ANALISIS TREN PERKEMBANGAN GAJI DATA SCIENTIST DALAM RENTANG WAKTU 2020-2024 MENGGUNAKAN PYTHON

Aji Sakti Saputra[1], Fadhli Jahfal Aufa Maulana[2], Rafid Farhan Zai[3], Riyan Putra Pratama[4]

Sains Data, Universitas Koperasi Indonesia, Jatinangor, Indonesia;
Sains Data, Universitas Koperasi Indonesia, Jatinangor, Indonesia;
Sains Data, Universitas Koperasi Indonesia, Jatinangor, Indonesia;
Sains Data, Universitas Koperasi Indonesia, Jatinangor, Indonesia;

[1]ajisakti554@gmail.com, [2]fadhlijahfal1@gmail.com, [2]bankbro24@gmail.com
[4]riyanputrapratama78@gmail.com

ABSTRAK

Penelitian ini menganalisis tren gaji Data Scientist global dari 2020-2024, memanfaatkan data dari Kaggle dan alat analisis Python. Tujuannya adalah memvisualisasikan tren gaji berdasarkan tahun kerja, pengalaman, jenis pekerjaan, jabatan, mata uang, lokasi, dan ukuran perusahaan, serta mengidentifikasi pola perubahannya. Metode kuantitatif digunakan, mencakup analisis deskriptif dan inferensial dengan library Python (Pandas, Seaborn). Hasilnya menunjukkan tren peningkatan gaji umum, namun dengan variasi signifikan antar kategori. Penurunan rata-rata gaji pada 2021 teramati, diduga akibat dampak ekonomi pandemi dan dinamika pasar. Faktor makroekonomi, industri, lokasi, dan kebijakan perusahaan berpengaruh terhadap gaji. Penelitian ini mengkonfirmasi efektivitas Python dalam analisis data gaji dan memberikan wawasan penting bagi stakeholder di bidang Data Science, serta menggarisbawahi perlunya penelitian lebih lanjut untuk memahami fluktuasi gaji.

Kata kunci: *Data Scientist, Analisis Data, Python.*

PENDAHULUAN

Perkembangan teknologi informasi dan digitalisasi telah menghasilkan peningkatan volume data secara eksponensial, yang dikenal sebagai big data. Seiring dengan perkembangan tersebut analisa terhadap data menjadi penting untuk menghasilkan sebuah keputusan. Sehingga *Data Scientist* akan menjadi sangat penting, dimana peran *Data Scientist* meliputi 3 (tiga) fase yaitu desain data, mengumpulkan data, dan analisis data (Muhajir & Widodo, 2021). *Data Scientist* akan memainkan peran krusial dalam berbagai sektor industri, termasuk e-commerce, keuangan, dan pemerintahan. Peran ini juga memiliki dampak pada masyarakat secara keseluruhan. Banyak akses membuka pintu inovasi dalam berbagai bidang, mulai dari pelayanan kesehatan hingga sistem transportasi yang efisien (Aliwijaya, A, 2023). Dengan memahami peran *Data Scientist* dalam perkembangan digital, kita dapat melihat betapa pentingnya disiplin ini dalam membentuk masa depan teknologi. Dengan demikian, investasi dalam data science dan teknologi terkait penting untuk masa depan yang lebih baik (Watajdid, N. I., Lathifah, A., Andini, D. S., & Fitroh, F. 2021).

Namun, ditengah tingginya permintaan akan *Data Scientist*, terdapat beberapa tantangan dan permasalahan, terutama terkait dengan aspek gaji. Informasi yang terstruktur dan komprehensif mengenai tren gaji *Data Scientist* di dunia masih terbatas. Meskipun terdapat informasi yang tersebar di berbagai platform, analisis mendalam yang mempertimbangkan berbagai faktor seperti tingkat pengalaman, spesialisasi keahlian, dan lokasi geografis masih diperlukan. Dinamika pasar tenaga kerja yang cepat berubah juga mempengaruhi tren gaji secara signifikan.

Pendekatan konvensional dalam menganalisis data gaji seringkali membutuhkan alokasi waktu dan sumber daya yang signifikan. Sebagai alternatif, studi ini mengadopsi data dengan memanfaatkan Python, sebuah bahasa pemrograman interpretative, berorientasi dan semantik yang dinamis (M Ridhoni, 2017). Data yang digunakan dalam penelitian ini diperoleh dari Kaggle, sebuah platform

penyedia dataset publik yang relevan. Ketersediaan data yang terstruktur di Kaggle memungkinkan penulis untuk melakukan analisis yang lebih mendalam.

Oleh karena itu, penelitian ini bertujuan untuk menganalisis tren perkembangan gaji *Data Scientist* di dunia dalam rentang waktu 2020-2024 menggunakan *Python*. Tujuan spesifik dari penelitian ini adalah memvisualisasikan tren gaji *Data Scientist* secara keseluruhan berdasarkan kategori tertentu seperti tahun kerja, level pengalaman, jenis pekerjaan, jabatan, gaji, mata uang, lokasi, dan ukuran perusahaan; menganalisis perubahan tren gaji dari tahun 2020 hingga 2024 untuk mengidentifikasi pola pertumbuhan atau penurunan; dan mendemonstrasikan penggunaan *Python* sebagai alat analisis data yang efektif.

Hasil penelitian ini diharapkan dapat memberikan manfaat bagi berbagai pihak, antara lain: calon *Data Scientist* yang memberikan pemahaman yang lebih baik mengenai tren gaji dan membantu perencanaan karir; memberikan kontribusi pada studi empiris mengenai pasar kerja *Data Science*; serta memberikan pengenalan *Python* sebagai alat analisis data.

TINJAUAN PUSTAKA

Data Science adalah cabang statistik yang berfokus pada penggunaan teknologi komputer dan pemahaman statistik untuk mengekstraksi pengetahuan dari data (William S. Cleveland). Pengetahuan ini tidak hanya berfokus pada pengolahan data, tetapi juga pada pemahaman mendalam terhadap data tersebut dan aplikasinya dalam pengambilan keputusan. *Data Science* melibatkan pemahaman tentang data, cara data diproses, dianalisis, serta diinterpretasikan untuk mendapatkan informasi yang bermakna. Peran *Data Scientist* sangat penting dalam analisis tersebut. Seorang *Data Scientist* berperan dalam perencanaan hingga interpretasi hasil analisis.

Peran *Data Science* dalam era digital saat ini tidak dapat diabaikan. Perusahaan sangat bergantung pada *Data Science* untuk mengambil keputusan, mengembangkan produk dan layanan, meningkat kan efisiensi operasional, dan memahami pola pelanggan lebih baik. Ketersediaan data yang banyak dan kemampuan analisis yang canggih memungkinkan perusahaan untuk mendapatkan keunggulan dari pesaing. Oleh karena itu, investasi dalam *Data Science* menjadi sangat penting bagi keberlanj-utan dan pertumbuhan organisasi.

2.1. Konsep dan Teori Gaji dalam Lingkungan Kerja

Gaji merupakan pembayaran serta balas jasa yang diberikan kepada karyawan, tatausaha dan mana -jer sebagai konsekuensi dari sumbangan yang diberikannya dalam pencapaian tujuan perusahaan (As'ad, 2005:136). Faktor yang biasanya menjadi acuan dari gaji meliputi tingkat pengalaman, spesialisasi, keahlian, dan performa kerja. Sementara itu, Faktor pasar tenaga kerja meliputi lokasi geografis, ukuran dan jenis perusahaan, serta dinamika permintaan dan penawaran tenaga kerja. Oleh karena itu, pemahaman terhadap kedua jenis factor ini penting dalam menganalisis tren gaji.

Teori permintaan dan penawaran tenaga kerja menjelaskan bahwa tingkat gaji ditentukan oleh interaksi antara permintaan dan penawaran tenaga kerja. Ketika permintaan terhadap *Data Scientist* tinggi sementara penawarannya terbatas, gaji akan cenderung naik, dan sebaliknya. Keseimbangan antara permintaan dan penawaran tenaga kerja akan menciptakan tingkat gaji yang wajar. Oleh karena itu, fluktuasi permintaan dan penawaran tenaga kerja dapat mempengaruhi tren gaji secara signifikan.

2.2. Peran Python dalam Analisis Data

Python merupakan sebuah Bahasa pemrograman tingkat tinggi yang dibuat oleh Guido Van Rossum dan dirilis pada tahun 1991. Python juga merupakan Bahasa yang sangat populer belakangan ini. Selain itu, Python juga merupakan Bahasa pemrograman yang multi fungsi salah satunya pada bidang Big Data (Karimah Tauhid, Volume 2 Nomor 1 (2023)). Kepopuleran Python dalam analisis data didukung oleh ketersediaan berbagai pustaka (library) yang kuat, seperti Pandas untuk manipulasi data, NumPy untuk komputasi numerik, Seaborn untuk visualisasi data.

Kemudahan penggunaan dan fleksibilitas Python menjadikannya pilihan ideal untuk analisis data. Sintaks Python yang mudah dibaca dan dipahami membuatnya mudah dipelajari dan digunakan. Python menjadi alat yang sangat penting bagi *Data Scientist* dalam melakukan analisis data. Implementasi Python dalam analisis gaji melibatkan library seperti Pandas untuk mengolah data gaji dari Kaggle, dan Seaborn untuk memvisualisasikan tren gaji. Dengan demikian, Python menjadi alat yang sangat efektif untuk seorang *Data Scientist* melakukan analisis data.

2.3. Perumusan Hipotesis Penelitian

Hipotesis pertama (H1) adalah bahwa terdapat tren peningkatan gaji *Data Scientist* secara keseluruhan dari tahun 2020 hingga 2024, seiring dengan meningkatnya permintaan di pasar tenaga kerja. Hipotesis kedua (H2) adalah bahwa terdapat perbedaan signifikan dalam tren gaji *Data Scientist* berdasarkan kategori seperti tahun kerja, level pengalaman, jenis pekerjaan, jabatan, gaji, mata uang, lokasi, dan ukuran perusahaan. Kategori ini dipilih karena dianggap memiliki pengaruh signifikan pada variasi gaji *Data Scientist*. Hipotesis ketiga (H3) adalah bahwa penggunaan Python sebagai alat analisis data efektif dalam memvisualisasikan dan menganalisis tren gaji *Data Scientist*, dengan memanfaatkan data yang terstruktur dari Kaggle.

METODE PENELITIAN

Penelitian ini menggunakan metode kuantitatif menggunakan metode Analisis Deskriptif dan Analisis Inferensial. Analisis deskriptif dilakukan untuk menggambarkan karakteristik data gaji *Data Scientist*. Statistik deskriptif seperti Mean, Median, dan Standar Deviasi akan dihitung untuk data gaji secara keseluruhan pada tahun 2020 hingga 2024. Analisis inferensial dilakukan untuk menguji hipotesis penelitian, hipotesis pertama yaitu apakah terdapat tren peningkatan gaji yang signifikan. Hipotesis kedua yaitu apakah terdapat perbedaan signifikan dalam tren gaji berdasarkan kategori. Visualisasi data akan dibuat untuk menyajikan hasil analisis secara normative dan mudah dipahami. Seluruh analisis data dilakukan menggunakan Python dengan library seperti Pandas dan Seaborn.

Data gaji *Data Scientist* diperoleh dari Kaggle yang merupakan platform penyedia dataset publik yang relevan. Dataset yang digunakan adalah dataset yang telah dikumpulkan oleh pengguna Kaggle dan mencakup informasi yang dibutuhkan untuk penelitian ini. Pemilihan Kaggle didasarkan pada ketersediaan data terstruktur serta akses yang mudah dan gratis. Data yang digunakan adalah data sekunder, sehingga kualitas dan kelengkapan data bergantung pada sumber data tersebut. Data dari Kaggle akan diunduh dalam format yang kompatibel. Selanjutnya, data akan dibersihkan dan diproses untuk menghilangkan data yang tidak relevan. Data yang telah diproses kemudian akan diubah kedalam format yang siap dianalisis.

Penelitian ini bertujuan untuk menganalisis tren perkembangan gaji Data Scientist di dunia dalam rentang waktu 2020 hingga 2024. Penelitian ini didasari oleh meningkatnya peran *Data Scientist* dalam era big data dan digitalisasi, serta adanya celah informasi yang signifikan terkait tren gaji di bidang ini. Penelitian ini diharapkan dapat memberikan pemahaman yang lebih baik mengenai tren gaji, berkontribusi pada studi empiris tentang lingkungan kerja *Data Science*, dan mendemonstrasi -kan penggunaan *Python* sebagai alat analisis yang relevan.

HASIL PENELITIAN DAN PEMBAHASAN

Hasil analisis data penelitian ini menunjukkan bahwa kami dapat menemukan berbagai elemen yang berkaitan dengan kompensasi para ilmuwan data di seluruh dunia. Aspek-aspek tersebut termasuk rata-rata gaji, kisaran gaji, dan tren perkembangan gaji. Mereka juga mencakup indikator statistik penting lainnya, seperti standar deviasi dan persentil, yang memberikan gambaran lengkap tentang distribusi data. Data yang digunakan untuk melakukan analisis ini mencakup posisi *Data Scientist* dengan berbagai tingkat pengalaman dari tahun 2020 hingga 2024. Temuan ini memberikan gambaran mendalam tentang perubahan pasar tenaga kerja bagi pekerjaan data scientist selama

rentang waktu tersebut. Hasil ini dapat digunakan sebagai referensi untuk studi lebih lanjut atau pe -mbuatan kebijakan di bidang *Data Science*. Berikut adalah hasil penelitian dan pembahasan kami:

1. Meng import Library yang Akan digunakan

```python
#menggunakan Library Pandas dan Seaborn
import pandas as pd
import seaborn as sns
```

Dalam penelitian ini, kami menggunakan library *Python* Pandas dan Seaborn untuk analisis data dan visualisasi. Berikut penjelasannya:

a. Pandas adalah library open-source yang menawarkan struktur data seperti DataFrame yang memungkinkan manipulasi dan analisis data secara efisien, seperti pembersihan, transformasi, dan perhitungan statistik.

b. Seaborn adalah library visualisasi berbasis statistik yang memudahkan membuat grafik informatif seperti boxplot, scatterplot, dan heatmap. Kombinasi kedua library ini menduk -ung analisis mendalam dan penyajian hasil yang lebih mudah dipahami.

2. Menginput Data/File Kedalam Program

```python
df = pd.read_csv('https://docs.google.com/spreadsheets/d/e/2PACX-[illegible]')
```

Dengan library Pandas, sintaks df = pd.read_csv() digunakan untuk membaca file dalam format Comma-Separated Values (CSV) dan mengonversinya menjadi DataFrame. Salah satu format data berbasis teks adalah format CSV, yang menyimpan data dalam bentuk tabel dengan satu entri per baris dan tanda koma (`,`) yang memisahkan kolom.

Data yang digunakan dalam penelitian ini berasal dari platform Kaggle, yang biasanya menyediakan dataset dalam format CSV atau Excel. Meskipun data awalnya dalam format Excel, format CSV dipilih karena kompatibilitasnya yang tinggi, ukurannya yang lebih kecil, dan kemudahan membaca dan memprosesnya langsung menggunakan library Pandas dengan menggunakan fungsi read_csv(). Untuk mendukung analisis mendalam, ini memungkinkan manipulasi data yang lebih efisien.

3. Mengecek hasil dari input (apakah berhasil/tidak)

```python
df.head()
```

Fungsi bawaan Pandas df.head() digunakan untuk secara otomatis menampilkan lima baris pertama dari sebuah DataFrame. Fungsi ini sangat bermanfaat selama tahap awal analisis data, terutama untuk memastikan bahwa file data telah dimuat dan ditampilkan dengan benar.

Dengan menggunakan df.head(), kami dapat memastikan struktur dataset, termasuk nama kolom, tipe data, dan beberapa baris awal yang menunjukkan isi dataset. Untuk mencegah kesalahan lebih lanjut selama proses analisis, langkah ini sangat penting. Dengan menambahkan parameter, seperti df.head(10) untuk menampilkan sepuluh baris pertama, jumlah baris yang ditampilkan dapat disesuaikan jika diperlukan. Hasilnya adalah sebagai berikut:

index	work_year	experience_level	employment_type	job_title	salary	salary_currency	salary_in_usd	employee_residence	remote_ratio	company_location	company_siz
0	2021	Mid	Full Time	Data Scientist	30400000	CLP	40038	CL	100	CL	L
1	2021	Mid	Full Time	BI Data Analyst	11000000	HUF	36259	HU	50	US	L
2	2020	Mid	Full Time	Data Scientist	11000000	HUF	35735	HU	50	HU	L
3	2021	Mid	Full Time	ML Engineer	8500000	JPY	77364	JP	50	JP	S
4	2022	Senior	Full Time	Lead Machine Learning Engineer	7500000	INR	95386	IN	50	IN	L

Berikut ini adalah penjelasan tentang tiap kolom yang memiliki korelasi dengan baris atau nilainya:

- **Index:** Merupakan nomor urut yang diberikan secara otomatis untuk membedakan setiap baris dalam dataset. Hanya digunakan untuk pelacakan baris dan tidak memiliki nilai analitik.
- **Work_year:** Tahun di mana pekerjaan dilakukan, digunakan untuk menganalisis tren atau perkembangan data secara waktu.
- **Experience_level:** Tingkat pengalaman karyawan, seperti tingkat menengah atau senior, yang berkaitan dengan bagaimana gaji atau peran berubah sesuai dengan tingkat pengalaman.
- **Employement_type:** Jenis pekerjaan yang dipegang, seperti pekerjaan penuh waktu Jenis pekerjaan memengaruhi gaji dan faktor lainnya, dan informasi ini dapat membantu.
- **Job_title:** Nama pekerjaan, seperti Data Scientist atau BI Data Analyst. Ini menghubungkan pengalaman dan tanggung jawab kerja.
- **Salary:** Analisis gaji dan perbandingan pekerjaan berdasarkan gaji dalam mata uang lokal salary_money
- **Salary_currency:** Mata uang seperti CLP (Peso Chili), HUF (Forint Hongaria), JPY (Yen Jepang), dll. digunakan untuk mencatat gaji. Sebelum melakukan konversi, kolon ini sangat penting untuk menginterpretasikan data sesuai dengan konteks lokal.
- **Salary_in_usd:** Gaji yang telah dikonversi ke dalam mata uang USD (United States Dollar). Mempermudah perbandingan gaji secara global tanpa terpengaruh oleh nilai tukar mata uang.
- **Employee_residence:** Lokasi tempat tinggal karyawan diwakili oleh kode negara, misalnya CL (Chili), HU (Hongaria), dan JP (Jepang). Kolom ini membantu memahami pola pekerja di wilayah tersebut.
- **Remote_ratio:** Persentase pekerjaan yang dilakukan secara jarak jauh, juga dikenal sebagai remote, seperti seratus persen (sepenuhnya remote) atau lima puluh persen (hybrid). Pengaruh kolo ini terhadap gaji dan pekerjaan dapat dianalisis.
- **Company_location:** Kode negara menunjukkan lokasi perusahaan tempat pekerjaan dilakukan. Sangat berguna untuk menganalisis perbedaan gaji berdasarkan lokasi bisnis.
- **Company_size:** Ukuran perusahaan berdasarkan skala, seperti L, yang berarti besar, dan S, yang berarti kecil. Ukuran perusahaan seringkali berkorelasi dengan gaji dan jenis pekerjaan.

4. Merubah Data Kolom/Baris Agar Lebih Mudah di Analisis

```
df.rename(columns={ #Mengubah nama data di dalam kolom, dengan tujuan agar mudah dibaca dan dianalisis
    'Unnamed: 0':'Index',
    'work_year':'Tahun_kerja',
    'experience_level':'Level_pengalaman',
    'employment_type':'Jenis_pekerjaan',
    'job_title':'Jabatan',
    'salary':'Gaji',
    'salary_currency':'Mata_uang',
    'salary_in_usd':'Gaji_dalam_usd',
    'employee_residence':'Residence_karyawan',
    'remote_ratio':'Remoteness_ratio',
    'company_location':'Lokasi_perusahaan',
    'company_size':'Ukuran_perusahaan'
}, inplace=True)
```

Index	Tahun_kerja	Level_pengalaman	Jenis_pekerjaan	Jabatan	Gaji	Mata_uang	Gaji_dalam_usd	Residence_karyawan	Remoteness_ratio	Lokasi_perusahaan	Ukuran_perusahaan
0	2021	Mid	Full Time	Data Scientist	30400000	CLP	40038	CL	100	CL	L
1	2021	Mid	Full Time	BI Data Analyst	11000000	HUF	36259	HU	50	US	L
2	2020	Mid	Full Time	Data Scientist	11000000	HUF	35735	HU	50	HU	L
3	2021	Mid	Full Time	ML Engineer	8000000	JPY	77364	JP	50	JP	S
4	2022	Senior	Full Time	Lead Machine Learning Engineer	7500000	INR	99360	IN	50	IN	L

Kami melakukan renaming (penggantian nama) pada kolom dan nilai di setiap baris dataset selama proses analisis data, yang bertujuan untuk mempermudah interpretasi data, meningkatkan keterbac -aan, dan mengurangi kesalahan yang mungkin terjadi. Untuk memenuhi kebutuhan penelitian,

istilah yang lebih deskriptif digunakan untuk menggantikan nama kolom asli yang biasanya mengg -unakan singkatan atau kode.

Kami mengganti nama kolom dan mengganti nilai pada beberapa baris dataset. Tujuan dari langkah ini adalah untuk menyederhanakan data, menjamin konsistensi, dan meningkatkan efisiensi analisis. Misalnya, nilai-nilai di kolom pengalaman_level seperti Mid diubah menjadi Menengah, dan nilai Full Time di kolom employment_type diubah menjadi Penuh Waktu. Untuk mencegah data menjadi ambigu atau tidak jelas, proses penggantian ini sangat penting, terutama ketika kumpulan data menggunakan istilah teknis atau singkatan yang tidak biasa. Analisis menjadi lebih mudah dipahami dengan data yang lebih seragam dan mudah dibaca, dan hasilnya dapat lebih mudah diinterpretasikan oleh berbagai pihak. Hasilnya adalah sebagai berikut:

```
[74] df['level_pengalaman'].replace({ #Mengubah nama data di dalam baris, dengan tujuan agar lebih mudah di analisis
        'SE':'Senior',
        'MI':'Mid',
        'EN':'Entry',
        'EX':'Executive'
    }, inplace=True)
```

```
[75] df['Jenis_pekerjaan'].replace({ #Mengubah nama data di dalam baris, dengan tujuan agar lebih mudah di analisis
        'FT':'Full Time',
        'PT':'Part Time',
        'CT':'Contract',
        'FT':'Freelance'
    }, inplace=True)
```

Mengecek kembali apakah sudah berhasil/tidak. Berikut hasilnya:

	Tahun_kerja	Level_pengalaman	Jenis_pekerjaan	Jabatan	Gaji	Mata_uang	Gaji_dalam_usd	Residence_karyawan	Rasetrasi_ratio	Lokasi_perusahaan	Ukuran_perusahaan
0	2021	Mid	Full Time	Data Scientist	30400000	CLP	40038	CL	100	CL	L
1	2021	Mid	Full Time	BI Data Analyst	11000000	HUF	36259	HU	50	US	L
2	2020	Mid	Full Time	Data Scientist	11000000	HUF	35735	HU	50	HU	L
3	2021	Mid	Full Time	ML Engineer	8000000	JPY	77364	JP	50	JP	S
4	2022	Senior	Full Time	Lead Machine Learning Engineer	7500000	INR	85300	IN	50	IN	L

5. Memfilter Data dengan Jabatan Khusus untuk *Data Scientist*, Sesuai dengan Tujuan Analisis Ini

```
data_scientist= df.loc[ (df['Jabatan']=='Data Scientist'),
```

- Filter Data

```
#MEMFILTER DATA DENGAN JABATAN KHUSUS UNTUK DATA SCIENTIST, SESUAI DENGAN TUJUAN ANALISIS INI

data_scientist= df.loc[ (df['Jabatan']=='Data Scientist'),#SINTAKS (loc [])DIGUNAKAN UNTUK MENGAKSES BARIS DAN KOLOM BERDASARKAN LABEL(NAMABARIS/NAMAKOLOM)
 ['Tahun_kerja',                                         #DENGAN KRITERIA TERTENTU(SESUAI TUJUAN)
  'Jabatan',
  'Jenis_pekerjaan',
  'Level_pengalaman',
  'Gaji_dalam_usd',
  'Mata_uang',
  'Lokasi_perusahaan',
  #MEMFILTER  KOLOM/BARIS YG HANYA DIPERLUKAN SAJA UNTUK ANALISIS
].sort_values(by='Gaji_dalam_usd', ascending=False)#SINTAKS (SORT_VALUES()) PERUNTUKAN MENG SORTIR DATA /MENGURUTKAN DATA DARI YANG TERTINGGI MENUJU
#KE PALING KECIL, KARENA (ASCENDING=FALSE) ARTINYA DIURUTKAN DARI YG TERBESAR KE YG TERKECIL
#JIKA (ASCENDING=TRUE)MAKA DIURUTKAN DARI NILAI YG TERKECIL KE YG TERBESAR
data_scientist #PRINT
```

- Syntax ini menggunakan filtering pada DataFrame (df) untuk mendapatkan data khusus dari jabatan "Data Scientist".
- Metode indeksasi berbasis label (.loc[]) mengakses data berdasarkan label atau nama kolo m Pemilihan Kolom.

```
['Tahun_kerja',
 'Jabatan',
 'Jenis_pekerjaan',
 'Level_pengalaman',
 'Gaji_dalam_usd',
 'Mata_uang',
 'Lokasi_perusahaan',
 #MEMFILTER  KOLOOM/BARIS YG HANYA DIPERLUKAN SAJA
]].sort_values(by='Gaji_dalam_usd', ascending=False)
```

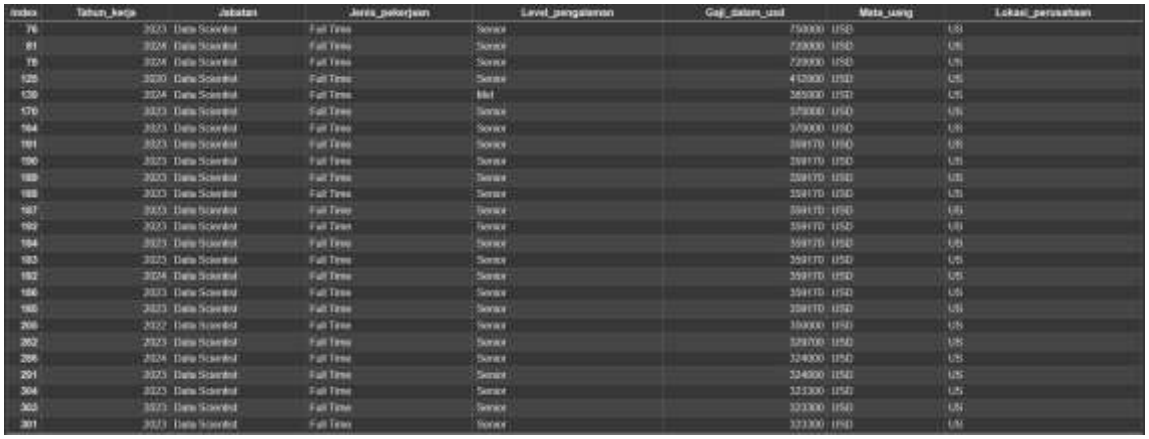
- Ini adalah daftar kolom yang telah dipilih untuk diambil dari dataset. Hanya kolom-kolom yang disebutkan ini yang akan muncul dalam hasil filtering, untuk membantu analisis berkonsentrasi pada variabel yang relevan saja.
- Pengurutan Data

```
.sort_values(by='Gaji_dalam_usd', ascending=False)#
```

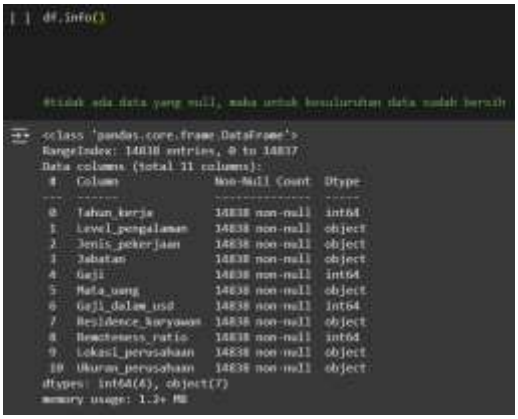
  - Menggunakan syntax (**.sort_values()**) agar bisa mengurutkan data.
  - Parameter **(by='Gaji_dalam_usd'**) menentukan kolom yang dijadikan dasar pen -gurutan.
  - (**ascending=False**) menjadikan pengurutan dimulai dari nilai tertinggi ke nilai terendah (descending).
  - Jika (**ascending=True**), maka urutan akan dari nilai terendah ke tertinggi.

```
data_scientist #PRINT
```

Hasil Outputnya sebagai berikut:



6. Mengecek Apakah Ada dari Dataset yang Berada dalam Keaadan Kotor



a. df.isnull().sum()

Fungsi ini memberikan penjelasan singkat tentang DataFrame, seperti:Membantu memahami struktur dataset.

- Membantu memaham sturktur dataset.
- Membantu menemukan masalah kualitas data.
- Membantu menemukan nilai yg null dan memastikan tipe data setiap kolom sesuai.
- Memberikan gambaran umum tentang ukuran dan fitur dataset.

b. df.info()

- Jumlah total baris (entries).
- Range index.
- Nama di setiap kolom.
- Total nilai non-null di setiap kolom.
- Tipe data (dtype) masing-masing kolom.
- Penggunaan memori.

7. Menganalisis Data Gaji sebagai *Data Scientist* dengan Jenis Pekerjaan Full Time dan Leve l Pengalaman Kerja EXECUTIVE

a. Penjelasan syntax

- **data_scientist.loc :** Penulisan **(.loc)** digunakan untuk mengakses baris atau kolomberdasarkan label atau kondisi tertentu.
- **Kondisi Filtering:**
  - **(data_scientist['Jenis_pekerjaan'] == 'Full Time')** Memfilter baris dengan nilai kolom '**Jenis_pekerjaan' yang sama dengan 'Full Time'.**
  - **(data_scientist['Level_pengalaman'] == 'Executive')** Memfilter baris dengan nilai kolom '**Level_pengalaman**' yang sama dengan '**Executive'.**

Kedua kondisi tersebut dihubungkan oleh operator & (logika "AND"), sehingga hanya baris yang memenuhi kedua kriteria ini yang akan dipilih.

- Daftar kolom yang ingin diambil dari dataset:
  - **'Tahun_kerja'**: Tahun pekerjaan dilakukan.
  - **'Jabatan'**: Jabatan atau posisi pekerjaan.
  - **'Level_pengalaman'**: Tingkat pengalaman kerja.
  - **'Gaji_dalam_usd'**: Besaran gaji dalam USD.
  - **'Mata_uang':** Mata uang asli sebelum konversi ke USD.
  - **'Lokasi_perusahaan':** Lokasi geografis perusahaan.
- **.sort_values(by='Gaji_dalam_usd'):** Mengurutkan data berdasarkan kolom 'Gaji_dalam_usd'.
- **ascending=False:** Mengurutkan data dari nilai terbesar ke yang terkecil (urutan menurun).

Hal ini dilakukan untuk memprioritaskan tampilan data dengan gaji tertinggi terlebih dahulu. Lalu print variabelnya yg memiliki nama **(Executive)**, dan hasil outputnya sebagai berikut:

	Tahun_kerja	Jabatan	Level_pengalaman	Gaji_dalam_usd	Mata_uang	Lokasi_perusahaan
464	2023	Data Scientist	Executive	300000	USD	US
474	2023	Data Scientist	Executive	299500	USD	US
942	2023	Data Scientist	Executive	258750	USD	US
943	2023	Data Scientist	Executive	258750	USD	US
1150	2023	Data Scientist	Executive	250000	USD	US
1191	2023	Data Scientist	Executive	249600	USD	US
1296	2024	Data Scientist	Executive	246100	USD	US
1770	2023	Data Scientist	Executive	228856	USD	GB
1769	2023	Data Scientist	Executive	228856	USD	GB
1898	2023	Data Scientist	Executive	225000	USD	US
2122	2023	Data Scientist	Executive	220000	USD	US
2146	2023	Data Scientist	Executive	220000	USD	US
3109	2023	Data Scientist	Executive	200000	USD	US
3092	2023	Data Scientist	Executive	200000	USD	US
3086	2023	Data Scientist	Executive	200000	USD	US
3966	2023	Data Scientist	Executive	185000	USD	US
3952	2023	Data Scientist	Executive	185000	USD	US

```
correlation = df['Gaji_dalam_usd'].corr(df['Tahun_kerja'])
print(correlation)
```

```
0.09087347570331308
```

b. Menghitung Hubungan Linier Antara Gaji dalam USD (Gaji_dalam_usd) dan Tahun Kerja (Tahun_kerja).

- **df['Gaji_dalam_usd']:** Mengambil kolom yang berisi data gaji dalam dolar AS dari DataFrame df.
- **df['Tahun_kerja']:** Mengambil kolom yang berisi data tahun kerja dari DataFrame df.
- **.corr():** Metode ini digunakan untuk menghitung korelasi antara dua variabel.
- Korelasi ini mengukur hubungan linier antara kedua variabel: Nilai korelasi berkisar antara -1 dan 1:
  - 1 : Hubungan positif sempurna (ketika satu variabel naik, variabel lainnya juga naik).
  - -1 : Hubungan negatif sempurna (ketika satu variabel naik, variabel lainnya turun).
  - 0 : Tidak ada hubungan linier antara kedua variabel.

Hubungan antara kedua variabel sangat kuat, karena nilai 0,9 hampir sama dengan 1. Artinya, tahun kerja memiliki pengaruh yang signifikan terhadap gaji dalam dolar di dataset ini.

c. Menghitung rata-rata gaji dari tahun 2020-2024 sebagai data scientist dengan jenis pekerjaan full time dan level pengalaman kerja executive

- **Executive.groupby('Tahun_kerja'):** Metode **groupby()** digunakan untuk mengelompok kan data berdasarkan kolom tertentu, dalam hal ini adalah kolom 'Tahun_kerja'. Data akan dikelompokkan ke dalam kategori unik dari kolom 'Tahun_kerja'.
- **['Gaji_dalam_usd']:** Setelah data dikelompokkan berdasarkan 'Tahun_kerja', kolom 'Gaji_dalam_usd' dipilih untuk analisis lebih lanjut.
- **.mean():** Fungsi mean() menghitung rata-rata dari nilai di kolom 'Gaji_dalam_usd' untuk setiap kelompok (setiap nilai unik di 'Tahun_kerja').

Sehingga hasil outputnya:

Tahun_kerja	Gaji_dalam_usd
2022	144500.000000
2023	188562.344828
2024	165500.000000

dtype: float64

Memvisualisasikan table diatas dengan menggunakan library Seaborn:

```
[ ] sns.barplot(
        data=Executive,
        x='Tahun_kerja',
        y='Gaji_dalam_usd',
        hue='level_pengalaman',
       ci=None

    ).set(title=' GAJI DATA SCIENTIST DENGAN JENIS PEKERJAAN FULL TIME (LEVEL PENGLAMAN KERJA EXECUTIVE)')#judul
    #VISUALISASI DATA DENGAN MENGGUNAKAN LIBRARY BARPLOT DENGAN DATA EXECUTIVE
```

Penjelasan syntax:

- **Sns.barplot() :** Menggunakan library Seaborn untuk membuat plot..
- **Data :** Dataset yang digunakan bernama '**Executive'.**
- **X :** Variabel tahun kerja ditunjukkan dalam sumbu X,
- **y :** variabel gaji ditunjukkan dalam sumbu Y.
- **hue :** mengurutkan bar berdasarkan tingkat pengalaman.
- **ci=none :** Interval keyakinan (error bars) tidak ditunjukkan.
- **.set(title='GAJI DATA SCIENTIST DENGAN JENIS PEKERJAAN FULL TIME (LEVEL PENGALAMAN KERJA EXECUTIVE)' :** Menambahkan judul grafik.

Sehingga mendapatkan output sebagai berikut:

Berdasarkan hasil visualisasi data gaji Data Scientist dengan tingkat pengalaman kerja eksekutif dari tahun 2022 hingga 2024, terlihat tren peningkatan yang signifikan. Analisis menunjukkan bahwa rata rata gaji pada tahun 2022 berada pada kisaran 145,000 USD, kemudian meningkat secara signi fikan pada tahun 2023, mencapai sekitar 185,000 USD, dan tren ini berlanjut hingga tahun 2024, ketika rata rata gaji mencapai kisaran 170,000 USD.Terjadi peningkatan valuasi yang signifikan te rhadap posisi Data Scientist level Executive,terutama pada periode 2022 2023 dengan pertumbuha n sekitar 27.6%. Fenomena ini mengindikasikan beberapa implikasi penting:

1. Fluktuasi gaji yang terlihat dapat menunjukkan dinamika pasar tenaga kerja dalam industri data science, khususnya untuk posisi tingkat eksekutif.
2. Data menunjukkan bahwa pekerjaan ilmuwan data di tingkat eksekutif masih menerima kompensasi yang kompetitif, mengingat tingginya permintaan untuk keahlian analitis dan kepemimpinan dalam industri *Data Science*.

Hasil visualisasi ini memberikan gambaran menyeluruh tentang tren kompensasi untuk posisi eksekutif dalam bidang *Data Science*. Ini dapat menjadi referensi yang bermanfaat untuk studi tentang perkembangan pasar tenaga kerja di industri TI.

8. Menganalisis Gaji Sebagai Data Scientist dengan Jenis Pekerjaan Full time dan Level Pengalaman Kerja SENIOR

```
#DATA GAJI DI DATA SCIENTIST DENGAN JENIS PEKERJAAN FULL TIME (LEVEL PENGLAMAN KERJA SENIOR)

Senior=data_scientist.loc[(data_scientist['Jenis_pekerjaan'] == 'Full Time') &(data_scientist['Level_pengalaman']== 'Senior'),
 ['Tahun_kerja',
  'Jabatan',
  'Level_pengalaman',
  'Gaji_dalam_usd',
  'Mata_uang',
  'Lokasi_perusahaan',]
].sort_values(by='Gaji_dalam_usd', ascending=False)
Senior
#MEMFILTER DENGAN SYARAT JENIS PEKERJAAN=FULLTIME, LEVEL PENGALAMAN=senior
```

Syntax ini sama dengan yang ditunjukkan pada poin 7, tetapi hanya variable data yang disebut senior dan syntax (**data_scientist[‘Level_pengalaman]=’senior).** Karena kami membutuhkan tingkat pengalaman senior pada saat ini. Jadi, hasilnya adalah sebagai berikut:

	Tahun_kerja	Jabatan	Level_pengalaman	Gaji_dalam_usd	Mata_uang	Lokasi_perusahaan
76	2023	Data Scientist	Senior	750000	USD	US
78	2024	Data Scientist	Senior	720000	USD	US
81	2024	Data Scientist	Senior	720000	USD	US
125	2020	Data Scientist	Senior	412000	USD	US
170	2023	Data Scientist	Senior	370000	USD	US
...	...	...	...	...	...	...
14645	2022	Data Scientist	Senior	37824	EUR	ES
14642	2022	Data Scientist	Senior	37824	EUR	ES
14641	2022	Data Scientist	Senior	37824	EUR	ES
14775	2024	Data Scientist	Senior	31250	GBP	GB
4367	2021	Data Scientist	Senior	20171	TRY	TR

a. Menghitung rata- rata
Input :

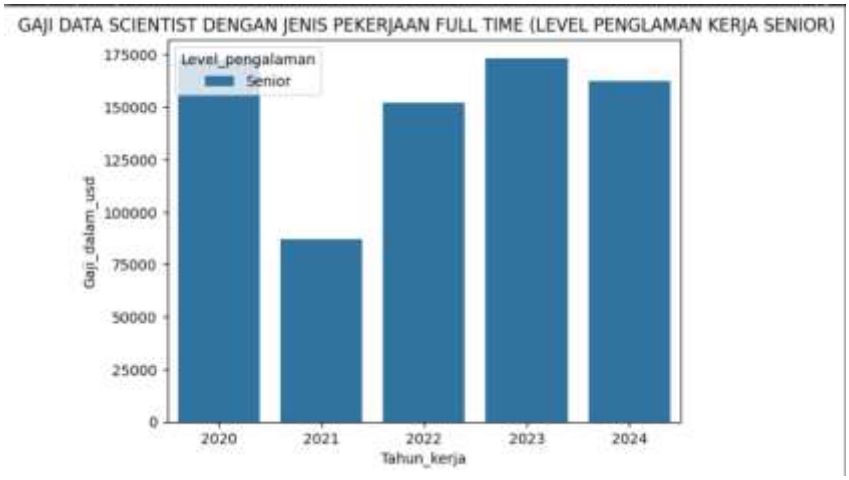
```
#Menghitung Rata-rata dari Level pengalaman kerja

Senior.groupby('Tahun_kerja')['Gaji_dalam_usd'].mean()
```

Output:

Tahun_kerja	Gaji_dalam_usd
2020	172916.250000
2021	87071.250000
2022	151931.518519
2023	173480.984352
2024	162411.981595

dtype: float64

b. Visualisasi data:



Berdasarkan visualisasi tersebut, saya dapat memberikan interpretasi berikut:

- Tren Fluktuatif:
  - Pada Tahun 2020 gaji sekitar 155,000 USD dimulai.
  - Terjadi penurunan signifikan di tahun 2021 menjadi sekitar 85,000 USD.
  - Tahun 2022 mengalami kenaikan kembali ke level 150,000 USD.
  - Puncak tertinggi terjadi pada 2023 dengan gaji mencapai 175,000 USD.
  - Sedikit penurunan di 2024 menjadi sekitar 165,000 USD.
- Analisis Perubahan:
  - Penurunan drastis dari 2020 hingga 2021 (sekitar 45%) mungkin dipengaruhi oleh kondisi pandemi global.
  - Pemulihan yang kuat terjadi dari 2021 ke 2022 (kenaikan sekitar 76%).
  - Stabilisasi dan pertumbuhan moderat dari 2022 hingga 2024 .
- Insights Penting:
  - Meskipun ada perubahan, posisi Data Scientist tingkat senior tetap menerima gaji di atas 150.000 USD dalam tiga tahun terakhir.Terjadi tren pemulihan dan stabilisasi setelah masa krisis di tahun 2021.
  - Ada tren pemulihan dan stabilisasi setelah krisis di tahun 2021, dan gaji cenderung stabil di antara 150.000 dan 175.000 USD selama periode 2022–2024..
- Implikasi:
  - Meskipun ada guncangan pasar, posisi data scientist level senior menunjukkan ketahanan dalam hal kompensasi.
  - Dalam beberapa tahun terakhir, ada indikasi stabilitas dan kematangan pasar untuk posisi level senior.

```
Mid=data_scientist.loc[(data_scientist['Jenis_pekerjaan'] == 'Full Time') &(data_scientist['Level_pengalaman']== 'Mid'),
 ['Tahun_kerja',
  'Jabatan',
  'Level_pengalaman',
  'Gaji_dalam_usd',
  'Mata_uang',
  'Lokasi_perusahaan',
]].sort_values(by='Gaji_dalam_usd', ascending=False)
Mid
```

9. Menganalisis Data Gaji Sebagai Data Scientist Dengan Jenis Pekerjaan Full Time dan Level Pengalaman Kerja MID

Syntax ini sama dengan syntax pada point 7, namun yg berbeda hanya variabel data yang Bernama Mid, dan syntax **(data_scientist[‘Level_pengalaman]==’Mid’).** Karena di point ini yang kami c ari yaitu level pengalaman Mid. Sehingga memiliki output sebagai berikut:

	Tahun_kerja	Jabatan	Level_pengalaman	Gaji_dalam_usd	Mata_uang	Lokasi_perusahaan
139	2024	Data Scientist	Mid	385000	USD	US
347	2024	Data Scientist	Mid	310000	USD	US
348	2023	Data Scientist	Mid	310000	USD	US
503	2023	Data Scientist	Mid	296980	USD	US
851	2023	Data Scientist	Mid	260000	USD	US
...	...	...	...	...	...	...
14804	2021	Data Scientist	Mid	25532	EUR	DE
14780	2022	Data Scientist	Mid	25000	USD	TR
65	2023	Data Scientist	Mid	24613	THB	TH
62	2023	Data Scientist	Mid	17025	INR	IN
60	2021	Data Scientist	Mid	16904	INR	IN

701 rows × 6 columns

a. Menghitung rata-rata
Input:

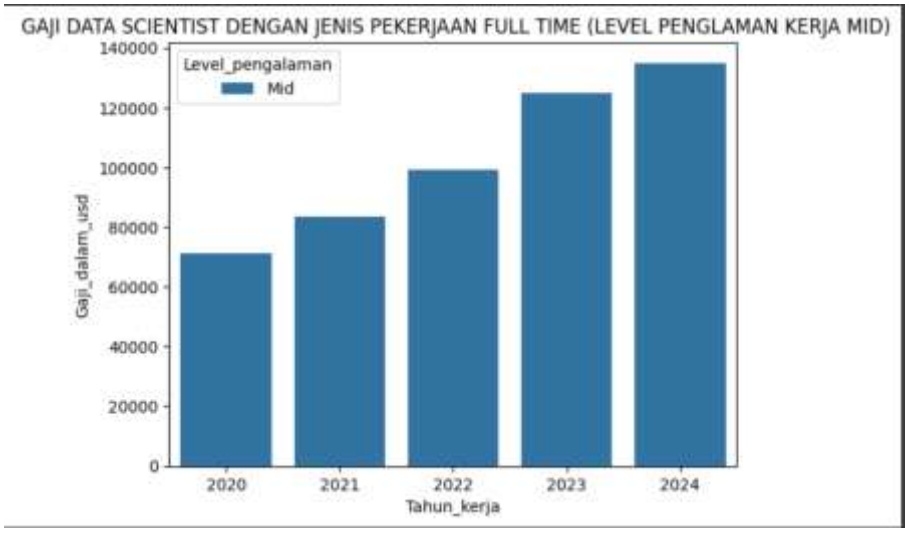
```
Mid.groupby('Tahun_kerja')['Gaji_dalam_usd'].mean()
```

Output:

Tahun_kerja	Gaji_dalam_usd
2020	71256.000000
2021	83348.173913
2022	99188.440000
2023	124881.494545
2024	135024.984227

dtype: float64

b. Visualisasi data



Sebagai kesimpulan dari visualisasi gaji Data Scientist level Mid ini, kami dapat mengidentifikasi tren berikut:

- Tren Kenaikan Konsisten:
  - 2020: sekitar 70,000 USD
  - 2024: mencapai 135,000 USD
  - Menunjukkan peningkatan hampir 2 kali lipat dalam waktu 5 tahun.
- Pola Pertumbuhan:
  - kenaikan gaji yang stabil setiap tahun.
  - Tidak ada penurunan di periode 2020-2024.
  - Kenaikan tertinggi terjadi antara 2022-2023
- Kesimpulan Penting:
  Posisi Data Scientist level Mid memiliki prospek yang sangat positif; peningkatan gaji yang konsisten menunjukkan permintaan yang terus meningkat, dan level Mid menjadi posisi yang menjanjikan dengan peningkatan kompensasi yang stabil. Ini menunjukkan bahwa posisi Data Scientist level Mid memiliki jalur karir yang jelas dengan peningkatan kompensasi yang dapat diprediksi setiap tahunnya.

10. Menganalisis Data Gaji Sebagai Data Scientist Dengan Jenis Pekerjaan Full Time dan Level Pengalaman Kerja ENTRY

Input:

```
#DATA GAJI DI DATA SCIENTIST DENGAN JENIS PEKERJAAN FULL TIME (LEVEL PENGALAMAN KERJA Entry)

Entry=data_scientist.loc[(data_scientist['Jenis_pekerjaan'] == 'Full Time') &(data_scientist['Level_pengalaman']== 'Entry'),
 ['Tahun_kerja',
  'Jabatan',
  'Level_pengalaman',
  'Gaji_dalam_usd',
  'Mata_uang',
  'Lokasi_perusahaan',
 ]].sort_values(by='Gaji_dalam_usd', ascending=False)
Entry
```

Syntax ini sama dengan yang digunakan pada poin 7, tetapi hanya ada satu variable yang berbeda, yaitu Entry, dan syntax (data_scientist[‘Level_pengalaman] ==’Entry’). Ini karena level pengalaman Entry adalah yang kami cari di sini. Sehingga memiliki output sebagai berikut:

	Tahun_kerja	Jabatan	Level_pengalaman	Gaji_dalam_usd	Mata_uang	Lokasi_perusahaan
3629	2023	Data Scientist	Entry	190000	USD	US
4350	2022	Data Scientist	Entry	180000	USD	US
4335	2022	Data Scientist	Entry	180000	USD	US
4556	2024	Data Scientist	Entry	175100	USD	US
4557	2024	Data Scientist	Entry	175100	USD	US
...	...	...	...	...	...	...
11156	2023	Data Scientist	Entry	19910	BRL	BR
14824	2023	Data Scientist	Entry	19434	EUR	GR
54	2022	Data Scientist	Entry	17805	INR	IN
7	2022	Data Scientist	Entry	17684	HUF	HU
14830	2023	Data Scientist	Entry	16000	USD	EC

127 rows × 6 columns

a. Menghitung Rata-Rata

Input:

```
Entry.groupby('Tahun_kerja')['Gaji_dalam_usd'].mean()
```

Output:

Tahun_kerja	Gaji_dalam_usd
2020	61646.200000
2021	63604.333333
2022	79514.851852
2023	93537.952381
2024	92477.000000

dtype: float64

b. Visualisasi Data
Input:

```
sns.barplot(
    data=Entry,
    x='Tahun_kerja',
    y='Gaji_dalam_usd',
    hue='Level_pengalaman',
    ci=None
).set(title=' GAJI DATA SCIENTIST DENGAN JENIS PEKERJAAN FULL TIME (LEVEL PENGLAMAN KERJA Entry)')
```

Output:

Berdasarkan visualisasi gaji tingkat entri Data Scientist, kami dapat memperkirakan hal-hal berikut:

- Tren Gaji:
  - 2020-2021: Relatif stabil di sekitar 60,000 USD.
  - 2022: Kenaikan besar menjadi 80,000 USD.
  - 2023-2024: Stabil di kisaran 90,000 USD.
- Pola Pertumbuhan:
  - Kenaikan terbesar terjadi antara 2021-2022.
  - Kenaikan total sekitar 30,000 USD dari tahun 2020 hingga 2024.
  - Tidak ada penurunan gaji selama 5 tahun.
- Insight Penting:
  Posisi Entry Level menunjukkan gaji awal yang cukup kompetitif; tren kenaikan positif menunjukkan industri yang berkembang; stabilitas gaji di tahun terakhir menunjukkan standarisasi posisi entry level. Visualisasi ini menunjukkan bahwa karir awal sebagai data scientist menawarkan kompensasi yang menjanjikan dengan tren peningkatan yang positif, meskipun tidak sedramatis posisi Mid atau Senior.

11. Menganalisis Data Gaji Sebagai Data Scientist dengan Jenis Pekerjaan CONTRACT dan Level Pe-ngalaman Kerja SENIOR

Input:

```
#DATA GAJI DI DATA SCIENTIST DENGAN JENIS PEKERJAAN CONTRACT (LEVEL PENGLAMAN KERJA Senior)

Senior=data_scientist.loc[(data_scientist['Jenis_pekerjaan'] == 'Contract') &(data_scientist['Level_pengalaman']== 'Senior'),
 ['Tahun_kerja',
  'Jabatan',
  'Jenis_pekerjaan',
  'Level_pengalaman',
  'Gaji_dalam_usd',
  'Mata_uang',
  'Lokasi_perusahaan',
 ]].sort_values(by='Gaji_dalam_usd', ascending=False)
Senior
```

Output:

	Tahun_kerja	Jabatan	Jenis_pekerjaan	Level_pengalaman	Gaji_dalam_usd	Mata_uang	Lokasi_perusahaan
3612	2024	Data Scientist	Contract	Senior	191027	USD	US
9519	2024	Data Scientist	Contract	Senior	120869	USD	US

a. Menghitung Rata-Rata

Tahun_kerja	Gaji_dalam_usd
2024	155948.0

dtype: float64

Berdasarkan data di atas, dapat disimpulkan bahwa data untuk jenis pekerjaan kontrak memiliki jumlah tenaga kerja yang masih minimal, yaitu 2, jauh berbeda dengan jenis pekerjaan full time, yang rata-rata gajinya hanya 160.000 USD pada tahun 2024.

12. Data Gaji di Data Scientist dengan Jenis Pekerjaan Contract, Freelance, Partime dengan Level Pe-nglaman Kerja Executive, Mid, Entry

Hanya dua data scientist dengan pengalaman kerja senior dan kontrak, menunjukkan bahwa *Data Scientist* sebagian besar bekerja fulltime.

13. Statistik Rata-rata Gaji Data Scientist Berdasarkan Level Pengalaman dari Tahun 2020-2024

Input:

```
rata_rata= data_scientist.groupby('Level_pengalaman')['Gaji_dalam_usd'].mean().sort_values(ascending=False)
rata_rata
#menghitung rata rata gaji sebagai data scientist dari tahun 2020-2024 di seluruh dunia, dan diurutkan dari yg terbesar
```

Penjelasan Syntax:

- **data_scientist.groupby('Level_pengalaman'):**data dikelompokkan berdasarkan Level_pengalaman (Entry, Mid, Senior, Executive).
- **['Gaji_dalam_usd'].mean():** Menghitung rata-rata gaji dalam dolar untuk setiap level pengalaman.
- **sort_values(ascending=False**): Mengurutkan nilai rata-rata dari yang tertinggi ke yang terendah.

Sehingga didapatkan output sebagai berikut:

Level_pengalaman	Gaji_dalam_usd
Executive	182440.756757
Senior	167625.925303
Mid	124480.467236
Entry	87028.373134

dtype: float64

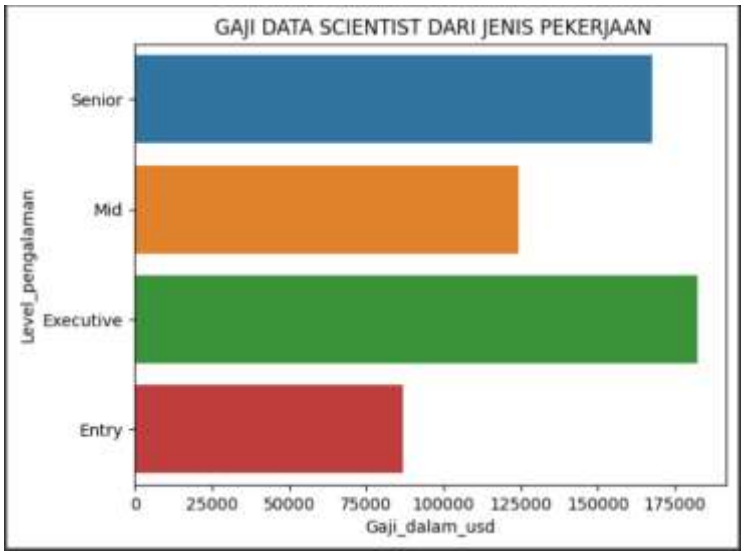
a. Visualisasi data
Input:

```
sns.barplot(
    data=data_scientist,
    y='Level_pengalaman',
    x='Gaji_dalam_usd',
    hue='Level_pengalaman',
    ci=None

).set(title=' GAJI DATA SCIENTIST DARI JENIS PEKERJAAN')
#hasil dari data sebelmnya, di visualisasikan
```

Penjelasan syntax:

- **Sns.barplot():** Membuat plot menggunakan library Seaborn.
- **Data:** Dataset yang dipakai bernama **'Data_scientist'.**
- **x:** Sumbu X menampilkan variable bernama **'Gaji_dalam_usd'.**
- **y:** Sumbu Y menampilkan variable bernama **'Level_pengalaman'.**
- **hue:** membedakan warna bar perdasarkan level_pengalaman.
- **ci=none:** Tidak menampilkan confidence interval (error bars).
- **.set(title= GAJI DATA SCIENTIST DARI JENIS PEKERJAAN):** Menambahkan judul grafik.

Sehingga outputnya sebagai berikut:



Sehingga dapat disimpulkan sebagai berikut: Berdasarkan Visualisasi Gaji Data Scientist berdasarkan level pengalaman, dapat diinterpretasikan sebagai berikut:

1. LevelExecutive
   - Gaji tertingginya sekitar 175,000 USD.
   - Merupakan posisi tertinggi dengan lebih banyak tanggung jawab dan pengalaman.
   - Menunjukan puncak karir data science
2. Level Senior
   - Gaji rata-rata sekitar 165,000 USD.
   - Sedikit lebih tinggi daripada level eksekutif.
   - Dan menghargai pengalaman dan keahlian sebagai senior.
3. Level Mid
   - Gaji rata-rata sekitar 125,000 USD.
   - Posisi tengah yang menunjukkan peningkatan signifikan dari entry level.
   - dan ada jarak yang cukup besar dengan level senior, yang kira-kira 40,000 USD.
4. Level Entry
   - Gaji terendah sekitar 80,000 USD.
   - Meski terendah, tetap menunjukkan kompensasi yang kompetitif untuk posisi awal.
   - ini juga menunjukkan banyak potensi pertumbuhan ke level berikutnya.

Sehingga dapat disimpulkan bahwa Industri *Data Science* memiliki jenjang karir yang jelas denga n peningkatan kompensasi yang signifikan; perbedaan gaji antar tingkat menunjukkan nilai pengal-aman dalam industri,gaji yang menarik bahkan untuk tingkat masuk dan struktur gaji menunjukkan industri yang matang dengan jalur karir yang jelas.

14. Menghitung Distribusi Frekuensi

Tujuan dari menghitung distribusi frekuensi ini antara lain untuk mengetahui seberapa banyak frekuensi tenaga kerja sebagai data scientist.

Input:

```
frequency_table = df['Tahun_kerja'].value_counts().reset_index()
frequency_table.columns = ['Tahun_kerja', 'Frekuensi']
print(frequency_table)# MENGHITUNG JUMLAH PEKERJA SEBAGAI DATA SCIENTIST DARI TAHUN 2020-2024
```

Penjelasan syntax:

- **df['Tahun_kerja']:** Untuk mengakses kolom "Tahun_kerja" dari DataFrame df, df['Tahun_kerja'] digunakan. Kolom ini dianggap mengandung data numerik yang menunjukkan tahun-tahun kerja individu.
- **Hitung Frekuensi:** Selanjutnya, fungsi value_counts() diterapkan pada kolom tersebut. Fungsi ini secara otomatis menghitung jumlah nilai unik yang muncul dalam kolom tersebut, menghasilkan sebuah seri (seperti kolom) dengan indeksnya adalah nilai unik dan nilainya adalah frekuensi kemunculannya.
- **.reset_index():** Ubah Indeks: Setelah proses penghitungan frekuensi selesai, metode.reset_index() digunakan untuk mengubah indeks dari Seri hasil value_counts() menjadi kolom biasa. Ini dilakukan karena kami ingin menyajikan hasil dalam bentuk DataFrame dengan dua kolom, yaitu 'Tahun_kerja' dan 'Frekuensi'. Frequency_table.colu mns = ['Tahun_kerja', 'Frekuensi'].
- **frequency_table.columns = ['Tahun_kerja', 'Frekuensi']:** Nama yang lebih spesifik diberikan pada kedua kolom hasil DataFrame dalam langkah terakhir ini. Kolom pertama berisi "Tahun_kerja", yang menunjukkan nilai khusus tahun kerja, dan kolom kedua berisi "Frekuensi", yang menunjukkan jumlah kali nilai tahun kerja tersebut muncul.

Output:

	Tahun_kerja	Frekuensi
0	2023	8519
1	2024	4374
2	2022	1652
3	2021	218
4	2020	75

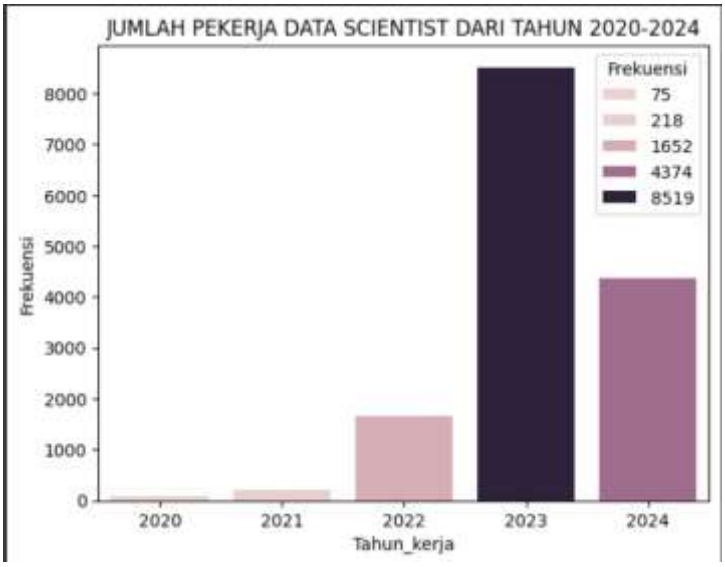
a. Visualisasi data

```
#visualisasi jumlah pekerja data scientist dari tahun 2020-2024

sns.barplot(
    data=frequency_table,
    x='Tahun_kerja',
    y='Frekuensi',
    hue='Frekuensi',
    ci=None
).set(title='JUMLAH PEKERJA DATA SCIENTIST DARI TAHUN 2020-2024')
```

Penjelasan syntax:

- **Sns.barplot():** Membuat plot menggunakan library seaborn.
- **Data:** Dataset yang digunakan bernama **'frequency_table'.**
- **x:** Sumbu X menampilkan variabel **'Tahun_kerja'.**
- **y:** Sumbu Y menampilkan variabel **'Frekuensi'.**
- **hue:** Warna bar dibedakan berdasarkan frekuensi
- **ci=none:** Tidak menampilkan confidence interval (error bars).
- . **set (title= 'JUMLAH PEKERJA DATA SCIENTIST DARI TAHUN 2020-2024):** Menambahkan judul grafik.

Sehingga hasil outputnya sebagai berikut:



Seperti yang ditunjukkan di atas, jumlah pekerja data scientist meningkat dari tahun ke tahun, terutama pada tahun 2023. Sehingga dapat disimpulkan bahwa:

- Tren Meningkat Dari tahun 2020 hingga 2024: Jumlah karyawan data scientist menigkat dari tahun 2020 hingga tahun 2024. Yang menunjukan bahwa permintaan yang meningkat untuk pekerjaan ini.
- Lonjakan Besar di Tahun 2023: pada tahun tersebut, jumlah pekerja data scientist jauh melampaui tahun-tahun sebelumnya. Ini menunjukkan peningkatan yang signifikan dalam permintaan akan pekerjaan ini.
- Perlambatan Pertumbuhan di Tahun 2024: Lonjakan jumlah pekerja data scientist pada tahun 2024 lebih lambat daripada lonjakan pada tahun 2023.
- Kemungkinan penyebab peningkatan: Perkembangan Teknologi: Perkembangan pesat teknologi, terutama dalam bidang big data dan kecerdasan buatan (AI), meningkatkan kebutuhan akan tenaga kerja yang mampu mengolah dan menganalisis data.
- Digitalisasi Bisnis: Karena lebih banyak perusahaan beralih ke model bisnis digital, data scientist diperlukan untuk membantu pengambilan keputusan berbasis data.
- Pandemi COVID-19: Pandemi COVID-19 meningkatkan digitalisasi di banyak sektor dan meningkatkan permintaan *Data Scientist*.

Implikasi:

- Peluang karir: Karena permintaan ilmuwan data semakin meningkat, ada banyak peluang karir bagi mereka yang memiliki keahlian di bidang ini.
- Pentingnya Keterampilan Data: Kemampuan untuk mengolah dan menganalisis data semakin penting di berbagai bidang
- Tantangan dalam Perekrutan: Pertumbuhan yang pesat ini dapat membuat sulit bagi perusahaan untuk merekrut dan mempertahankan data scientist berkualitas tinggi.

Kesimpulan dari Visualisasi ini menunjukkan peningkatan permintaan untuk pekerjaan data scientist dalam beberapa tahun terakhir. Tren ini menunjukkan bahwa data scientist memiliki masa depan yang cerah.

15. Menghitung Percentile

Input:

```
percentile=data_scientist.groupby('Level_pengalaman')['Gaji_dalam_usd'].quantile([0.1,0.25,0.5,0.75,0.9]) #menghitung percentile
percentile
```

Penjelasan syntax:

- **Percentile:** Variabel baru yang disebut percentile dibuat di bagian ini. Ini akan digunakan untuk menyimpan hasil perhitungan yang akan kita lakukan.
- **Data_scientist:** Ini dianggap sebagai "DataFrame", atau tabel data, yang berisi informasi tentang ilmuwan data. DataFrame ini mungkin berisi kolom-kolom seperti "Level_pengalaman", yang menunjukkan tingkat pengalaman, dan "Gaji_dalam_USD", yang menunjukkan gaji dalam dolar AS.
- **. groupby ('Level_pengalaman'):** Metode.groupby() digunakan untuk mengelompokkan data berdasarkan nilai pada kolom 'Level_pengalaman'. Ini berarti bahwa data akan dibagi menjadi beberapa kelompok berdasarkan berbagai tingkat pengalaman.
- **['Gaji_dalam_USD']:** Kami hanya ingin berkonsentrasi pada kolom "Gaji_dalam_USD" setelah data dikelompokkan berdasarkan tingkat pengalaman. Bagian ini menunjukkan bahwa kami akan melakukan perhitungan pada kolom gaji ini untuk setiap kelompok.
- **. quantile ([0.1, 0.25, 0.5, 0.75, 0.9]):** Metode.quantile() dapat digunakan untuk menghitung kuartil dan persentil dari data. Angka-angka dalam list [0.1, 0.25, 0.5, 0.75, 0.9] mewakili persentil yang ingin kita hitung.
    - 0.1: Persentil ke-10 (10% data berada di bawah nilai ini).
    - 0.25: Kuartil pertama (25% data berada di bawah nilai ini).
    - 0.5: Median (50% data berada di bawah nilai ini).
    - 0.75: Kuartil ketiga (75% data berada di bawah nilai ini).
    - 0.9: Persentil ke-90 (90% data berada di bawah nilai ini).

Sehingga hasil Outputnya:

Level_pengalaman		Gaji_dalam_usd
Entry	0.10	32974.00
	0.25	51141.00
	0.50	81000.00
	0.75	119800.00
	0.90	150000.00
Executive	0.10	112160.40
	0.25	144540.00
	0.50	181900.00
	0.75	225000.00
	0.90	253500.00
Mid	0.10	61472.30
	0.25	86144.25
	0.50	120000.00
	0.75	160000.00
	0.90	194080.00
Senior	0.10	101050.00
	0.25	130000.00
	0.50	160350.00
	0.75	198440.00
	0.90	239000.00

a. Visualisasi Data
Input:

```
[ ] percentile = percentile.reset_index()# mengubah Pandas Series menjadi Pandas DataFrame. karena seaborn membutuhkan data dalam bentuk dataframe, bukan series.
```

Penjelasan syntax:

- **Percentile:** Bagian ini mendefinisikan ulang variabel percentile. Operasi di sebelah kanan tanda sama dengan akan disimpan ke dalam variabel ini. percentile. Reset index ():
- **. reset index ()**: Ini digunakan untuk mengubah indeks dari DataFrame percentile menjadi kolom biasa. Ini berguna jika indeks sebelumnya memiliki makna khusus, seperti tanggal atau kategori, dan kita ingin memperlakukannya sebagai data biasa.

Kami mengubah seri Pandas menjadi Pandas DataFrame agar dapat dilihat karena Seaborn membutuhkan data dalam bentuk dataframe daripada seri.
Input:

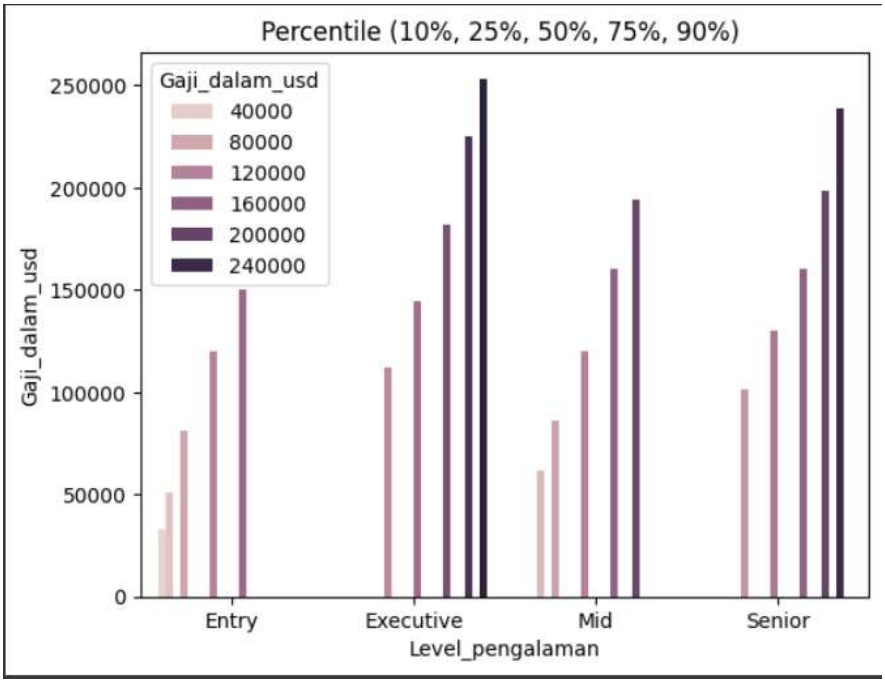
```
sns.barplot(
    data=percentile,
    x='Level_pengalaman',
    y='Gaji_dalam_usd',
    hue='Gaji_dalam_usd',
    ci=None
).set(title='Percentile (10%, 25%, 50%, 75%, 90%) ')
```

Penjelasan Syntax:

- **Sns.barplot():** Membuat plot menggunakan library seaborn.
- **Data:** Dataset yang digunakan bernama **'percentile.**
- **x:** Sumbu X menampilkan variabel **'level_pengalaman'.**
- **y:** Sumbu Y menampilkan variabel **'gaji_dalam_usd'.**

- **hue:** membedakan warna bar berdasarkan gaji_dalam_usd
- **ci=none:** Tidak menampilkan confidence interval (error bars).
- . **set (title= 'JUMLAH PEKERJA DATA SCIENTIST DARI TAHUN 2020-2024):** Menambahkan judul grafik.

Sehingga hasil outputnya:



Kami dapat mengambil kesimpulan dari visualisasi di atas bahwa:

- Peningkatan Gaji Dengan Pengalaman: Ini adalah fakta umum bahwa gaji cenderung meningkat seiring dengan pengalaman. Grafik dengan batang-batang untuk level Entry dan Mid biasanya lebih rendah daripada batang-batang untuk level Senior dan Executive, menunjukkan bahwa gaji yang diterima seseorang lebih tinggi sesuai dengan tingkat senioritas mereka.
- Rentang Gaji yang Luas: Ada banyak opsi gaji untuk setiap tingkat pengalaman dan pengalaman. Ini ditunjukkan oleh perbedaan yang signifikan antara batang-batang yang mewakili persentil ke-10 dan ke-90. Artinya, individu dalam kelompok tersebut sangat berbeda meskipun rata-rata gaji pada suatu tingkat mungkin sama.
- Perbedaan Gaji Antara Tingkat: Gaji karyawan baru di tingkat masuk sangat berbeda dari karyawan senior atau eksekutif. Ini menunjukkan bahwa ada perbedaan yang signifikan antara gaji karyawan yang lebih berpengalaman dan yang baru memulai karir.

16. Menghitung Ekstremum Minimal

Input:

```
#menghitung data ekstremum dari data diatas
minmax=data_scientist.groupby('Level_pengalaman')['Gaji_dalam_usd'].agg(['min','max']).sort_values(by='min', ascending=False)
minmax
```

Penjelasan syntax:

- **minmax:** Komponen yang menyimpan hasil perhitungan ke dalam variabel minmax. DataFrame baru akan dibuat berdasarkan level pengalaman dengan nilai minimum dan maksimum gaji.

- **Data scientist. Groupby ('Level_pengalaman'):** DataFrame yang berisi data scientist disebut data_scientist.
- **. groupby ('Level_pengalaman')**: Digunakan untuk mengelompokkan data berdasarkan kolom "Level_pengalaman", yang berarti bahwa data akan dibagi menjadi banyak kelompok yang masing-masing mewakili satu level pengalaman.
- **['Gaji_dalam_usd']:** Kita hanya ingin melihat kolom "Gaji_dalam_usd" setelah data dikelompokkan, jadi kita memilih kolom ini untuk dihitung
- **. agg 'min', 'max']):** Fungsi untuk menerapkan fungsi agregasi pada data yang telah dikelompokkan adalah agg (). "min", "max" menunjukkan bahwa kami ingin menghitung nilai minimum dan maksimum dari kolom "Gaji_dalam_usd" untuk masing-masing kelompok.
- **. sortvalues (by='min', ascending=False):** Kita ingin mengurutkan hasil setelah mendapatkan nilai minimum dan maksimum.
- **. sortvalues(by='min')**: Digunakan untuk mengurutkan berdasarkan kolom yang disebut sebagai "min", yang berarti nilai minimum.
- **ascending=False**: Maksudnya diurutkan dari yang terbesar ke yang terkecil. Karena itu, kelompok yang memiliki gaji minimum tertinggi akan berada di atas.

Sehingga outputnya sebagai berikut:

Level_pengalaman	min	max
Executive	78000	300000
Senior	20171	750000
Mid	16904	385000
Entry	16000	190000

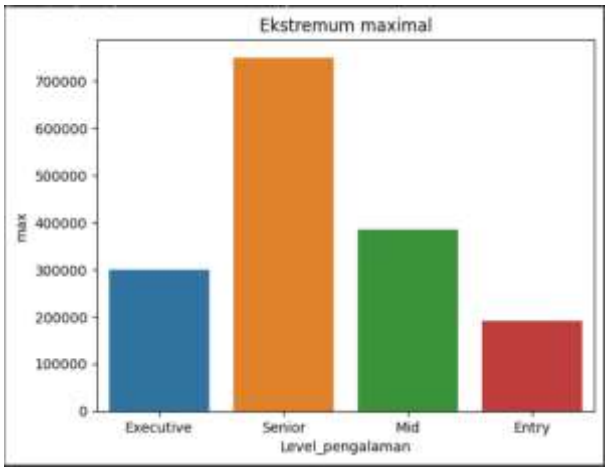
a. Visualisasi data
Input:

```
#visualisasi dalam  data ekstremum minimal

sns.barplot(
    data=minmax,
    x='Level_pengalaman',
    y='min',
    hue='Level_pengalaman',
    ci=None
).set(title='Ekstremum minimal')
```

Penjelasan Syntax:

- **Sns.barplot():** Membuat plot menggunakan library Seaborn.
- **Data:** Dataset yang digunakan bernama **'minmax.**
- **x:** Sumbu X menampilkan variabel **'level_pengalaman'.**
- **y:** Sumbu Y menampilkan variabel **'min'.**
- **hue:** Memedakan warna bar berdasarkan Level_pengalaman.
- **ci=none:** Tidak menampilkan confidence interval (error bars).
- . **set (title= 'Ekstremum nilai Data Scientist'):** Menambahkan judul grafik.

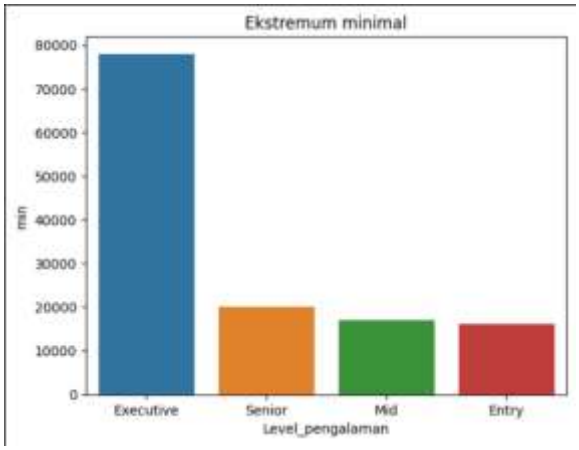
Sehingga hasil Output:



b. Visualisasi untuk ekstremum maksimal
Input:

```
[41] #visualisasi dalam  data ekstremum maximal
     sns.barplot(
         data=minmax,
         x='Level_pengalaman',
         y='max',
         hue='Level_pengalaman',
         ci=None
     ).set(title='Ekstremum maximal')
```

Output:



Dari visualisasi diatas kami dapat menyimpulkan bahwa:

- **Interpretasi exstremum min:** Dibandingkan dengan tingkat pengalaman lainnya, posisi executive memiliki gaji dasar yang paling tinggi. Gaji minimum biasanya lebih tinggi untuk data scientist yang lebih senior. Gaji minimum untuk masing-masing level pengalaman sangat berbeda.
- **Interpretasi ekstremum max:** Grafik ini menunjukkan gaji tertinggi yang dapat dicapai seorang data scientist pada setiap level pengalamannya. Kita dapat melihat bahwa tingkat senior menerima gaji maksimal yang jauh lebih tinggi daripada tingkat lainnya; ini menunjukkan bahwa data scientist senior memiliki potensi penghasilan yang jauh lebih

besar. Dengan demikian, tingkat eksekutif menerima gaji maksimal yang cukup tinggi, meskipun tidak setinggi tingkat senior, menunjukkan bahwa posisi eksekutif juga merupakan pekerjaan yang sangat menjanjikan dari segi finansial. Selain itu, tingkat masuk dan tingkat eksekutif menerima gaji maksimal yang lebih rendah dibandingkan dengan tingkat senior dan eksekutif, yang menunjukkan bahwa pengalaman dan tanggung jawab yang lebih besar pada tingkat yang lebih tinggi dihargai dengan gaji yang lebih tinggi.

- **Kesimpulan Keseluruhan Dengan menggabungkan kedua grafik ini:** Kita dapat melihat struktur gaji untuk data scientist berdasarkan tingkat pengalaman: Level Senior adalah level yang paling menguntungkan, baik dari segi gaji minimum maupun maksimum; Level Executive juga sangat menjanjikan dari segi finansial; dan Level Mid dan Entry memiliki potensi pertumbuhan gaji yang cukup besar jika mereka terus belajar.

17. Menghitung Standar Deviasi

a. Mengitung Standar Deviasi Tahun 2020

Penjelasan syntax:

- **Standard_deviation:** Deviasi standar dihitung dengan baris ini dan disimpan sebagai variabel deviasi standar.
- **Data_scientist.loc[...]:** Bagian ini digunakan untuk mengumpulkan data dari data_scientist dari DataFrame.
- **Data_scientist['Tahun_kerja'] == 2020:** Data yang disaring dalam kondisi ini hanya untuk tahun kerja 2020. Kolom 'Gaji_dalam_usd' adalah kolom yang akan kita gunakan untuk menghitung deviasi standarnya.
- **.std():** Deviasi standar dari data yang telah diseleksi dihitung dengan menggunakan fungsi.std(). Deviasi standar adalah ukuran sebaran data dari rata-rata. Nilai deviasi standar yang lebih besar menunjukkan bahwa data tersebar lebih jauh dari rata-rata.

Sehingga dapat disimpulkan bahwa jumlah 81436.64871089616 adalah deviasi standar dari gaji Data Scientist pada tahun 2020. Ini berarti bahwa gaji rata-rata mereka sekitar 81.436,65 USD lebih rendah dari gaji rata-rata.

b. Menghitung Standar Deviasi Tahun 2021

Kami dapat mengetahui bahwa jumlah 40573.319274 adalah deviasi standar dari gaji dalam dolar untuk *Data Scientist* yang bekerja pada tahun 2021. Dengan kata lain, gaji rata-rata *Data Scientist* pada tahun 2021 adalah sekitar 40.573,32 dolar.

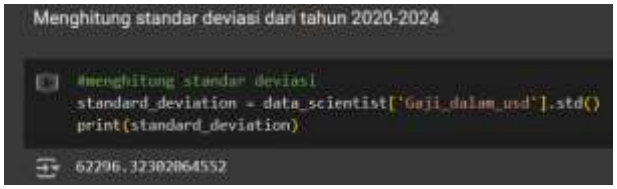
c. Menghitung Standar Deviasi Tahun 2022

Jadi, kami dapat mengetahui bahwa jumlah 53191.035555 adalah deviasi standar dari gaji *Data Scientist* pada tahun 2023. Artinya, gaji rata-rata *Data Scientist* pada tahun 2023 sekitar 53.191.55 USD lebih rendah dari gaji rata-rata mereka.

d. Menghitung Standar Deviasi Tahun 2024

Jadi kami dapat mengetahui bahwa gaji sebagai *Data Scientist* pada tahun 2024 rata-rata 62.228.70 USD lebih rendah dari gaji rata-rata.

e. Menghitung Standar Deviasi Tahun 2020-2024



Jadi, kami dapat menyimpulkan bahwa: Dalam konteks ini, nilai standar deviasi sebesar 62296.32 USD menunjukkan variasi yang signifikan dalam gaji para *Data Scientist* dalam data. Artinya, ada beberapa *Data Scientist* yang menerima gaji yang jauh di atas rata-rata, dan ada juga yang menerima gaji yang jauh dibawah rata-rata.

Dapat disimpulkan dari perhitungan ini bahwa dalam sampel data yang digunakan, gaji *Data Scientist* memiliki distribusi yang cukup luas. Perbedaan gaji yang signifikan ini dapat disebabkan oleh variabel seperti pengalaman, keterampilan, perusahaan tempat bekerja, lokasi, dan tanggung jawab pekerjaan.

KESIMPULAN

Hasil analisis tentang peningkatan gaji *Data Scientist* dari tahun 2020 hingga 2024 menunjukkan tren peningkatan gaji umum di semua level pengalaman (Executive, Senior, Mid, dan Entry) karena kebutuhan akan pengolahan data yang kompleks dan pengambilan keputusan berbasis data semakin meningkat.

Sebaliknya, penelitian menemukan bahwa ada perbedaan; beberapa kelompok mengalami variasi gaji selama periode waktu tertentu, baik naik maupun turun. Salah satu contoh yang signifikan adalah penurunan rata-rata gaji pada tahun 2021. Penurunan ini diduga disebabkan oleh sejumlah variabel, seperti dampak ekonomi pandemi COVID-19, perubahan yang dibuat oleh perusahaan mengenai investasi teknologi, dan dinamika pasar tenaga kerja. Kondisi makroekonomi, industri, dan lokasi geografis adalah beberapa variabel yang menunjukkan ketidakpastian tren ini.

Ketidakpastian dalam tren ini menunjukkan bahwa banyak variabel, termasuk kondisi makro ekonomi, industri, lokasi geografis, kebijakan perusahaan, dan tingkat permintaan pasar untuk tenaga kerja di bidang *Data Science*, memengaruhi gaji *Data Scientist*. Akibatnya, analisis tambahan diperlukan untuk mengidentifikasi sumber fluktuasi.

Secara keseluruhan, temuan analisis ini memberikan gambaran penting tentang bagaimana kompensasi pekerjaan *Data Scientist* berkembang. Selain itu, mereka memiliki kemampuan untuk membantu bisnis, tenaga kerja, dan pemangku kebijakan memahami dinamika pasar tenaga kerja di bidang ini.

DAFTAR PUSTAKA

[1]. Alfarizi, M, R, Z, Dkk. 2023. *"Penggunaan Python Sebagai Bahasa Pemrograman Untuk Machine Learning dan Deep Learning".* Karimah Tauhid, Volume 2 Nomor 1 (2023), e-ISSN 2963-590X.

[2]. Aliwijaya, A. 2023. "*Peluang Pemanfaatan Big Data di Perpustakaan: Sebuah Kajian Literatur.".* Media Informasi, 32(2), 215-223. sian *(Journal of Marketing Science)*, 20(2), 163-179.

[3]. As' ad. Moh. (2005). *"Manajemen Sumber Daya Manusia, Galia Indonesia".* Yogyakarta.

[4]. Muhajir, S. Widodo 2021. *"Peran Data Science dan Data Scientist Untuk Mentransformasi Data dalam Industri 4.0".* E-ISSN: 2797 – 4111.

[5]. Ridhoni, M. 2017. *"Rekayasa WEB MAP Dinamis Dengan Tile Based MAP Menggunakan Framework Django dan Mapnik".* Skripsi thesis, STMIK AKAKOM Yogyakarta. Diakses pada https://eprints.akakom.ac.id/4003/

[6]. Samuel, D. 2023. *"Pengantar Data Science".* Diakses pada https://raharja.ac.id/2023/10/21/pengantar-data-science/