

Data Engineering Pipeline - University Rank

Kelompok 4 :

- Rizky Alif Ramadhan 19/446785/TK/49890
- Roby Attoillah 19/444068/TK/49264
- Muhammad Farrel R 19/444062/TK/49258
- Harry Krisna D 19/446781/TK/49886



Pendahuluan



Latar Belakang

- Ranking Universitas dijadikan alasan dalam pengambilan keputusan bagi seorang Mahasiswa/Siswa untuk melanjutkan studi ke jenjang berikutnya
- Banyak versi perankingan yang ada di Internet, 2 diantaranya adalah dari Times Higher Education (THE) dan QS World Ranking
- Keinginan untuk menjadikan UGM sebagai universitas terbaik di dunia





Permasalahan

- Mana saja universitas yang memiliki skor tertinggi di dunia?
- Berapakah skor rata-rata universitas di dunia?
- Mana saja negara dengan skor rata-rata universitas tertinggi di dunia?
- Apa saja komponen penilaian yang penting untuk menjadi universitas terbaik di dunia?



Tujuan

- Membuat skor rata-rata dari 2 penyedia ranking universitas dan membuat ranking baru dari hasil skor tersebut
- Mengetahui *Top Universities* di dunia, Asia, dan Indonesia
- Mengetahui skor rata-rata setiap negara atau benua
- Mengetahui komponen penilaian terpenting yang menentukan ranking dari universitas tersebut

Proses ETL



Proses Extract



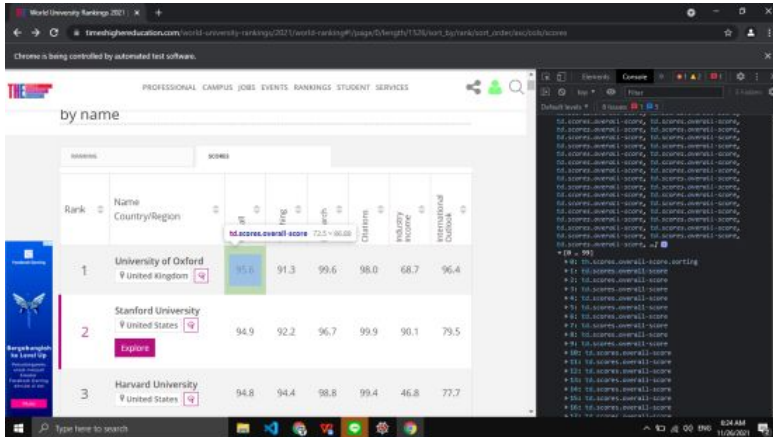


Metode Ekstraksi Data

- Menggunakan metode *web scraping*
- Website sumber data
 - <https://www.topuniversities.com/university-rankings/world-university-rankings/2021> (QS World University Rankings 2021)
 - <https://www.timeshighereducation.com/world-university-rankings/2021/world-ranking> (THE World University Rankings 2021)
- Alat yang digunakan :
 - Python
 - Selenium
 - Spyder (Anaconda3)

Proses Web Scrapping

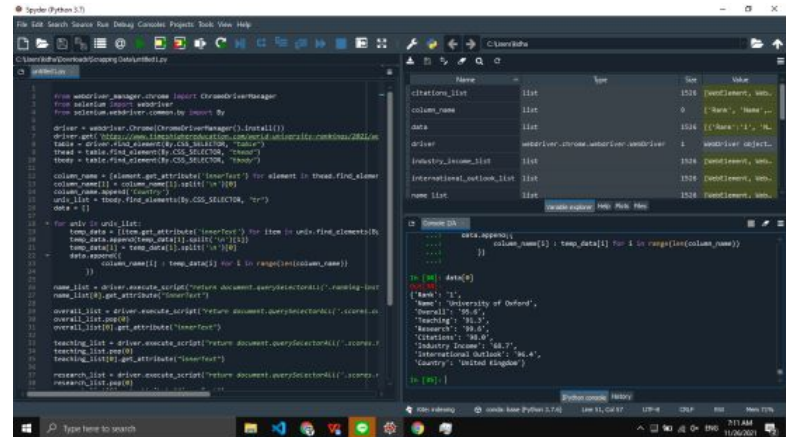
- Dokumentasi gambar



World University Rankings 2021

by name

Rank	Name	Country/Region	Overall Score	Academic Score	Teaching Score	Research Score	International Outlook Score
1	University of Oxford	United Kingdom	95.6	91.3	99.6	98.0	96.4
2	Stanford University	United States	94.9	92.2	96.7	99.9	79.5
3	Harvard University	United States	94.8	94.4	98.8	99.4	77.7



```
from selenium.webdriver.chrome import ChromeDriverManager
from selenium.webdriver.common.by import By
from selenium.webdriver.chrome.service import Service
from selenium.webdriver.support.ui import WebDriverWait
from selenium.webdriver.support import expected_conditions as EC
from bs4 import BeautifulSoup

driver = webdriver.Chrome(ChromeDriverManager().install())
driver.get('https://www.timeshighereducation.com/world-university-rankings/2021')
table = driver.find_element_by_css_selector('table')
tbody = table.find_element_by_css_selector('tbody')

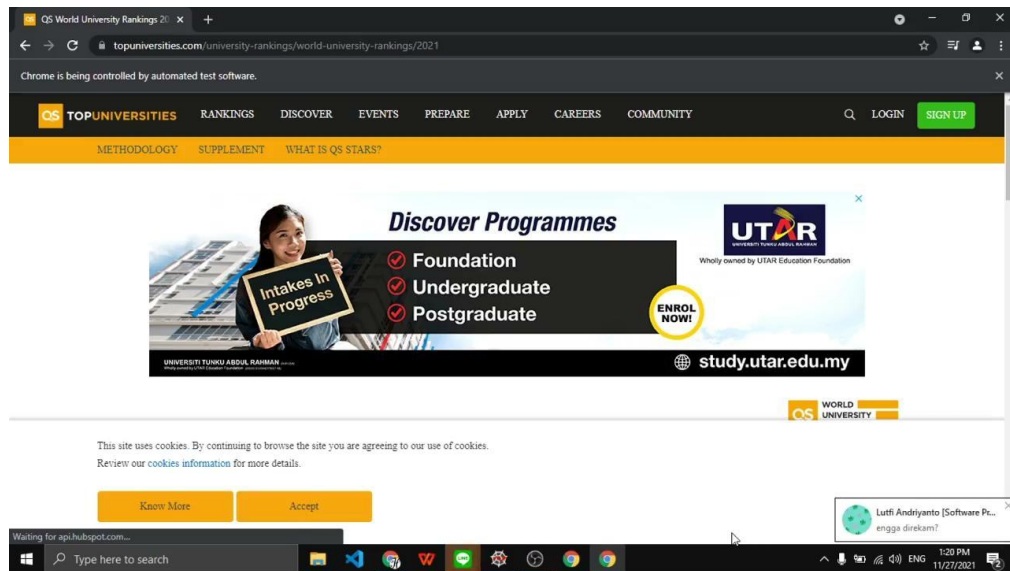
column_name = element.get_attribute('innerText') for element in tbody.find_element_by_css_selector('tr').find_elements_by_css_selector('th')
column_name_append('country')
tbody_list = tbody.find_elements_by_css_selector('tr')
data = []

for entry in tbody_list:
    temp_data = [item.get_attribute('innerText') for item in entry.find_elements_by_css_selector('td')]
    temp_data.append('country')
    temp_data[0] = temp_data[0].replace(' ', '')
    column_name[1] = temp_data[0] for i in range(len(column_name))
    data.append(temp_data)

name_list = driver.execute_script('return document.querySelector(\".ranking-list\").querySelectorAll(\".name\");')
name_list[0].get_attribute('innerText')
overall_list = driver.execute_script('return document.querySelector(\".score-overall\").querySelectorAll(\".score-overall\");')
overall_list[0].get_attribute('innerText')
teaching_list = driver.execute_script('return document.querySelector(\".score-teaching\").querySelectorAll(\".score-teaching\");')
teaching_list[0].get_attribute('innerText')
research_list = driver.execute_script('return document.querySelector(\".score-research\").querySelectorAll(\".score-research\");')
research_list[0].get_attribute('innerText')
```

Proses *Web Scrapping*

- Dokumentasi video



Hasil Ekstraksi Data

- Dataset hasil, source code, dan dokumentasi tambahan web scrapping dapat diakses di

https://drive.google.com/file/d/1N9cYLxEzL3luW4pDf8ChOJAK1BMd_xnB/view

Rank	University	Overall	Sci	Internation	Internation	Faculty	Stu	Citations	Academic	Employer	Location
0	1 Massachusetts	100	91.9	100	100	99.1	100	100	100	100	Cambridge,United States
1	2 Stanford Un	98.4	63.6	99.7	100	98.1	100	100	100	100	Stanford,United States
2	3 Harvard Un	97.9	69.9	85.2	98.6	99.1	100	100	100	100	Cambridge,United States
3	4 California I	97	88.2	100	100	99.9	97	82.8	100	100	Pasadena,United States
4	5 University o	96.7	98.3	99.4	100	81.3	100	100	100	100	Oxford,United Kingdom
5	6 ETH Zurich	95	97.9	100	80.8	96.4	98.7	96.6	100	100	Zürich,Switzerland
6	7 University o	94.3	97.4	100	100	69.2	100	100	100	100	Cambridge,United Kingdom
7	8 Imperial Co	93.6	100	100	99.9	68.6	98.5	99.8	100	100	London,United Kingdom
8	9 University o	93.1	82.6	67.1	94.4	86.3	99.4	91.3	100	100	Chicago,United States
9	10 UCL	92.9	100	99.3	98.4	65.4	99.4	98.3	100	100	London,United Kingdom
10	11 National U	91.5	71.4	100	90.7	72.9	99.7	98.4	100	100	Singapore,Singapore
11	12 Princeton U	91	65.6	71.6	68.6	100	99.9	99	100	100	Princeton,United States
12	13 Nanyang Te	89.9	67.6	100	91.5	89	89.8	89.8	100	100	Singapore,Singapore
13	14 EPFL	89.6	100	100	96.3	98.5	80.4	80	100	100	Lausanne,Switzerland
14	15 Tsinghua U	89.2	29.7	55.3	93.3	83.2	98.2	98.6	100	100	Beijing,China (Mainland)
15	16 University o	88.6	65.3	88.7	100	63.8	96.1	91.5	100	100	Philadelphia,United States
16	17 Yale Univer	88	54.5	85.3	100	52.8	99.9	100	100	100	New Haven,United States
17	18 Cornell Uni	87.6	70	93.7	63.7	88.6	98.5	90.9	100	100	Ithaca,United States
18	19 Columbia U	86.5	96.3	37.2	100	48.5	99.7	97.3	100	100	New York City,United States
19	20 The Univer	85.8	99	98.2	83.1	50.2	98	95.4	100	100	Edinburgh,United Kingdom
20	21 University o	84.6	41.3	74.8	89.4	58	98.9	92.7	100	100	Ann Arbor,United States
21	22 The Univer	83.7	98.9	100	83.8	48	98.1	76.3	100	100	Hong Kong,Hong Kong SAR

Rank	Name	Overall	Teaching	Research	Citations	Industry In	Internation	Country
0	1 University o	95.6	91.3	99.6	98	68.7	96.4	United Kingdom
1	2 Stanford Un	94.9	92.2	96.7	99.9	90.1	79.5	United States
2	3 Harvard Un	94.8	94.4	98.8	99.4	46.8	77.7	United States
3	4 California I	94.5	92.5	96.9	97	92.7	83.6	United States
4	5 Massachus	94.4	90.7	94.4	99.7	90.4	90	United States
5	6 University o	94	90.3	99.2	95.6	52.1	95.7	United Kingdom
6	7 University o	92.2	85.8	97.2	99.1	84.3	72.3	United States
7	8 Yale Univer	91.6	91.9	93.8	97.9	56.1	68.4	United States
8	9 Princeton U	91.5	88.8	92.5	98.9	58	80.2	United States
9	10 The Univer	90.3	88.9	90.5	98.6	54.9	74	United States
10	11 Imperial Co	89.4	82.3	88.2	97.2	69.6	97.4	United Kingdom
11	12 Johns Hopk	89.2	81.6	91.8	97.7	93.4	73.9	United States
12	13 University o	88.9	85.4	89.9	98.1	77.9	66.3	United States
13	14 ETH Zurich	87.9	80.4	92.3	90.5	62.8	98	Switzerland
14	15 University o	87.1	82.5	90.2	96.5	57.6	65.3	United States
15	16 UCL	86.9	76.6	89.4	96.2	42.1	96.5	United Kingdom
16	17 Columbia U	86.8	85.1	82.9	97.7	45	79.8	United States
17	18 University o	86	75.4	90.9	94.5	50	87.2	Canada
18	19 Cornell Uni	85.3	78.8	86.7	97.2	36.3	73.7	United States
19	20 Duke Unive	84.8	80.7	80.4	96.9	99.7	65.5	United States
20	20 Tsinghua U	84.8	87.7	94.9	78.8	100	51.1	China
21	22 University o	84	79	86.9	95.4	47.2	59.6	United States



Proses Transformasi





Tahapan Transformasi

1. Transformasi dataset Times Higher Education
2. Transformasi dataset QS World Ranking
3. Integrasi dataset (outer join)
4. Transformasi dataset hasil integrasi



Transformasi Dataset Times Higher Education

Pada dataset ini dilakukan beberapa tahapan transformasi:

1. Mengubah nama kolom untuk memudahkan proses *join*
 - a. Kolom **Name** menjadi **University**
 - b. Kolom **Country** menjadi **Country_THE**
2. Memfilter kolom yang tidak diperlukan
 - a. Kolom **Unnamed: 0**, **Overall** dan **Rank** akan dihapus
 - b. Kolom **Overall** akan diganti dengan **Average_score** di akhir proses
 - c. Rank universitas akan dibuat baru setelah dataset digabungkan
3. Mengubah beberapa *value* agar lebih konsisten (sinkronisasi *value*)
 - a. Contoh: Universitas **École Polytechnique Fédérale de Lausanne** pada dataset lainnya dinamakan **EPFL**.
 - b. Terdapat universitas yang pada kata akhirnya singkatan, misalnya **Bandung Institute of Technology (ITB)** kita akan menghapus kata **(ITB)**
4. *Handling missing value*
 - a. Mengecek apakah ada *missing value* pada dataset. Baris yang memiliki *missing value* akan dihapus jika memungkinkan



Transformasi Dataset QS World Ranking

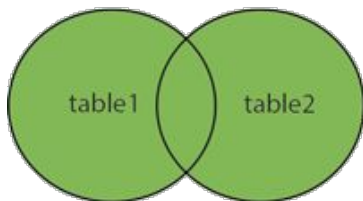
Pada dataset ini dilakukan beberapa tahapan transformasi:

1. Menghapus nama kota pada kolom Location
 - a. Pada kolom **Location** yang berisi [namakota,negara]. Yang diperlukan hanya nama negara saja
 - b. Dibuat kolom baru **Country_qsworld** untuk menampung nama negara
2. Memfilter kolom yang tidak diperlukan
 - a. Kolom **Unnamed: 0**, **Overall Score**, **Rank**, dan **Location** akan dihapus
 - b. Kolom **Location** dihapus karena sudah diwakili kolom **Country_qsworld**
3. Mengubah beberapa *value* dataset agar lebih konsisten (sinkronisasi *value*)
 - a. Terdapat universitas yang pada kata akhirnya singkatan,
 - b. Contoh: "**Bandung Institute of Technology (ITB)**" kita akan menghapus kata "**(ITB)**"
 - c. Pada "**KAIST - Korea Advanced Institute of Science and Technology**" kita hanya mengambil **KAIST** nya saja.
4. *Handling missing value*
 - a. Mengecek apakah ada *missing value*. Jika pada kolom yang bersifat numerik, kita akan menggantinya dengan 0. Jika terdapat pada kolom yang bersifat kategorik maka akan dihapus.



Integrasi Dataset

FULL OUTER JOIN



- Pada tahapan integrasi, dilakukan Outer Join dari dua dataset yang telah ditransformasi menggunakan kolom **University** sebagai primary key
- Digunakan outer join karena tidak semua nama universitas pada tabel 1 ada pada tabel 2.
- Konsekuensinya adalah munculnya nilai NaN
- Untuk mencari Average_Score nilai NaN tidak dihandle karena tidak akan dimasukan dalam perhitungan rata-rata.
- Nilai NaN akan dihandle setelah mendapat Average_Score

	University	Teaching	Research	Citations	Industry Income	International Outlook	Country_THE	International Students Ratio	International Faculty Ratio	Faculty Student Ratio	Citations per Faculty	Academic Reputation	Employer Reputation	Country_qsworld
0	University of Oxford	91.3	99.6	98.0	68.7	96.4	United Kingdom	98.3	99.4	100.0	81.3	100.0	100.0	United Kingdom
1	Stanford University	92.2	96.7	99.9	90.1	79.5	United States	63.6	99.7	100.0	98.1	100.0	100.0	United States
2	Harvard University	94.4	98.8	99.4	46.8	77.7	United States	69.9	85.2	98.6	99.1	100.0	100.0	United States
3	California Institute of Technology	92.5	96.9	97.0	92.7	83.6	United States	88.2	100.0	100.0	99.9	97.0	82.8	United States
4	Massachusetts Institute of Technology	90.7	94.4	99.7	90.4	90.0	United States	91.9	100.0	100.0	99.1	100.0	100.0	United States
...
2010	Université de Technologie de Compiègne	NaN	NaN	NaN	NaN	NaN	NaN	32.1	0.0	38.3	22.5	0.0	18.0	France
2011	University of California, San Francisco	NaN	NaN	NaN	NaN	NaN	NaN	0.0	44.3	100.0	48.7	42.2	0.0	United States
2012	University of the Arts London	NaN	NaN	NaN	NaN	NaN	NaN	100.0	48.3	0.0	0.0	0.0	0.0	United Kingdom



Transformasi Dataset Hasil Integrasi

Pada dataset ini dilakukan beberapa tahapan transformasi:

1. Membuat kolom baru **Country** untuk menampung value dari kolom **Country_THE** dan kolom **Country_qsworld** yang saling melengkapi
2. Menghapus kolom **Country_THE** dan kolom **Country_qsworld**
3. Menghitung **Average_Scoring** dan menambahkan kolom **World_Rank**
4. Menambahkan kolom **Continent** dari kolom **Country** menggunakan pycountry-convert
5. Mengubah nilai NaN dengan angka 0
6. Membuat kolom **Continent_Rank** dan **National_Rank** yang menunjukkan ranking universitas tersebut pada wilayah benua dan negara



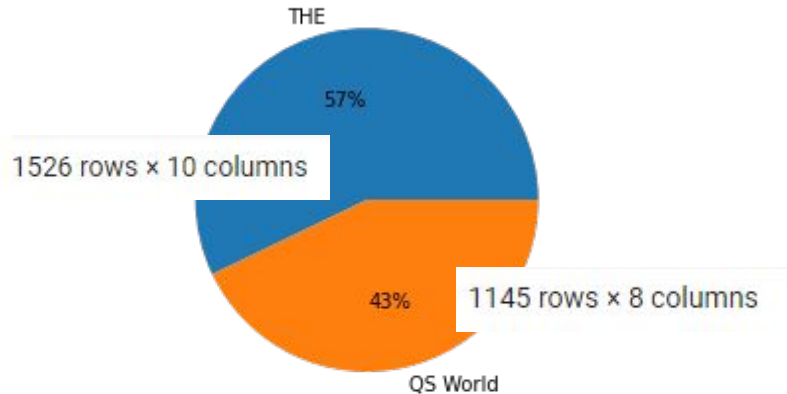
Exploratory Data Analysis





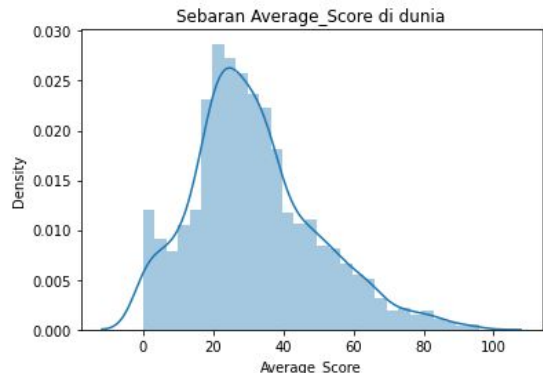
Komposisi Dataset

Banyaknya Universitas Setiap Data Set



Dari kedua website sumber dataset, data yang disediakan memiliki perbedaan pada jumlah universitas dan komponen penilaian yang digunakan. Perbandingan jumlahnya adalah 57:43.

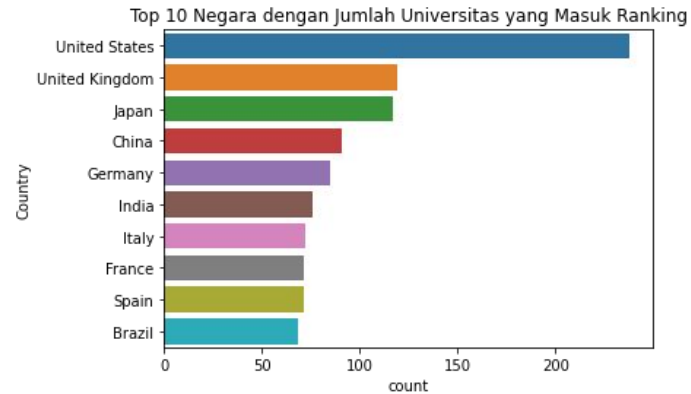
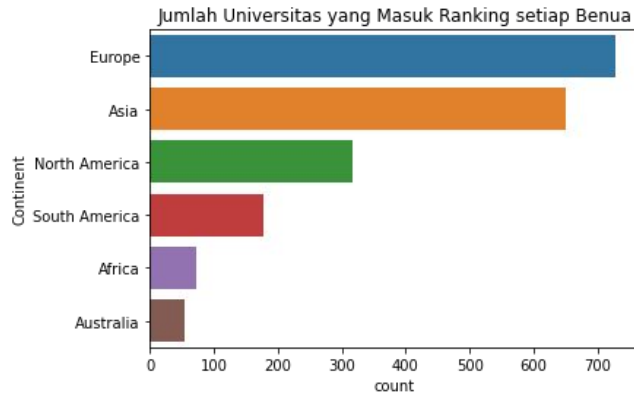
Sebaran Average_Score



```
count    2000.000000
mean      31.676581
std       17.916977
min        0.000000
25%       19.947500
50%       29.117424
75%       41.325000
max       96.018182
Name: Average_Score, dtype: float64
```

Dapat dilihat bahwa kita mendapat mean dari average score = 31.67 dan dengan standar deviasi = 17.91. Standar deviasi yang cukup besar menunjukkan bahwa kualitas universitas di dunia belum merata.

Perbandingan Antar Benua dan Antar Negara

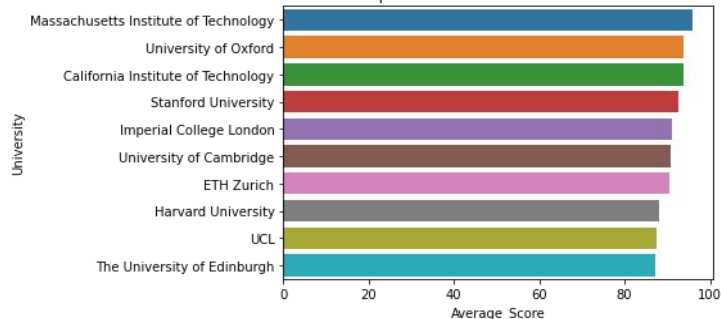


- Benua Eropa merupakan benua dengan universitas terbanyak yang masuk ranking
- United States memiliki jauh lebih banyak universitas yang masuk ranking daripada negara-negara lainnya

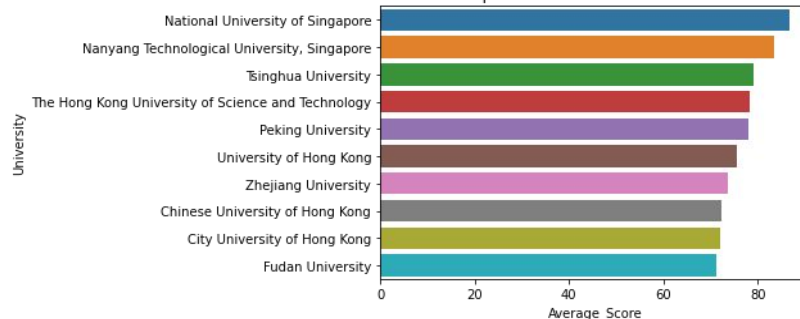


Top 10 Universitas

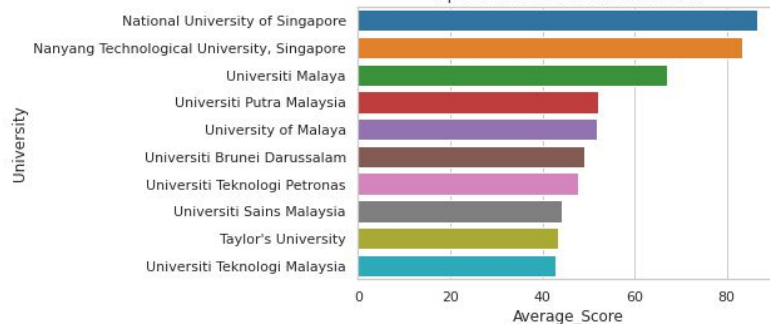
Top 10 Universitas Terbaik di Dunia



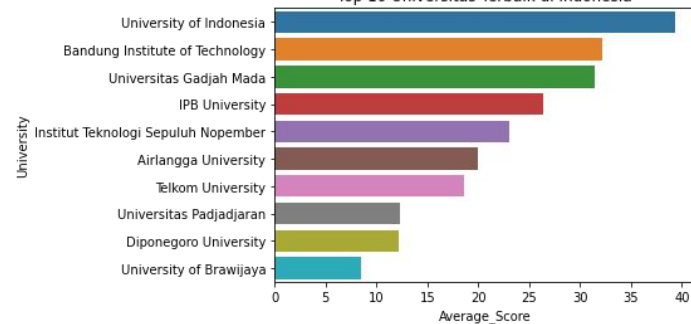
Top 10 Universitas Terbaik di Asia



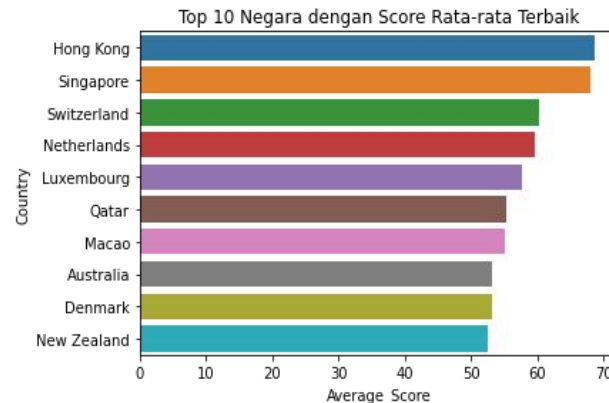
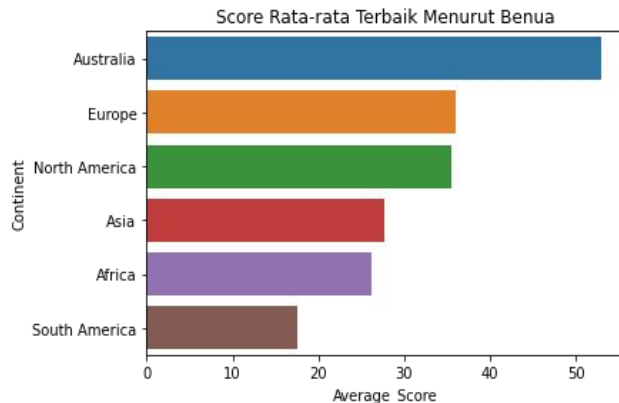
Top 10 Universitas Terbaik di ASEAN



Top 10 Universitas Terbaik di Indonesia

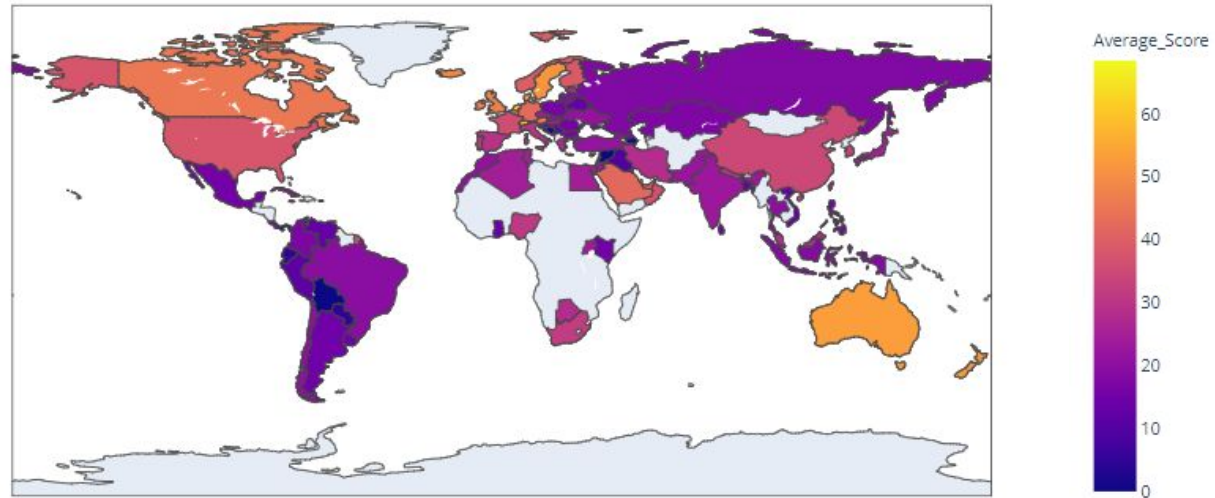


Skor Rerata terbaik



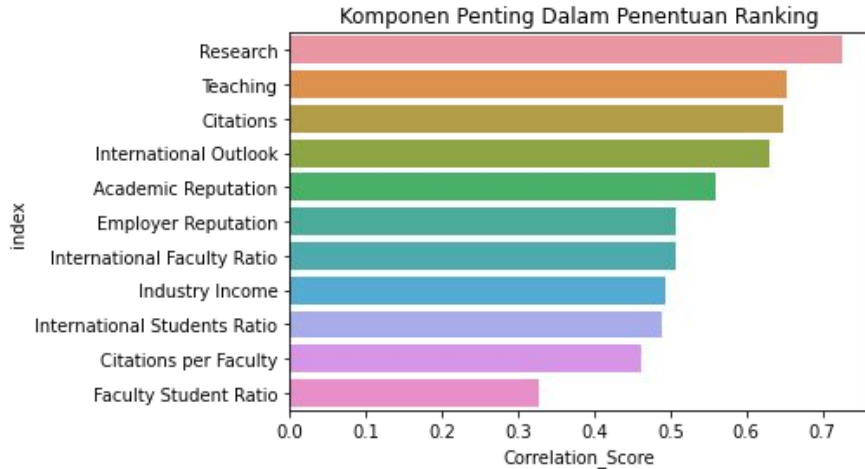
- Benua Australia memiliki skor rerata yang tertinggi (>50 poin) dibandingkan benua lainnya. Disusul oleh benua Eropa dan Amerika Utara
- Hongkong merupakan negara dengan skor rerata universitas tertinggi

Rerata Skor berdasarkan letak negara





Komponen Penting dalam Penentuan Ranking



- Dilakukan pengecekan korelasi antara kolom `average_score` dengan kolom-kolom lainnya
- Hasil korelasi menunjukkan kontribusi kolom dalam penentuan ranking universitas
- Kolom Research merupakan kolom dengan nilai korelasi tertinggi. Diikuti oleh kolom Teaching, Citations, dan International Outlook



Kesimpulan





Kesimpulan

- Pada proyek ini digunakan ELT (Extract, Load, dan Transform) data pipeline model. Data disimpan di Google Drive dalam bentuk raw data (belum ditransformasikan).
- Pada pipeline ELT, Transformasi data hanya dilakukan ketika dibutuhkan saja.



Kesimpulan

- Top 3 universitas di dunia :
 - Massachusetts Institute of Technology (MIT)
 - University of Oxford
 - California Institute of Technology (Caltech)
- Top 3 universitas di Asia :
 - National University of Singapore (NUS)
 - Nanyang Technological University, Singapore (NTU)
 - Tsinghua University
- Top 3 universitas di Indonesia :
 - University of Indonesia
 - Bandung Institute of Technology (ITB)
 - Universitas Gadjah Mada



Kesimpulan

- Skor rata-rata terbaik
 - Benua Australia memiliki skor rerata yang tertinggi (>50 poin) dibandingkan benua lainnya. Disusul oleh benua Eropa (~35 poin) dan Amerika Utara (~35 poin)
 - Hongkong merupakan negara dengan skor rerata universitas tertinggi di dunia (~69 poin), disusul oleh Singapura (~68 poin) dan Switzerland (~60 poin)
- Komponen penilaian terpenting yang menentukan ranking dari universitas yaitu :
 - Research
 - Teaching
 - Citations



Saran

Agar memperbaiki hasil re-scoring, akan lebih baik jika menggunakan weighted scoring. Weighted scoring dapat dilihat pada situs Times Higher Education maupun QS World Ranking.

Pillar	Metric	% weighting
1. Teaching	Reputation survey	15.00
	Academic staff-to-student ratio	4.50
	Doctorates awarded-to-bachelor's degrees awarded ratio	2.25
	Doctorates awarded-to-academic staff ratio	6.00
	Institutional income	2.25
2. Research	Reputation survey	18.00
	Research income	6.00
	Research productivity	6.00
3. Citations	Citations	30.00
4. International outlook	Proportion of international students	2.50
	Proportion of international staff	2.50
	International collaboration	2.50
5. Industry income	Industry income	2.50
		100