

**TUGAS BESAR PENGENALAN KOMPUTASI
ANALISIS DATA GAJI DATA-SCIENTIST 2023
DI DUNIA MELALUI VISUALISASI DATA**

LAPORAN

*Disusun untuk memenuhi tugas besar mata kuliah KU1102 Pengenalan
Komputasi yang dibimbing oleh Bapak Dr. Fadhil Hidayat, S.Kom., MT.*



DISUSUN OLEH:

Nadia Apsarini Baizal (19623122)

Alisha Tazkia Anugraha (16523066)

Muhammad Falih Rosyid (16523108)

Abdul Hakim Yafi (16523206)

**SEKOLAH TEKNIK ELEKTRO DAN INFORMATIKA
INSTITUT TEKNOLOGI BANDUNG
BANDUNG
DESEMBER 2023**

KATA PENGANTAR

Puji syukur kami panjatkan ke hadirat Tuhan Yang Maha Esa. Atas rahmat dan hidayah-Nya, penulis bisa menyelesaikan Laporan Tugas Besar 2 yang berjudul “Analisis Data Gaji Data-scientist di Dunia Melalui Visualisasi Data”

Laporan ini dibuat untuk memenuhi persyaratan Tugas Besar mata kuliah KU1102 Pengenalan Komputasi yang dibimbing oleh Dr. Fadhil Hidayat, S.Kom., MT. Pembuatan laporan ini membutuhkan kerja keras dan tidak dapat diselesaikan tanpa dukungan seluruh pihak yang terlibat. Penulis mengucapkan terima kasih sebesar-besarnya kepada:

1. Orang tua penulis yang memberikan dukungan, baik moral maupun material.
2. Semua pihak yang tidak dapat penulis perinci satu per satu yang telah membantu proses pembuatan laporan ini.

Penulis menyadari bahwa laporan ini masih memiliki kekurangan dan keterbatasan. Oleh karena itu, penulis mengharapkan saran dan kritik yang membangun dari pembaca untuk menyempurnakan laporan ini.

Sumedang, Desember 2023

Penulis

ABSTRAK

Laporan ini secara terstruktur disusun untuk menyajikan hasil dari tugas besar 2 yang telah dilaksanakan oleh kelompok 12 dalam mata kuliah Pengenalan Komputasi. Tugas besar 2 ini berfokus pada analisis statistik dan visualisasi data terkait gaji pekerjaan di bidang data science. Data yang kami teliti disajikan dalam format .csv yang terdiri dari 11 kolom dan 3755 baris, mencakup informasi seperti tahun pembayaran gaji, tingkat jabatan, jenis pekerjaan, posisi pekerjaan, besaran gaji, mata uang gaji, gaji dalam USD, lokasi tempat tinggal pekerja, jumlah pekerjaan jarak jauh, lokasi perusahaan, ukuran perusahaan, dan gaji dalam IDR.

Dalam pelaksanaan tugas besar ini, kelompok kami menggunakan berbagai algoritma statistik dan teknik visualisasi data untuk menggali informasi terkait kelompok data tertentu. Analisis ini bertujuan untuk memahami pengaruh, perubahan, atau korelasi di antara data yang telah kami telaah.

DAFTAR ISI

KATA PENGANTAR	2
ABSTRAK	3
DAFTAR ISI.....	4
BAB I	6
PENDAHULUAN	6
1. Latar Belakang	6
BAB II	7
HASIL PENELITIAN	7
2.1 Deskripsi Data dan File	7
2.2 Karakteristik Data.....	7
2.2.1 Deskripsi Atribut Data	7
2.3 Statistik Data.....	9
2.3.1 Sampel Data.....	9
2.3.2 Statistik	9
2.4 Visualisasi Data	10
2.4.1 Rata-Rata dan Gaji Berdasarkan Tahun Kerja	10
2.4.2 Rata-Rata dan Gaji Berdasarkan Pengalaman Kerja.....	11
2.4.3 5 Rata-Rata Gaji Tertinggi Berdasarkan Jenis Pekerjaan	12
2.4.4 Perbandingan Jumlah Pekerja Berdasarkan Pengalaman kerja	12
2.4.5 Rata-Rata Gaji Berdasarkan Ukuran Perusahaan	13
2.4.6 Rata-Rata Gaji Berdasarkan Cara Kerja	13
2.4.7 Perbandingan Jumlah Pekerja Berdasarkan Tahun Kerja.....	14
2.4.8 Rata-Rata Gaji Berdasarkan Tempat Perusahaan	14
2.5 Korelasi.....	15
2.5.1 Korelasi Keseluruhan	15
2.5.2 Korelasi Antara Jabatan dan Gaji.....	15
2.5.3 Korelasi Antara Ukuran Perusahaan dan Gaji	16
2.5.4 Korelasi Antara Cara Kerja dan Gaji.....	16
2.6 Data Cleansing	16
2.6.1 Penjabaran Kolom Remote-Ratio	16
2.6.2 Penjabaran Kolom Employment-Type.....	17
2.6.3 Konversi Gaji USD ke Rupiah	17
BAB III	17
PENUTUP.....	17
3.1 Kesimpulan	17
3.2 Lesson Learned.....	17

LAMPIRAN.....	18
4.1 Source Code Program:	18
4.2 Raw Data:	18

BAB I

PENDAHULUAN

1. Latar Belakang

Seiring dengan berkembangnya ilmu pengetahuan dan teknologi di seluruh dunia khususnya di Indonesia, kebutuhan akan data analis dan data scientist semakin meningkat pesat di berbagai perusahaan. Hal ini juga didorong oleh pandemi COVID-19 yang mempercepat penggunaan teknologi digital. Menurut laporan kontan.co.id, pengguna aktif teknologi digital setiap bulan meningkat dari 90 juta pada awal pandemi menjadi 100 juta pada Mei 2022. Mahasiswa STEI yang ingin bekerja sebagai data analis atau data scientist harus memperhatikan distribusi gaji di bidang data science dan faktor-faktor yang mempengaruhi tinggi rendahnya gaji. Harapannya dengan menganalisis data tersebut, mahasiswa STEI akan lebih siap dalam merencanakan karir data analis.

BAB II

HASIL PENELITIAN

2.1 Deskripsi Data dan File

Penelitian tugas besar 2 kami didasarkan pada data gaji di bidang *data science* yang didapatkan dari situs <https://www.kaggle.com/datasets/data-science-salaries-2023>. Data berformat .csv dan berukuran 26 KB ini merupakan data yang berisi 11 kolom dan 3755 baris data, serta terdiri dari data kategorik dan data numerik. Dari data tentang gaji profesional *data scientist*, kita dapat mengetahui faktor faktor apa saja yang menentukan perbedaan gaji seorang profesional *data scientist* dan berapa gaji seorang data scientist dalam gambaran besar.

2.2 Karakteristik Data

2.2.1 Deskripsi Atribut Data

Column	Description
work_year	Tahun ketika gaji dibayarkan (kuantitatif)
experience_level	Jabatan ketika data diambil. EN Entry-level / Junior MI Mid-level / Intermediate SE Senior-level / Expert EX Executive-level / Director (kategorikal)
employment_type	Jenis perjanjian kerja, (kategorikal) PT Part-time FT Full-time CT Contract FL Freelance
job_title	Posisi yang diisi ketika data diambil, (kategorikal)

salary	Gaji kotor, (kuantitatif)
salary_currency	mata uang yang dibayarkan dengan kode ISO 4217, (kategorikal)
salaryinusd	Gaji dalam konversi US Dollar, (kuantitatif)
employee_residence	Domisili pekerja ketika data diambil dengan kode ISO 3166 country code. (kategorikal)
remote_ratio	Bagian pekerjaan yang dimungkinkan untuk dikerjakan secara online atau from home. 0 No remote work (less than 20%) 50 Partially remote 100 Fully remote (more than 80%) (kuantitatif)
company_location	Domisili dari kantor pekerja, (kategorikal)
company_size	Ukuran kantor pemberi pekerjaan. S <50 pegawai (small). M, 50 hingga 250 pegawai (medium) L >250 pegawai (large) (kategorikal)

2.3 Statistik Data

2.3.1 Sampel Data

```
[ ] display(dataset[:5])
print("Sampel 5 data Awal")
```

index	work_year	experience_level	employment_type	job_title	salary	salary_currency	salary_in_usd	employee_residence	remote_ratio	company_location	company_size
0	2023	SE	FT	Principal Data Scientist	80000	EUR	85847	ES	100	ES	L
1	2023	MI	CT	ML Engineer	30000	USD	30000	US	100	US	S
2	2023	MI	CT	ML Engineer	25500	USD	25500	US	100	US	S
3	2023	SE	FT	Data Scientist	175000	USD	175000	CA	100	CA	M
4	2023	SE	FT	Data Scientist	120000	USD	120000	CA	100	CA	M

Show 25 per page
Sampel 5 data Awal

sampel data 5 baris teratas

2.3.2 Statistik

```
desc= dataset.describe()
display(desc)
percent = pd.DataFrame({
    "Tahun Bekerja": [dataset["work_year"].quantile(0.1),dataset["work_year"].quantile(0.25),dataset["work_year"].quantile(0.5),dataset["work_year"].quantile(0.75),dataset["work_year"].quantile(0.9)],
    "Gaji": [dataset["salary"].quantile(0.1),dataset["salary"].quantile(0.25),dataset["salary"].quantile(0.5),dataset["salary"].quantile(0.75),dataset["salary"].quantile(0.9)],
    "Gaji dalam USD": [dataset["salary_in_usd"].quantile(0.1),dataset["salary_in_usd"].quantile(0.25),dataset["salary_in_usd"].quantile(0.5),dataset["salary_in_usd"].quantile(0.75),dataset["salary_in_usd"].quantile(0.9)],
    "Rasio Remote": [dataset["remote_ratio"].quantile(0.1),dataset["remote_ratio"].quantile(0.25),dataset["remote_ratio"].quantile(0.5),dataset["remote_ratio"].quantile(0.75),dataset["remote_ratio"].quantile(0.9)]
},index = ["10%", "25%", "50%", "75%", "90%"])
display(percent)
```

index	work_year	salary	salary_in_usd	remote_ratio
count	3755.0	3755.0	3755.0	3755.0
mean	2022.3736351531293	190695.57177097205	137570.38988015978	48.271637816245004
std	0.6914482342671989	671676.5005079068	63055.6252762241	48.589050470587566
min	2020.0	6000.0	5132.0	0.0
25%	2022.0	100000.0	95000.0	0.0
50%	2022.0	138000.0	135000.0	0.0
75%	2023.0	180000.0	175000.0	100.0
max	2023.0	3040000.0	450000.0	100.0

Show 25 per page

index	Tahun Bekerja	Gaji	Gaji dalam USD	Rasio Remote
10%	2022.0	60000.0	59537.000000000002	0.0
25%	2022.0	100000.0	95000.0	0.0
50%	2022.0	138000.0	135000.0	0.0
75%	2023.0	180000.0	175000.0	100.0
90%	2023.0	228599.99999999999	219000.0	100.0

Show 25 per page

Dari data tersebut, dapat diketahui bahwa gaji dalam USD memiliki rata - rata sebesar \$13.750 dengan standar deviasi \$63.055

```
print("Gaji terbesar adalah 450000 USD")
print("Gaji terkecil adalah 5132 USD")
print("Range gaji adalah (450000-5132) USD")
print()
print("Tahun kerja terlama adalah 2020")
print("Tahun kerja terbaru adalah 2023")
```

Gaji terbesar adalah 450000 USD
Gaji terkecil adalah 5132 USD
Range gaji adalah 444868 USD

Tahun kerja terlama adalah 2020
Tahun kerja terbaru adalah 2023

```

Jumlah pekerja berdasarkan tahun kerja
2020: 76
2021: 230
2022: 1664
2023: 1785

Jumlah pekerja berdasarkan pengalaman kerja
SE: 2516
MI: 805
EN: 320
EX: 114

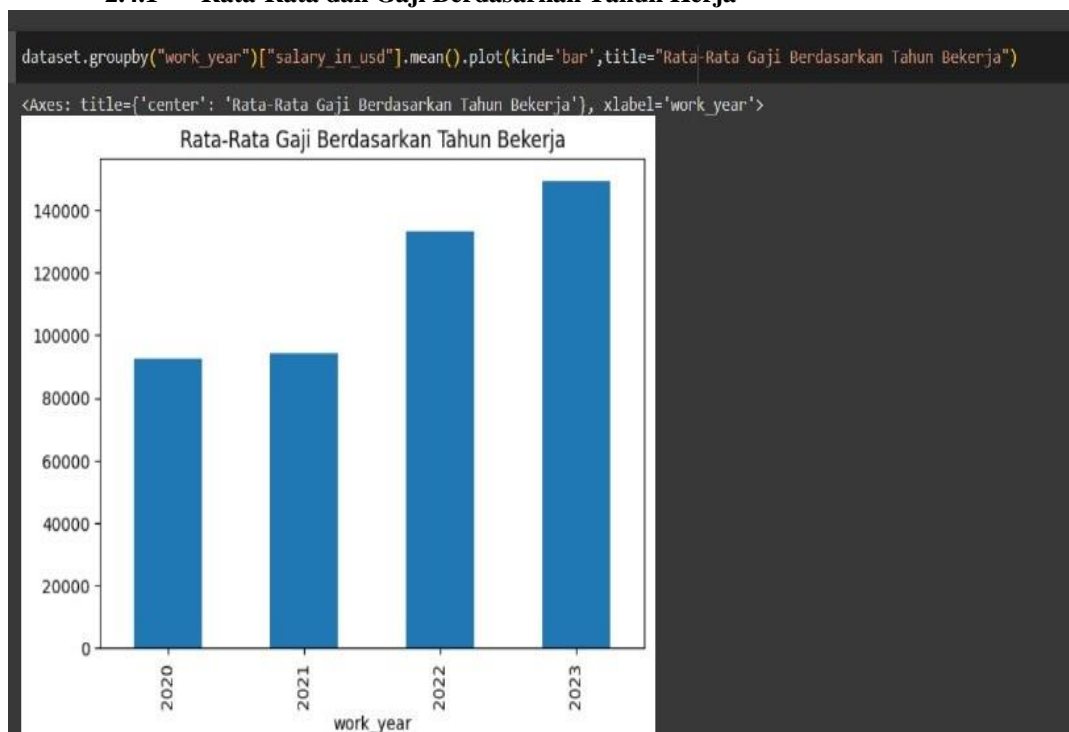
Jumlah pekerja berdasarkan ukuran perusahaan
S: 148
M: 454
L: 3153

```

Dari data tersebut dapat dilihat, gaji terbesar, gaji terkecil, range gaji, tahun kerja terlama, dan terbaru. Kita juga dapat melihat jumlah pekerja berdasarkan tahun kerjanya, berdasarkan pengalaman dan juga berdasarkan ukuran pekerjaannya.

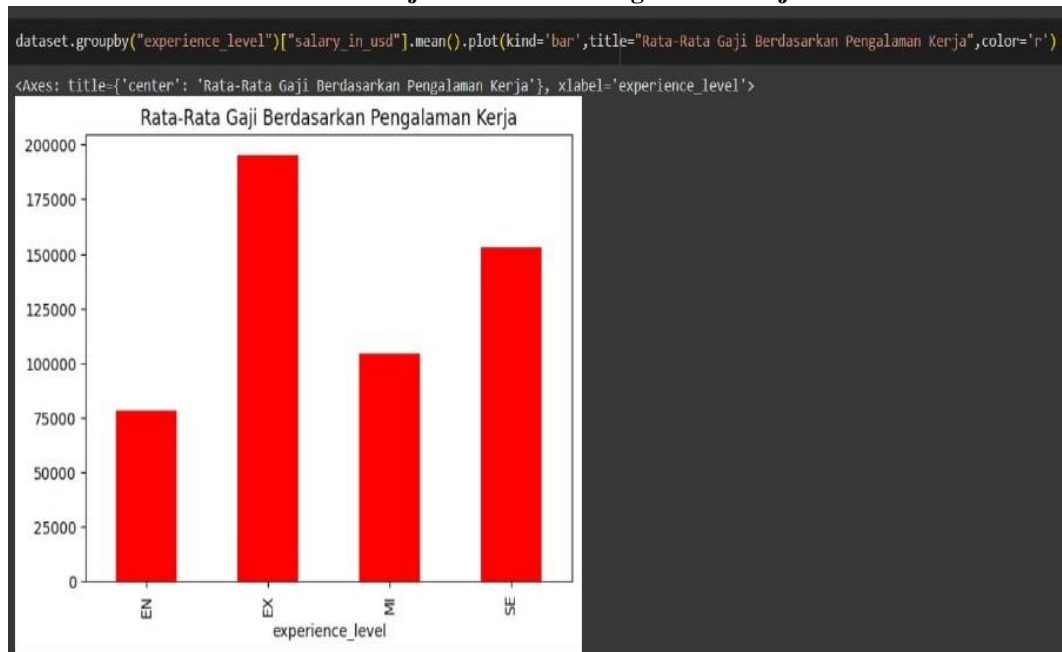
2.4 Visualisasi Data

2.4.1 Rata-Rata dan Gaji Berdasarkan Tahun Kerja



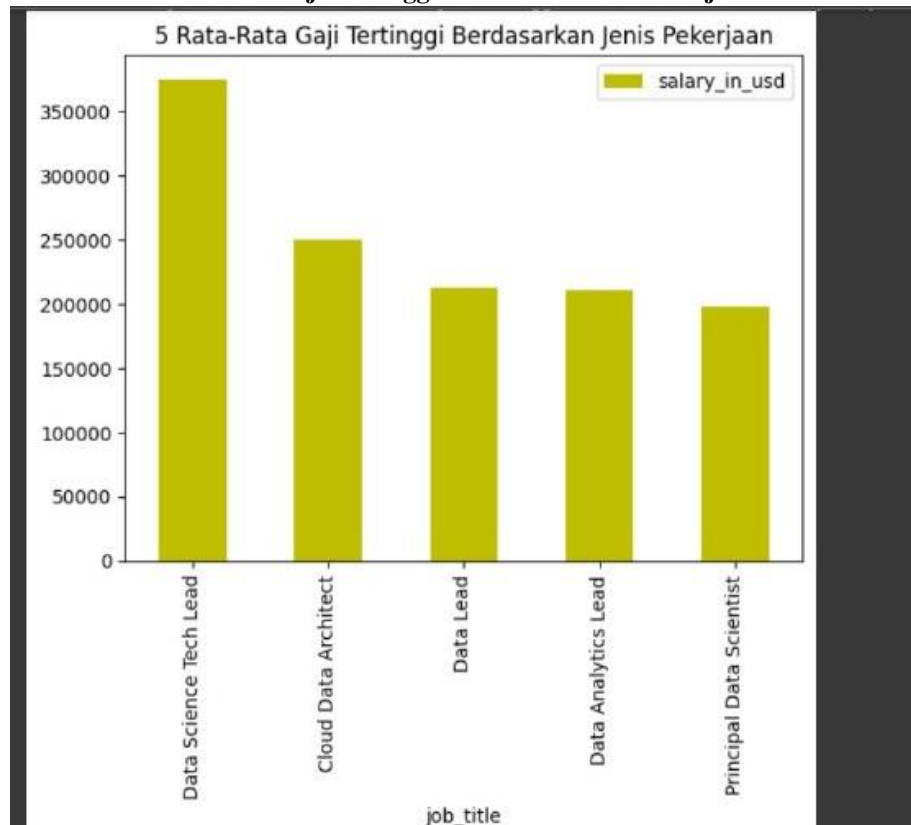
Rata-rata gaji yang dimiliki oleh Data Analyst termasuk tinggi. Berdasarkan tahun kerjanya, data analyst memiliki gaji yang akan selalu meningkat dari tahun ke tahunnya. Sehingga ini sangat menjanjikan untuk mahasiswa STEI untuk memilih pekerjaan ini.

2.4.2 Rata-Rata dan Gaji Berdasarkan Pengalaman Kerja



Dari data ini dapat dilihat bahwa, semakin tinggi jabatan dari seorang data analys, gaji juga semakin meningkat. Untuk entry saja rata-rata gaji dari data analys nya saja sudah tinggi, ini sangat menjanjikan untuk siswa STEI ITB untuk memilih data analys sebagai pekerjaan nantinya.

2.4.3 5 Rata-Rata Gaji Tertinggi Berdasarkan Jenis Pekerjaan



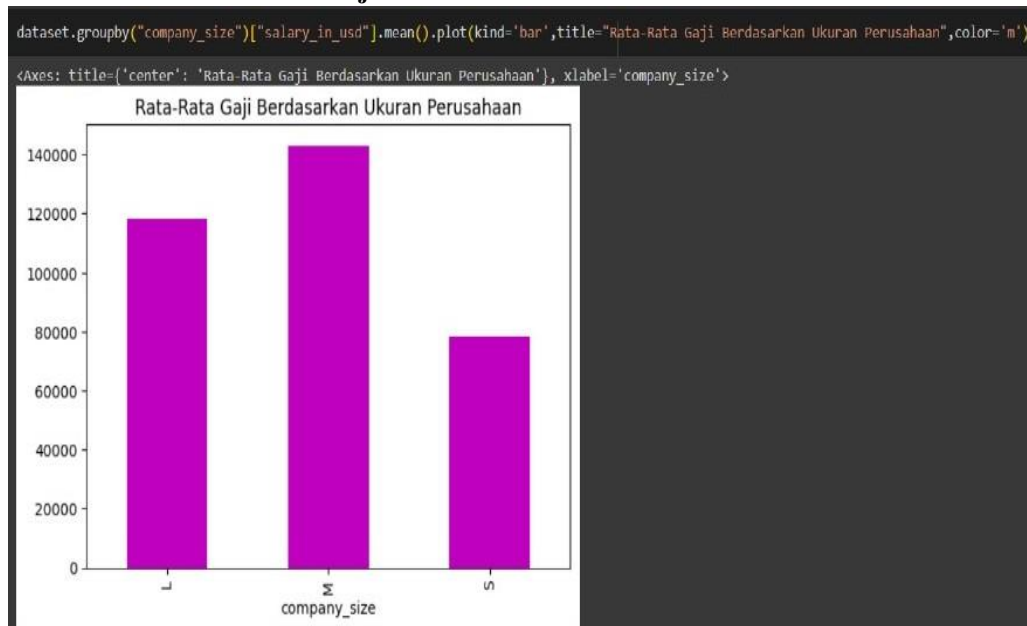
Dari data di atas dapat dilihat bahwa, *Data Science Tech Lead* memiliki rata-rata gaji tertinggi dibanding yang lainnya. Untuk rata-rata jenis lainnya hampir sama dengan range \$200.000- \$250.000.

2.4.4 Perbandingan Jumlah Pekerja Berdasarkan Pengalaman kerja



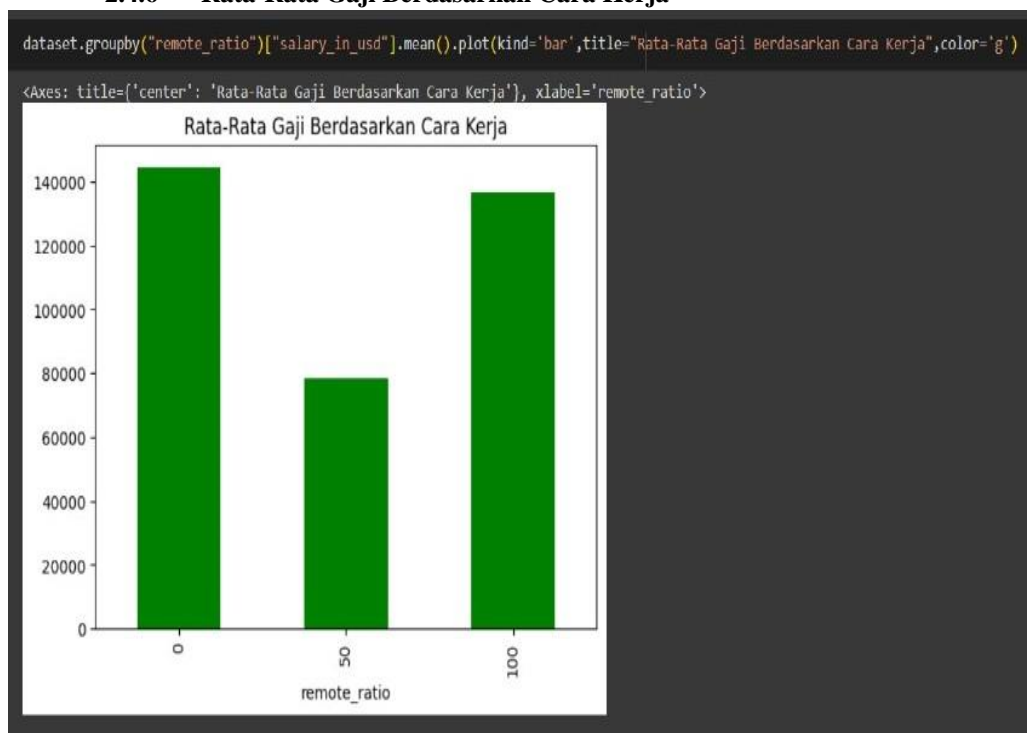
Dari grafik di atas didapati bahwa jumlah pekerja pada level *senior* yang paling besar, dan pada *executive* yang paling sedikit.

2.4.5 Rata-Rata Gaji Berdasarkan Ukuran Perusahaan



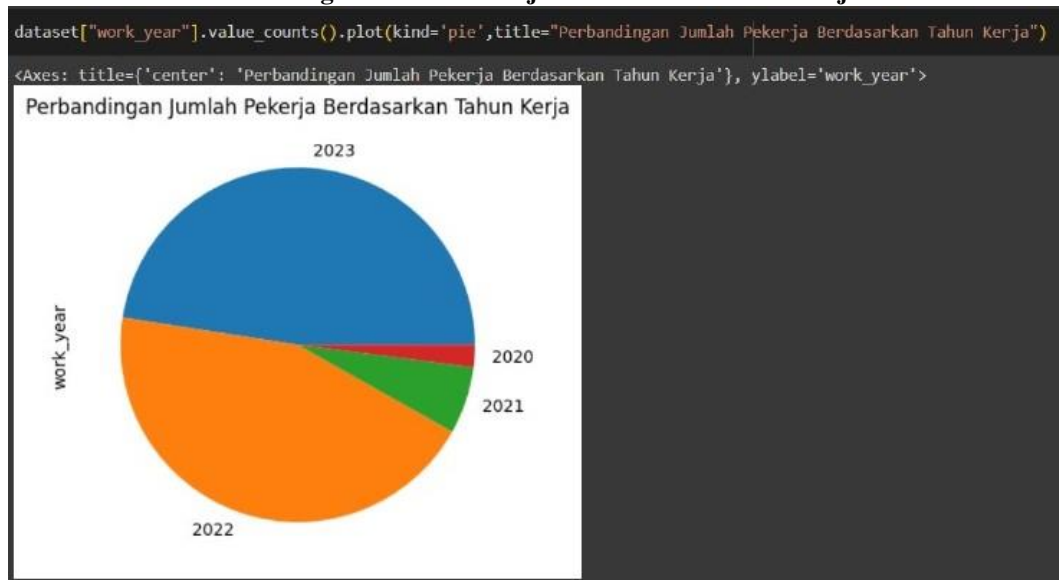
Dari grafik di atas didapatkan bahwa rata - rata gaji tertinggi berada pada perusahaan dengan skala *large* (pegawai > 250). Berdasarkan analisis, semakin besar ukuran perusahaan maka akan semakin meningkat juga rata - rata gajinya, tapi untuk rata-rata gaji terbesar saat ini pada perusahaan *medium*.

2.4.6 Rata-Rata Gaji Berdasarkan Cara Kerja



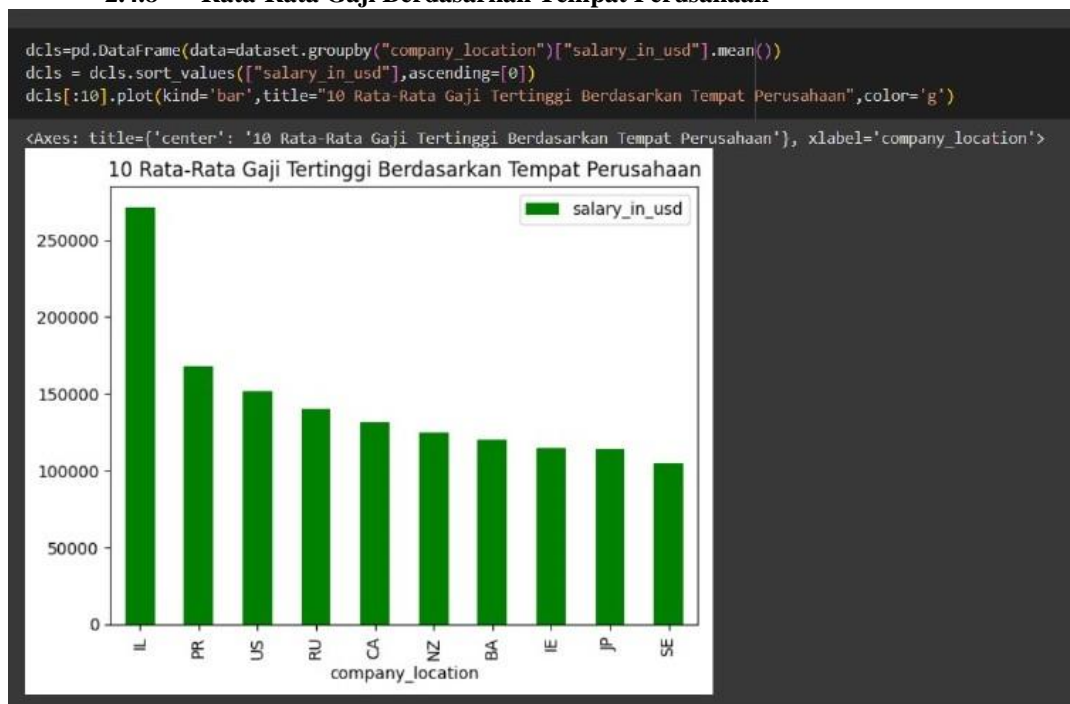
Berdasarkan grafik di atas, didapatkan bahwa pekerjaan yang dilakukan dengan *work from home* (0) memiliki opsi gaji rata - rata tertinggi. Sementara pekerjaan yang dilakukan secara *hybrid* (50) memiliki rata rata gaji yang paling rendah, ini menandakan bahwasanya pekerjaan ini memiliki fleksibilitas yang sangat baik.

2.4.7 Perbandingan Jumlah Pekerja Berdasarkan Tahun Kerja



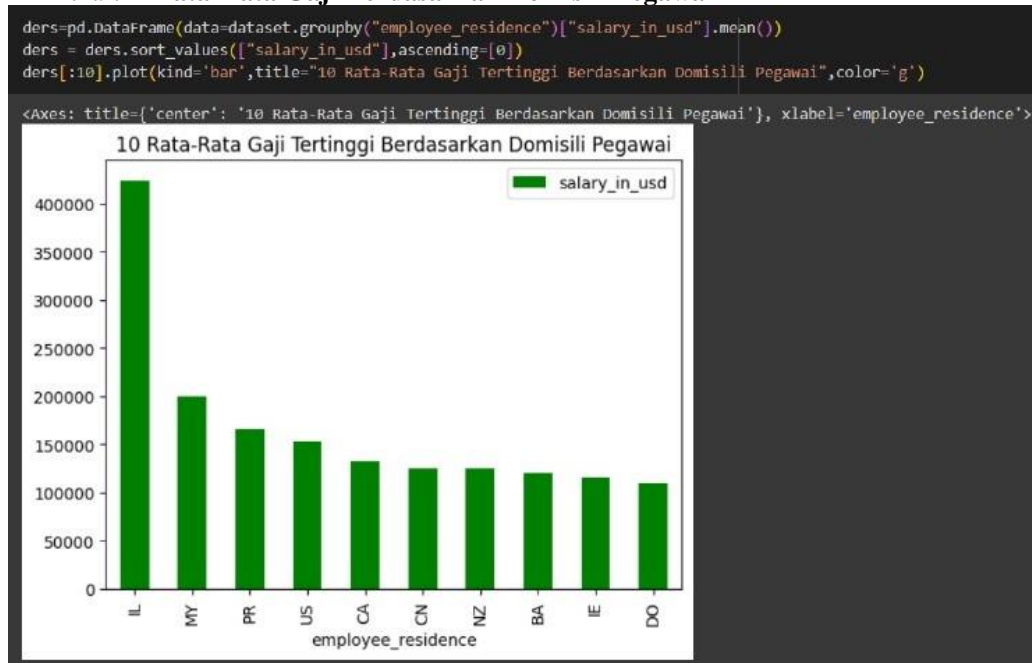
Berdasarkan grafik diatas jumlah pekerja paling banyak pada tahun 2023, ini menandakan semakin banyaknya kebutuhan pada zaman ini, terhadap data *analys*.

2.4.8 Rata-Rata Gaji Berdasarkan Tempat Perusahaan



Berdasarkan grafik diatas, tiga negara dengan perusahaan yang memiliki gaji *data scientist* tertinggi adalah Israel, Poerto Rico dan Amerika.

2.4.9. Rata-Rata Gaji Berdasarkan Domisili Pegawai



Berdasarkan grafik diatas, tiga negara dengan *data scientist* yang memiliki gaji tertinggi adalah Israel, Malaysia, Poerto Rico.

2.5 Korelasi

2.5.1 Korelasi Keseluruhan

Berikut ini data korelasi antar semua atribut kuantitatif yang ada pada tabel:

Korelasi

```
corrdata = dataset[["work_year", "salary", "salary_in_usd"]].corr()
corrdata
```

	work_year	salary	salary_in_usd
work_year	1.0	-0.09472388764106004	0.22629002243291174
salary	-0.09472388764106004	1.0	-0.02367581398124942
salary_in_usd	0.22629002243291174	-0.02367581398124942	1.0

Show 25 per page

Berdasarkan tabel diatas, nilai korelasi antara tahun bekerja dengan pendapatan dalam USD hampir mendekati nol. Artinya tidak terdapat hubungan Antara keduanya sama sekali.

Keterangan: Untuk korelasi yang bernilai 1 tidak perlu dijelaskan karena sesungguhnya memang merupakan elemen yang sama.

2.5.2 Korelasi Antara Jabatan dan Gaji

Dari grafik sebelumnya, Anda dapat melihat bahwa semakin tinggi jabatan Anda, semakin tinggi rata-rata gaji yang Anda terima. Tapi kita bisa mendapatkan wawasan baru. Artinya,

semakin tinggi gelar seseorang, maka semakin tinggi pula gaji maksimum untuk jabatan tersebut, namun tidak seluruhnya. Faktanya, kami menemukan bahwa gaji maksimum untuk posisi tingkat menengah lebih tinggi daripada gaji maksimum untuk posisi senior.

2.5.3 Korelasi Antara Ukuran Perusahaan dan Gaji

Dari grafik sebelumnya terlihat bahwa semakin besar perusahaan tempat Anda bekerja, maka semakin tinggi juga rata-rata gaji yang akan kita terima. Sebaliknya, dari grafik sebelumnya kita dapat menyimpulkan bahwa semakin besar perusahaan tempat kita bekerja, maka semakin tinggi juga gaji yang akan kita terima.

2.5.4 Korelasi Antara Cara Kerja dan Gaji

Berdasarkan grafik 7.1. dapat ditarik kesimpulan bahwa semakin besar remote-ratio seseorang, maka rata-rata gajinya cenderung menaik walau pada *hybrid* tidak.

2.6 Data Cleansing

```
[ ] print(f"Cek ketidakadaan sebuah data di dataset")
print(dataset.isnull().values.any(),":Karena hasilnya 'False' maka tidak ada nilai Nan di data")

Cek ketidakadaan sebuah data di dataset
False :Karena hasilnya 'False' maka tidak ada nilai Nan di dataset
```

2.6.1 Penjabaran Kolom Remote-Ratio

Kolom remote_ratio di data mentah yang kami dapatkan masih berupa angka 0, 50, 100

```
print(f"Karena di columns 'remote_ratio' itu kurang cocok dalam bentuk quantitative")
print(f"Maka sebaiknya di ubah ke categorical nominal yakni:")
print(f"0 = WFO")
print(f"50 = Hybrid")
print(f"100 = WFH")
dataset["remote_ratio"].replace(to_replace=0,value="WFO",inplace=True)
dataset["remote_ratio"].replace(to_replace=50,value="Hybrid",inplace=True)
dataset["remote_ratio"].replace(to_replace=100,value="WFH",inplace=True)
display(dataset)
```

Karena di columns 'remote_ratio' itu kurang cocok dalam bentuk quantitative
Maka sebaiknya di ubah ke categorical nominal yakni:
0 = WFO
50 = Hybrid
100 = WFH

Untuk memudahkan pembacaan data, kami mengonversi data kolom ini menjadi string yang mudah dipahami. Perubahan yang kami lakukan adalah sebagai berikut:

Dilakukan looping untuk setiap elemen kolom remote_ratio, mengganti yang nilai 0 menjadi "WFO", nilai 50 menjadi "Hybrid", dan sisanya menjadi "WFH".

BAB III

PENUTUP

3.1 Kesimpulan

Dari kumpulan data yang diolah menggunakan dataframe dari library pandas dan Matplotlib, kita dapat menyimpulkan bahwa perbedaan gaji data scientist dipengaruhi oleh ukuran perusahaan, jabatan, dan bagaimana cara scientist bekerja. Untuk mendapatkan gaji yang lebih tinggi, Anda dapat mencari perusahaan yang lebih besar, posisi yang lebih tinggi, dan lebih gaya bekerja dari rumah. Working experience juga mempengaruhi tingkat tinggi atau rendahnya data scientist.

3.2 Lesson Learned

Dalam proses pengerjaan tugas besar ke-2 ini, kami mendapatkan pemahaman mengenai penggunaan pandas dan Matplotlib dalam analisis data statistik melalui analisis data gaji *data scientists* di dunia.

LAMPIRAN

Pembagian Tugas dalam Kelompok
Tabel 1.2 Pembagian Tugas Dalam Kelompok

No.	NIM	NAMA LENGKAP	TUGAS
1	19623122	Nadia Apsarini Baizal	Membuat PPT
2	16523066	Alisha Tazkia Anugraha	Membuat laporan
3	16523108	Muhammad Falih Rosyid	Menyusun kode pemograman
4	16523206	Abdul Hakim Yafi	Membuat laporan

4.1 Source Code Program:

<https://colab.research.google.com/drive/1-6KE54C3AxicIXeADcclEr0obHxtq0Y3?usp=sharing>

4.2 Raw Data:

<https://www.kaggle.com/datasets/arnabchaki/data-science-salaries-2023/>