Assigment 22 Web Scrapping

Web Scraping Data Lowongan Kerja dari JobStreet Menggunakan Selenium & BeautifulSoup

Fuad Hasyim

WRITE TO US

Mengotomatisasi Pengumpulan Data Lowongan Kerja untuk Analisis Data

Tujuan Proyek

- *Mengotomatisasi proses pengumpulan data lowongan kerja dari JobStreet Indonesia.*
- *Mengekstrak informasi penting seperti:*
  - *Judul Pekerjaan*
  - *Nama Perusahaan*
  - *Lokasi*
  - *Gaji*
  - *Tautan Lowongan*
- *Membangun dataset untuk analisis tren pekerjaan dan kebutuhan pasar kerja.*

Tools & Library

Library	Fungsi
Selenium	Mengontrol browser dan mengambil data dari website dinamis
BeautifulSoup	Memproses dan mengekstrak data dari HTML
Pandas / NumPy	Mengelola dan membersihkan data
Matplotlib / Seaborn	Membuat visualisasi data
WordCloud	Menampilkan frekuensi kata dari deskripsi pekerjaan

Alur Kerja (Workflow)

1

Inisialisasi Selenium WebDriver (Chrome).

2

Mengakses URL untuk tiap posisi pekerjaan

3

Melakukan scroll agar seluruh konten termuat.

4

Parsing HTML dengan BeautifulSoup

5

Mengambil detail pekerjaan (title, company, location, salary, link).

6

Menyimpan hasil ke dalam list.

7

Mengubah hasil ke DataFrame dan menyimpannya ke CSV.

8

Menyimpan ke database untuk analisis lanjutan dan visualisasikan

Logika Web Scraping

```
option = webdriver.ChromeOptions()
driver = webdriver.Chrome(options=option)
```

Setup Selenium webdriver.Chrome() berperan seperti “browser otomatis”.

```
for page in range(1, MAX_PAGES + 1):
    url = f"https://id.jobstreet.com/id/data-scientist-jobs?page={page}"
    driver.get(url)
    time.sleep(3)
    driver.execute_script("window.scrollTo(0, document.body.scrollHeight);")
```

Mengambil data dari setiap halaman dan memastikan semua card lowongan termuat sebelum diproses.

Ekstraksi Informasi Pekerjaan

```
soup = BeautifulSoup(driver.page_source, 'html.parser')
jobs = soup.find_all('article')

for job in jobs:
    title = job.find('a', {'data-automation': 'jobTitle'})
    company = job.find('a', {'data-automation': 'jobCompany'})
    salary = job.find('span', {'data-automation': 'jobSalary'})
```

Parsing HTML dengan BeautifulSoup

Mengambil data dari setiap halaman dan memastikan semua kartu lowongan termuat sebelum diproses.

```
job_info = {
  'title': title.text.strip(),
  'company': company.text.strip(),
  'location': location.text.strip(),
  'salary': salary.text.strip() if salary else 'NaN',
  'link': link
}
results.append(job_info)
```

Semua hasil disimpan dalam list of dictionaries,
lalu diubah menjadi DataFrame Pandas.

Ekspor ke CSV

```python
df_job = pd.DataFrame(results)
df_job.to_csv('jobstreet_data_engineer_all_pages.csv', index=False)
print(f"Total lowongan terkumpul: {len(df_job)}")
```

In [39]:

```python
# Convert to DataFrame
df_job = pd.DataFrame(results)
df_job.head()
```

Out[39]:

	title	company	location	salary	link
	Database Finance Administrator	KSO Sucofindo-Surveyor Indonesia	Jakarta Selatan	NaN	https://id.jobstreet.com/id/job/86382708?type=standard&ref=search-standalone&origin=cardTitle#sol=8b6c3ec88ab4bac7d3372d0d8e4f6f19a8577d0f
	Senior Database Administrator Engineer - Engineering Platform	PT GOTO GOJEK TOKOPEDIA TBK	Jakarta Raya	NaN	https://id.jobstreet.com/id/job/86494164?type=standard&ref=search-standalone&origin=cardTitle#sol=9802c8e825f33fadea5bb7aac9e0c9cb8df20038
	SENIOR DATABASE ENGINEER	PT JAYA AGUNG TEKNOLOGI	Jakarta Utara	NaN	https://id.jobstreet.com/id/job/86535509?type=standard&ref=search-standalone&origin=cardTitle#sol=a8c4d2d4973d3504bd66916030c7bee2d743fe9c

Menyimpan hasil scraping ke file CSV llalu gabungkan semua hasil web scrapping menjadi 1, agar bisa dianalisis lebih lanjut atau diunggah ke database.

Contoh Hasil

- *Data Posisi yang Diambil*
- *Business Analyst — 30 halaman*
- *Data Scientist — 9 halaman*
- *Data Engineer — 30 halaman*
- *Database Administrator — 4 halaman*

Total: ±70 halaman lowongan dikumpulkan secara otomatis

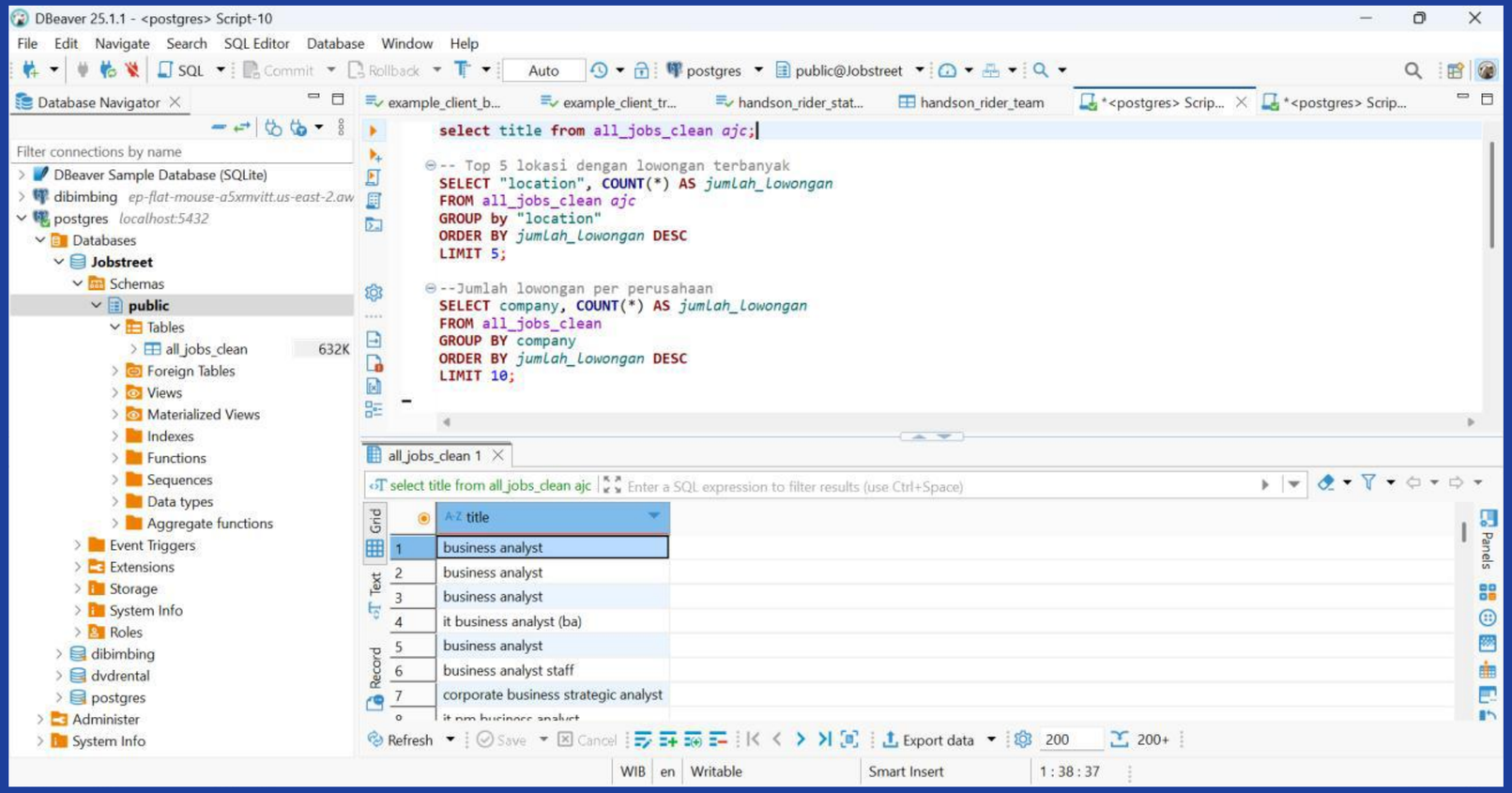
Data Cleaning dan selection

1. *Handling Duplicate*
2. *Kategorisasi dan Menghapus Pekerjaan yang Tidak Relevan*

Meskipun sudah menggunakan kata kunci "Data Scientist, Data Engineer, Data Base Administrator dan Bussines analyst", ternyata tidak semua pekerjaan yang muncul benar-benar terkait dengan keempat hal tersebut.

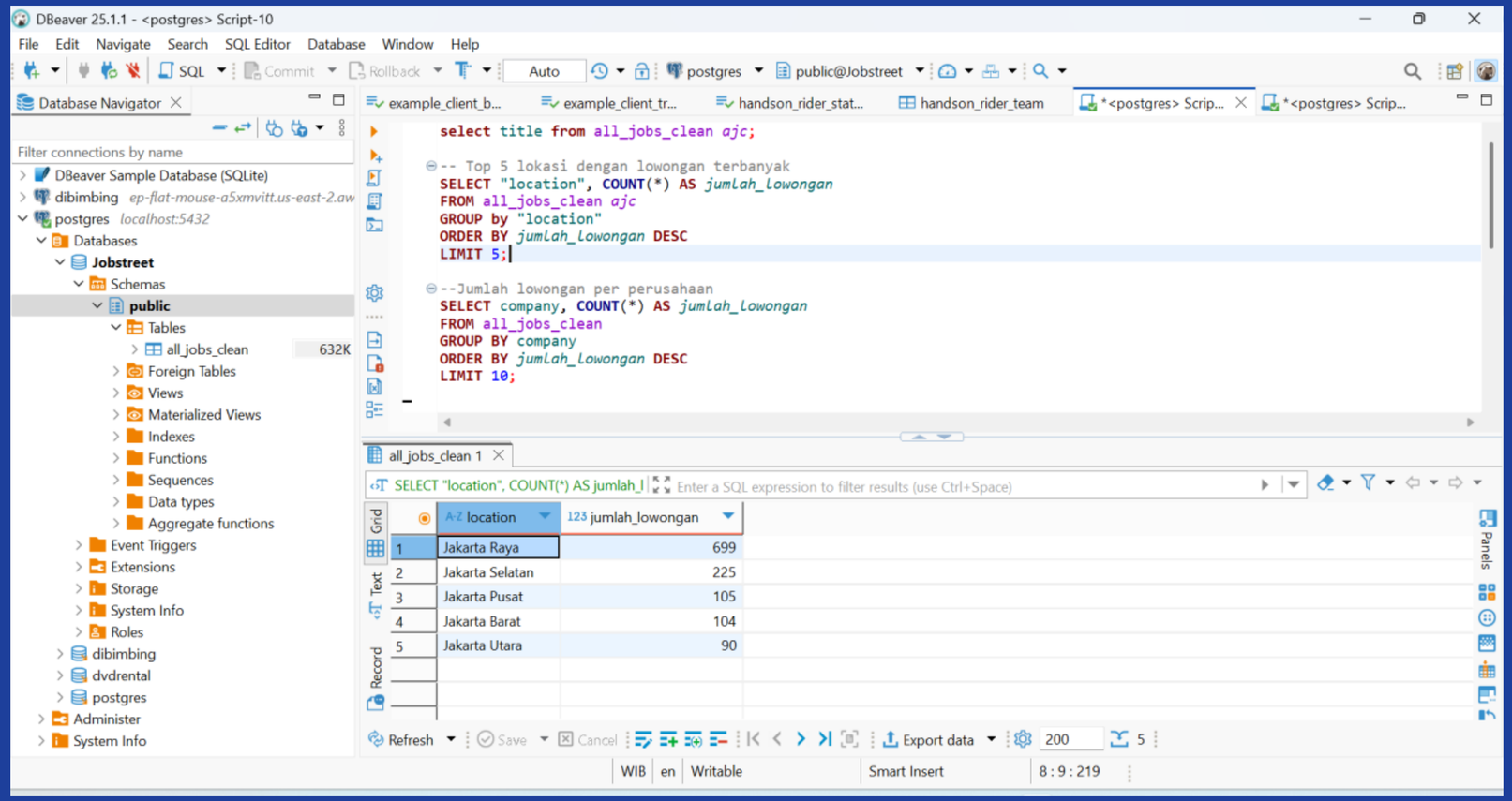
```
ds_list = ['data scientist', 'data science', 'machine learning','artificial intelligence', 'ai/ml', 'ml engineer','ai engineer']
de_list = ['data engineer', 'big data engineer', 'etl engineer','data pipeline', 'data warehouse', 'hadoop', 'spark engineer']
dba_list = ['database administrator', 'dba', 'sql server','oracle dba', 'mysql administrator', 'postgresql administrator','db administrator']
ba_list = ['business analyst', 'bisnis analis', 'system analyst','business intelligence','bi analyst']
length = len(df['title'])
df_title = df['title']
```

Ekploratory Data



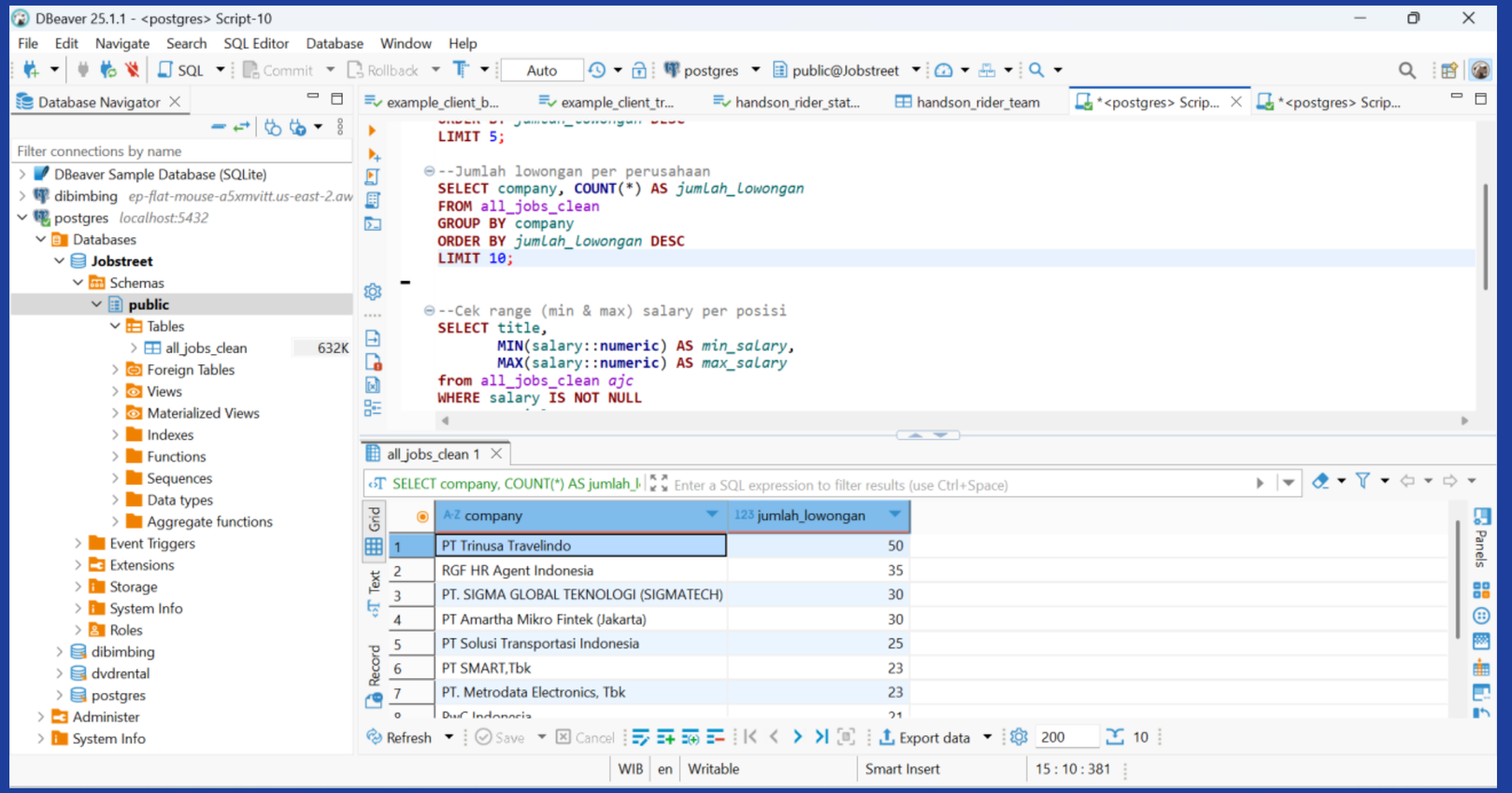
Ekploratory Data

TOP 5 job by Location



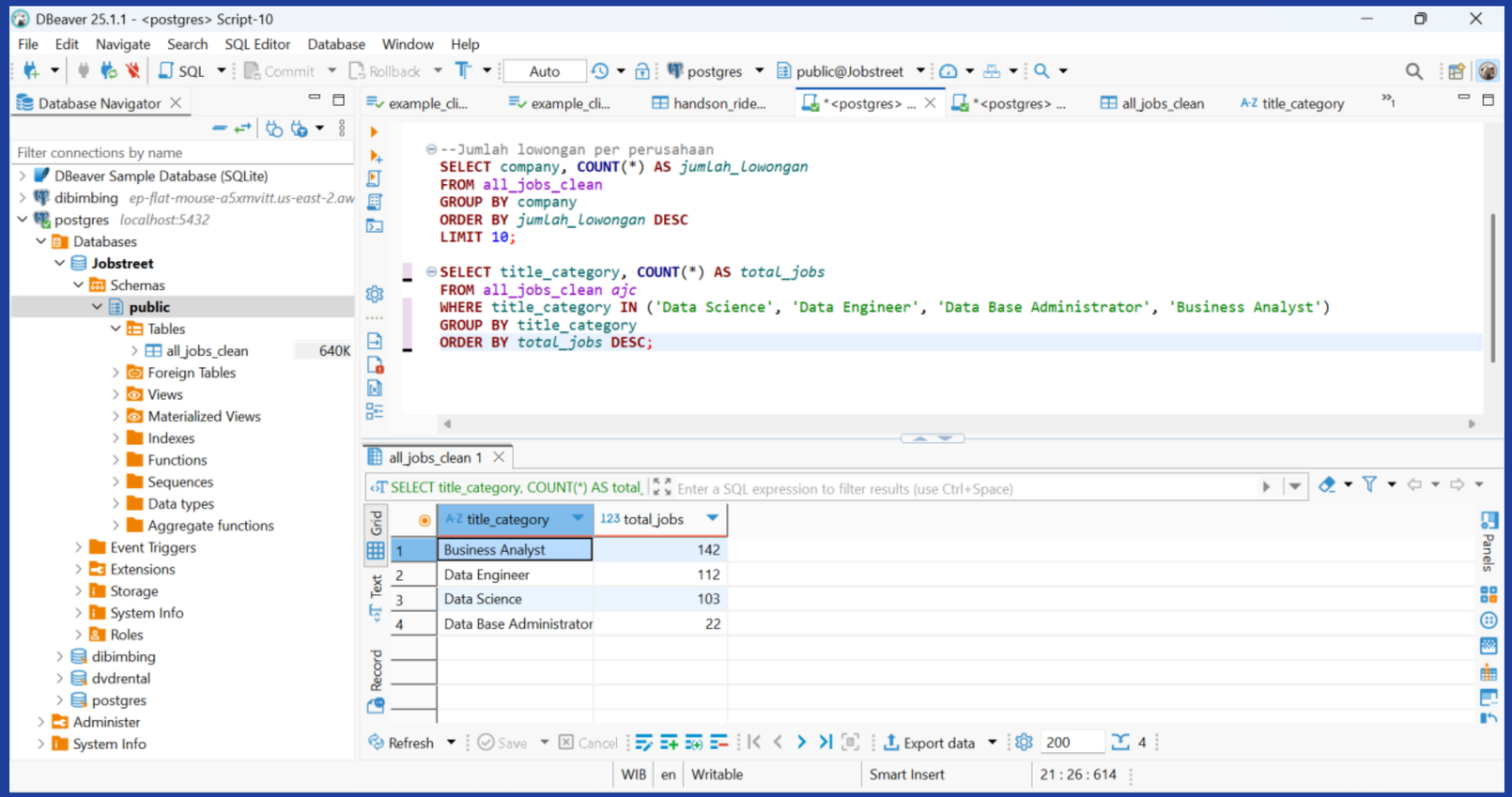
Ekploratory Data

Job Title by Company

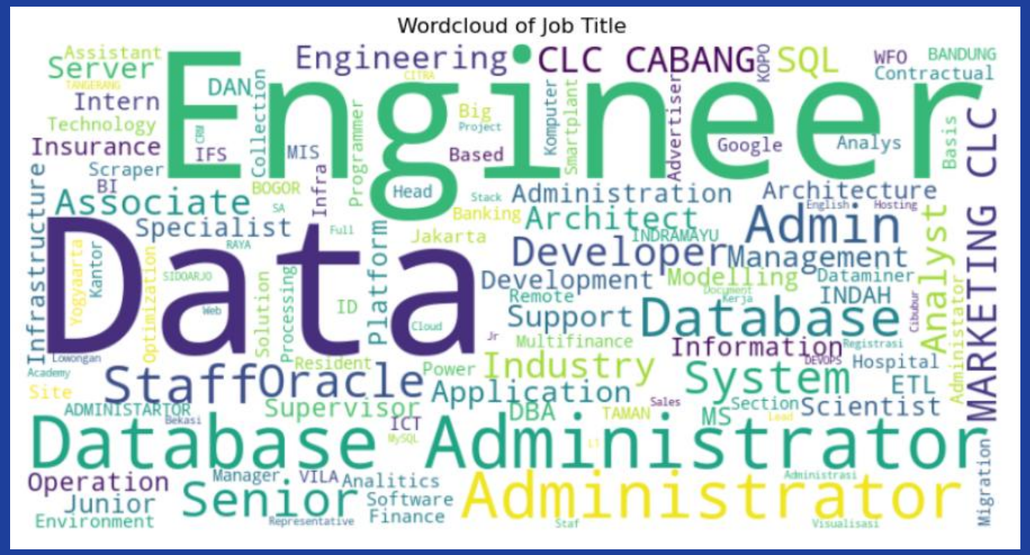


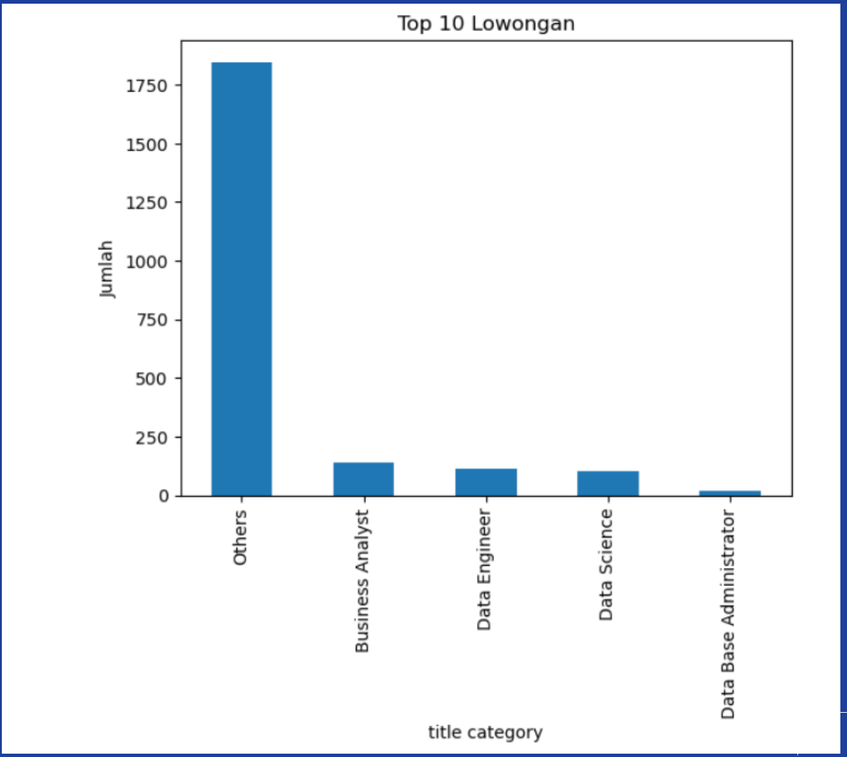
Ekploratory Data

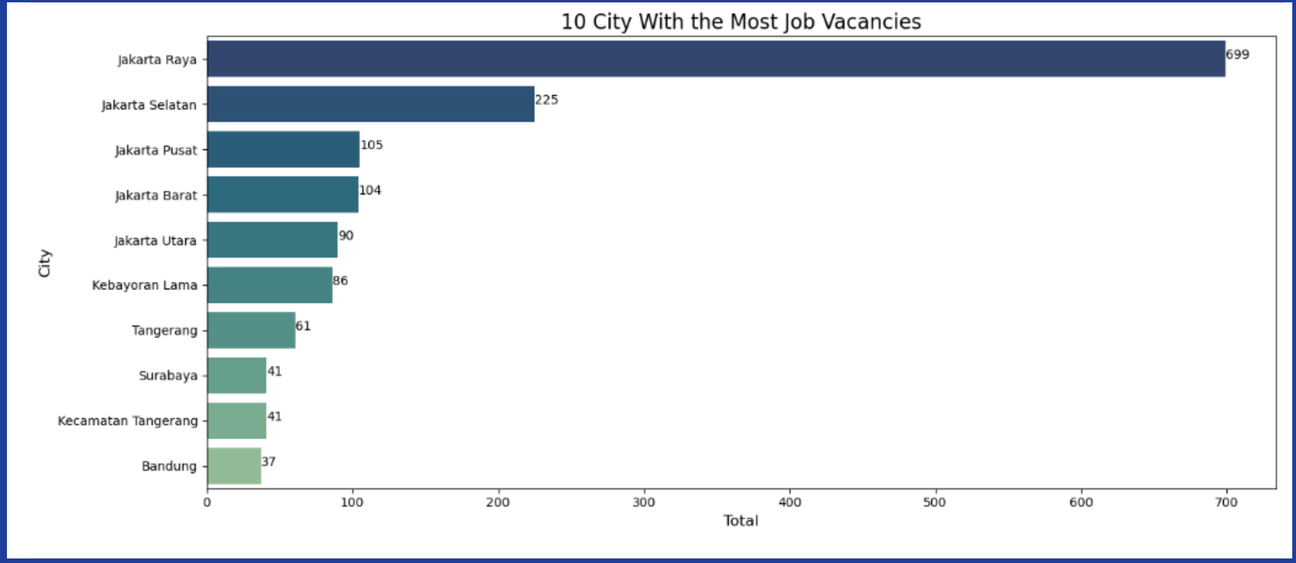
Job by Title Category



Vizualization







Conclusion

Top 5 job Data by location

- *Jakarta Raya : 699 job*
- *Jakarta Selatan : 225 job*
- *Jakarta Pusat : 105 job*
- *Jakarta Barat : 104 job*
- *Jakarta Utara : 90 job*

Top 5 job Data by Company

- *PT Trinusa Travelindo : 50 job*
- *RGF HR Indonesia : 35 job*
- *PT Sigma Global Teknologi (SIGMATECH): 30 job*
- *PT Amartha mikro Fintech : 30 job*
- *PT Solusi Transportasi Indonesia: 25 job*

Top 5 job Data by Title Job

- *Data Analys : 699 job*
- *Data Engineer: 225 job*
- *Data Science : 105 job*
- *Database Administrator: 104 job*

Terima kasih.

Silahkan bertanya

Tapi, jangan yang susah” ya!!!!!