualisasi data menggunakan Power BI guna mempermudah pemantauan pendaftar dan menghasilkan insight yang actionable. Diskusi intensif dengan tim internal dilakukan untuk merumuskan KPI, menyusun daftar rekomendasi standarisasi kolom, serta



mendokumentasikan temuan awal anomali data sebagai dasar hasil akhir proyek.

### 3.3.1.3 Data understanding & dokumentasi variabel dataset pendaftar

Sebelum melakukan analisis lebih lanjut, langkah pertama yang perlu dilakukan adalah pemahaman data dan dokumentasi data. Pemahaman data merupakan tahap yang sangat penting untuk memastikan bahwa data yang digunakan untuk analisis adalah data yang akurat dan bebas dari kesalahan. Proses ini juga bertujuan untuk memastikan konsistensi data, mengurangi data yang tidak relevan, dan mempersiapkan dataset agar siap digunakan dalam visualisasi.



**DATA UNDERSTANDING**

```
[261]: # Membaca data
df1 = pd.read_csv('./Basiswa Diaspora Pendaftar.csv', sep=';')
df2 = pd.read_csv('./Hasil Test IQ & EQ Pendaftar.csv', sep=';', encoding='utf-8')
# Lihat beberapa baris pertama
df1.tail()
```

	Timestamp	Email address	Nama Lengkap (Sesuai KTP/Ijazah)	Tanggal Lahir	Unnamed: 4	Jenis Kelamin	Jenjang Pendidikan Terakhir / Sedang dijalani	Asal Provinsi	Asal Kota / Kabupaten	Email Aktif
1399	2025-09-14 23:48:32.209659	dirja.prasetyo@gmail.com	Dirja Prasetyo	22/04/09	NaN	Laki-laki	SMP	Bengkulu	Tangerang	dirja.prasetyo@gmail.com
1400	9/14/2025 10:19:53	fajariantiaulia@gmail.com	Aulia Fajarianti	02/10/06	NaN	Perempuan	S1 (Strata 1)	Daerah Khusus Ibukota Jakarta	Jakarta Timur	fajariantiaulia@gmail.com
1401	11/03/25 8.56	SAPUTRA6782@gmail.com	TRIYO AJI SAPUTRA	02/02/10	NaN	Laki-laki	SMA / SMK / MA Sederajat	Jawa Timur	Kab bilar	SAPUTRA6782@gmail.com

Gambar 3. 6 Data Understanding Project #1 Basiswa Diaspora

Setelah melakukan pembersihan data, tahap berikutnya adalah memastikan bahwa dataset yang digunakan untuk analisis sudah dalam format yang benar dan bebas dari kesalahan seperti pada gambar 3.6 ini. Kolom yang tidak relevan, seperti Nama Lengkap, Email address, dan kolom Unnamed telah dihapus untuk memudahkan proses analisis lebih lanjut. Kolom User\_ID yang merupakan identifikasi unik dipindahkan ke posisi paling kiri agar lebih mudah dalam penggabungan dataset nantinya.

Kemudian dilakukan proses pembacaan lima baris terakhir data untuk memberikan gambaran awal mengenai kondisi dan isi dataset yang digunakan. Gambar 3.7 menunjukkan hasil pengecekan lima

baris terakhir pada dataset Project 1 sebagai langkah verifikasi awal sebelum proses pembersihan dan integrasi data lebih lanjut.

df2.tail()

	User_ID	Nama	IQ_Score	Ketekunan_Kerja	Perhatian_Detail	Daya_Tahan_Kerja	Kedisiplinan	Motivasi_Kerja	Loyalitas	Stabilitas_Emosi	Kemampu
1399	USR01400	Dirja Prasetyo	91	2	2	3	2	3	1	3	
1400	USR01401	Aulia Fajarianti	101	5	5	5	5	4	5	5	
1401	USR01402	TRIYO AJI SAPUTRA	78	3	2	2	3	3	2	2	
1402	USR01403	Wahyu Zahendra	125	5	5	5	5	5	5	5	
1403	USR01404	Rizka Dlandra Firdaus	127	4	4	5	4	5	4	4	

Gambar 3. 7 Mengecek Dataframe 5 terbawah project #1

Pada tahap ini, dapat melihat beberapa baris terakhir dari dataset untuk memverifikasi data yang terkandung di dalamnya. Dataset pertama Beasiswa Diaspora Pendaftar berisi kolom-kolom seperti Nama Lengkap, Tanggal Lahir, Email address, dan Asal Kota/Kabupaten, sedangkan dataset kedua Hasil Test IQ & EQ Pendaftar berisi kolom-kolom seperti User\_ID, IQ\_Score, dan Penilaian Soft Skill.

```
# (termasuk Nama Lengkap (Sesuai KTP/Ijazah) dan Nama)
columns_to_drop = [
    'Nama Lengkap (Sesuai KTP/Ijazah)', 'Nama', 'Email Aktif', 'Nomor WhatsApp (62***)',
    'Username Instagram anda (@username)', 'Username Tiktok anda (@username)',
    'Upload bukti bayar / transfer pendaftaran Beasiswa pada Lynk.id', 'REF ID Pembayaran',
    'Kategori -', 'Status Ekonomi', 'Apakah anda masih menjadi Siswa Aktif?', 'Masukkan rata-rata nilai rapor',
    'Upload KTP/Kartu Pelajar', 'Upload Ijazah terakhir', 'Upload Sertifikat Prestasi', 'Upload Pasfoto',
    'Upload SKTM / Slip gaji Orang Tua', 'Upload Surat Pernyataan Tidak Mampu', 'Penghasilan Orang tua',
    'Upload Esai Beasiswa', 'Upload Link Video Instagram', 'Upload screenshot bukti follow Instagram @KawanStudy',
    'Upload screenshot postingan IG', 'Saya menyatakan data valid', 'Saya menyetujui UU PDP',
    'Saya bersedia mengikuti program', 'Saya menjaga nama baik KawanStudy',
    'Unnamed: 34', 'Unnamed: 35', 'Unnamed: 36', 'Unnamed: 37', 'Unnamed: 38', 'Unnamed: 39',
    'Unnamed: 40', 'Unnamed: 41', 'Unnamed: 42', 'Unnamed: 43', 'Unnamed: 44', 'Unnamed: 45',
    'Unnamed: 46', 'Unnamed: 47', 'Unnamed: 48', 'Email address'
]

df1_cleaned = df1.drop(columns=columns_to_drop, errors='ignore')
df2_cleaned = df2.drop(columns=columns_to_drop, errors='ignore')

Removing Unnamed constained

df1_cleaned = df1_cleaned.loc[:, ~df1_cleaned.columns.str.contains(r'^Unnamed', na=False)]
df2_cleaned = df2_cleaned.loc[:, ~df2_cleaned.columns.str.contains(r'^Unnamed', na=False)]
```

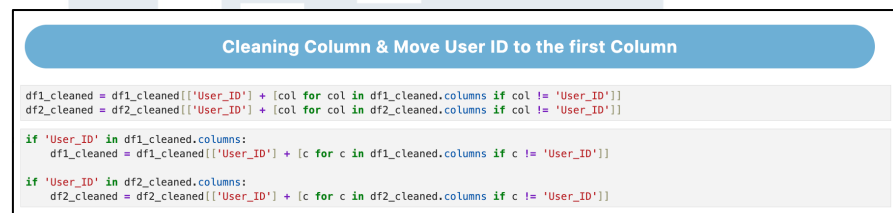
Gambar 3. 8 Mendrop semua kolom yang tidak dibutuhkan

Pada tahap pembersihan data, dapat dilihat pada gambar diatas yaitu gambar 3.8 merupakan langkah pertama yang dilakukan adalah menghapus kolom-kolom yang tidak relevan dari dataset. Kolom-kolom yang tidak relevan, seperti Nama Lengkap (Sesuai KTP/Ijazah), Nama, Email Aktif, dan beberapa kolom lain yang

mengandung informasi sensitif atau tidak diperlukan untuk analisis lebih lanjut, telah dihapus untuk menjaga privasi dan memperbaiki konsistensi data.

Pada tahap pembersihan data, langkah penting yang dilakukan adalah memindahkan kolom User\_ID ke posisi paling kiri dalam dataset. Kolom User\_ID adalah identifikasi unik untuk setiap pendaftar dan sangat penting untuk penggabungan data atau analisis lebih lanjut.

#### 3.3.1.4 Data Preparation, Cleaning, validasi, seleksi kolom relevan, dan finalisasi dataset pendaftar



Gambar 3. 9 Membersihkan kolom dan memindahkan User ID

Gambar 3.9 diatas menunjukkan proses penataan ulang kolom dan pembersihan lanjutan pada dataset pendaftar. Untuk memastikan User\_ID berada di urutan pertama, dengan memeriksa apakah kolom tersebut ada dalam dataset, dan jika ada, kolom tersebut dipindahkan ke posisi pertama, sementara kolom lainnya tetap berurutan setelahnya. Langkah ini tidak hanya memastikan konsistensi struktur data, tetapi juga memudahkan proses pengolahan data lebih lanjut, seperti saat menggabungkan dataset berdasarkan User\_ID. Integrasi Data Psikotes dan Psikometrik.

Setelah dataset pendaftar dinyatakan valid, melanjutkan ke tahap integrasi data antar-sumber, yaitu penggabungan data pendaftar dengan hasil psikotes dan psikometrik dari para pendaftar beasiswa diaspora Merdeka Kawan Study. Proses ini melibatkan penyesuaian

relasi antar tabel seperti ID peserta, jalur pendaftaran, dan hasil tes untuk memastikan kesesuaian antar-dataset.

Merge Both dataset by USER ID with JOIN

```
[228]: final_data = pd.merge(df1_cleaned, df2_cleaned, on='User_ID', how='left')

print("final_data shape:", final_data.shape)
final_data.head()
```

final\_data shape: (1404, 21)

```
[228]:
```

	User_ID	Timestamp	Tanggal Lahir	Jenis Kelamin	Jenjang Pendidikan Terakhir / Sedang dijalani	Asal Provinsi	Asal Kota / Kabupaten	IQ_Score	Ketekunan_Kerja	Perhatian_Detail	...	Kedisiplinan	Motiva
0	USR00001	2025-09-17 13:09:56.929844	27/06/05	Laki-laki	D3	Gorontalo	Palembang	77	2	2	...	2	
1	USR00002	2025-10-08 23:38:56.446892	14/03/07	Laki-laki	SMK	Jawa Barat	Tidore Kepulauan	122	4	3	...	4	
2	USR00003	2025-10-10 19:40:28.508901	04/06/06	Perempuan	D3	Bali	Kotamobagu	104	5	4	...	4	
3	USR00004	2025-09-26 01:04:28.824425	22/01/03	Laki-laki	D1	Jawa Tengah	Palopo	85	3	4	...	3	
4	USR00005	2025-09-28 24:01:04	04/01/04	Perempuan	D3	Kepulauan	Damaralam	76	3	2	...	2	

Gambar 3. 10 Mengabungkan kedua dataset dengan User ID

Gambar 3.10 menunjukkan proses penggabungan dua dataset yang telah melalui tahap pembersihan dan validasi sebelumnya. Pada langkah ini, kedua dataset yang sudah dibersihkan dan diproses sebelumnya digabungkan menjadi satu dataset tunggal yang lebih lengkap dengan menggunakan User\_ID sebagai kunci penggabungan. Proses ini menggunakan fungsi `pd.merge()` dari Pandas, yang memungkinkan penggabungan data berdasarkan kunci yang sama, dalam hal ini adalah kolom User\_ID.

<pre>final_data.info()</pre>			
<pre>&lt;class 'pandas.core.frame.DataFrame'&gt; RangeIndex: 1404 entries, 0 to 1403 Data columns (total 33 columns): #   Column                                     Non-Null Count  Dtype ---  ---                                     - 0   User_ID                                   1404 non-null   object 1   Timestamp                                1353 non-null   object 2   Tanggal Lahir                            1404 non-null   datetime64[ns] 3   Jenis Kelamin                            1369 non-null   object 4   Jenjang Pendidikan Terakhir / Sedang dijalani  1367 non-null   object 5   Asal Provinsi                            1366 non-null   object 6   Asal Kota / Kabupaten                    1360 non-null   object 7   IQ_Score                                 1404 non-null   int64 8   Ketekunan_Kerja                          1404 non-null   int64 9   Perhatian_Detail                         1404 non-null   int64 10  Daya_Tahan_Kerja                         1404 non-null   int64 11  Kedisiplinan                             1404 non-null   int64 12  Motivasi_Kerja                           1404 non-null   int64 13  Loyalitas                                1404 non-null   int64 14  Stabilitas_Emosi                         1404 non-null   int64 15  Kemampuan_Adaptasi                       1404 non-null   int64 16  Kerjasama                                1404 non-null   int64 17  Persentase_Kepribadian                   1404 non-null   int64 18  Kesehatan_Jiwa                           1404 non-null   object 19  Nilai_Wawasan_Kebangsaan                 1404 non-null   float64 20  Rekomendasi_Asesmen                      1404 non-null   object 21  Usia (Tahun)                             1404 non-null   int64 22  age_group                                1404 non-null   category 23  birth_month                              1404 non-null   int32 24  birth_dayofweek                          1404 non-null   int32 25  is_minor                                 1404 non-null   int64 26  education_level                          1362 non-null   float64 27  softskill_total                          1404 non-null   int64 28  softskill_avg                            1404 non-null   float64 29  high_potential                           1404 non-null   int64 30  mental_health_level                     1404 non-null   int64 31  rec_above_standard                       1404 non-null   int64 32  engagement_index                         1362 non-null   float64 dtypes: float64(1), int64(1), category(1), datetime64[ns](1), float64(3), int32(2), int64(16), object(8) memory usage: 344.5+ KB</pre>			

Gambar 3. 11 Mengecek Nama Kolom dan tipe data

Gambar 3.11 menunjukkan hasil pengecekan nama kolom dan tipe data pada dataset hasil penggabungan. Dataset ini memiliki 1404 baris yang menunjukkan jumlah pendaftar yang tercatat, dengan 33 kolom yang masing-masing berisi informasi yang berbeda mengenai pendaftar. Kolom-kolom dalam dataset ini meliputi informasi dasar seperti User\_ID, Timestamp, Tanggal Lahir, IQ\_Score, serta penilaian terkait soft skills seperti Ketekunan Kerja, Perhatian Detail, dan Kemampuan Adaptasi.

Setiap kolom dalam dataset memiliki jenis data yang berbeda, dengan tipe data seperti object untuk data kategori atau string seperti nama, jenis kelamin, dan asal provinsi, int64 untuk data numerik tanpa desimal seperti skor IQ atau nilai dalam tes soft skills, dan float64 untuk data numerik dengan nilai desimal seperti Persentase Kepribadian dan Nilai Wawasan Kebangsaan. Selain itu, kolom dengan tipe datetime64 digunakan untuk data tanggal, seperti Tanggal Lahir dan Timestamp.

Namun, beberapa kolom masih memiliki nilai null atau kosong yang perlu ditangani lebih lanjut. Kolom-kolom seperti Jenis Kelamin, Asal Provinsi, dan beberapa kolom lainnya memiliki nilai kosong yang perlu diperbaiki agar dataset menjadi lebih konsisten dan siap untuk dianalisis lebih lanjut.

Untuk memberikan gambaran yang lebih sistematis mengenai struktur dataset, Tabel 3.2 dibawah menyajikan daftar nama kolom, tipe data, serta deskripsi singkat dari masing-masing variabel yang digunakan dalam Project 1.

Tabel 3. 2 Nama Kolom dan deskripsi data project #1

No	Nama Kolom	Tipe Data	Deskripsi
1	User_ID	object	ID unik pelanggan
2	Timestamp	object	Waktu pendaftaran pelanggan
3	Tanggal Lahir	datetime64	Tanggal lahir pelanggan
4	Jenis Kelamin	object	Jenis kelamin pelanggan
5	Jenjang Pendidikan Terakhir / Sedang dijalani	object	Tingkat pendidikan pelanggan
6	Asal Provinsi	object	Provinsi asal pelanggan
7	Asal Kota / Kabupaten	object	Kota/Kabupaten asal pelanggan
8	IQ_Score	int64	Nilai IQ pelanggan
9	Ketekunan_Kerja	int64	Skor ketekunan kerja
10	Perhatian_Detail	int64	Skor perhatian detail
11	Daya_Tahan_Kerja	int64	Skor daya tahan kerja
12	Kedisiplinan	int64	Skor kedisiplinan
13	Motivasi_Kerja	int64	Skor motivasi kerja
14	Loyalitas	int64	Skor loyalitas
15	Stabilitas_Emosi	int64	Skor stabilitas emosi
16	Kemampuan_Adaptasi	int64	Skor kemampuan adaptasi
17	Kerjasama	int64	Skor kerjasama
18	Persentase_Kepribadian	int64	Skor persentase kepribadian

No	Nama Kolom	Tipe Data	Deskripsi
19	Kesehatan_Jiwa	object	Tingkat kesehatan jiwa pelanggan
20	Nilai_Wawasan_Kebangsaan	float64	Nilai wawasan kebangsaan pelanggan
21	Rekomendasi_Asesmen	object	Rekomendasi asesmen berdasarkan data yang ada
22	Usia (Tahun)	Int64	Usia pelanggan
23	age_group	category	Kelompok umur pelanggan
24	birth_month	int32	Bulan kelahiran pelanggan
25	birth_dayofweek	int32	Hari kelahiran pelanggan
26	is_minor	int64	Apakah pelanggan masih di bawah umur
27	education_level	float64	Tingkat pendidikan pelanggan
28	softskill_total	int64	Total skor softskill
29	softskill_avg	float64	Rata-rata skor softskill
30	high_potential	int64	Potensi tinggi pelanggan
31	mental_health_level	int64	Level kesehatan mental pelanggan
32	rec_above_standard	int64	Rekomendasi di atas standar
33	engagement_index	Float64	Indeks keterlibatan pelanggan

Tabel di atas menyajikan struktur data yang terdiri dari berbagai kolom yang dikelompokkan berdasarkan fungsinya dalam analisis. Kolom-kolom dalam kategori Identitas & Demografi mencakup informasi dasar pelanggan seperti ID unik, tanggal lahir, jenis kelamin, tingkat pendidikan, dan lokasi asal. Sementara itu, kategori Riwayat Pembiayaan & Transaksi berfokus pada skor softskill dan karakteristik psikologis pelanggan, seperti ketekunan kerja, perhatian detail, motivasi kerja, serta stabilitas emosi, yang semuanya memberikan wawasan tentang potensi dan perilaku pelanggan dalam



konteks transaksi atau pembiayaan. Di sisi lain, kolom-kolom dalam kategori Soft Skills memberikan skor untuk berbagai aspek kemampuan pribadi pelanggan, seperti kerjasama dan kemampuan adaptasi. Dengan informasi yang lengkap ini, dataset ini memungkinkan untuk analisis yang mendalam mengenai profil pelanggan, yang dapat digunakan untuk segmentasi pasar dan perencanaan strategi bisnis yang lebih tepat.

Langkah identifikasi nilai hilang ini menjadi krusial untuk memastikan bahwa proses analitik yang dilakukan selanjutnya dibangun di atas fondasi data yang bersih dan valid.

MISSING VALUE	
<pre>missing_values = final_data.isnull().sum() print("Missing Values (NaN):") print(missing_values)</pre>	
Missing Values (NaN):	
User_ID	0
Timestamp	47
Tanggal Lahir	0
Jenis Kelamin	35
Jenjang Pendidikan Terakhir / Sedang dijalani	37
Asal Provinsi	38
Asal Kota / Kabupaten	44
IQ_Score	0
Ketekunan_Kerja	0
Perhatian_Detail	0
Daya_Tahan_Kerja	0
Kedisiplinan	0
Motivasi_Kerja	0
Loyalitas	0
Stabilitas_Emosi	0
Kemampuan_Adaptasi	0
Kerjasama	0
Persentase_Kepribadian	0
Kesehatan_Jiwa	0
Nilai_Kawasan_Kebangsaan	0
Rekomendasi_Asesmen	0
dtype: int64	

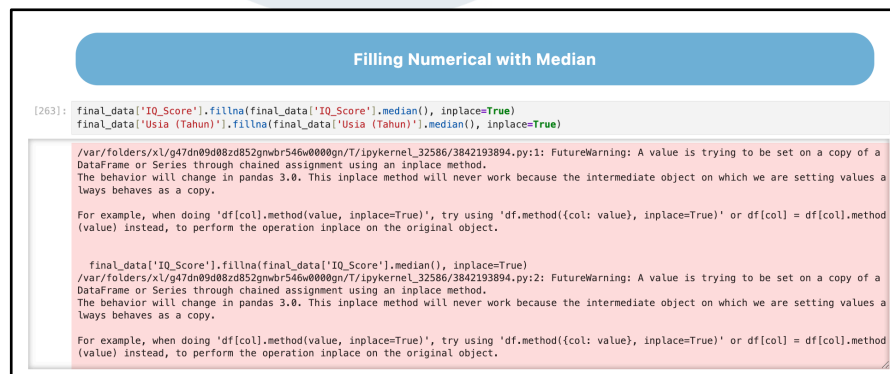
Gambar 3. 12 Mengecek Missing Value

Langkah identifikasi pada gambar 3.12 nilai hilang ini menjadi krusial untuk memastikan bahwa proses analitik yang dilakukan selanjutnya dibangun di atas fondasi data yang bersih dan valid. Seperti yang ditunjukkan pada Gambar dibawah ini, analisis terhadap nilai hilang dilakukan dengan memeriksa jumlah missing values pada setiap kolom dalam dataset. Proses ini membantu untuk mengidentifikasi kolom-kolom yang memerlukan penanganan lebih lanjut.

Berdasarkan hasil pemeriksaan, terdapat beberapa kolom yang memiliki nilai hilang, di antaranya Timestamp, Jenis Kelamin,

Jenjang Pendidikan Terakhir / Sedang dijalani, Asal Provinsi, dan Asal Kota / Kabupaten, dengan jumlah nilai kosong yang bervariasi. Atribut lainnya seperti IQ\_Score, Ketekunan\_Kerja, dan beberapa atribut lainnya tidak memiliki nilai hilang. Untuk mengatasi masalah nilai hilang, langkah selanjutnya adalah melakukan imputasi dengan menggunakan median untuk atribut numerik dan mode untuk atribut kategorikal. Dengan demikian, dataset dapat disiapkan untuk analisis lebih lanjut tanpa mengorbankan kualitas data.

Langkah pertama dalam menangani nilai hilang adalah dengan mengisi missing values pada atribut numerik. Sebagaimana yang terlihat pada Gambar 3.13 dibawah mengenai atribut IQ\_Score dan Usia yang memiliki nilai kosong diisi menggunakan median dari masing-masing kolom. Penggunaan median sangat berguna dalam hal ini karena median lebih robust terhadap outlier dibandingkan dengan mean, sehingga menjaga distribusi data tetap stabil.



```
[263]: final_data['IQ_Score'].fillna(final_data['IQ_Score'].median(), inplace=True)
      final_data['Usia (Tahun)'].fillna(final_data['Usia (Tahun)'].median(), inplace=True)

/var/folders/xl/g47dn89d88zd852gnvbr546w000gn/T/ipykernel_32586/3842193894.py:1: FutureWarning: A value is trying to be set on a copy of a
DataFrame or Series through chained assignment using an inplace method.
The behavior will change in pandas 3.0. This inplace method will never work because the intermediate object on which we are setting values a
lways behaves as a copy.

For example, when doing 'df[col].method(value, inplace=True)', try using 'df.method({col: value}, inplace=True)' or df[col] = df[col].method
(value) instead, to perform the operation inplace on the original object.

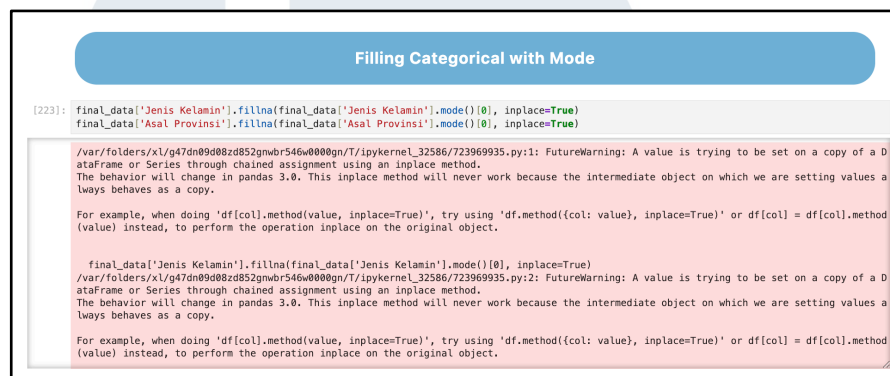
final_data['IQ_Score'].fillna(final_data['IQ_Score'].median(), inplace=True)
/var/folders/xl/g47dn89d88zd852gnvbr546w000gn/T/ipykernel_32586/3842193894.py:2: FutureWarning: A value is trying to be set on a copy of a
DataFrame or Series through chained assignment using an inplace method.
The behavior will change in pandas 3.0. This inplace method will never work because the intermediate object on which we are setting values a
lways behaves as a copy.

For example, when doing 'df[col].method(value, inplace=True)', try using 'df.method({col: value}, inplace=True)' or df[col] = df[col].method
(value) instead, to perform the operation inplace on the original object.
```

Gambar 3. 13 Mengisi Numerikan Fitur dengan Median

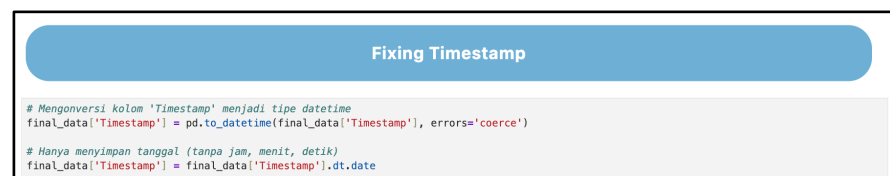
Setelah menangani atribut numerik, langkah berikutnya adalah menangani nilai hilang pada atribut kategorikal. Seperti yang ditunjukkan pada Gambar 3.14, atribut Jenis Kelamin dan Asal Provinsi yang memiliki nilai kosong diisi dengan mode, yaitu nilai yang paling sering muncul dalam kolom tersebut.

Dengan menggunakan mode, data kategorikal tetap terjaga konsistensinya dan dapat diproses lebih lanjut untuk analisis. Pada tahap ini, perlu menangani kolom Timestamp yang ada dalam dataset, karena kolom tersebut kemungkinan masih dalam format yang kurang tepat untuk analisis lebih lanjut. Kolom Timestamp ini awalnya mungkin berupa string atau format lain yang tidak dapat langsung digunakan untuk analisis waktu. Oleh karena itu, perlu dilakukan konversi ke tipe data datetime agar dapat diolah dengan mudah.



Gambar 3. 14 Mengisi kategorikal Fitur dengan mode

Seperti yang terlihat pada Gambar 3.14, mengisi nilai kosong pada kolom Jenis Kelamin dan Asal Provinsi dengan menggunakan mode, yang merupakan nilai paling sering muncul dalam kolom tersebut. Metode ini memastikan bahwa data kategorikal yang memiliki missing values tetap terjaga konsistensinya.

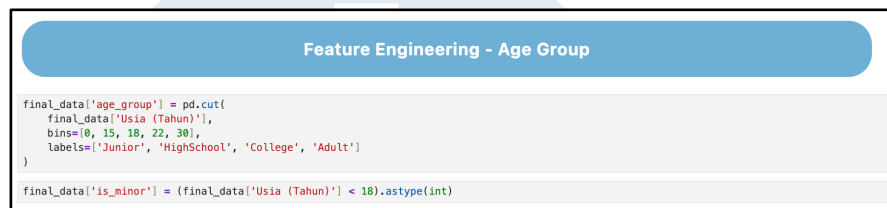


Gambar 3. 15 Memperbaiki Timestamp

Gambar 3.15 menunjukkan proses perbaikan kolom Timestamp melalui konversi ke format datetime yang valid. Pada tahap ini, nilai waktu yang sebelumnya masih dalam format teks atau format tidak seragam diubah menjadi tipe data datetime, kemudian disederhanakan

menjadi komponen tanggal saja. Langkah ini dilakukan untuk memastikan konsistensi struktur data berbasis waktu serta memudahkan analisis temporal pada tahap selanjutnya.

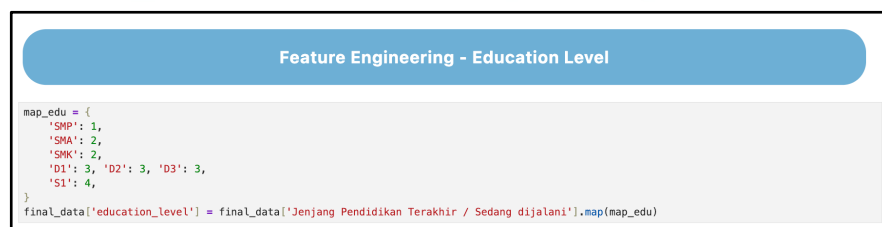
Langkah selanjutnya dalam tahap ini adalah membangun fitur `age_group`, yaitu segmentasi usia yang akan mempermudah analisis berdasarkan kelompok umur peserta.



Gambar 3.16 Fitur Engineering - Age Group

Gambar 3.16 memperlihatkan pembentukan fitur `age_group` menggunakan `pd.cut()`, yang mengelompokkan peserta menjadi beberapa kategori seperti *Junior*, *HighSchool*, *College*, dan *Adult*. Fitur tambahan `is_minor` juga ditambahkan untuk menandai peserta berusia di bawah 18 tahun.

Dengan fitur usia yang kini lebih terstruktur, proses *feature engineering* dapat dilanjutkan ke transformasi variabel pendidikan menjadi fitur numerik `education_level`, yang akan digunakan pada tahap berikutnya.



Gambar 3.17 Feature Engineering - Education Level

Gambar 3.17 menampilkan proses transformasi kolom pendidikan menggunakan mapping dictionary yang mengubah kategori seperti SMP, SMA, SMK, Diploma, hingga Sarjana menjadi

nilai numerik bertingkat. Konversi ini dibutuhkan agar tingkat pendidikan dapat dianalisis secara kuantitatif dan dijadikan variabel model.

Setelah fitur pendidikan terstruktur dalam format angka, tahap feature engineering dilanjutkan pada atribut soft skill, di mana delapan variabel perilaku peserta akan dirangkum menjadi indikator yang lebih representatif.

```
Feature Engineering Soft Skill

soft_cols = ['Ketekunan_Kerja', 'Perhatian_Detail', 'Daya_Tahan_Kerja',
             'Kedisiplinan', 'Motivasi_Kerja', 'Loyalitas', 'Kerjasama']

final_data['softskill_total'] = final_data[soft_cols].sum(axis=1)

final_data['softskill_avg'] = final_data[soft_cols].mean(axis=1)

final_data['high_potential'] = (
    (final_data['IQ_Score'] > 110) &
    (final_data['softskill_total'] > 25)
).astype(int)
```

*Gambar 3. 18 Feature engineering - Soft Skill*

Gambar 3.18 memperlihatkan proses agregasi delapan atribut soft skill menjadi dua fitur utama, yaitu softskill\_total dan softskill\_avg, serta pembentukan indikator high\_potential yang menandai peserta dengan potensi tinggi berdasarkan skor IQ dan soft skill di atas ambang tertentu.

Dengan fitur soft skill yang kini telah terstruktur dan lebih ringkas, tahap feature engineering dapat dilanjutkan ke bidang kesehatan mental untuk mengonversi kategori stres menjadi representasi numerik yang dapat dibandingkan dengan variabel lainnya.

```
Feature Engineering - Mental health

mental_map = {
    'NORMAL': 0,
    'STRES RINGAN': 1,
    'STRES SEDANG': 2,
    'STRES BERAT': 3
}

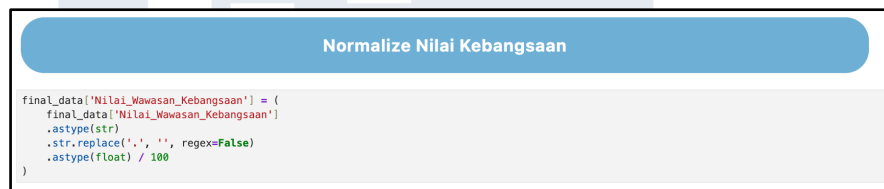
final_data['mental_health_level'] = final_data['Kesehatan_Jiwa'].map(mental_map)

final_data['rec_above_standard'] = final_data['Rekomendasi_Asesmen'].str.contains('Atas').astype(int)
```

*Gambar 3. 19 Feature Engineering - Mental Health*

Gambar 3.19 menunjukkan proses pemetaan kategori kesehatan mental seperti NORMAL, STRES RINGAN, STRES SEDANG, dan STRES BERAT menjadi nilai numerik 0–3 menggunakan mapping dictionary. Selain itu, fitur tambahan `rec_above_standard` dibangun untuk menandai rekomendasi asesmen kategori “Atas”.

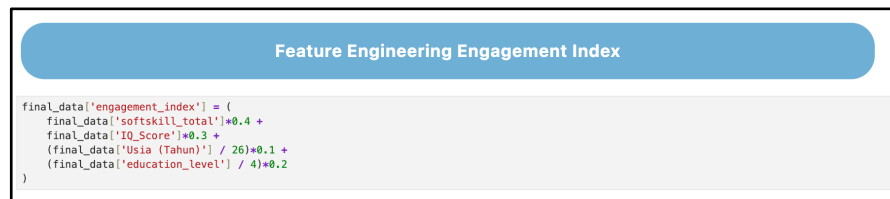
Dengan kesehatan mental yang kini direpresentasikan secara numerik, tahap berikutnya adalah melakukan normalisasi pada `Nilai_Wawasan_Kebangsaan`, yang masih berada dalam format teks persen dan harus diubah ke format numerik agar terintegrasi dengan variabel lainnya.



*Gambar 3. 20 Normalize Nilai Kebangsaan*

Gambar 3.20 memperlihatkan pembersihan nilai kebangsaan dengan menghapus tanda titik dan mengubahnya ke format numerik menggunakan tipe float melalui pembagian dengan 100, sehingga nilainya berada pada skala 0–1. Normalisasi ini memastikan variabel ini memiliki skala yang sejalan dengan variabel numerik lain.

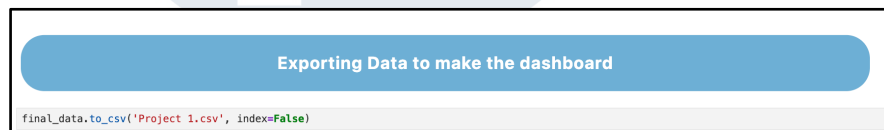
Setelah seluruh variabel inti telah distandarkan, langkah berikutnya adalah menyusun fitur komposit `engagement_index`, sebuah metrik integratif yang menggambarkan tingkat keterlibatan peserta berdasarkan faktor-faktor penting seperti IQ, soft skill, pendidikan, dan usia.



*Gambar 3. 21 Feature Engineering - Engagement Index*

Gambar 3.21 menunjukkan rumus perhitungan engagement\_index, yang menggabungkan sejumlah variabel relevan menjadi satu indikator komposit dengan bobot tertentu seperti soft skill, IQ, tingkat pendidikan, dan usia. Fitur ini memberikan gambaran holistik mengenai keterlibatan peserta secara keseluruhan dalam proses asesmen.

Setelah seluruh fitur penting selesai direkayasa, dataset sudah memasuki tahap akhir yaitu proses ekspor ke format CSV untuk kebutuhan dashboard dan analisis lanjutan.



*Gambar 3. 22 Export final CSV*

Gambar 3.22 memperlihatkan proses ekspor dataset menggunakan .to\_csv() sebagai tahap akhir dari proses data preparation. Dengan seluruh data yang telah dibersihkan, diimputasi, dan direkayasa menjadi fitur-fitur baru yang relevan, file CSV ini siap digunakan untuk pembuatan dashboard visualisasi dan analisis mendalam pada tahap berikutnya.

### **3.3.1.5 Dashboard Overview Project 1 dan validasi KPI & insight**

Pada tahap implementasi ini, dashboard yang dibuat bertujuan untuk memberikan visualisasi yang informatif dan mudah diakses oleh pengguna. Sebelum masuk ke dalam visualisasi data yang lebih detail, mulai dengan Homepage Kawan Study, yang menyediakan



akses mudah ke berbagai halaman terkait peserta dan analisis akademik.

Halaman ini memudahkan pengguna untuk mengarahkan fokus mereka ke data yang relevan. Setelah ini, akan masuk ke Dashboard IQ & EQ, yang memberikan gambaran lebih mendalam mengenai metrik peserta.



Gambar 3. 23 Dashboard Peserta Pendaftaran

Pada tahap implementasi ini yang ditampilkan oleh gambar 3.23, dashboard yang dibuat bertujuan untuk memberikan visualisasi yang informatif dan mudah diakses oleh pengguna. Sebelum masuk ke dalam visualisasi data yang lebih detail, mulai dengan Homepage Kawan Study, yang menyediakan akses mudah ke berbagai halaman terkait peserta dan analisis akademik. Halaman ini memudahkan pengguna untuk mengarahkan fokus mereka ke data yang relevan.

Pada gambar sebelumnya, sudah melihat bagaimana peserta di kategorikan berdasarkan tingkat stres mereka dan bagaimana distribusi asesmen mempengaruhi hasil tersebut. Sekarang, akan melanjutkan ke gambar yang lebih mendalam, yang menunjukkan hubungan antara IQ dan kesehatan mental, serta pengaruh engagement index terhadap kinerja peserta. Ini memberikan gambaran yang lebih

luas tentang seberapa besar pengaruh IQ terhadap kemampuan peserta dalam mengelola stres dan kesehatan emosional mereka.

### 3.3.1.6 Pembuatan, penyempurnaan, dan finalisasi Psychometric Dashboard



Gambar 3. 24 Dashboard IQ & EQ Peserta

Gambar 3.24 ini menampilkan hubungan antara IQ dan kesehatan mental, serta bagaimana peserta dengan IQ lebih tinggi cenderung memiliki tingkat engagement index yang lebih baik. Visualisasi scatter plot di bagian bawah memberikan wawasan lebih lanjut mengenai korelasi antara skor IQ dan kesehatan mental peserta, serta bagaimana engagement index mereka berubah seiring usia.

### 3.3.1.7 Final review, dokumentasi akhir, backup, Homepage dan handover Project #1

Tahap penutup Project 1 mencakup final review seluruh hasil dashboard, penyempurnaan akhir, dokumentasi proses kerja secara lengkap termasuk alur ETL, feature engineering, dan visualisasi, backup data serta file proyek, serta handover resmi kepada tim Kawan Study.



*Gambar 3. 25 Home Page Kawan Study Dashboard*

Gambar 3.25 menampilkan Homepage Kawan Study Dashboard yang berfungsi sebagai halaman navigasi utama. Homepage ini menyediakan akses yang terstruktur ke berbagai dashboard, seperti Overview Peserta Pendaftaran dan Psychometric IQ & EQ, sehingga memudahkan pengguna internal dalam menavigasi seluruh fitur analisis peserta dan asesmen psikotes. Desain homepage yang interaktif bertujuan untuk meningkatkan user experience serta mendukung pemanfaatan dashboard secara optimal dalam pengambilan keputusan berbasis data.

Homepage ini dirancang untuk meningkatkan user experience dengan menu interaktif, sehingga memudahkan pengguna tim internal Kawan Study dalam menavigasi seluruh fitur analisis peserta dan asesmen psikotes.

### **3.3.2 Project 2 - Pemetaan dan Visualisasi distribusi diaspora**

#### **3.3.2.1 Overview Project**

Data mengenai tujuan menetap Warga Negara Indonesia di luar negeri merupakan salah satu sumber informasi penting yang dapat digunakan untuk memahami pola migrasi, kebutuhan tenaga kerja global, serta sebaran komunitas diaspora Indonesia [23]. Informasi ini

relevan bagi berbagai lembaga pemerintah maupun platform pendidikan seperti Kawan Study, terutama dalam merancang program layanan dan kebijakan yang selaras dengan kebutuhan masyarakat Indonesia di luar negeri.

Namun, data yang diperoleh dari laporan umumnya masih berbentuk tabel statis dengan struktur yang kompleks, terdiri dari banyak kategori seperti “Bekerja”, “Belajar”, “Magang”, “Mendampingi Suami/Istri”, dan “Anggota Keluarga/Pengikut”. Format data yang tidak rapi serta adanya kolom-kolom tanpa header (Unnamed) menyulitkan proses analisis awal. Selain itu, beberapa baris agregat seperti "Grand Total" turut muncul dalam data mentah sehingga berpotensi menimbulkan ketidakakuratan apabila tidak dibersihkan terlebih dahulu.

Kendala tersebut menyebabkan proses pembacaan data menjadi kurang efisien dan sulit diolah secara langsung. Tim membutuhkan cara yang lebih efektif untuk menyajikan data tersebut dalam bentuk visual yang mudah dipahami, sehingga insight mengenai distribusi WNI di berbagai negara dapat diperoleh secara cepat dan akurat [23]. Melalui visualisasi data yang tepat, platform dapat melihat negara mana yang memiliki sebaran WNI terbanyak, membandingkan kategori tujuan, serta memantau dinamika perpindahan berdasarkan kawasan.

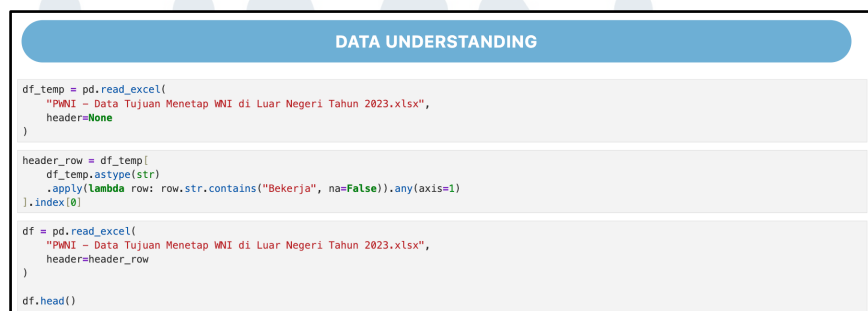
Oleh karena itu, tujuan utama dari pengembangan proyek ini adalah menyiapkan dataset yang bersih dan terstruktur, kemudian membangun dashboard visualisasi interaktif menggunakan Power BI [34]. Dashboard ini dirancang untuk mempermudah pemangku kepentingan dalam memahami pola diaspora WNI, memetakan lokasi-lokasi dengan populasi signifikan, serta memberikan insight berbasis data yang dapat digunakan untuk mendukung pengambilan keputusan strategis. Melalui penyajian visual yang informatif, proses

pemantauan dan analisis data dapat dilakukan dengan lebih efektif dan transparan.

### 3.3.2.2 Data understanding & cleaning awal dataset WNI

Proses pertama dalam menangani data adalah memuat file Excel yang berisi informasi penting terkait peserta dan kategori mereka. Di sini, mulai dengan membaca file yang berisi data mengenai tujuan WNI yang menetap di luar negeri, termasuk kategori pekerjaan seperti "Bekerja," "Belajar," dan lainnya. Langkah pertama adalah membaca data dan mendeteksi baris pertama yang berisi header yang tepat untuk memastikan konsistensi data yang dibaca.

Gambar 3.26 menunjukkan proses data understanding awal melalui pembacaan file Excel menggunakan library pandas. Pada tahap ini, data dibaca tanpa menetapkan header secara langsung, kemudian dilakukan identifikasi baris yang berisi nama kolom yang relevan. Setelah baris header yang tepat ditemukan, struktur kolom ditetapkan kembali agar dataset dapat diproses lebih lanjut dengan format yang sesuai dan terstruktur.



```
DATA UNDERSTANDING

df_temp = pd.read_excel(
    "PwNI - Data Tujuan Menetap WNI di Luar Negeri Tahun 2023.xlsx",
    header=None
)

header_row = df_temp[
    df_temp.astype(str)
    .apply(lambda row: row.str.contains("Bekerja", na=False)).any(axis=1)
].index[0]

df = pd.read_excel(
    "PwNI - Data Tujuan Menetap WNI di Luar Negeri Tahun 2023.xlsx",
    header=header_row
)

df.head()
```

Gambar 3. 26 Data Understanding Project #2 Tujuan Diaspora Menetap

Setelah file dibaca dan header ditetapkan dengan benar, langkah berikutnya adalah memastikan bahwa semua kolom data telah diberi label yang sesuai. Misalnya, kolom "Unnamed: 0" dan "Unnamed: 1" akan diubah namanya menjadi kolom yang relevan, seperti "Kawasan" dan "Negara" untuk mempermudah analisis.

	Unnamed: 0	Unnamed: 1	Unnamed: 2	Bekerja	Belajar	Magang	Lain-lain	Mendampingi Suami/Istri	Anggota Keluarga/Pengikut
0	Asia Pasifik dan Afrika - Asean	BRUNEI DARUSSALAM	21441	19309	153	0	202	1239	538
1	Asia Pasifik dan Afrika - Asean	FILIPINA	6379	3751	386	8	390	428	1416
2	Asia Pasifik dan Afrika - Asean	KAMBOJA	14157	10722	21	197	1930	424	863
3	Asia Pasifik dan Afrika - Asean	LAOS	238	153	2	0	1	33	49
4	Asia Pasifik dan Afrika - Asean	MALAYSIA	567311	309634	8343	107	238057	10591	579

Gambar 3. 27 Membaca 5 baris pertama project #2

Gambar 3.27 menunjukkan hasil pembacaan lima baris pertama (head) pada dataset Project 2 sebagai langkah verifikasi awal struktur dan isi data setelah proses penentuan header. Melalui pengecekan ini, penulis memastikan bahwa kolom-kolom utama telah terbaca dengan benar serta nilai data berada pada posisi yang sesuai.

Setelah struktur awal dataset diverifikasi, tahap selanjutnya adalah melakukan penyesuaian nama kolom agar lebih representatif dan mudah dipahami. Kolom-kolom yang sebelumnya masih menggunakan label tidak informatif, seperti Unnamed, diubah menjadi nama kolom yang relevan, yaitu Kawasan, Negara, dan Jumlah\_Total. Penyesuaian ini dilakukan untuk meningkatkan keterbacaan data serta mempermudah proses analisis dan visualisasi menggunakan Microsoft Power BI pada tahap akhir project.

```
Renaming Unnamed to the relevant column name

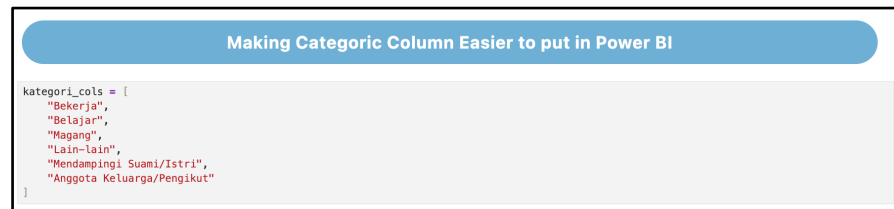
df = df.rename(columns={
    "Unnamed: 0": "Kawasan",
    "Unnamed: 1": "Negara",
    "Unnamed: 2": "Jumlah_Total"
})
```

Gambar 3. 28 Mengubah nama unnamed menjadi column yang relevant

Gambar 3.28 menunjukkan proses pengubahan nama kolom yang sebelumnya masih bersifat tidak informatif (Unnamed) menjadi nama kolom yang lebih relevan dan representatif. Penamaan ulang kolom ini dilakukan untuk memastikan setiap atribut memiliki makna yang jelas, sehingga memudahkan proses analisis dan interpretasi data pada tahap selanjutnya.

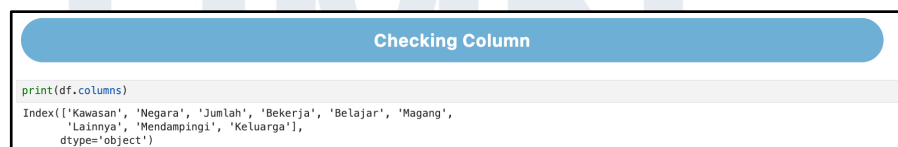
Setelah memverifikasi struktur kolom, langkah berikutnya adalah melakukan pembersihan lebih lanjut untuk memastikan bahwa data

yang digunakan dalam analisis adalah data yang bersih dan terstruktur dengan baik. Setelah data dibersihkan, proses dapat dilanjutkan dengan proses analisis yang lebih dalam, seperti membuat distribusi peserta berdasarkan kategori mereka tujuan diaspora.



Gambar 3. 29 Membuat kategorikal - tujuan menetap

Gambar 3.29 menunjukkan proses pembentukan variabel kategorikal pada dataset Project 2, khususnya untuk atribut tujuan menetap WNI di luar negeri. Dengan struktur kolom yang telah memiliki penamaan yang jelas, data kemudian dikelompokkan ke dalam kategori-kategori tujuan seperti Bekerja, Belajar, Magang, serta kategori lainnya. Pembentukan variabel kategorikal ini bertujuan untuk mempermudah proses analisis dan visualisasi data pada Microsoft Power BI, terutama dalam menyajikan distribusi diaspora berdasarkan tujuan keberadaan mereka.



Gambar 3. 30 Memeriksa kolom data project #2

Hasil dari pemeriksaan kolom menunjukkan bahwa dataset memiliki kolom-kolom yang relevan seperti "Kawasan", "Negara", "Jumlah\_Total", "Kategori", dan "Jumlah". Langkah berikutnya



adalah memeriksa 5 baris pertama untuk mengetahui hasil setelah mengecek kolom tadi.

Checking the 5 first Data					
df_long.head()					
	Kawasan	Negara	Jumlah_Total	Kategori	Jumlah
0	Asia Pasifik dan Afrika - Asean	BRUNEI DARUSSALAM	21441	Bekerja	19309
1	Asia Pasifik dan Afrika - Asean	FILIPINA	6379	Bekerja	3751
2	Asia Pasifik dan Afrika - Asean	KAMBOJA	14157	Bekerja	10722
3	Asia Pasifik dan Afrika - Asean	LAOS	238	Bekerja	153
4	Asia Pasifik dan Afrika - Asean	MALAYSIA	567311	Bekerja	309634

Gambar 3. 31 Memeriksa 5 rows pertama pada data project #2

Gambar 3.31 menunjukkan hasil pemeriksaan lima baris pertama pada dataset Project 2 menggunakan perintah `df_long.head()`. Pengecekan ini dilakukan untuk memastikan bahwa struktur data telah terbentuk dengan baik setelah proses pembersihan dan transformasi. Hasilnya menunjukkan bahwa dataset telah tersusun secara rapi dengan kolom-kolom utama, seperti Kawasan, Negara, Jumlah\_Total, Kategori, dan Jumlah, yang masing-masing terisi dengan nilai yang relevan. Melalui pengecekan ini, diperoleh gambaran awal mengenai distribusi kategori tujuan diaspora, seperti kategori Bekerja, yang menjadi salah satu komponen utama dalam analisis. Setelah struktur dan isi data terverifikasi, proses dilanjutkan ke tahap pemeriksaan missing value untuk memastikan tidak terdapat kolom dengan nilai kosong yang berpotensi memengaruhi akurasi visualisasi dan analisis data pada dashboard Power BI.

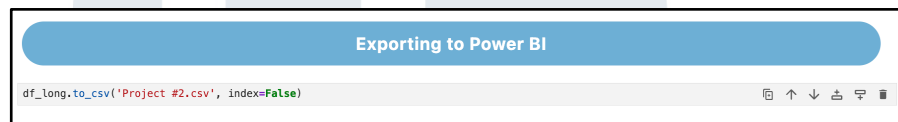
```
missing_values = df_long.isnull().sum()
print("Missing Values (NaN):")
print(missing_values)

Missing Values (NaN):
Kawasan      0
Negara        6
Jumlah_Total  0
Kategori      0
Jumlah        0
dtype: int64
```

Gambar 3. 32 Memeriksa missing value pada project #2

Gambar 3.32 menunjukkan hasil pemeriksaan missing value pada dataset Project 2. Berdasarkan hasil analisis, hanya kolom Negara yang memiliki enam nilai kosong, sementara kolom lainnya, seperti Kawasan, Jumlah\_Total, Kategori, dan Jumlah, tidak menunjukkan

adanya nilai yang hilang. Temuan ini menunjukkan bahwa secara umum dataset telah berada dalam kondisi yang cukup baik untuk digunakan pada tahap analisis dan visualisasi. Pemeriksaan missing value ini dilakukan untuk memastikan bahwa dataset yang digunakan tidak mengandung kekosongan data yang dapat memengaruhi akurasi hasil visualisasi pada dashboard. Setelah memastikan kondisi data relatif bersih dan siap digunakan, proses pengolahan dilanjutkan ke tahap akhir, yaitu mengekspor dataset ke dalam format CSV agar dapat diimpor dan divisualisasikan menggunakan Microsoft Power BI.



Gambar 3. 33 Export CSV project #2

Gambar 3.33 ini menampilkan perintah `df_long.to_csv('Project #2.csv', index=False)`, dataset yang telah diproses akan disimpan dalam format CSV. File CSV ini kemudian dapat diimpor ke Power BI untuk dilakukan analisis lebih lanjut dan visualisasi data. Namun penting untuk memahami isi dari detail kolom pada project #2

Tabel 3. 3 Nama Kolom dan deskripsi data project #2

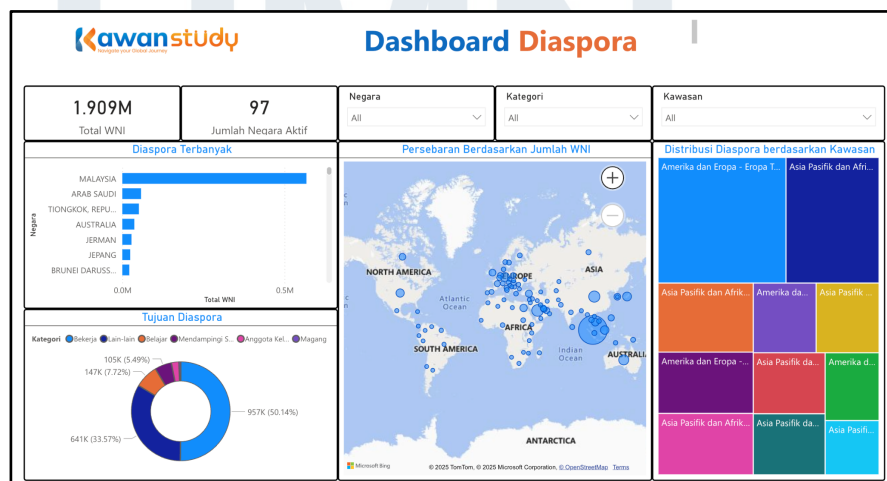
No	Nama Kolom	Deskripsi
1	Kawasan	Kawasan geografis tempat WNI berada
2	Negara	Negara tujuan tempat WNI menetap
3	Jumlah Total	Total jumlah WNI di negara tersebut
4	Kategori	Kategori tujuan
5	Jumlah	Jumlah WNI per kategori

Dataset pada tabel 3.3 ini terdiri dari 5 kolom utama yang menggambarkan distribusi Warga Negara Indonesia yang menetap di

luar negeri. Kolom Kawasan dan Negara bertipe object yang menyimpan informasi geografis dalam bentuk teks, sementara kolom Jumlah\_Total dan Jumlah bertipe int64 yang berisi data numerik untuk keperluan agregasi dan analisis kuantitatif. Kolom Kategori mencatat tujuan WNI menetap di luar negeri seperti "Bekerja", "Belajar", "Magang", "Mendampingi Suami/Istri", dan "Anggota Keluarga/Pengikut". Struktur data ini memudahkan proses visualisasi dan analisis sebaran diaspora Indonesia berdasarkan lokasi dan tujuan keberadaan mereka.

### 3.3.2.3 Dokumentasi final Project 2, QA KPI, arsip, dan wrap-up magang

Dashboard pada gambar 3.34 ini menyajikan data mengenai diaspora WNI di luar negeri dalam tampilan yang informatif dan interaktif. Di bagian atas dashboard, informasi mengenai total WNI di luar negeri dan jumlah negara aktif yang memiliki diaspora WNI disajikan untuk memberikan gambaran umum. Total diaspora WNI yang tercatat mencapai 1.909 juta, dengan 97 negara aktif yang menampung diaspora tersebut.



Gambar 3. 34 Dashboard Tujuan Diaspora Menetap

Di bagian atas dashboard, ditampilkan total jumlah WNI yang berada di luar negeri, yaitu 1.909 juta, serta jumlah negara aktif yang

memiliki diaspora WNI, yang mencapai 97 negara. Di bawah informasi ini, terdapat grafik batang yang menunjukkan negara-negara dengan diaspora terbanyak. Dari grafik tersebut, terlihat bahwa Malaysia memiliki jumlah diaspora WNI yang paling banyak, diikuti oleh negara-negara seperti Arab Saudi, Tiongkok, Australia, Jerman, dan Jepang.

Di sebelah grafik batang, terdapat peta dunia yang menunjukkan penyebaran diaspora WNI berdasarkan jumlah. Peta ini menampilkan titik-titik yang ukurannya mencerminkan jumlah WNI di masing-masing negara, memberikan visualisasi yang jelas tentang konsentrasi diaspora WNI di berbagai wilayah dunia.

Bagian kanan dashboard menunjukkan distribusi diaspora berdasarkan kawasan. Kawasan-kawasan ini dibedakan dengan warna yang berbeda, seperti Amerika dan Eropa serta Asia Pasifik dan Afrika, yang memberikan informasi mengenai sebaran diaspora WNI di masing-masing kawasan geografis.

Di bagian bawah, terdapat diagram lingkaran yang menggambarkan tujuan diaspora WNI berdasarkan kategori. Kategori yang ditampilkan antara lain Bekerja, yang merupakan kategori terbesar dengan jumlah 641K WNI (33,57%), diikuti oleh kategori Mendampingi yang mencakup 957K WNI (50,14%). Kategori lainnya seperti Belajar dan Lain-lain masing-masing mencakup 147K (7,72%) dan 105K (5,49%) WNI.

Dashboard ini memberikan tampilan yang sangat interaktif dan informatif mengenai keberadaan dan distribusi diaspora WNI di seluruh dunia, mempermudah analisis lebih lanjut terkait pola migrasi dan tujuan diaspora tersebut.

### **3.3.3 Kendala yang Ditemukan**

1. Pengumpulan data pendaftar melalui sheets mengalami keterlambatan karena target responden belum tercapai sesuai timeline yang direncanakan. Hal ini menyebabkan tertundanya proses cleaning data, integrasi dataset psikotes dengan data akademik, serta validasi kelengkapan informasi pendaftar. Keterlambatan ini berdampak pada mundurnya jadwal pembuatan dashboard dan analisis lanjutan.
2. Inkonsistensi Format dan Kualitas Data yang diekspor dari sheets ke CSV raw mengalami beberapa masalah struktur, antara lain: perbedaan encoding file UTF dengan ANSI, format delimiter yang tidak konsisten, serta adanya kolom tanpa header bahkan setelah analisis (Unnamed). Selain itu, ditemukan anomali seperti duplikasi data, variasi kapitalisasi nama dan kota, format tanggal lahir yang tidak seragam, serta nilai usia yang tidak valid. Kondisi ini memerlukan waktu ekstra untuk melakukan standarisasi dan validasi ulang sebelum data dapat digunakan.
3. Keterbatasan Interpretasi Data Mentah Data dalam bentuk tabel CSV mentah sulit diinterpretasikan secara langsung, terutama untuk kolom-kolom penting seperti hasil psikotes, distribusi provinsi, jalur pendaftaran, dan kategori ekonomi. Tanpa visualisasi yang memadai, proses identifikasi pola, perbandingan antar kelompok, dan pengambilan insight menjadi kurang efisien dan membutuhkan waktu lebih lama.

### **3.3.4 Solusi atas Kendala yang Ditemukan**

Penyajian data dalam bentuk tabel mentah pada tahap awal membuat proses interpretasi menjadi kurang efisien. Beberapa kolom penting seperti kategori ekonomi, hasil psikotes, dan nilai rapor sulit dianalisis tanpa visualisasi yang representatif, sehingga diperlukan pengembangan dashboard Business Intelligence agar hasil analisis dapat disajikan secara interaktif dan mudah dipahami untuk mendukung pengambilan keputusan berbasis data.

1. Koordinasi Rutin dan Penyesuaian Timeline dengan melakukan koordinasi berkala dengan pembimbing lapangan dan tim internal Kawan Study untuk memantau progres pengumpulan data. Sebagai langkah mitigasi, timeline pengerjaan disesuaikan dengan memprioritaskan aktivitas yang tidak bergantung langsung pada kelengkapan data, dokumentasi alur kerja, dan penyusunan template dashboard. Pendekatan ini memastikan bahwa proses kerja tetap produktif meskipun data belum lengkap sepenuhnya.
2. Otomasi Pembersihan Data dengan menggunakan script Python menggunakan library *pandas* dan *numpy* di Jupyter Notebook untuk melakukan standarisasi data secara otomatis. Script tersebut mencakup: konversi encoding ke UTF-8, unifikasi format tanggal, normalisasi kapitalisasi teks, penghapusan duplikasi berdasarkan User\_ID, serta deteksi dan penanganan missing values. Proses otomasi ini mempercepat tahap preprocessing dan memastikan konsistensi struktur data sebelum masuk ke tahap integrasi.
3. Pengembangan Dashboard Interaktif dengan Power BI Sebagai solusi atas keterbatasan interpretasi data mentah, merancang dashboard Business Intelligence menggunakan Power BI yang menyajikan visualisasi interaktif seperti grafik distribusi peserta, peta sebaran provinsi, skor psikotes, dan perbandingan jalur pendaftaran. Dashboard ini dilengkapi dengan fitur filter dinamis yang memungkinkan pengguna mengeksplorasi data berdasarkan kategori tertentu, sehingga insight dapat diperoleh dengan cepat dan mendukung pengambilan keputusan yang lebih efektif.