**MODEL PEMBELAJARAN DAN LAPORAN AKHIR**

**PROJECT BASED LEARNING**

**MATA KULIAH DATA WRANGLING**

**KELAS D**



**“KLASIFIKASI DATA PENGHASILAN DATA SCIENTIST DENGAN METODE DATA WRANGLING”**

**DISUSUN OLEH KELOMPOK “IV” :**

1. KANESSA JASMINE PRISHEILA A. S. ( 22083010016 )
2. MUHAMMAD NASHIF FARID ( 22083010024)
3. DEVINA PUTRI YULIANI ( 22083010028)
4. VANNESA NATHANIA ( 22083010044)
5. MUHAMMAD GHINAN NAVSIH ( 22083010057)

**DOSEN PENGAMPU:**

KARTIKA MAULIDA HINDRAYANI, S.KOM, M.KOM

(NIP. 199209092022032009 )

**PROGRAM STUDI SAINS DATA**

**FAKULTAS ILMU KOMPUTER**

**UNIVERSITAS PEMBANGUNAN NASIONAL “VETERAN”**

**JAWA TIMUR**

**2023**

**BAB I**

**PENDAHULUAN**

* 1. **Deskripsi Dataset dan Metadata**

Dataset “ds\_salaries.csv” yang kelompok kami ambil merupakan file csv yang berisi data pekerjaan data scientist dan tingkat pengalamannya. Dataset ini menyimpan berbagai informasi tentang penghasilan data scientist di berbagai perusahaan yang ada di penjuru Dunia. Dalam file ini terdapat beberapa kolom yang berisi berbagai atribut terkait waktu bekerja, pengalaman, tingkat Pendidikan, job, gaji, lokasi perusahaan, dll.

Metadata dalam dataset “ds\_salaries.csv” adalah sebagai berikut:

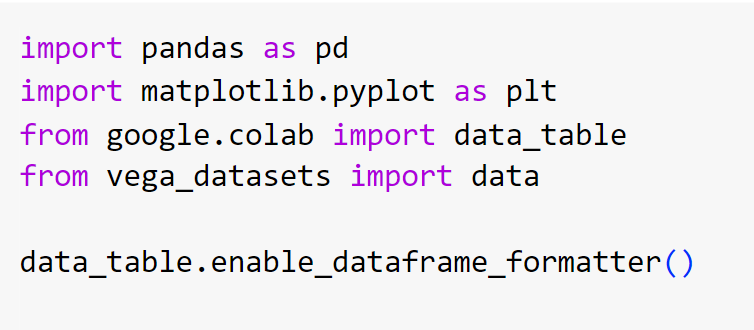
* + - 1. Nama Dataset: ds\_salaries.csv
      2. Penjelasan Dataset: Dataset ds\_salaries.csv merupakan file csv yang berisi data gaji data scientist dan pengalamannya di seluruh dunia, data ini dapat digunakan untuk membandingkan gaji di berbagai lokasi, memprediksi gaji, dan menganalisis faktor-faktor yang mempengaruhi gaji seorang data scientist
      3. Sumber Data: aijobs.net
      4. Tanggal Pembuatan: Terakhir diperbarui satu bulan yang lalu (April 2023)
      5. Jumlah Baris dan Kolom: Dataset ini berisi 11 kolom dengan 3755 baris.
      6. Nama Kolom:

1. work\_year : Merupakan kolom berisi tahun gaji dibayarkan.
2. experience\_level : Kolom berisi tingkat pengalaman kerja selama tahun tersebut.
3. employment\_type : Berisi jenis pekerjaan dalam peran tersebut.
4. job\_title : Berisi jabatan yang dijabat selama tahun tersebut.
5. salary : Berisi jumlah total gaji kotor yang didapatkan.
6. salary\_currency : Berisi mata uang gaji yang dibayarkan sebagai kode mata uang ISO 4217.
7. salary\_in\_usd : Berisi besarnya gaji dalam mata uang USD.
8. employee\_residence : Berisi negara tempat tinggal karyawan selama bekerja sebagai kode negara ISO 3166.
9. remote\_ratio : rasio jumlah pekerjaan yang dilakukan dari jarak jauh.
10. company\_location : Berisi negara dimana kantor berada.
11. company\_size : Berisi jumlah rata-rata orang yang bekerja dalam kantor tersebut.
    * + 1. Jenis Data
12. work\_year : berisi data dengan jenis data integer (int64)
13. experience\_level : berisi data dengan jenis data object
14. employment\_type : berisi data dengan jenis data object
15. job\_title : berisi data dengan jenis data object
16. salary : berisi data dengan jenis data int (int64)
17. salary\_currency : berisi data dengan jenis data object
18. salary\_in\_usd : berisi data dengan jenis data integer (int64)
19. employee\_residence : berisi data dengan jenis data object
20. remote\_ratio : berisi data dengan jenis data integer (int64)
21. company\_location : berisi data dengan jenis data object
22. company\_size : berisi data dengan jenis data object
    * + 1. Lisensi: Database: Open Database, Contents: Database Contents

**BAB II**

**METODE DATA WRANGLING**

* 1. **Import dan Read Data** 
     1. **Import Library**

****

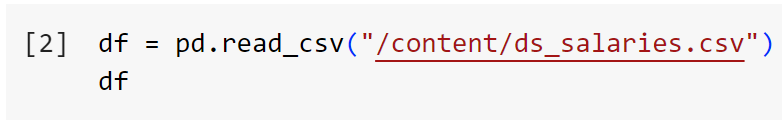
Dalam pengelolaan data pendapatan pekerjaan yang berkecimpung pada bidang Data Science, terdapat beberapa library dan package yang sangat penting, yaitu pandas, matplotlib.pyplot, google.colab, dan vega\_datasets. Pandas digunakan untuk analisis dan manipulasi data. Library ini menyediakan struktur data dan alat yang efisien untuk pemrosesan data. Dalam proses data wrangling, Pandas digunakan untuk membaca data, mengidentifikasi nilai null, serta melakukan operasi subset data.

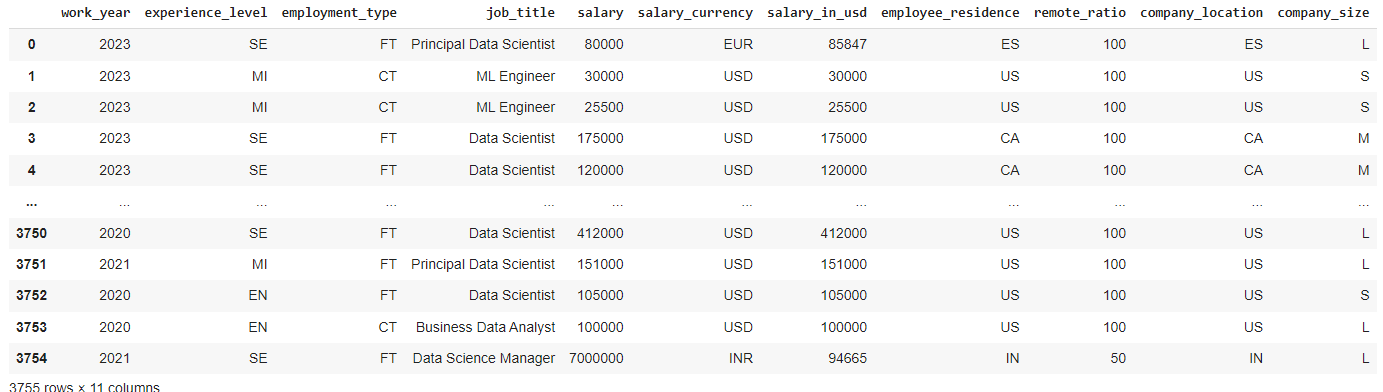
Selanjutnya, terdapat library Matplotlib, khususnya submodul pyplot, yang memungkinkan pembuatan visualisasi grafik. Matplotlib.pyplot menyediakan antarmuka yang mirip dengan matlab untuk membuat plot, histogram, dan boxplot. Dalam tahap ini, visualisasi data merupakan langkah terpisah setelah proses Data Wrangling selesai. Visualisasi data menggunakan Matplotlib.pyplot digunakan untuk mempresentasikan data pendapatan pekerjaan dalam bentuk grafik yang informatif dan mudah dipahami.

Data tabel dalam package google.colab adalah modul yang menyediakan tampilan dan manipulasi data dalam format tabel. Dalam modul ini memiliki fitur tampilan tabel interaktif di *environtment* notebook Google Colab. Sedangkan vega\_datasets merupakan library pada Python yang dapat memvisualisasikan kumpulan data. Package ini dapat melakukan visualisasi, machine learning, dan eksplorasi data.

Dengan menggunakan kombinasi pandas, matplotlib.pyplot, google.colab, dan vega\_datasets dapat melakukan analisis data, memanipulasi data, serta memvisualisasikan hasilnya dalam bentuk grafik yang relevan dan berguna dalam memahami data pendapatan pekerjaan di bidang Data Science.

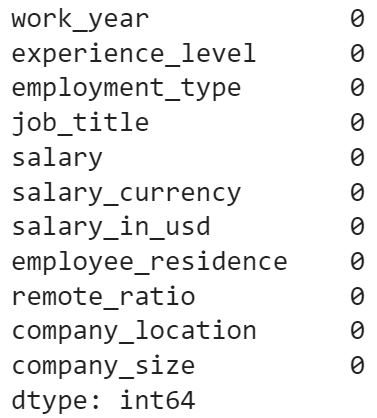
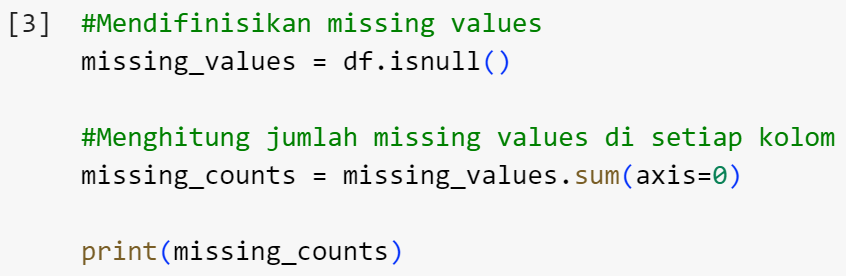
* + 1. **Read Data**

****

Dalam metode proses data wrangling penting dijalankannya metode ini sebagai pengambilan sumber data yang akan diolah bersama metode lainnya. Dengan pembelajaran ini seorang user terdapat mengakses beberapa sumber data dengan beberapa tipe yang berbeda juga diantaranya file csv, excel, txt, pdf hingga file berupa html yang biasa disebut dengan metode web scraping. Dataset yang diakses melalui website kaggle dengan pembahasan Data Science Salaries. Langkah pertama yang harus dilakukan yakni mengimport library pandas (import pandas as pd) sebagai pemanfaatan fitur bawaan yang sudah terdapat dalam pandas, yakni salah satunya untuk pemanggilan dan menampilkan data. Kemudian setelah itu, pemanggilan dataset dengan menggunakan syntax pd.read\_csv(…) dengan memasukkan nama dataset di dalam tanda kurung. Setelah pemanggilan selesai maka data dapat ditampilkan dengan menuliskan variabel data yang sudah diinputkan yakni variabel dataset.

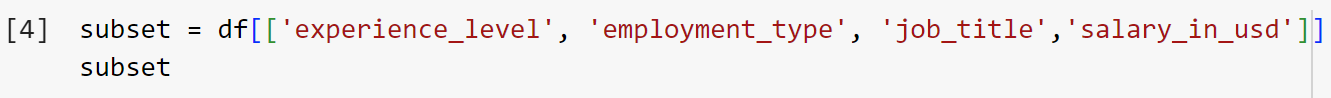
* 1. **Data Cleaning**

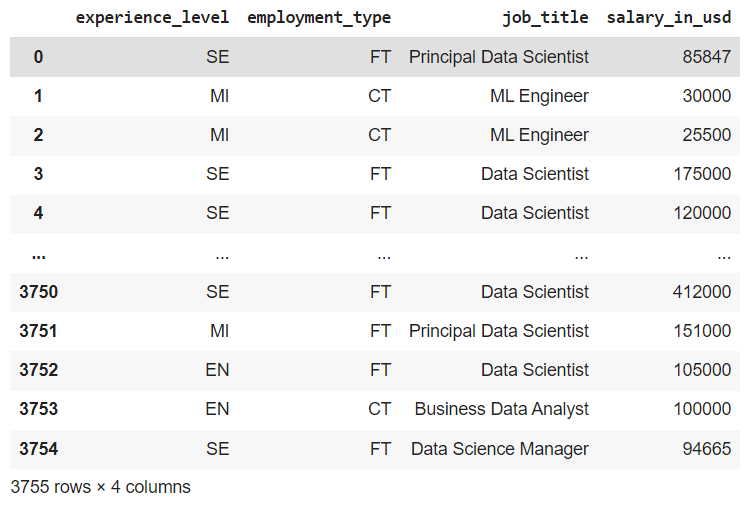
**2.2.1 Mencari Missing Value**

****

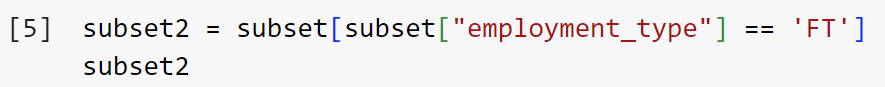
NumPy menyediakan alat yang kuat untuk mengelola nilai yang hilang dalam dataframe Pandas. Alat-alat ini termasuk fungsi untuk memeriksa, mengisi, membuang, menggabungkan, dan menginterpolasi nilai yang hilang. Dalam kode script di samping, tujuannya adalah untuk memeriksa apakah terdapat kolom yang mengandung nilai 'NaN' dalam dataframe. Dengan menggunakan syntax isnull() pada output yang ditampilkan, kita dapat mengidentifikasi apakah terdapat kekosongan nilai pada data. Dalam kasus ini, dapat disimpulkan bahwa tidak ada data yang memiliki kekosongan nilai.

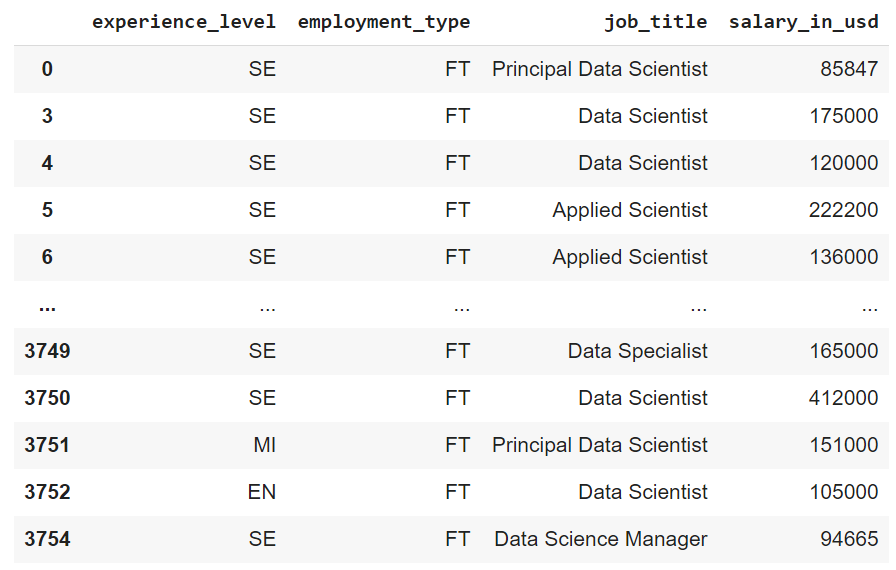
* + 1. **Subsetting**

****

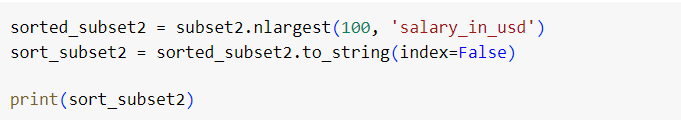
****

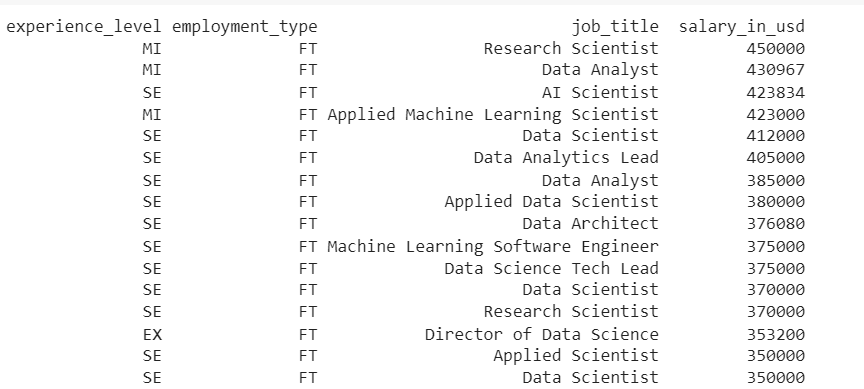
Subsetting merupakan salah satu cara yang dapat dilakukan python untuk mengambil sebagian objek. Dalam program ini, dengan bentuk data frame dapat melakukan penghapusan kolom ataupun baris yang tidak diperlukan. Dalam subset pertama, kita tampilkan 4 kolom yang semula berjumlah 11 kolom yang ada pada dataframe awal untuk dilakukan analisa. 4 kolom tersebut adalah: experience\_level, employment\_type, job\_title, dan salary\_in\_usd.

****

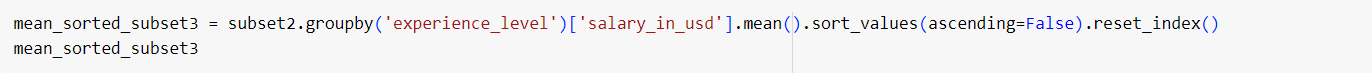
****

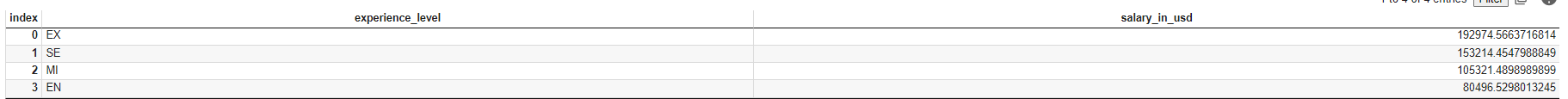
Pada subset kedua, kita hanya menampilkan baris yang mengandung string 'FT' pada kolom 'employment\_type', yang merupakan kepanjangan dari Full Time. Hal ini dilakukan agar analisis yang dilakukan hanya berfokus pada data scientist yang bekerja secara penuh waktu.

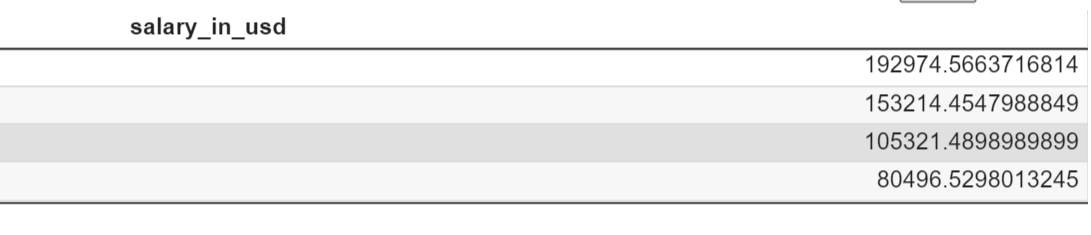




Pada variabel sorted\_subset2 akan memproses penampilan data frame baru yang berisi data pada subset2. Perbedaanya terletak pada urutan baris, sorted\_subset2 akan menampilkan 100 baris dengan nilai yang diurutkan berdasarkan salary\_in\_usd (gaji pendapatan). Salary\_in\_usd diurutkan secara descending; pengurutan data dari nilai terbesar menuju nilai yang lebihkecil.



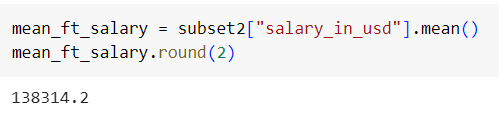




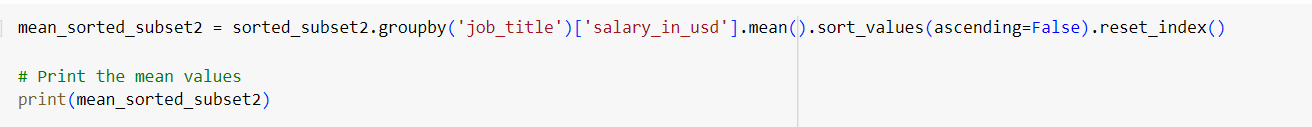
Pada subset ketiga, akan memproses pengelompokan experience\_level dengan salary\_in\_usd. Output yang akan diproses adalah urutan level pengalaman yang dihubungkan pada gaji dalam dollar. Sehingga, pada variabel mean\_sorted\_subset3 akan menampilkan 3 index yang terdiri dari “EX”, “SE”, “MI”, “EN”. Ex adalah level pengalaman tertinggi yang disebut *Expert,* “SE” adalah inisialisasi dari *Senior,*”MI” adalah inisialisasi dari  *Mid Level,* sedangkan “EN” adalah inisialisasi dari  *Entry Level.*

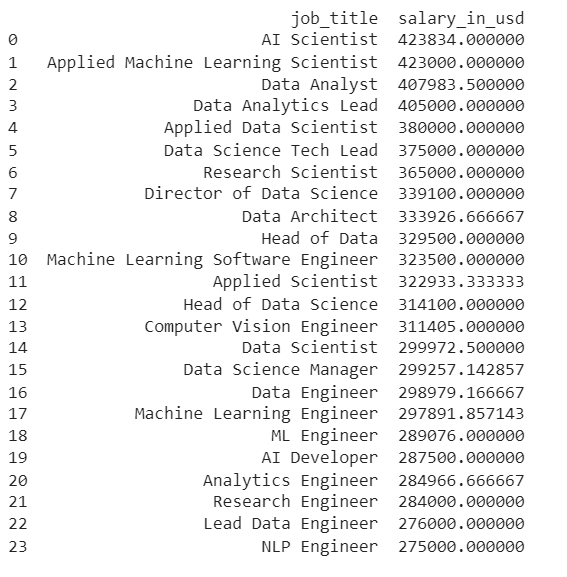
* 1. **Processing Data**

**2.3.1 Mean**



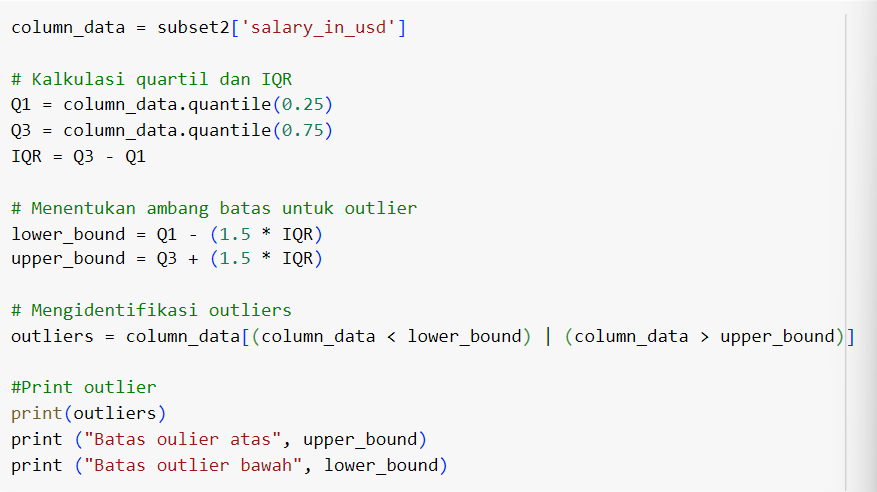
Pada syntax tersebut dapat memberikan perhitungan rata-rata dari kolom “salary\_in\_usd” pada data di variabel subset2. Dari output dapat diketahui bahwa pendapatan memiliki rata-rata nilai $138314.2 atau dalam rupiah berkisar pada Rp2.047.621.203.



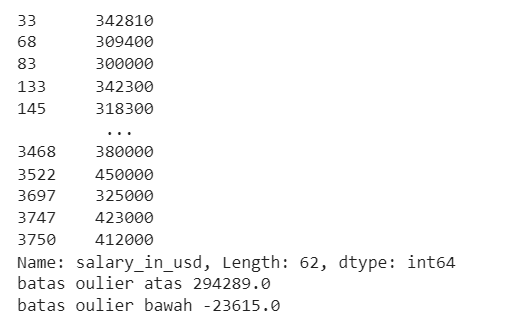


Pada variabel mean\_sorted\_subset2 akan memproses perhitungan dan pengelompokkan. Pengelompokkan yang dilakukan berdasarkan job\_title pada data frame sorted\_subset2. Perhitungan yang dilakukan adalah menghitung mean dari salary\_in\_usd dengan menggunakan urutan dari yang terbesar menuju nilai yang lebih kecil atau disebut urutan descending. Output yang akan diproses adalah urutan pekerjaan dengan rata-rata gaji tertinggi menuju gaji yang lebih rendah dengan serta menampilkan nama pekerjaan.

* + 1. **Outlier**



Pada syntax ini menjelaskan perhitungan nilai-nilai kuartil pada data subset2. Penggunaan library Pandas akan mempermudah perhitungan kuartil. Kuartil 1 adalah nilai yang membagi data menjadi 25% kebawah. Kuartil 3 adalah nilai yang membagi data menjadi 75%. Dalam penentuan batas diperlukannya perhitungan selisih kuartil dengan 1.5 dikali rentang antar kuartil. Sehingga output yang akan diberikan adalah sebagai berikut:



Batas atas pada data kolom salary\_in\_usd diidentifikasi pada nilai $294289. Sedangkan pada batas bawah diidentifikasi pada nilai $-293615. Berdasarkan output, outlier tersebar pada data ke-33, 68, 83, 3747, 3750, dan banyak lainnya. Dengan memperhatikan “length: 62” maka dapat diketahui bahwa jumlah data yang termasuk outlier berjumlah 62 data dari 3754 total data.

**BAB III**

**HASIL DAN PEMBAHASAN**

**3.1 Informasi Data**

Metode data wrangling merupakan proses yang bertujuan untuk mempersiapkan dan mengolah data mentah menjadi data yang sudah terstruktur dan siap untuk di analisis lebih lanjut serta untuk menggali informasi lebih dalam dari dataset tersebut sebagai insight. Berdasarkan hal tersebut, setelah melalui beberapa metode data wrangling yang sudah dikerjakan sebelumnya, data ini dihasilkan dengan tujuan untuk mengetahui seberapa tinggi penghasilan seorang data scientist berdasarkan kategori waktu bekerja (Full Time). Sehingga dapat dihasilkan data seperti halnya tabel berikut.

Tabel Penghasilan Full Time Data Scientist

|  |  |  |
| --- | --- | --- |
| NO | JOB TITLE | SALARY IN USD |
| 1 | AI Scientist | 423834.000000 |
| 2 | Applied Machine Learning Scientist | 423000.000000 |
| 3 | Data Analyst | 407983.500000 |
| 4 | Data Analytics Lead | 405000.000000 |
| 5 | Applied Data Scientist | 380000.000000 |
| 6 | Data Science Tech Lead | 375000.000000 |
| 7 | Research Scientist | 365000.000000 |
| 8 | Director of Data Science | 339100.000000 |
| 9 | Data Architect | 333926.666667 |
| 10 | Head of Data | 329500.000000 |
| 11 | Machine Learning Software Engineer | 323500.000000 |
| 12 | Applied Scientist | 322933.333333 |
| 13 | Head of Data Science | 314100.000000 |
| 14 | Computer Vision Engineer | 311405.000000 |
| 15 | Data Scientist | 299972.500000 |
| 16 | Data Science Manager | 299257.142857 |
| 17 | Data Engineer | 298979.166667 |
| 18 | Machine Learning Engineer | 297891.857143 |
| 19 | ML Engineer | 289076.000000 |
| 20 | AI Developer | 287500.000000 |
| 21 | Analytics Engineer | 284966.666667 |
| 22 | Research Engineer | 284000.000000 |
| 23 | Lead Data Engineer | 276000.000000 |
| 24 | NLP Engineer | 275000.000000 |

Tabel diatas menunjukkan rata-rata penghasilan seorang data scientist berdasarkan title masing-masing. Hal ini bertujuan sebagai informasi serta acuan bagi seseorang dalam merencanakan dan menentukan posisi data scientist yang diinginkan berdasarkan penghasilan. Selain hal tersebut, tabel diatas juga bertujuan untuk analisis lanjutan dalam menggali informasi faktor utama posisi tersebut dapat memiliki penghasilan begitu besar serta pertimbangan dan acuan seseorang dalam mendalami salah satu subbab materi yang ada didalamnya.

Selain itu, dari metode data wrangling juga didapati perbandingan gaji seorang data scientist berdasarkan pengalaman kerja. Terdapat 4 kategori yakni Expert, Senior, Mid level serta Entry level yang telah kami ketahui rata-rata penghasilannya.

Tabel Penghasilan Data Scientist berdasarkan Pengalaman

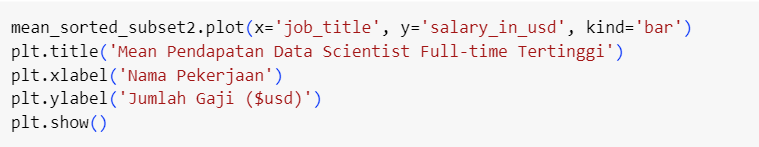
|  |  |  |
| --- | --- | --- |
| NO | EXPERIENCE LEVEL | SALARY IN USD |
| 1 | EXPERT | 19297.45663716810 |
| 2 | SENIOR | 15321.44547988840 |
| 3 | MID LEVEL | 10532.14898989890 |
| 4 | ENTRY LEVEL | 80496.5298013245 |

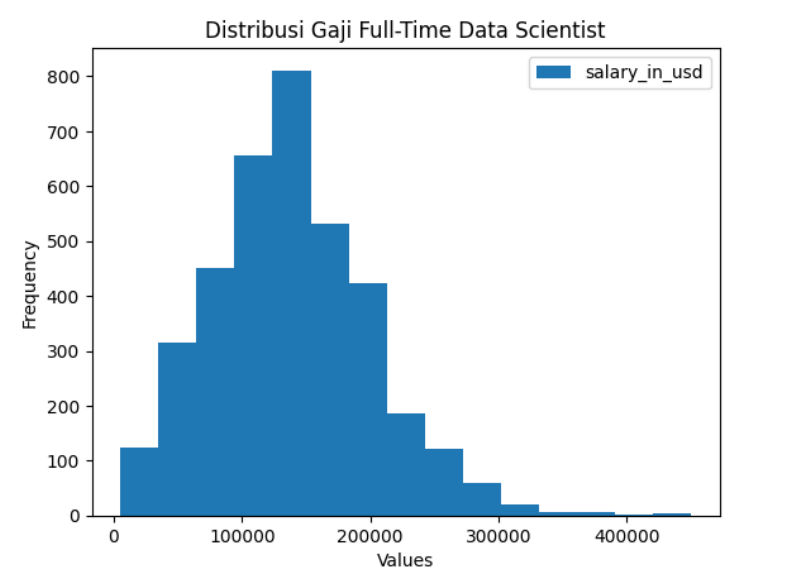
Informasi diatas dapat diambil kesimpulan yakni semakin tinggi tingkat pengalaman dalam bekerja maka semakin tinggi juga penghasilan yang didapatkan.

**3.2 Presentasi Hasil**

Dengan hasil data yang diperoleh, langkah selanjutnya presentasi serta interpretasi hasil dengan menggunakan beberapa visualisasi dengan tujuan memperoleh informasi lebih mendalam dari data yang dimiliki.

**3.2.1 Grafik Histogram Distribusi Penghasilan Full Time Data Scientist**



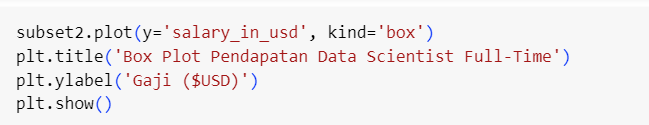


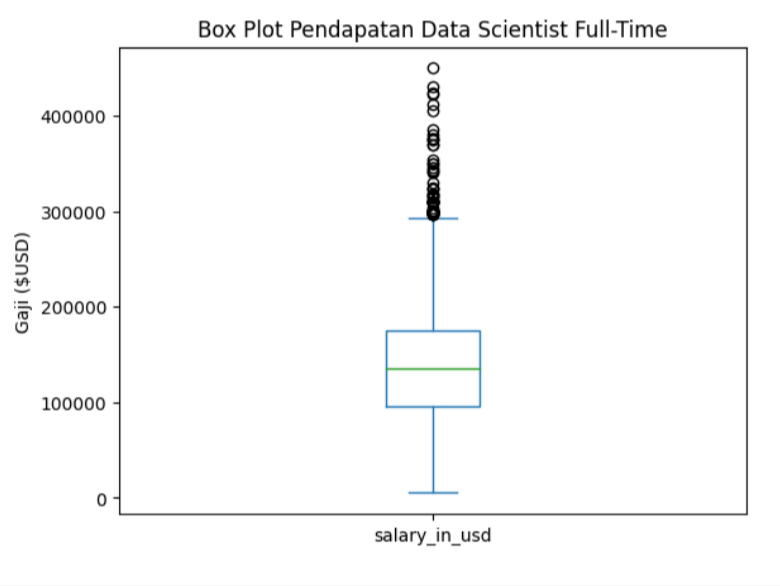
Dari grafik histogram Distribusi Gaji Full-Time Data Science dapat diketahui bahwa distribusi histogram tersebut adalah distribusi data skewness dimana nilainya tidak berdistribusi normal dan cenderung positif skewness atau histogram yang memiliki ekor. Sebelum membuat interpretasi dari histogram, ada langkah yang harus dilakukan yaitu mengecek skala yang ditunjukkan pada sumbu x dan y.

Dari gambar tersebut skala pada sumbu x adalah 100.000 sedangkan sumbu y adalah 100 dan tidak ditemukannya nilai yang tidak sebanding sehingga dapat diinterpretasikan lebih akurat. Menurut sumbu x, persebaran nilai tersebar pada 0 sampai dengan 400.000. Sedangkan pada sumbu y, frekuensi data tersebar pada 0 sampai dengan 600.

Frekuensi tertinggi ada pada nilai di antara 100.000 dan 200.000 atau ada rentang 130.000 dengan total frekuensi 800. Sedangkan frekuensi terendah ada pada nilai lebih dari 400.000. Sehingga, dapat disimpulkan bahwa dominasi nilai dari gaji adalah $130.000 atau setara dengan Rp1.942.669.500.

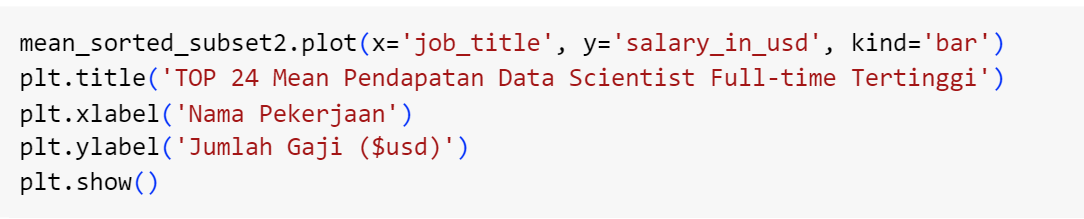
**3.2.2 Grafik Box-Plot Penghasilan Full Time Data Scientist**

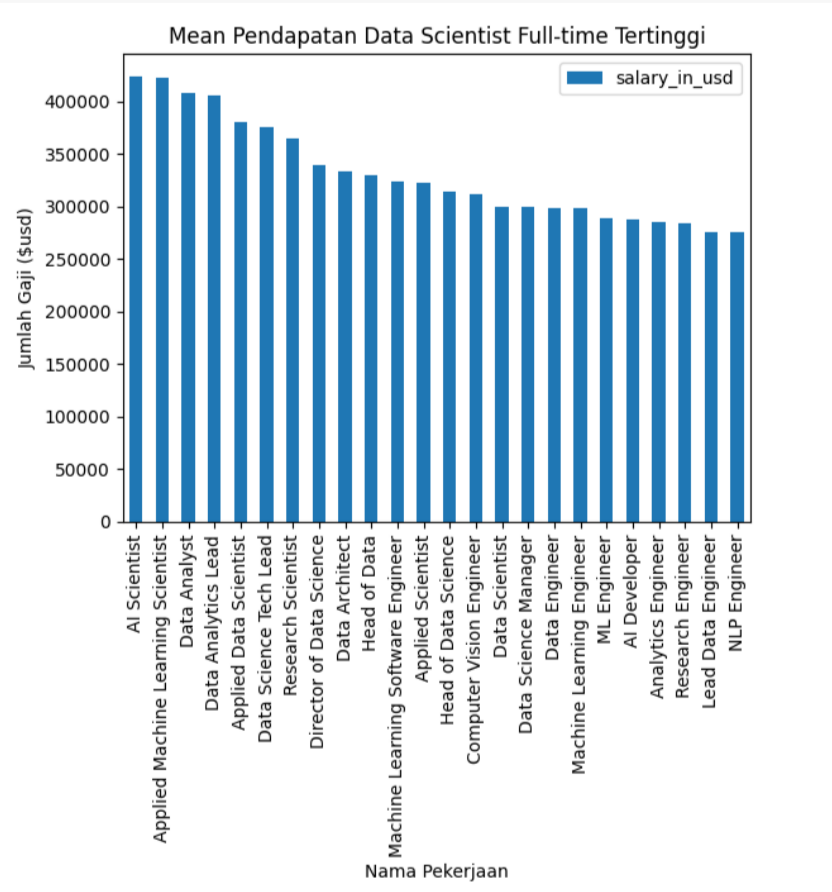




Selain menggunakan perhitungan dengan metode IQR (Interquartile Range), dapat menggunakan perhitungan secara visual dengan menggunakan box plot. Dalam syntax ini akan memproses perhitungan subset2 atau variabel yang menyimpan data dengan pekerjaan Data Science secara full time job. Dari output yang diberikan dapat diketahui bahwa gaji dengan pekerjaan Data Science yang bekerja secara Full Time memiliki gaji di atas rata-rata dan tersebar di atas batas atas. Sehingga, dapat disimpulkan bahwa subset2 merupakan kumpulan data outlier.

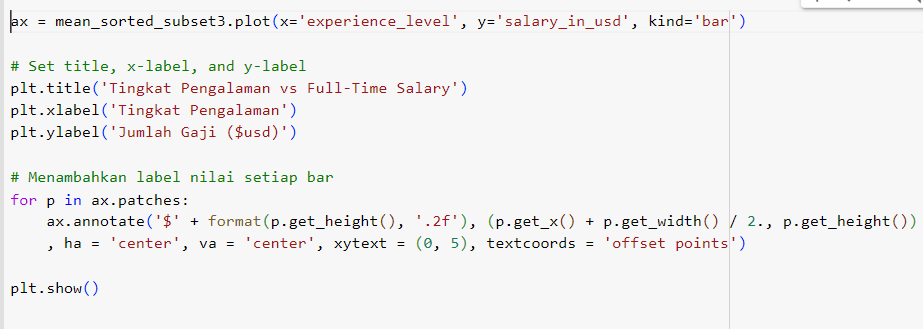
**3.2.3 Bar Graph Penghasilan Full Time Data Scientist**

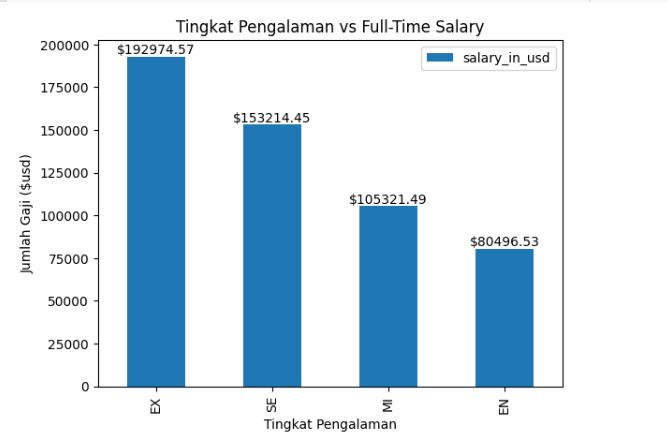
****



Menurut tabel Mean Pendapatan Data Science Full-Time Tertinggi dapat disimpulkan bahwa rata-rata penghasilan terbesar adalah pekerjaan AI Scientist dengan total gaji lebih dari $400.000. Sedangkan rata-rata penghasilan terkecil ada pada pekerjaan NLP Engineer dengan total gaji kurang dari $300.0000.

**3.2.4 Grafik Bar Graph Pengalaman dengan Salary**



****

Bar Graph merupakan grafik yang mewakili kategori atau variabel yang berbeda. Beberapa fungsi bar graph adalah membandingkan nilai atau frekuensi antara kategori yang berbeda, memvisualisasikan perubahan nilai, dan menampilkan pola atas perbedaan dari 2 variabel. Dari output yang dihasilkan setiap bar pada sumbu x mewakili tingkat pengalaman yang terdiri dari Expert, Senior, Mid Level, dan Entry Level. Sedangkan sumbu y mewakili jumlah gaji (salary).

Dari grafik dapat diketahui bahwa Tingkat Pengalaman sangat berpengaruh terhadap gaji. Gaji paling tinggi dimiliki oleh level pengalaman Expert dan gaji terendah dimiliki oleh level pengalaman Entry Level. Sehingga dapat disimpulkan bahwa, tiap kenaikan level pengalaman sebanding dengan meningkatnya gaji pendapatan.

**BAB IV**

**KESIMPULAN**

Distribusi histogram penghasilan menunjukkan bahwa penghasilan full time data scientist berdistribusi data skewness dimana nilainya tidak berdistribusi normal dan cenderung positif skewness Frekuensi dimana frekuensi tertinggi ada pada nilai di antara 100.000 dan 200.000 atau ada rentang 130.000 dengan total frekuensi 800. Sedangkan frekuensi terendah ada pada nilai lebih dari 400.000. Sehingga, dapat disimpulkan bahwa dominasi nilai dari gaji adalah $130.000 atau setara dengan Rp1.942.669.500.

Penghasilan seorang data scientist juga terdapat perbedaan bergantung pada tingkat pengalaman seorang data scientist. Semakin expert pengalaman maka semaki tinggi juga penghasilan yang akan didapatkan.

Kesimpulan akhir dari proses data wrangling ini yakni memberikan informasi penting tentang penghasilan Data Scientist berdasarkan kategori waktu bekerja, pengalaman kerja, dan distribusi penghasilan. Informasi ini dapat digunakan sebagai acuan bagi individu yang ingin mengeksplorasi karir dalam bidang Data Science dan mengambil keputusan berdasarkan faktor penghasilan.

**LAMPIRAN**

