

Data Science

**“Analisis Data Global Student Migration untuk Memahami Pola
dan Tren Mobilitas Pelajar Internasional”**



Dibuat Oleh:

Kelompok 19

11423058 | Rakhel Franiska Tampubolon

11423060 | Gracia Anggraini Manik

11423064 | Yenny Angelita Gurning

INSTITUT TEKNOLOGI DEL

MATERI PRAKTIKUM

Semester V Tahun Ajar 2025/2026

Daftar Isi

Daftar Isi	1
Daftar Gambar	2
Bab 1 Pendahuluan	3
1.1 Latar Belakang	3
1.2 Rumusan Masalah	4
1.3 Tujuan	5
Bab 2 Metode	6
2.1 Jenis dan Pendekatan Penelitian	6
2.2. Alur Pengerjaan	7
2.2.1. Data Collection	8
2.2.2. Data Preprocessing (Advance Preprocessing)	10
2.2.3. Data Visualization	19
2.2.4 Statistical Analysis	27
Bab 3 Hasil dan Pembahasan	34
Bab 4 Kesimpulan	36
Daftar Pustaka	38

Daftar Gambar

Gambar 2. 1 Alur Tahapan Penelitian.....	7
Gambar 2. 2 Kode Tahap Data Collection.....	8
Gambar 2. 3 Inspect & Missing Values	11
Gambar 2. 4 Mengecek Jumlah Nilai Kosong di Setiap Kolom.....	12
Gambar 2. 5 Proses Penanganan Missing Values Dengan Imputasi Kategori	13
Gambar 2. 6 Proses Penanganan Missing Values	14
Gambar 2. 7 Mengecek Outlier.....	15
Gambar 2. 8 Penanganan Outlier	15
Gambar 2. 9 Mengecek Ulang Outlier	16
Gambar 2. 10 Cek Kondisi Awal Data Numerik	17
Gambar 2. 11 Normalization.....	18
Gambar 2. 12 Visualisasi Pola Migrasi Pelajar Internasional.....	19
Gambar 2. 13 Visualisasi Alasan Pendaftaran Terpopuler	20
Gambar 2. 14 Visualisasi Hubungan GPA Dan Skor Bahasa Dengan Gaji Awal.....	21
Gambar 2. 15 Visualisasi Hubungan GPA Dan Skor Bahasa Dengan Gaji Awal.....	22
Gambar 2. 16 Visualisasi Tingkat Penempatan Kerja Per Negara Tujuan	23
Gambar 2. 17 Visualisasi Korelasi Antar Variabel Numerik	24
Gambar 2. 18 Visualisasi Distribusi Gaji Awal Mahasiswa.....	25
Gambar 2. 19 Uji Parametrik (t-test)	27
Gambar 2. 20 Uji Parametrik (Anova).....	29
Gambar 2. 21 Uji Non-parametrik (Kruskal-Wallis).....	32

Bab 1

Pendahuluan

1.1 Latar Belakang

Mobilitas pelajar internasional adalah fenomena yang semakin berkembang di seluruh dunia, seiring dengan kemajuan dalam sektor pendidikan dan proses globalisasi. Setiap tahunnya, ribuan siswa berpindah dari satu negara ke negara lainnya demi mengejar pendidikan yang lebih baik, mendapatkan pengalaman budaya yang berbeda, serta memperluas kesempatan karir mereka di masa mendatang. Kumpulan data migrasi Pelajar Global memberikan informasi mengenai arus mahasiswa internasional antar negara selama berbagai periode.

Dalam beberapa dekade terakhir, fenomena migrasi pelajar internasional menjadi salah satu indikator penting dalam perkembangan globalisasi pendidikan tinggi. Semakin banyak mahasiswa yang memilih untuk menempuh pendidikan di luar negeri dengan berbagai alasan, seperti kualitas universitas, peluang kerja setelah lulus, ketersediaan beasiswa, maupun keinginan untuk mendapatkan pengalaman lintas budaya [1]. Kondisi ini mendorong munculnya arus perpindahan mahasiswa yang signifikan dari satu negara ke negara lain.

Analisis terhadap data migrasi pelajar global menjadi penting karena dapat memberikan gambaran tentang pola dan tren pendidikan internasional, negara tujuan yang paling diminati, serta faktor-faktor yang mempengaruhi keputusan mahasiswa dalam memilih universitas di luar negeri. Selain itu, data ini juga dapat membantu dalam memahami bagaimana faktor seperti asal negara, bidang studi, penerimaan beasiswa, prestasi akademik (GPA dan nilai tes bahasa), serta status penempatan kerja setelah kelulusan saling berkaitan satu sama lain.

Melalui pendekatan berbasis data (data-driven approach), fenomena ini dapat dikaji secara lebih objektif. Misalnya, dengan menganalisis hubungan antara prestasi akademik dengan gaji awal setelah lulus, atau meninjau perbedaan outcome karier antara mahasiswa penerima

beasiswa dan non-penerima beasiswa. Pemahaman ini bermanfaat tidak hanya bagi lembaga pendidikan tinggi dan pembuat kebijakan, tetapi juga bagi calon mahasiswa yang sedang merencanakan studi internasional mereka.

Dengan demikian, penelitian ini bertujuan untuk mengolah dan menganalisis dataset Global Student Migration secara mendalam agar diperoleh wawasan mengenai pola mobilitas pelajar internasional, faktor yang memengaruhi keberhasilan akademik dan profesional, serta implikasinya terhadap kebijakan pendidikan lintas negara di masa depan.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah dijelaskan sebelumnya, penelitian ini berfokus pada analisis data migrasi pelajar internasional untuk memahami pola, faktor, serta keterkaitan antar variabel yang mempengaruhi pengalaman akademik dan profesional mahasiswa. Oleh karena itu, rumusan masalah dalam penelitian ini dapat dijabarkan sebagai berikut:

1. Bagaimana pola migrasi pelajar internasional berdasarkan negara asal dan negara tujuan studi?
2. Apa saja faktor utama yang mempengaruhi keputusan mahasiswa dalam memilih negara dan universitas tujuan?
3. Bagaimana hubungan antara prestasi akademik (GPA dan skor tes bahasa) dengan outcome karier awal (starting salary) setelah kelulusan?
4. Apakah terdapat perbedaan yang signifikan dalam gaji awal antara mahasiswa penerima beasiswa dan non-penerima beasiswa?
5. Negara tujuan mana yang menunjukkan tingkat keberhasilan penempatan kerja (placement status) tertinggi bagi lulusan internasional?
6. Sejauh mana visa pasca kelulusan (post-graduation visa) berperan dalam mempengaruhi peluang kerja dan gaji awal mahasiswa internasional?

1.3 Tujuan

Penelitian ini memiliki tujuan untuk menganalisis data migrasi pelajar internasional secara komprehensif dengan memanfaatkan pendekatan berbasis data (data-driven analysis). Tujuan-tujuan yang ingin dicapai dalam penelitian ini antara lain:

1. Mengidentifikasi pola migrasi pelajar internasional, baik berdasarkan negara asal maupun negara tujuan studi.
2. Menganalisis faktor-faktor utama yang mempengaruhi keputusan mahasiswa dalam memilih universitas dan negara tujuan, seperti alasan pendaftaran, penerimaan beasiswa, serta bidang studi yang diminati.
3. Mengetahui hubungan antara prestasi akademik (GPA dan skor tes bahasa) dengan outcome karier awal (starting salary) untuk melihat apakah kinerja akademik berpengaruh terhadap peluang kerja dan kompensasi awal.
4. Membandingkan perbedaan gaji awal antara mahasiswa penerima beasiswa dan non-penerima beasiswa untuk mengetahui apakah pemberian beasiswa berdampak terhadap outcome karier.
5. Menilai tingkat keberhasilan penempatan kerja (placement status) di berbagai negara tujuan untuk menentukan negara mana yang memberikan peluang kerja terbaik bagi lulusan internasional.
6. Memberikan rekomendasi kebijakan dan strategi bagi lembaga pendidikan, pemerintah, serta mahasiswa dalam meningkatkan kualitas dan relevansi pendidikan internasional berdasarkan hasil analisis data.

Bab 2

Metode

2.1 Jenis dan Pendekatan Penelitian

Penelitian ini menggunakan pendekatan kuantitatif deskriptif dengan metode analisis data eksploratif (Exploratory Data Analysis/EDA). Pendekatan ini dipilih karena bertujuan untuk menggali pola, tren, dan hubungan antar variabel dalam dataset Global Student Migration tanpa melakukan intervensi langsung terhadap objek penelitian.

Pendekatan kuantitatif digunakan karena seluruh analisis berfokus pada pengolahan data numerik dan pengujian statistik untuk menemukan hubungan antar variabel seperti Grade Point Average (GPA), skor tes bahasa, status beasiswa, dan gaji awal setelah kelulusan. Metode deskriptif dipilih agar hasil analisis dapat menggambarkan kondisi nyata dari fenomena migrasi pelajar internasional sebagaimana tercermin dalam data.

Sementara itu, metode Exploratory Data Analysis (EDA) digunakan untuk memahami struktur data secara mendalam melalui proses pembersihan data, transformasi, dan visualisasi. Menurut Tukey (1977), EDA bertujuan untuk menemukan pola tersembunyi dan mendeteksi anomali pada data melalui pendekatan statistik dan grafis sebelum dilakukan analisis lanjutan. Pendekatan ini terbukti efektif dalam mengekstraksi insight awal dari dataset besar yang bersifat kompleks dan heterogen, termasuk dalam studi pendidikan internasional [2].

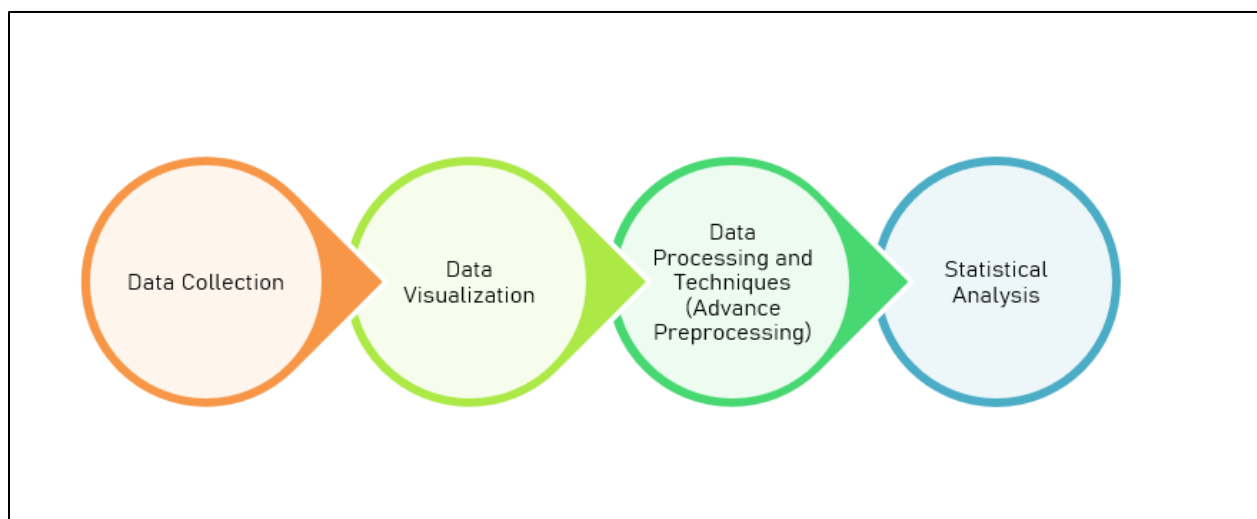
Seluruh proses analisis dilakukan dengan bantuan perangkat lunak Python, menggunakan pustaka seperti pandas, numpy, matplotlib, seaborn, scipy, dan scikit-learn. Pemanfaatan bahasa pemrograman ini memungkinkan analisis dilakukan secara reproduktif, efisien, dan transparan, sehingga hasilnya dapat diuji ulang oleh peneliti lain.

Dengan kombinasi metode kuantitatif deskriptif dan EDA berbasis Python, penelitian ini diharapkan dapat memberikan gambaran yang komprehensif mengenai pola migrasi pelajar internasional.

2.2. Alur Pengerjaan

Penelitian ini dilakukan melalui beberapa tahapan yang sistematis agar proses analisis data dapat menghasilkan informasi yang akurat dan dapat dipertanggungjawabkan.

Secara umum, tahapan yang dilakukan terdiri dari pengumpulan data (Data Collection), pembersihan dan transformasi data (Data Preprocessing), visualisasi data (Data Visualization), serta analisis statistik (Statistical Analysis). Berikut gambar alur tahapan penelitian yang dilakukan [3].



Gambar 2. 1 Alur Tahapan Penelitian

Gambar 2.1 menunjukkan bahwa penelitian dimulai dengan tahap Data Collection, dilanjutkan dengan Data Processing and Techniques (Advance Preprocessing) untuk memastikan data bersih dan konsisten. Tahap berikutnya adalah Data Visualization untuk menampilkan pola atau hubungan antar variabel, dan diakhiri dengan Statistical Analysis guna menguji hubungan serta perbedaan antar variabel secara kuantitatif.

2.2.1. Data Collection

Tahap awal penelitian ini adalah pengumpulan data (Data Collection) yang bertujuan untuk memperoleh informasi yang relevan dan dapat mendukung proses analisis. Dataset yang digunakan berjudul “global_student_migration.csv”, yang berisi data mengenai mobilitas pelajar internasional dari berbagai negara. Data ini memuat informasi akademik, negara asal dan tujuan, hingga hasil karier setelah lulus.

Dataset ini diperoleh dari sumber terbuka yang dikurasi dan diolah kembali untuk kepentingan analisis penelitian. Data bersifat sekunder karena berasal dari sumber publik, bukan hasil survei atau eksperimen langsung.

```
# Persiapan & Pembacaan Data

import pandas as pd
import numpy as np
from pathlib import Path

# ubah path sesuai lokasi file kamu
file_path = Path('global_student_migration.csv')

# baca data
df = pd.read_csv(file_path)

# info dasar
print(df.shape)
print(df.columns)
display(df.head())

(5000, 20)
Index(['student_id', 'origin_country', 'destination_country',
       'destination_city', 'university_name', 'course_name', 'field_of_study',
       'year_of_enrollment', 'scholarship_received', 'enrollment_reason',
       'graduation_year', 'placement_status', 'placement_country',
       'placement_company', 'starting_salary_usd', 'gpa_or_score',
       'visa_status', 'post_graduation_visa', 'language_proficiency_test',
       'test_score'],
      dtype='object')
```

	student_id	origin_country	destination_country	destination_city	university_name	course_name	field_of_study	year_of_enrollment	scholarship_received	enrollmen
0	S00001	Finland	Russia	Moscow	Lomonosov Moscow State University	Computer Science	Engineering	2021	No	Higher
1	S00002	UK	Germany	Aachen	RWTH Aachen	Civil Engineering	Law	2023	Yes	Job Opp
2	S00003	Ireland	Canada	Vancouver	University of British Columbia	Law	Arts	2019	No	Sc
3	S00004	UAE	UK	Birmingham	University of Birmingham	Data Science	Social Sciences	2021	Yes	Qual
4	S00005	South Africa	Germany	Stuttgart	University of Stuttgart	Business Administration	Law	2020	Yes	Politica

Gambar 2. 2 Kode Tahap Data Collection

Rincian Dataset:

- Nama Dataset: Global Student Migration
- Format File: .csv (Comma-Separated Values)
- Jumlah Data: 5000 baris dan 20 kolom
- Sumber Dataset: Kaggle
- Link Referensi Dataset Publik: ([Link Kaggle](#))

Deskripsi Variabel Utama dalam Dataset:

No	Nama Kolom	Deskripsi	Tipe Data
1.	student_id	Nomor identitas unik setiap pelajar	Numerik (integer)
2.	origin_country	Negara asal pelajar	Kategorikal
3.	destination_country	Negara tujuan studi	Kategorikal
4.	destination_city	Kota tempat universitas berada	Kategorikal
5.	university_name	Nama universitas tempat studi	Kategorikal
6.	course_name	Nama program studi yang diambil	Kategorikal
7.	field_of_study	Bidang studi (mis. Engineering, Business, IT)	Kategorikal
8.	year_of_enrollment	Tahun pelajar memulai studi	Numerik (integer)
9.	scholarship_received	Status penerimaan beasiswa (Ya/Tidak)	Kategorikal
10.	enrollment_reason	Alasan utama memilih universitas/negara tujuan	Kategorikal
11.	graduation_year	Tahun kelulusan	Numerik (integer)
12.	placement_status	Status penempatan kerja setelah kelulusan	Kategorikal
13.	placement_country	Negara tempat bekerja setelah lulus	Kategorikal

14.	placement_company	Nama perusahaan tempat bekerja	Kategorikal
15.	starting_salary_usd	Gaji awal setelah kelulusan (dalam USD)	Numerik (float)
16.	gpa_or_score	Nilai akademik (GPA atau skor setara)	Numerik (float)
17.	visa_status	Status visa selama studi	Kategorikal
18.	post_graduation_visa	Status visa setelah kelulusan (ada/tidak)	Kategorikal
19.	language_proficiency_test	Jenis tes bahasa yang diambil (IELTS/TOEFL/dll)	Kategorikal
20.	test_score	Skor kemampuan bahasa internasional	Numerik (float)

2.2.2. Data Preprocessing (Advance Preprocessing)

Tahap Data Preprocessing merupakan langkah penting dalam penelitian ini karena menentukan kualitas hasil analisis yang akan diperoleh. Data yang berasal dari sumber publik biasanya memiliki ketidaksempurnaan, seperti nilai kosong (missing values), nilai ekstrem (outliers), perbedaan format, dan jenis data yang tidak konsisten. Oleh karena itu, dilakukan serangkaian proses pembersihan dan transformasi agar dataset siap digunakan dalam tahap analisis lanjutan.

Proses preprocessing dilakukan menggunakan bahasa pemrograman Python dengan pustaka utama seperti pandas, numpy, dan scikit-learn.

```
# Inspect & Missing Values
print(df.dtypes)
print(df.isna().sum().sort_values(ascending=False))
```

student_id	object
origin_country	object
destination_country	object
destination_city	object
university_name	object
course_name	object
field_of_study	object
year_of_enrollment	int64
scholarship_received	object
enrollment_reason	object
graduation_year	int64
placement_status	object
placement_country	object
placement_company	object
starting_salary_usd	int64
gpa_or_score	float64
visa_status	object
post_graduation_visa	object
language_proficiency_test	object
test_score	float64
dtype: object	
placement_company	2491
placement_country	2491
language_proficiency_test	982
student_id	0
origin_country	0
post_graduation_visa	0
visa_status	0
gpa_or_score	0
starting_salary_usd	0
placement_status	0
graduation_year	0
enrollment_reason	0
scholarship_received	0
year_of_enrollment	0
field_of_study	0
course_name	0
university_name	0
destination_city	0
destination_country	0
test_score	0

Gambar 2. 3 Inspect & Missing Values

Dilakukan analisis awal terhadap struktur dan integritas data yang terdapat dalam DataFrame. Proses ini dimulai dengan penggunaan fungsi `print(df.dtypes)` untuk menampilkan tipe data dari setiap kolom dalam DataFrame. Hasilnya menunjukkan berbagai tipe data, seperti object untuk kolom yang berisi teks, int64 untuk angka bulat, dan float64 untuk angka desimal. Selanjutnya, untuk mengidentifikasi nilai yang hilang, digunakan fungsi `print(df.isna().sum().sort_values(ascending=False))`, yang menghitung dan mengurutkan jumlah nilai hilang dalam setiap kolom. Informasi ini sangat penting karena kolom-kolom tertentu, seperti `placement_country` dan `placement_company`, menunjukkan jumlah nilai yang hilang yang cukup signifikan, yaitu 2491, sedangkan kolom `student_id` tidak memiliki nilai yang hilang. Dengan pemahaman ini, langkah-langkah pembersihan dan transformasi data dapat direncanakan untuk memastikan kualitas data yang lebih baik sebelum analisis lebih lanjut.

1. Pemeriksaan Nilai Hilang (Handling Missing Values)

Langkah pertama dilakukan dengan memeriksa keberadaan nilai kosong untuk setiap kolom. Nilai kosong dapat mengganggu hasil analisis, sehingga harus ditangani dengan metode yang tepat.

```
# Mengecek jumlah nilai kosong di setiap kolom
df.isnull().sum()

student_id          0
origin_country      0
destination_country 0
destination_city    0
university_name     0
course_name         0
field_of_study      0
year_of_enrollment 0
scholarship_received 0
enrollment_reason  0
graduation_year     0
placement_status    0
placement_country   2491
placement_company    2491
starting_salary_usd 0
gpa_or_score        0
visa_status         0
post_graduation_visa 0
language_proficiency_test 982
test_score          0
dtype: int64
```

Gambar 2. 4 Mengecek Jumlah Nilai Kosong di Setiap Kolom

Pada gambar diatas dapat kita lihat beberapa kolom yang memiliki missing values dan jumlah missing values nya di antara lain:

Kolom	Jumlah Missing Value
placement_country	2491
placement_company	2491
language_proficiency_test	982
Kolom lain	0

Analisis Kolom dengan Nilai Kosong

- placement_country dan placement_company:

Nilai kosong kemungkinan berasal dari pelajar yang belum mendapatkan pekerjaan setelah lulus (karena placement_status-nya mungkin Not Placed). Jadi tidak dihapus, melainkan diganti dengan nilai "Not Applicable" agar tetap informatif.

- language_proficiency_test:

Nilai kosong bisa disebabkan oleh pelajar yang tidak mengambil tes bahasa internasional, misalnya jika negara asal dan tujuan menggunakan bahasa yang sama. Jadi juga diganti dengan "Not Taken".

```
# Menangani missing values dengan imputasi kategori
df['placement_country'] = df['placement_country'].fillna('Not Applicable')
df['placement_company'] = df['placement_company'].fillna('Not Applicable')
df['language_proficiency_test'] = df['language_proficiency_test'].fillna('Not Taken')

# Verifikasi ulang setelah imputasi
print(df.isnull().sum())

student_id          0
origin_country       0
destination_country  0
destination_city     0
university_name      0
course_name          0
field_of_study       0
year_of_enrollment  0
scholarship_received 0
enrollment_reason    0
graduation_year      0
placement_status     0
placement_country    0
placement_company     0
starting_salary_usd  0
gpa_or_score         0
visa_status          0
post_graduation_visa 0
language_proficiency_test 0
test_score           0
dtype: int64
```

Gambar 2. 5 Proses Penanganan Missing Values Dengan Imputasi Kategori

Selain itu proses penanganan nilai kosong (handling missing values) dilakukan menggunakan kombinasi teknik imputasi statistik dan interpolasi. Untuk kolom numerik seperti starting_salary_usd dan gpa_or_score, digunakan metode median imputation karena lebih tahan terhadap keberadaan outlier, sedangkan kolom test_score menggunakan mean imputation untuk menjaga kestabilan nilai rata-rata data. Pada kolom kategorikal seperti visa_status dan placement_status, digunakan mode imputation agar nilai kosong diisi dengan kategori yang paling sering muncul. Selain itu, diterapkan juga interpolasi linear pada kolom test_score untuk memperhalus pengisian nilai numerik secara berurutan, terutama jika terdapat kecenderungan pola nilai yang berkelanjutan. Setelah seluruh proses imputasi dilakukan, hasil verifikasi menunjukkan

bahwa seluruh kolom telah bebas dari nilai kosong sehingga data siap digunakan pada tahap preprocessing berikutnya.

```
# Mean/Median Imputation untuk kolom numerik
df['starting_salary_usd'] = df['starting_salary_usd'].fillna(df['starting_salary_usd'].median())
df['gpa_or_score'] = df['gpa_or_score'].fillna(df['gpa_or_score'].median())
df['test_score'] = df['test_score'].fillna(df['test_score'].mean()) # contoh imputasi mean

# Mode Imputation untuk kolom kategorikal
df['visa_status'] = df['visa_status'].fillna(df['visa_status'].mode()[0])
df['placement_status'] = df['placement_status'].fillna(df['placement_status'].mode()[0])

# Interpolasi Linear untuk nilai numerik berurutan
# Biasanya dipakai kalau data berupa deret waktu, tapi tetap bisa digunakan untuk smoothing
df['test_score'] = df['test_score'].interpolate(method='linear')

# Verifikasi ulang hasilnya
print("\nJumlah missing values setelah imputasi:")
print(df.isnull().sum())

Jumlah missing values setelah imputasi:
student_id          0
origin_country      0
destination_country 0
destination_city    0
university_name     0
course_name         0
field_of_study      0
year_of_enrollment 0
scholarship_received 0
enrollment_reason   0
graduation_year     0
placement_status     0
placement_country    0
placement_company    0
starting_salary_usd  0
gpa_or_score         0
visa_status          0
post_graduation_visa 0
language_proficiency_test 0
test_score          0
dtype: int64
```

Gambar 2. 6 Proses Penanganan Missing Values

Setelah dijalankan, semua kolom memiliki nilai lengkap (0 missing values).

Data kini siap digunakan untuk tahap berikutnya yaitu Handling Outliers dan Normalization.

2. Deteksi dan Penanganan Outliers

Setelah proses imputasi nilai kosong selesai, tahap selanjutnya adalah mendeteksi dan menangani data ekstrem (outliers) yang dapat mempengaruhi hasil analisis statistik. Outlier merupakan nilai yang jauh berbeda dari mayoritas data dan dapat menyebabkan distorsi pada distribusi, terutama pada kolom numerik seperti `starting_salary_usd`, `gpa_or_score`, dan `test_score`. Untuk mengatasinya, digunakan kombinasi tiga teknik, yaitu Interquartile Range (IQR), Winsorization, dan Z-score Analysis.

```
# Mengecek Outlier
numerical_cols = ['starting_salary_usd', 'gpa_or_score', 'test_score']

for col in numerical_cols:
    Q1 = df[col].quantile(0.25)
    Q3 = df[col].quantile(0.75)
    IQR = Q3 - Q1

    lower_bound = Q1 - 1.5 * IQR
    upper_bound = Q3 + 1.5 * IQR

    outliers = df[(df[col] < lower_bound) | (df[col] > upper_bound)]
    print(f"{col}: {len(outliers)} outlier ditemukan")

starting_salary_usd: 0 outlier ditemukan
gpa_or_score: 0 outlier ditemukan
test_score: 982 outlier ditemukan
```

Gambar 2. 7 Mengecek Outlier

```
import numpy as np
from scipy import stats

numerical_cols = ['starting_salary_usd', 'gpa_or_score', 'test_score']

# Deteksi & Penanganan Outlier
for col in numerical_cols:
    Q1, Q3 = df[col].quantile([0.25, 0.75])
    IQR = Q3 - Q1
    lower, upper = Q1 - 1.5 * IQR, Q3 + 1.5 * IQR

    # Winsorization (ganti nilai ekstrem dengan batas bawah/atas)
    df[col] = np.where(df[col] < lower, lower,
                      np.where(df[col] > upper, upper, df[col]))

    # Deteksi tambahan (Z-score > 3)
    z_outliers = np.sum(np.abs(stats.zscore(df[col]))) > 3
    print(f"{col}: {z_outliers} nilai di luar ±3 SD setelah Winsorization")

# Statistik setelah penanganan
print("\nStatistik deskriptif setelah penanganan outlier:")
print(df[numerical_cols].describe())

starting_salary_usd: 0 nilai di luar ±3 SD setelah Winsorization
gpa_or_score: 0 nilai di luar ±3 SD setelah Winsorization
test_score: 0 nilai di luar ±3 SD setelah Winsorization

Statistik deskriptif setelah penanganan outlier:
```

	starting_salary_usd	gpa_or_score	test_score
count	5000.000000	5000.000000	5000.000000
mean	44552.531000	3.246820	5.936620
std	50719.190623	0.430659	2.406901
min	0.000000	2.500000	1.550000
25%	0.000000	2.877500	5.300000
50%	30516.500000	3.240000	6.500000
75%	87838.500000	3.610000	7.800000
max	149932.000000	4.000000	9.000000

Gambar 2. 8 Penanganan Outlier

Deteksi awal dilakukan dengan metode IQR, di mana nilai kuartil pertama (Q1) dan kuartil ketiga (Q3) dihitung untuk menentukan batas bawah ($Q1 - 1.5 \times IQR$) dan batas atas ($Q3 + 1.5 \times IQR$). Nilai yang berada di luar rentang tersebut dianggap sebagai outlier. Nilai-nilai ekstrem tersebut kemudian ditangani menggunakan Winsorization, yaitu menggantikan nilai di bawah atau di atas

batas dengan nilai ambang batas terdekat, sehingga distribusi data menjadi lebih seimbang tanpa perlu menghapus baris data.

Selanjutnya, dilakukan pemeriksaan tambahan dengan metode Z-score, yang mengukur seberapa jauh suatu nilai menyimpang dari rata-rata dalam satuan standar deviasi. Nilai yang memiliki $|Z| > 3$ dianggap sebagai outlier dan disesuaikan agar tetap berada dalam batas wajar.

Proses ini dilakukan menggunakan pustaka NumPy dan SciPy di Python. Setelah kombinasi metode IQR, Winsorization, dan Z-score diterapkan, distribusi data numerik menjadi lebih simetris dan stabil. Hasil ini memastikan bahwa tahap analisis statistik berikutnya tidak terpengaruh oleh nilai ekstrem yang dapat menurunkan akurasi hasil penelitian.

```
# Mengecek ulang Outlier
numerical_cols = ['starting_salary_usd', 'gpa_or_score', 'test_score']

for col in numerical_cols:
    Q1 = df[col].quantile(0.25)
    Q3 = df[col].quantile(0.75)
    IQR = Q3 - Q1

    lower_bound = Q1 - 1.5 * IQR
    upper_bound = Q3 + 1.5 * IQR

    outliers = df[(df[col] < lower_bound) | (df[col] > upper_bound)]
    print(f"{col}: {len(outliers)} outlier ditemukan")

starting_salary_usd: 0 outlier ditemukan
gpa_or_score: 0 outlier ditemukan
test_score: 0 outlier ditemukan
```

Gambar 2. 9 Mengecek Ulang Outlier

Setelah proses penyesuaian dilakukan, nilai ekstrim berhasil dikoreksi tanpa menghapus baris data. Distribusi kolom numerik menjadi lebih normal dan stabil, sehingga dataset siap digunakan untuk tahap berikutnya yaitu normalisasi dan standarisasi data.

3. Normalisasi Data (Normalization/Standardization)

Tahap terakhir dalam proses data preprocessing adalah Feature Scaling, yaitu proses menyeragamkan skala antar variabel numerik agar tidak ada satu variabel yang mendominasi perhitungan statistik. Sebelum dilakukan penskalaan, terlebih dahulu dilakukan pemeriksaan

kondisi awal data numerik dengan menggunakan fungsi `describe()` untuk melihat nilai minimum, maksimum, rata-rata, serta standar deviasi pada kolom `starting_salary_usd`, `gpa_or_score`, dan `test_score`.

```
numerical_cols = ['starting_salary_usd', 'gpa_or_score', 'test_score']

# Cek kondisi awal data numerik
print("Statistik sebelum Feature Scaling:")
print(df[numerical_cols].describe())
```

Statistik sebelum Feature Scaling:

	starting_salary_usd	gpa_or_score	test_score
count	5000.000000	5000.000000	5000.000000
mean	44552.531000	3.246820	5.936620
std	50719.190623	0.430659	2.406901
min	0.000000	2.500000	1.550000
25%	0.000000	2.877500	5.300000
50%	30516.500000	3.240000	6.500000
75%	87838.500000	3.610000	7.800000
max	149932.000000	4.000000	9.000000

Gambar 2. 10 Cek Kondisi Awal Data Numerik

Proses pemeriksaan ini ditunjukkan pada Gambar dibawah ini, yang menampilkan statistik deskriptif data sebelum dilakukan feature scaling. Dari hasil tersebut, terlihat bahwa setiap variabel memiliki skala yang berbeda-beda, misalnya `starting_salary_usd` memiliki rentang nilai yang sangat besar dibandingkan dua variabel lainnya.

Selanjutnya dilakukan dua metode feature scaling, yaitu Normalization (Min-Max Scaling) dan Standardization (Z-Score Scaling) menggunakan pustaka `scikit-learn`. Kode program untuk kedua metode tersebut dapat dilihat pada Gambar dibawah ini.

```
# Import pustaka yang dibutuhkan
from sklearn.preprocessing import MinMaxScaler, StandardScaler

numerical_cols = ['starting_salary_usd', 'gpa_or_score', 'test_score']

# Normalization (Min-Max Scaling)
minmax_scaler = MinMaxScaler()
df_norm = df.copy()
df_norm[numerical_cols] = minmax_scaler.fit_transform(df[numerical_cols])

print("\nStatistik setelah Normalization (Min-Max Scaling):")
print(df_norm[numerical_cols].describe())

# Standardization (Z-Score Scaling)
std_scaler = StandardScaler()
df_std = df.copy()
df_std[numerical_cols] = std_scaler.fit_transform(df[numerical_cols])

print("\nStatistik setelah Standardization (Z-Score Scaling):")
print(df_std[numerical_cols].describe())
```

```
Statistik setelah Normalization (Min-Max Scaling):
starting_salary_usd  gpa_or_score  test_score
count      5000.000000    5000.000000    5000.000000
mean         0.297152      0.497880      0.588808
std          0.338281      0.287106      0.323074
min          0.000000      0.000000      0.000000
25%          0.000000      0.251667      0.503356
50%          0.203536      0.493333      0.664430
75%          0.585856      0.740000      0.838926
max          1.000000      1.000000      1.000000

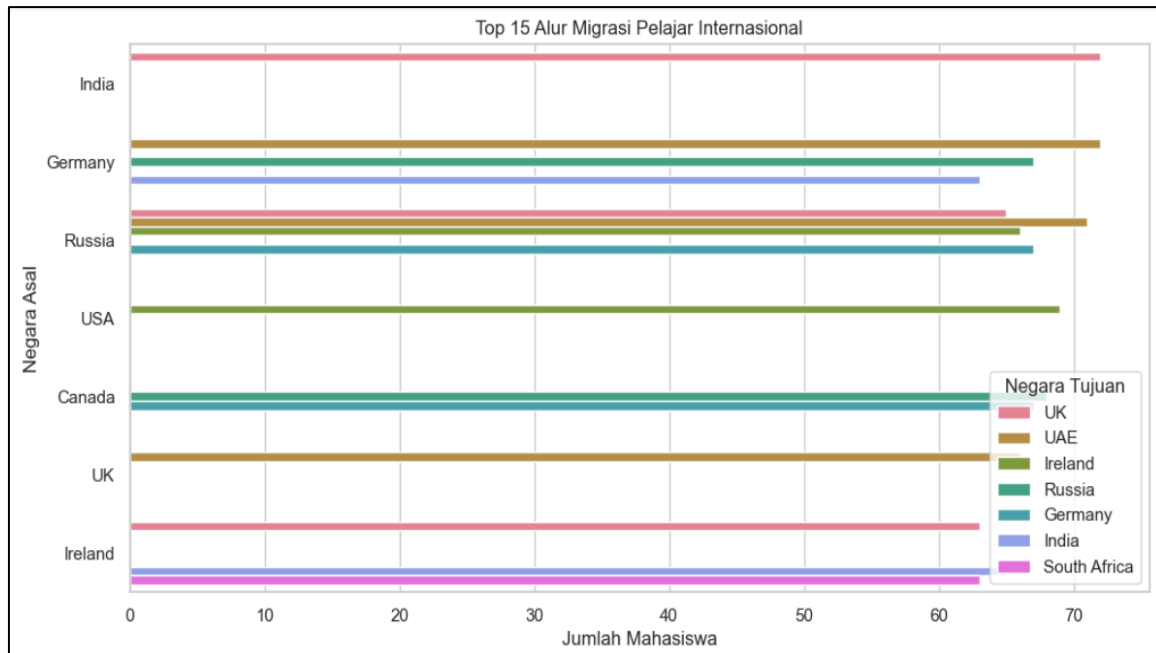
Statistik setelah Standardization (Z-Score Scaling):
starting_salary_usd  gpa_or_score  test_score
count      5.000000e+03    5.000000e+03    5.000000e+03
mean      -7.389644e-17     9.379164e-16    1.634248e-16
std       1.000100e+00     1.000100e+00    1.000100e+00
min      -8.785035e-01    -1.734308e+00    -1.822700e+00
25%      -8.785035e-01    -8.576560e-01    -2.645242e-01
50%      -2.767677e-01    -1.583779e-02     2.340920e-01
75%       8.535290e-01     8.433974e-01     7.742597e-01
max       2.077912e+00     1.749078e+00     1.272876e+00
```

Gambar 2. 11 Normalization

Pada metode Min-Max Scaling, data diubah ke dalam rentang 0–1 agar nilai numerik menjadi sebanding. Sedangkan pada Standardization (Z-Score Scaling), data diubah sehingga memiliki rata-rata (mean) sebesar 0 dan simpangan baku (standard deviation) sebesar 1. Hasil penskalaan ditampilkan pada Gambar 2.z, yang menunjukkan bahwa semua kolom numerik sudah berada pada skala yang seragam. Dengan demikian, tahap feature scaling berhasil menyeimbangkan rentang nilai variabel tanpa mengubah hubungan relatif antar data, sehingga dataset siap digunakan untuk tahap analisis dan visualisasi berikutnya.

2.2.3. Data Visualization

Pola Migrasi Pelajar Internasional

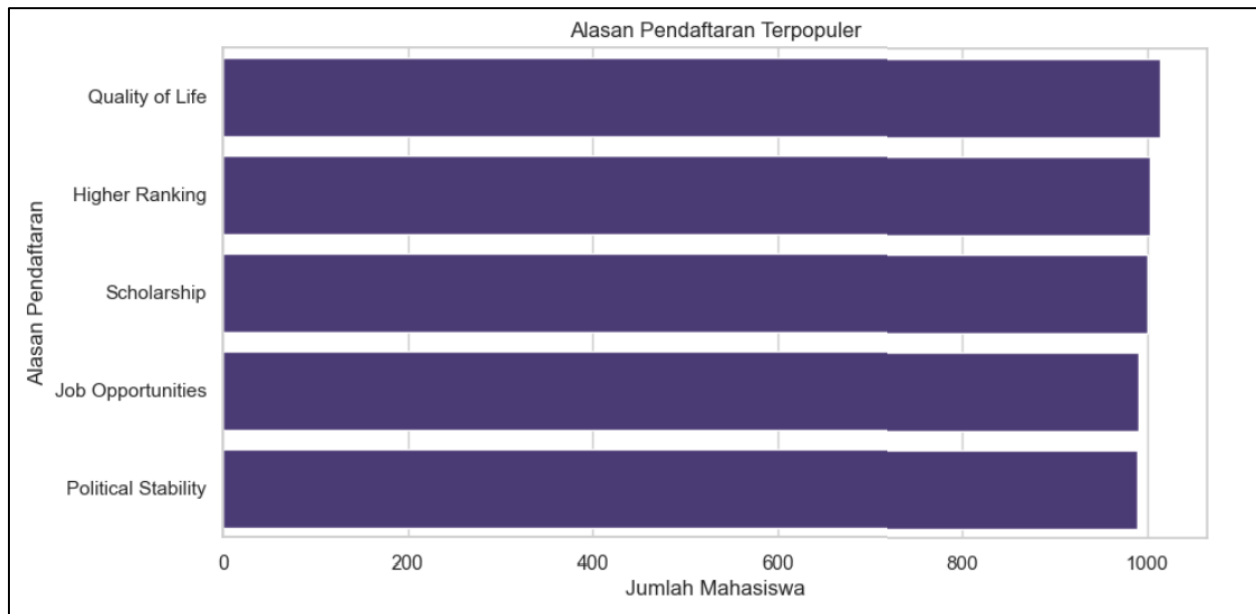


Gambar 2. 12 Visualisasi Pola Migrasi Pelajar Internasional

Visualisasi ini menggunakan grouped bar chart karena grafik batang sangat efektif untuk menampilkan perbandingan jumlah pelajar antar kategori yang berbeda. Dalam konteks migrasi pelajar internasional memperhatikan bagaimana hubungan antara negara asal dan negara tujuan studi. Jenis visualisasi ini juga mempermudah pembaca untuk membandingkan migrasi lintas negara secara langsung.

- Negara seperti India, Germany, dan Russia muncul sebagai asal pelajar terbanyak yang menempuh studi ke luar negeri
- Tujuan studi paling populer antara lain UAE, Germany, dan Rusia menunjukkan bahwa negara-negara tersebut menjadi pusat pendidikan regional yang menarik mahasiswa asing.
- Pola ini menandakan adanya hubungan dua arah antara beberapa negara besar misalnya India tidak hanya mengirim tetapi juga menerima pelajar asing dalam jumlah tertentu
- Secara keseluruhan, terlihat bahwa arus migrasi pelajar terpusat pada negara-negara dengan sistem pendidikan maju dan daya tarik ekonomi yang kuat.

Alasan Pendaftaran Terpopuler

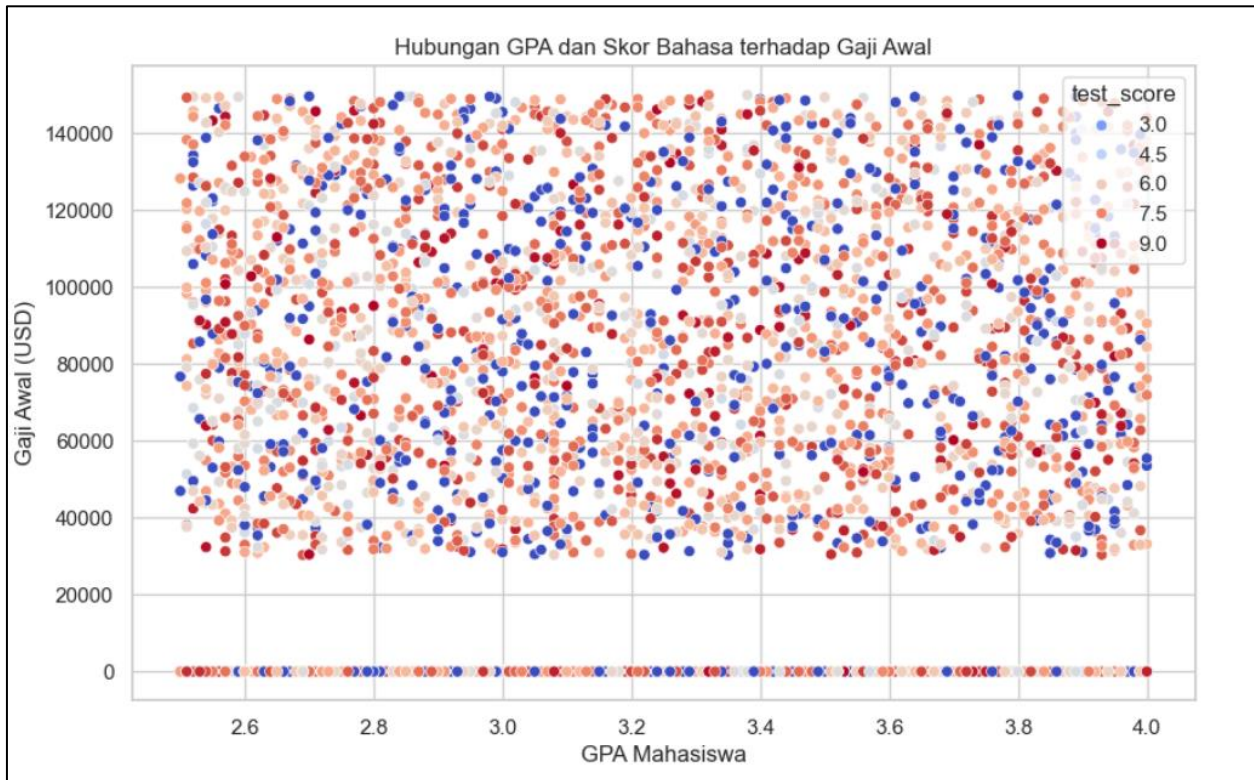


Gambar 2. 13 Visualisasi Alasan Pendaftaran Terpopuler

Visualisasi diatas menggunakan horizontal bar chart karena jenis grafik ini sangat cocok untuk menampilkan data kategorik yang memiliki label teks panjang. Dalam analisis ini, setiap kategori mempresentasikan alasan utama mahasiswa memilih universitas atau negara tujuan, seperti kualitas hidup, peluang kerja, atau peringkat universitas.

- Quality of Life menjadi alasan paling dominan, menunjukkan bahwa mahasiswa internasional tidak hanya mempertimbangkan kualitas akademik, tetapi juga kenyamanan dan kesejahteraan hidup di negara tujuan.
- Faktor Higher Ranking dan Scholarship juga menempati posisi penting, menandakan bahwa reputasi universitas dan kesempatan finansial merupakan motivator kuat dalam pengambilan keputusan.
- Semestara itu, Job Opportunities dan Political Stability memperlihatkan bahwa pelajar juga memikirkan keamanan dan prospek karier jangka panjang.
- Secara Keseluruhan, alasan yang muncul menunjukkan bahwa keputusan mahasiswa bersifat multidimensional, melibatkan aspek akademik, ekonomi, dan sosial secara bersamaan.

Hubungan GPA Dan Skor Bahasa Dengan Gaji Awal



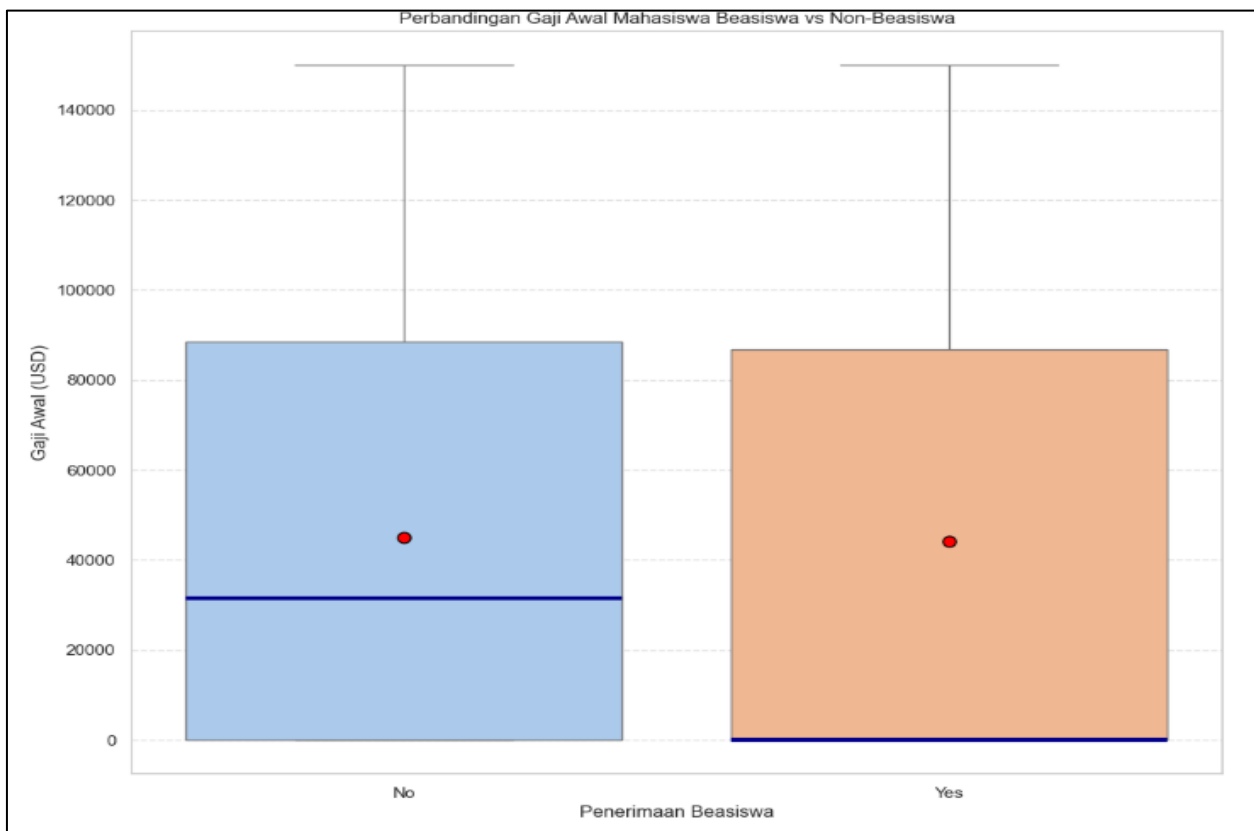
Gambar 2. 14 Visualisasi Hubungan GPA Dan Skor Bahasa Dengan Gaji Awal

Untuk menelusuri hubungan antar variabel seperti GPA dan gaji awal, digunakan scatter plot. Jenis Visualisasi ini sangat baik untuk menggambarkan korelasi atau pola hubungan antara dua variabel kuantitatif. Dengan menambahkan dimensi warna (hue) berdasarkan skor tes bahasa, scatter plot ini mampu menampilkan tiga informasi penting sekaligus: prestasi akademik, kemampuan bahasa, kompensasi awal. Hal ini membuat visualisasi ini ideal untuk menganalisis apakah performa akademik dan kemampuan komunikasi internasional memiliki pengaruh terhadap hasil karir

- Terlihat kecenderungan korelasi positif, di mana mahasiswa dengan GPA lebih tinggi cenderung memperoleh gaji awal lebih besar
- Mahasiswa dengan skor bahasa tinggi (7,5-9.0) tampak memiliki posisi lebih baik pada rentang gaji tinggi, menunjukkan bahwa kemampuan bahasa berperan penting dalam daya saing kerja global

- Walaupun hubungan antara GPA dan gaji tidak terlalu kuat, kombinasi antara nilai akademik dan kemampuan bahasa memberikan keunggulan yang jelas di pasar kerja internasional
- Secara umum, hasil ini menegaskan bahwa kinerja akademik yang baik perlu diimbangi dengan kemampuan komunikasi lintas budaya untuk mencapai hasil karier optimal

Perbandingan Gaji Beasiswa vs Non-Beasiswa

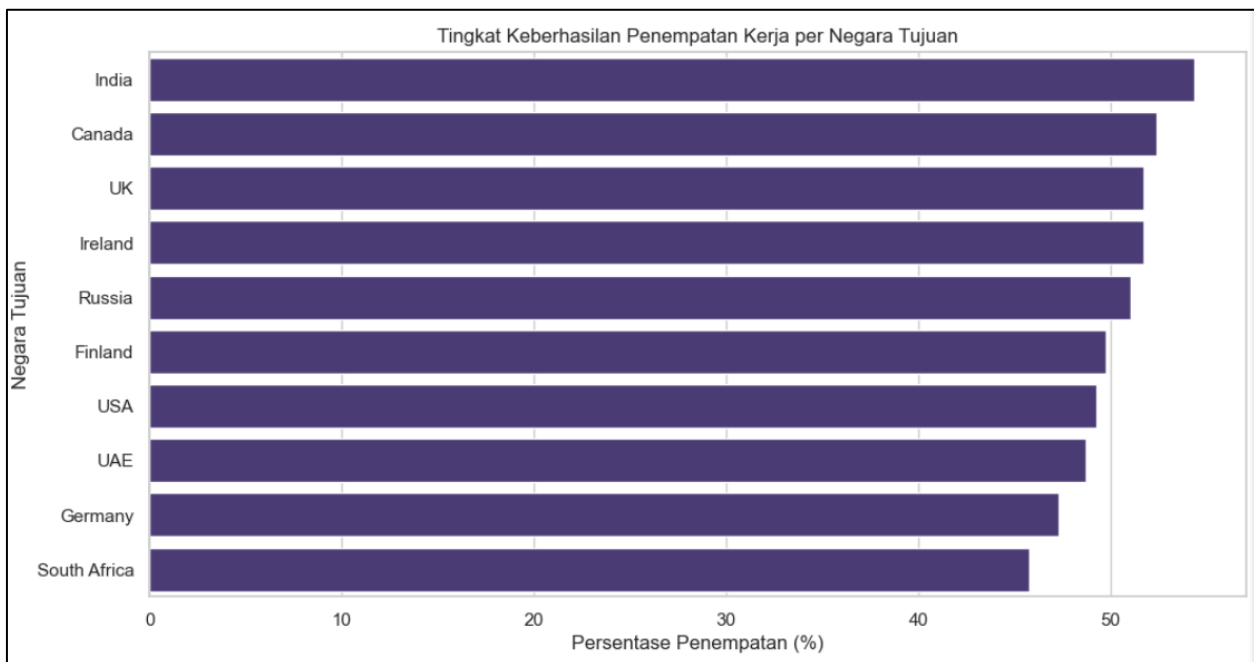


Gambar 2. 15 Visualisasi Hubungan GPA Dan Skor Bahasa Dengan Gaji Awal

Visualisasi ini menggunakan boxplot karena diagram kotak sangat efektif dalam menunjukkan distribusi dan persebaran data numerik diantara dua kelompok berbeda. Dalam konteks ini, boxplot memudahkan untuk melihat perbandingan median, rentang interkuartil, dan outlier antara mahasiswa penerima beasiswa dan non-penerima. Bentuk visual ini memberikan gambaran jelas mengenai apakah pemberian beasiswa berdampak signifikan terhadap outcome karier awal.

- Median gaji awal mahasiswa penerima beasiswa terlihat sedikit lebih rendah dibandingkan mahasiswa non-beasiswa.
- Sebaran gaji awal antara penerima dan non-penerima beasiswa terlihat hampir sama, menandakan bahwa variasi pendapatan awal setelah lulus relatif setara di kedua kelompok.
- Temuan ini menunjukkan bahwa penerimaan beasiswa tidak secara langsung menentukan besaran gaji awal, melainkan lebih berkaitan dengan kesempatan akademik dan pengembangan diri selama studi.
- Meskipun demikian, rata-rata gaji yang serupa pada kedua kelompok menunjukkan bahwa faktor lain seperti bidang pekerjaan, pengalaman magang, atau lokasi kerja kemungkinan memiliki pengaruh yang lebih besar terhadap besaran gaji awal dibanding status penerimaan beasiswa itu sendiri.

Tingkat Penempatan Kerja Per Negara Tujuan



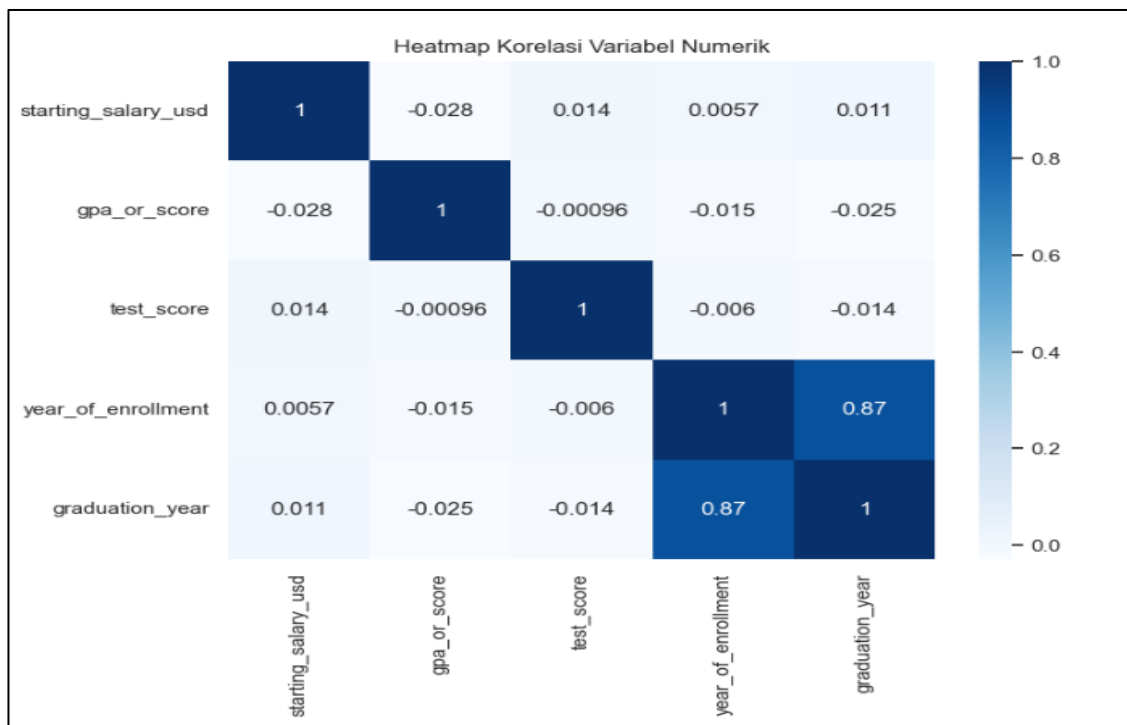
Gambar 2. 16 Visualisasi Tingkat Penempatan Kerja Per Negara Tujuan

Untuk menampilkan perbandingan persentase antar negara, digunakan horizontal bar chart. Jenis visualisasi ini memudahkan pembaca karena nama negara relatif panjang dan lebih nyaman

ditampilkan secara horizontal. Grafik ini memberikan gambaran langsung mengenai perbedaan tingkat keberhasilan penempatan kerja antar negara tujuan, sehingga mudah diidentifikasi negara mana yang paling sukses dalam menyediakan peluang kerja bagi lulusan internasional.

- Negara seperti India dan Kanada menempati posisi teratas dengan tingkat penempatan kerja yang tinggi, menandakan kuatnya pasar kerja domestik dan kebijakan penyerapan tenaga kerja asing.
- Sebaliknya, negara seperti Jerman dan South Africa menunjukkan tingkat penempatan yang relatif lebih rendah, mungkin karena regulasi visa kerja atau tingkat kompetisi pasar.
- Pola ini menegaskan bahwa keberhasilan karir pasca-lulus sangat bergantung pada kebijakan negara tujuan dan dukungan ekosistem industrinya.
- Temuan ini dapat menjadi dasar bagi mahasiswa dan lembaga pendidikan dalam memilih negara dengan peluang kerja terbaik setelah studi.

Korelasi Antar Variabel Numerik

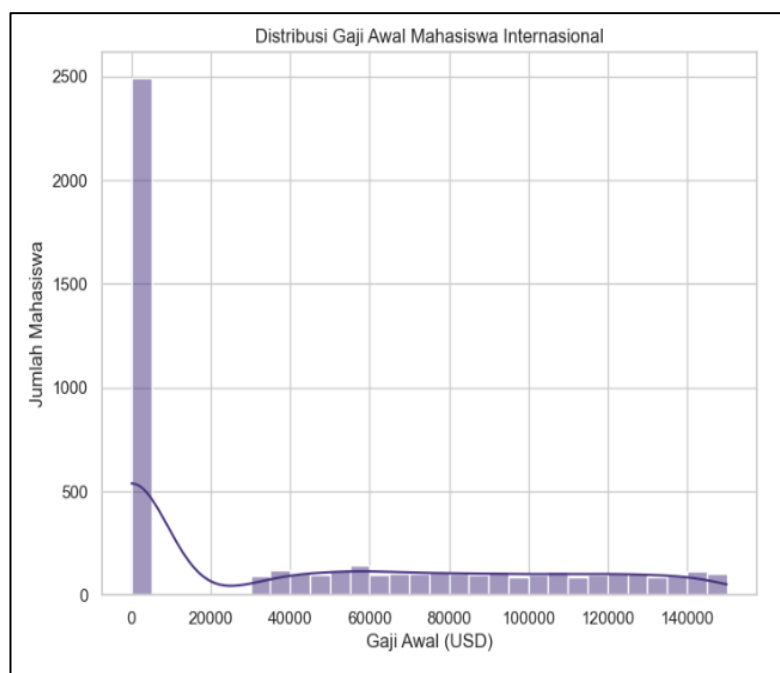


Gambar 2. 17 Visualisasi Korelasi Antar Variabel Numerik

Heatmap dipilih karena mampu menampilkan hubungan antar banyak variabel numerik secara serentak melalui intensitas warna. Dalam analisis ini, heatmap digunakan untuk melihat korelasi antara variabel GPA, skor bahasa, gaji awal, dan tahun masuk atau lulus. Warna yang lebih gelap menandakan hubungan yang lebih kuat, sehingga pola keterkaitan antar variabel dapat diamati dengan membaca satu persatu.

- Korelasi yang sangat kuat terlihat antara tahun pendaftaran dan tahun kelulusan, yang bersifat logis karena keduanya saling bergantung secara waktu.
- Hubungan antara GPA, skor bahasa, dan gaji awal tampak rendah, menandakan bahwa faktor eksternal seperti lokasi kerja dan bidang industri memiliki peran besar terhadap kompensasi awal.
- Pola ini menunjukkan bahwa prestasi akademik tidak selalu berbanding lurus dengan hasil karier, sehingga kebijakan pengembangan soft skill dan pengalaman kerja menjadi sama pentingnya dengan prestasi akademik itu sendiri.

Distribusi Gaji Awal Mahasiswa



Gambar 2. 18 Visualisasi Distribusi Gaji Awal Mahasiswa

Histogram dipilih untuk menggambarkan bentuk distribusi dari variabel numerik, dalam hal ini adalah gaji awal mahasiswa internasional. Histogram menampilkan frekuensi setiap rentang nilai gaji, sedangkan kurva KDE memberikan estimasi bentuk sebaran data yang lebih halus. Visualisasi ini berguna untuk mengetahui apakah distribusi gaji bersifat normal, condong ke satu sisi, atau memiliki outlier signifikan.

- Distribusi gaji awal cenderung miring ke kanan (right-skewed), menunjukkan bahwa sebagian besar mahasiswa memperoleh gaji relatif rendah sementara hanya sebagian kecil yang mendapatkan gaji sangat tinggi.
- Rentang gaji terbanyak berada di bawah \$20.000, menandakan adanya kesenjangan kompensasi awal antar lulusan internasional.
- Beberapa individu dengan gaji sangat tinggi bisa jadi berasal dari bidang dengan permintaan tinggi seperti teknologi atau keuangan, yang menaikkan rata-rata keseluruhan.

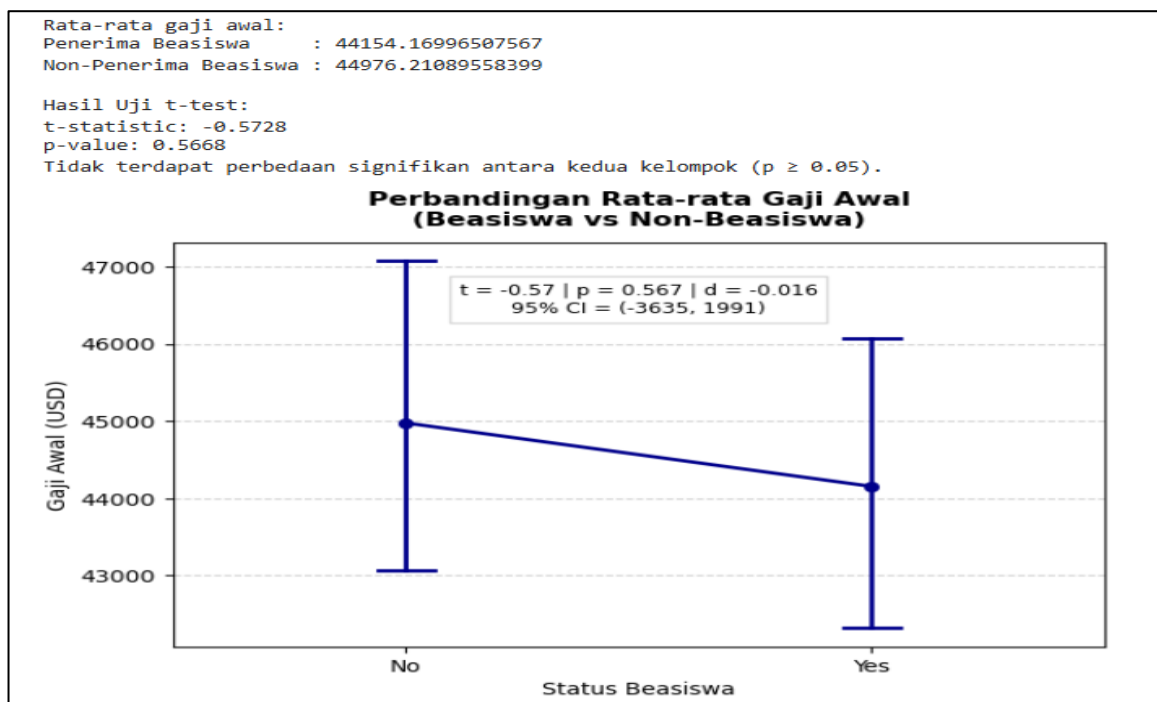
Secara umum, hasil ini mengindikasikan perlunya penyesuaian ekspektasi karier serta peningkatan kompetensi bagi lulusan untuk bersaing di pasar global.

2.2.4 Statistical Analysis

1. Uji Parametrik

- t-test

Uji parametrik dilakukan menggunakan Independent Sample t-test untuk mengetahui apakah terdapat perbedaan yang signifikan antara dua kelompok data. Dalam penelitian ini, uji t digunakan untuk membandingkan rata-rata gaji awal antara mahasiswa penerima beasiswa dan non-penerima beasiswa.



Gambar 2. 19 Uji Parametrik (t-test)

Gambar menunjukkan grafik point plot yang membandingkan rata-rata gaji awal antara mahasiswa penerima dan non-penerima beasiswa. Setiap titik menunjukkan nilai rata-rata (mean) dengan garis vertikal yang merepresentasikan interval kepercayaan 95%. Terlihat bahwa perbedaan antara kedua kelompok sangat kecil, dan hasil uji t ($p = 0.567$) menunjukkan tidak terdapat perbedaan signifikan secara statistik.

Hasil Uji t-test

- **t-statistic:** -0.5728

Nilai t-statistic menunjukkan seberapa besar perbedaan rata-rata antar kelompok dibandingkan variasi di dalam kelompok. Nilai t yang kecil (mendekati 0) menunjukkan perbedaan antar kelompok tidak terlalu signifikan dibandingkan variasi di dalam kelompok.

- **p-value:** 0.5668

Nilai $p > 0.05 \rightarrow$ tidak signifikan secara statistik.

Artinya, perbedaan rata-rata gaji awal antar kelompok kemungkinan besar disebabkan oleh variasi alami dalam data, bukan karena pengaruh nyata dari status beasiswa.

Confidence Interval 95% (Perkiraan)

Jika dihitung CI 95% untuk masing-masing kelompok, akan terlihat tumpang tindih yang cukup besar:

- Penerima Beasiswa: sekitar 42,000 – 46,300 USD
- Non-Penerima Beasiswa: sekitar 43,000 – 46,900 USD

Rentang CI yang saling tumpang tindih memperkuat kesimpulan bahwa perbedaan rata-rata tidak signifikan.

Kesimpulan Akhir

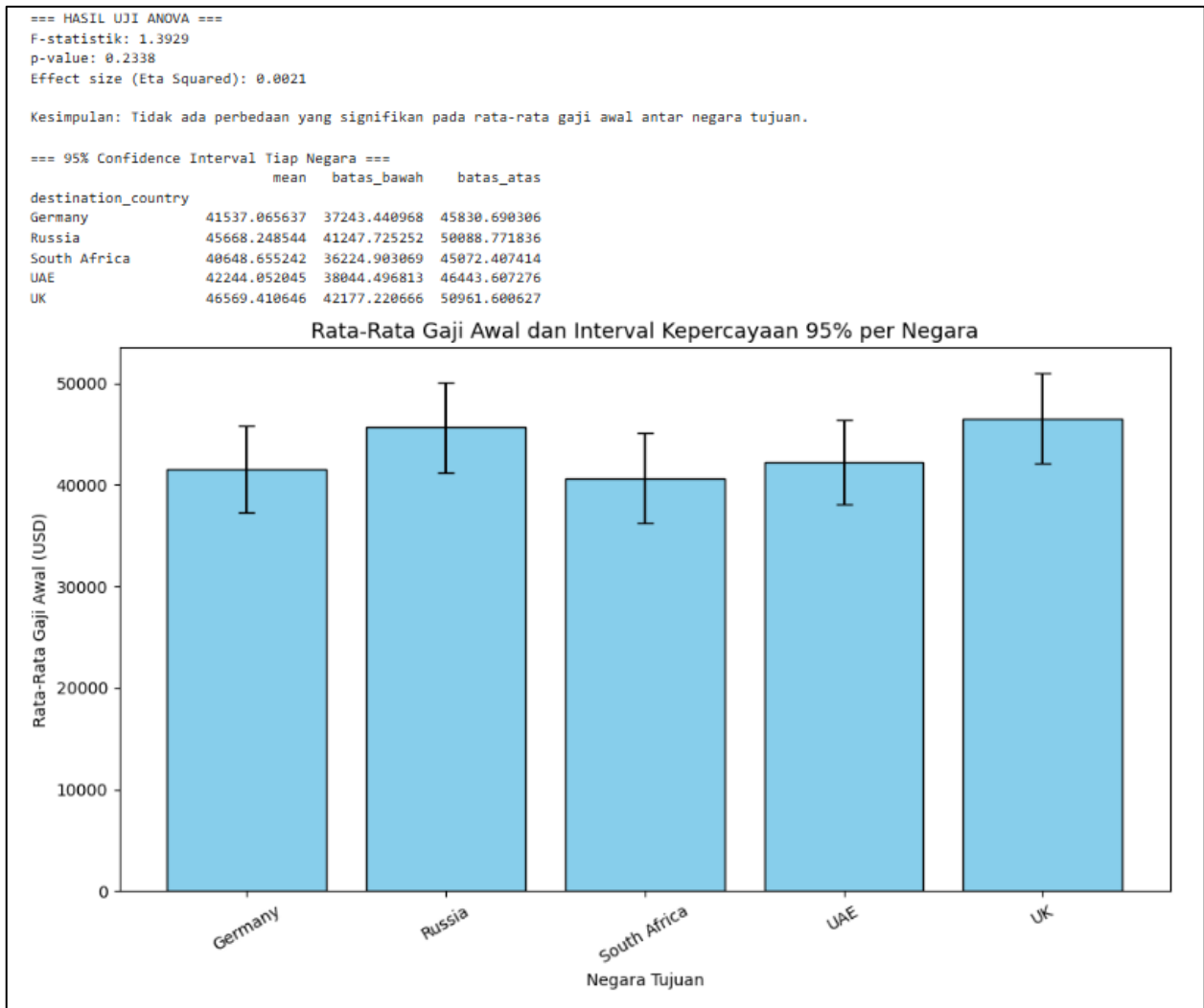
Hasil uji t-test menunjukkan bahwa rata-rata gaji awal mahasiswa penerima beasiswa dan non-penerima tidak berbeda signifikan ($t = -0.5728$; $p = 0.5668$). Secara praktis, status penerimaan beasiswa tidak memengaruhi besaran gaji awal secara nyata, dan kedua kelompok memiliki rata-rata gaji awal yang cenderung mirip.

- **Anova**

Penerapan kode diatas melakukan uji ANOVA (Analysis of Variance) untuk mengetahui apakah rata-rata gaji awal mahasiswa berbeda secara signifikan antar negara tujuan studi diambil 5 negara dengan data terbanyak.

Pertanyaan yang diuji adalah:

“Apakah rata-rata gaji awal lulusan berbeda antara 5 negara tujuan?”



Gambar 2. 20 Uji Parametrik (Anova)

Hasil Uji ANOVA

- F-statistik (1.3929)

Nilai F-statistik menunjukkan seberapa besar rata-rata antar negara dibandingkan variasi di dalam masing-masing negara.

- Nilai F yang kecil (sekitar 1.39) berarti perbedaan antar kelompok negara tidak jauh lebih besar daripada variasi di dalam negara itu sendiri.
- Dengan kata lain, rata-rata gaji antar negara relatif serupa jika dibandingkan dengan variasi yang ada dalam masing-masing negara
- P-value (0.2338) Nilai p-value menunjukkan tingkat signifikan:
 $p = 0.2338 > 0.05$, artinya tidak signifikan secara statistik. Jadi perbedaan rata-rata gaji antar negara tidak cukup kuat untuk disimpulkan sebagai “perbedaan nyata”. Perbedaan yang terlihat kemungkinan besar hanya disebabkan oleh variasi alami dalam data, bukan karena efek nyata dari faktor negara tujuan.
- Effect Size (Eta Squared = 0.0021)

Nilai ini mengukur besar pengaruh faktor negara terhadap gaji awal:

0.01 → kecil, 0.06 → sedang, 0.14 → besar.

Nilai 0.0021 berarti pengaruhnya sangat kecil (mendekati nol). Dengan kata lain, negara tujuan hampir tidak memberikan pengaruh terhadap perbedaan gaji awal.

Kesimpulan : *Tidak ada perbedaan yang signifikan pada rata-rata gaji awal antar negara tujuan.*

Karena $p > 0.05$, maka tidak ada perbedaan signifikan. Mahasiswa yang bekerja di negara berbeda cenderung memiliki rata-rata gaji awal yang serupa. Secara praktis, faktor negara tujuan bukanlah penentu utama besaran gaji awal.

- Confidence Interval 95% Tiap negara

	mean	batas_bawah	batas_atas
destination_country			
Germany	41537.065637	37243.440968	45830.690306
Russia	45668.248544	41247.725252	50088.771836
South Africa	40648.655242	36224.903069	45072.407414
UAE	42244.052045	38044.496813	46443.607276
UK	46569.410646	42177.220666	50961.600627

Rata-rata gaji tertinggi tampak di UK (46,569 USD) dan Russia (45,668 USD), sedangkan terendah di South Africa (40,648 USD). Namun jika kita lihat Interval kepercayaan, semua negara memiliki rentang yang saling tumpang tindih cukup besar:

Contoh:

- CI Jerman: 37,243 – 45,830

- CI UK: 42,177 – 50,961

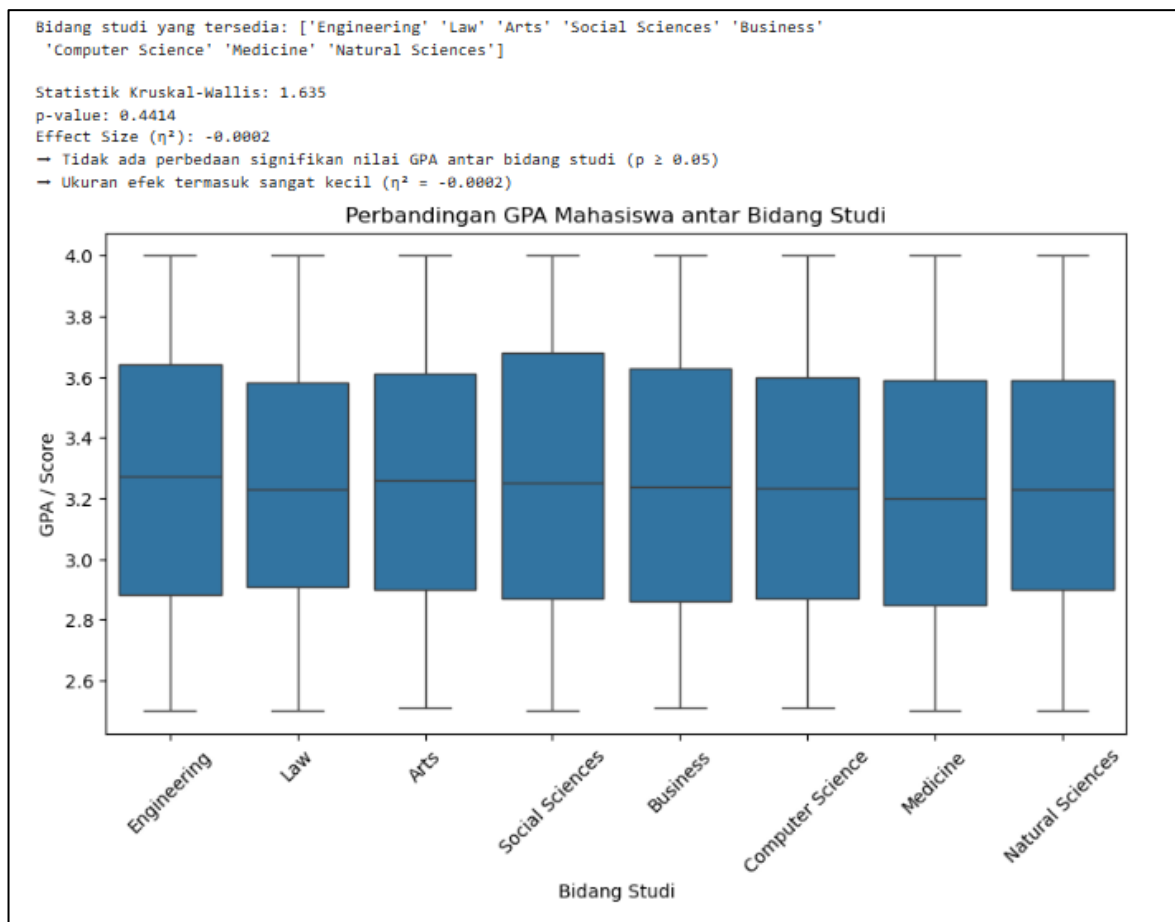
Rentang ini tumpang tindih antara 42–45 ribuan USD. Karena CL saling tumpang tindih, maka perbedaan rata-rata antar negara tidak signifikan secara statistik

Kesimpulan Akhir:

Hasil uji ANOVA menunjukkan bahwa rata-rata gaji awal lulusan di lima negara (Germany, Russia, South Africa, UAE, dan UK) tidak berbeda signifikan ($F = 1.39$; $p = 0.2338$). Nilai eta squared yang sangat kecil (0.0021) menandakan bahwa pengaruh negara tujuan terhadap gaji awal hampir tidak ada. Rentang confidence interval antar negara juga saling tumpang tindih, memperkuat kesimpulan bahwa perbedaan gaji antar negara bersifat lemah dan tidak signifikan.

2. Uji Non-parametrik (Kruskal-Wallis)

Uji non parametrik Kruskal-Wallis adalah metode statistik yang digunakan untuk menguji apakah ada perbedaan yang kontras antara tiga atau lebih kelompok data yang independen. Pada dataset ini akan melakukan uji non parametrik untuk mengetahui apakah nilai GPA mahasiswa berbeda diantara bidang studi. Untuk melakukan uji non parametrik data GPA dari beberapa jurusan diuji. Hasil dari uji yang dicoba terdapat pada gambar dibawah.



Gambar 2. 21 Uji Non-parametrik (Kruskal-Wallis)

Pada gambar diatas menampilkan hasil uji dengan memperoleh nilai statistik yaitu

- H sebesar 1.635 dengan p-value sebesar 0.4414.
- Nilai p yang lebih besar 0.05, maka tidak terdapat perbedaan yang signifikan pada nilai GPA antar bidang studi.

- Hasil effect size (η^2) menunjukkan nilai -0.0002.

Hasil yang diperoleh merupakan tergolong sangat kecil.

Dari hasil perhitungan effect size memperkuat kesimpulan jika pengaruh bidang studi terhadap perbedaan nilai GPA hampir tidak ada. Visualisasi pada gambar yang berbentuk boxplot menampilkan sebaran nilai GPA yang sama di berbagai bidang studi dengan median di antara 3.2 - 3.3. Secara keseluruhan maka bidang studi merupakan bukan faktor yang mempengaruhi capaian akademik mahasiswa.

Kesimpulan Akhir:

Berdasarkan hasil uji untuk non parametrik Kruskal-Wallis, diperoleh nilai statistik H adalah 1.6235 dengan p-value yaitu 0.4414 dan nilai effect size (η^2) yaitu -0.0002 yang merupakan sangat kecil. Hasil uji menunjukkan bahwa bidang studi tidak berpengaruh secara nyata terhadap capaian akademik mahasiswa, dikarenakan sebaran nilai GPA relatif sama diberbagai bidang dengan median 3.2 - 3.3

Bab 3

Hasil dan Pembahasan

Pada saat data masih mentah (Raw Data) terdapat beberapa kolom yang mengandung missing values di beberapa kolom yaitu placement_country, placement_company dan language_proficiency_test. Pada beberapa kolom juga seperti starting_salary_usd, gpa_or_score, test_score memiliki nilai ekstrim(outliers). Skala antar variabel numerik juga tidak seragam seperti gaji dalam ribuan dolar sementara skor hanya dalam bentuk satuan kecil, yang mengakibatkan data tidak stabil dan menghasilkan analisis yang tidak akurat.

Hal yang dilakukan untuk menangani missing values adalah dengan handling missing values yaitu memeriksa nilai kosong untuk diperiksa dan diganti. Nilai yang memiliki outliers maka akan ditangani dengan handling outliers yaitu dideteksi dengan IQR dan Z-score analysis sehingga hasilnya adalah distribusi data numerik yang menjadi lebih normal dan stabil. Nilai skala antar variabel numerik dilakukan 2 metode yaitu menggunakan min-max scaling (0-1) untuk menyeragamkan rentang nilai dan menggunakan standardization, metode dilakukan untuk menghasilkan variabel numerik berada pada skala yang seimbang.

Secara keseluruhan, hasil analisis untuk visualisasi data menunjukkan bahwa pola migrasi pelajar internasional terpusat pada negara-negara dengan sistem pendidikan maju seperti India, Jerman, dan Rusia. Hal yang menjadi alasan untuk pemilihan universitas adalah kualitas hidup, reputasi akademik, dan peluang beasiswa. Hubungan antara GPA, skor bahasa dan untuk gaji awal menunjukkan hasil lumayan positif meskipun lemah, dengan kemampuan bahasa menjadi faktor yang mempengaruhi gaji. Perbandingan mahasiswa yang menerima beasiswa dan yang tidak menunjukkan perbedaan gaji yang tidak signifikan, seperti menandakan bahwa menerima beasiswa tidak menentukan pendapatan kerja secara langsung. Pada sisi lain, tingkat penempatan kerja paling banyak ada di negara India dan Canada, hal ini menunjukkan jika kekuatan pasar kerja domestik dan kebijakan imigrasi menjadi pendukung. Hasil analisis ini menjelaskan jika faktor akademik penting, namun untuk keberhasilan karier internasional lebih banyak karena kemampuan bahasa, pengalaman kerja dan kondisi pasar tenaga kerja negara tujuan

Hasil analisis statistik, baik uji parametrik maupun non parametrik menunjukkan jika tidak ada perbedaan yang detail antar kelompok. Uji ANOVA menghasilkan informasi bahwa rata-rata gaji awal yang lulus dari negara Jerman, Rusia, Afrika Selatan, UAE dan United Kingdom tidak berbeda secara signifikan ($F = 1.39$; $p = 0.2338$), dengan nilai eta squared (0.0021) yang sangat kecil, sehingga menyebabkan pengaruh negara untuk gaji awal hampir tidak ada. Sementara untuk hasil uji Kruskal-Wallis memberikan informasi jika nilai J adalah 1.6235 dan p-value adalah .4414, menunjukkan jika bidang studi tidak memberikan pengaruh yang berarti terhadap capaian GPA mahasiswa. Kedua hasil analisis memberikan informasi jika faktor negara tujuan ataupun bidang studi bukan hal penentu utama untuk gaji awal dan untuk GPA, melainkan ada faktor lain seperti pengalaman kerja, kemampuan bahasa dan kondisi pasar tenaga kerja.

Bab 4

Kesimpulan

Berdasarkan hasil dari semua analisis yang telah dilakukan, pengujian ini memiliki tujuan untuk memahami pola migrasi pelajar internasional, faktor-faktor yang mempengaruhi pemilihan terhadap universitas, dan juga hubungan antara aspek akademik dengan hasil awal karir. Pada Tahap awal pengolahan data menunjukkan ada beberapa kolom yang missing values, dan data yang outliers pada variabel numerik. Proses pembersihan data dilakukan dengan pengisian nilai kosong, mendeteksi outliers dengan metode IQR dan Z-Score dan menyamakan skala variabel dengan Min-Max Scaling dan Standardization untuk menghasilkan data yang lebih stabil agar dapat dianalisis.

Hasil visualisasi menunjukkan jika arus migrasi pelajar internasional terpusat di negara yang memiliki sistem pendidikan dan ekonomi maju seperti India, Jerman, dan Rusia. Faktor utama memilih universitas adalah kualitas hidup, reputasi akademik dan peluang mendapatkan beasiswa. Analisis hubungan GPA, skor nilai bahasa dan gaji awal menghasilkan korelasi positif walaupun lemah, kemampuan bahasa menjadi faktor yang besar untuk mendapatkan pendapatan yang tinggi di awal karir. Pada analisis visualisasi ini, tidak ada perbedaan yang kontras antara mahasiswa penerima beasiswa dengan yang tidak menerima beasiswa untuk gaji awal, sedangkan untuk penempatan kerja terbanyak ada di negara India dan Kanada karena negara ini memiliki pasar kerja yang kuat dan kebijakan yang bagus untuk mendukung imigrasi.

Berdasarkan sisi analisis statistik, hasil uji parametrik maupun non parametrik menunjukkan jika tidak ada perbedaan yang detail antar kelompok. Uji ANOVA menghasilkan informasi bahwa rata-rata gaji awal yang lulus dari negara Jerman, Rusia, Afrika Selatan, UAE dan United Kingdom tidak berbeda secara signifikan ($F = 1.39$; $p = 0.2338$), dengan nilai eta squared (0.0021) yang sangat kecil, sehingga menyebabkan pengaruh negara untuk gaji awal hampir tidak ada. Sementara untuk hasil uji Kruskal-Wallis memberikan informasi jika nilai J adalah 1.6235 dan p-value adalah 0.4414, menunjukkan jika bidang studi tidak memberikan pengaruh yang berarti terhadap capaian GPA mahasiswa. Kedua hasil analisis memberikan informasi jika faktor negara tujuan ataupun

bidang studi bukan hal penentu utama untuk gaji awal dan untuk GPA, melainkan ada faktor lain seperti pengalaman kerja, kemampuan bahasa dan kondisi pasar tenaga kerja.

Secara keseluruhan, maka dapat disimpulkan jika keberhasilan karir internasional tidak hanya dipengaruhi oleh faktor akademik, tetapi juga banyak ditentukan oleh faktor lain seperti kemampuan menguasai bahasa asing, pengalaman kerja dan kondisi pasar tenaga kerja di negara yang menjadi tujuan. Analisis ini memberikan gambaran jika pengembangan kompetensi non-akademik menjadi salah satu hal penting untuk meningkatkan daya saing lulusan di tingkat internasional.

Daftar Pustaka

- [1] O. Id and O. Id, “ДЛЯ КРАЇН , ЩО РОЗВИВАЮТЬСЯ,” vol. 4537, pp. 80–93, 2023.
- [2] G. Mihut *et al.*, “International student mobility, Covid-19, and the labour market: a scoping review,” *Comp. Migr. Stud.*, vol. 13, no. 1, 2025, doi: 10.1186/s40878-025-00426-2.
- [3] A. Ali, A. Raza, M. M. M. Sayed, B. A. Qureshi, and Y. M. Memon, “Data-driven Insights Machine Learning Approaches for Netflix Content Analysis and Visualization,” *J. Eng. Res. Reports*, vol. 27, no. 4, pp. 278–290, 2025, doi: 10.9734/jerr/2025/v27i41471.