**LAPORAN TUGAS BESAR**

**ANALISIS DATA PEKERJAAN DAN PERUSAHAAN DI BIDANG TEKNIK INFORMATIKA**

**Mata Kuliah : Pengenalan Komputasi (KU1102) Stream Pemrograman**

# Dosen Pengampu : Dr. Nur Ulfa Maulidevi S.T., M. Sc.



Disusun Oleh:

Fedry Firman Anugerah 16522048

Fairuz Apuilla Rahagi 16522188

Raden Fransisco Trianto B. 19622078

Muhammad Rafi Dhiyaulhaq 19622158`

SEKOLAH TEKNIK ELEKTRO DAN INFORMATIKA

INSTITUT TEKNOLOGI BANDUNG

BANDUNG

NOVEMBER 2022

1. Latar Belakang

Data merupakan suatu hal yang tak pernah luput dari aspek kehidupan kita sehari-hari. Data adalah kenyataan yang mewakili suatu kejadian serta merupakan suatu bentuk yang masih mentah yang belum dapat bercerita banyak sehingga perlu diolah lebih lanjut melalui suatu model untuk menghasilkan informasi (Tata Sutabri, 2012). Data ada dimana-mana, mulai dari lembaga pemerintahan, kependudukan, institusi pendidikan, hingga lingkup yang paling kecil sekalipun misalnya identitas diri seperti nama, usia, dan alamat tempat tinggal. Data tak bisa dipisahkan dalam hidup manusia, karenanya sangat penting bagi kita untuk memahami apa dan bagaimana data tersebut bekerja.

Berdasarkan sifatnya, data dibagi menjadi dua yakni data kualitatif dan data kuantitatif. Data kualitatif ialah data atau informasi yang berbentuk deskriptif dan tidak bisa diukur dengan angka. Sementara itu, data kuantitatif adalah sekumpulan informasi yang dapat diukur, dihitung, dan dibandingkan dengan konteks numerikal. Adapun perbedaan antara data kualitatif dan kuantitatif secara garis besar disajikan dalam tabel berikut.

|  |  |
| --- | --- |
| Data Kualitatif | Data kuantitatif |
| Deskriptif, berhubungan dengan kata-kata | Dapat dihitung/diukur dengan angka |
| Menjawab pertanyaan “bagaimana” dan “mengapa” | Menjawab pertanyaan “apa” dan “berapa" |
| Dinamis, subjektif, dan mampu diinterpretasikan | Bersifat umum/universal dan faktual |
| Cara pengumpulan data dengan observasi dan *interview* | Cara pengumpulan data dengan pengukuran dan perhitungan |
| Dianalisis dengan mengelompokkan data hingga menjadi kategori | Dianalisis dengan analisis statistik |

Tabel 1.1: Perbedaan antara Data Kualitatif dan Data Kuantitatif

Pada laporan ini, digunakan data kuantitatif sebagai *source data* agar dapat dianalisis secara statistik dengan memanfaatkan module pandas sebagai library pada python. Nantinya, akan dijelaskan dan ditampilkan mengenai masing-masing bagian data dengan tujuan memahami bagaimana data tersebut mendeskripsikan informasi yang terkandung di dalamnya.

1. Isi Laporan
2. Deskripsi Dataset

Dataset yang digunakan ialah data mengenai informasi pekerjaan informatika dan perusahaannya di Amerika Serikat. Dataset ini diambil dari *Kaggle.com* berformat csv. Alasan pemilihan dataset ini sebagai dataset tugas besar dikarenakan tema dari data ini berhubungan dengan teknik informatika dan memenuhi syarat dataset harus digunakan. Selain itu, dataset ini bersifat kuantitatif yang mana akan mudah untuk dianalisis dengan metode perhitungan numerik serta dapat menjelaskan antar variabel yang terdapat di dalamnya.

* Nama file data: data\_2021.csv
* Deskripsi data: Data pekerja dan perusahaannya di bidang teknologi informatika
* Format data: csv
* Sumber dataset: Kaggle
* Dimensi data: 742 baris x 41 kolom
* Ukuran file: 3.051 kb

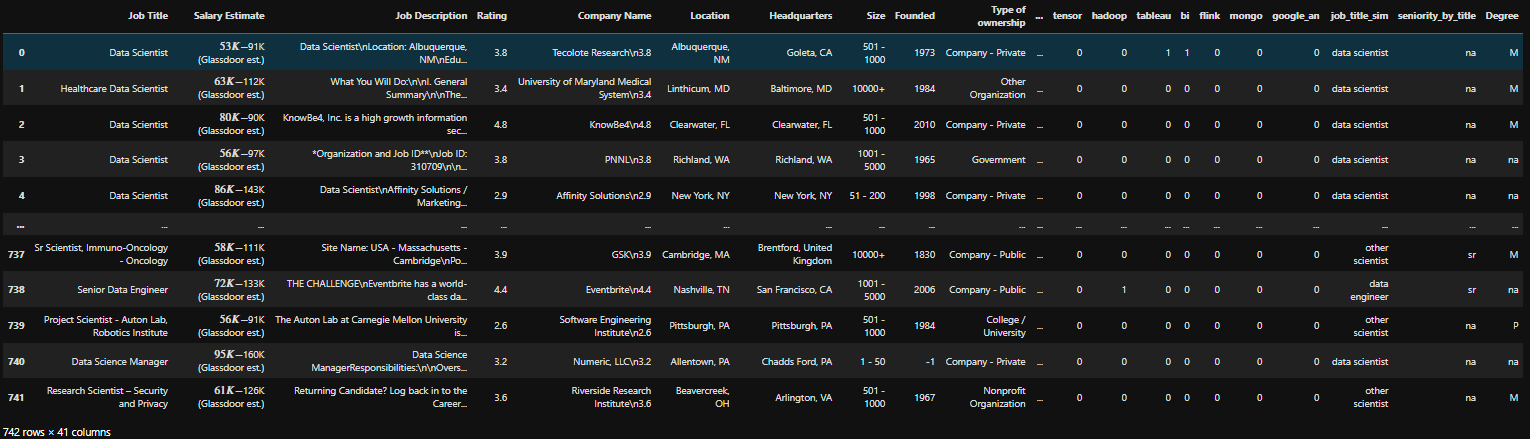
1. Analisis data dengan bahasa Python di Jupyter Notebook:
   1. Membaca / loading data

import pandas as pd

col = list(pd.read\_csv("data\_2021.csv",nrows = 1).columns)

df = pd.read\_csv("data\_2021.csv", usecols=[ i for i in col if i != "index"])

df



Deskripsi:

* Import pandas untuk melakukan pembacaan dan analisis data
* Mendapatkan list semua nama kolom sebagai col yang diperlukan karena pada kolom pertama terdapat index data yang sudah dibersihkan, berarti tidak berurut dan tidak berguna sehingga tidak perlu dibaca
* Membaca semua data kecuali colom index pada data

1. Dimensi data

len(df) # jumlah baris

# output

742

len(df.columns) # jumlah kolom

# output

41

1. Penjelasan isi atribut/kolom
   1. Job Title → nama pekerjaan di perusahaan
   2. Salary Estimate → perkiraan gaji tahunan
   3. Job Description → deskripsi tugas/pekerjaan tersebut
   4. Rating → Penilaian pekerja terhadap perusahaan tempat bekerja
   5. Company Name → nama perusahaan
   6. Location → lokasi perusahaan
   7. Headquarters → lokasi pusat perusahaan
   8. Size → skala banyak pekerja dalam perusahaan
   9. Founded → tahun dibangunnya perusahaan
   10. Type of ownership → tipe kepemilikan
   11. Industry → bidang yang didalami perusahaan
   12. Sector → sektor yang dialami dari bidang perusahaan
   13. Revenue → penghasilan per tahun perusahaan
   14. Competitors → kompetitor perusahaan (jika ada)
   15. Hourly → keterangan dibayar per jam (jika ada)
   16. Employer provided → karyawan yang disediakan (jika ada)
   17. Lower Salary → gaji paling rendah
   18. Upper Salary → gaji paling tinggi
   19. Avg Salary(K) → gaji rata-rata
   20. Company\_txt → nama perusahaan resmi
   21. Job Location → code lokasi (code state US, contoh: NY {New York})
   22. Age → umur perusahaan
   23. Job\_title\_sim → nama resmi pekerjaan
   24. Seniority\_by\_title → ada tidaknya senioritas (konsep senior-junior)
   25. Degree → gelar pendidikan

Untuk di bawah ini adalah bahasa pemrograman, aplikasi, library dan alat lainnya dalam bidang informatika, data dalam bentuk boolean(1 jika dipakai/iya, 0 jika tidak dipakai)

1. Python

aa. Spark

ab. Aws

ac. Excel

ad. Sql

ae. Sas

af. Keras

ag. Pytorch

ah. Scikit

ai. Tensor

aj. Hadoop

ak. Tableau

al. Bi

am. Flink

an. Mongo

ao. Google\_an

1. Bentuk data atribut/kolom
   1. **Categorical-Nominal:**

Job Title, Job Description, Company Name, Location, Headquarters, Founded, Type of ownership, Industry, Sector, Revenue, Competitors, company\_txt, Job Location,  job\_title\_sim, seniority\_by\_title, Degree

* 1. **Categorical-Ordinal**

Size

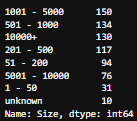
Size merupakan kisaran jumlah pegawai di perusahaan

* + - 1. 1 - 50 jumlah : 31
      2. 51 - 200 jumlah : 94
      3. 201 - 500 jumlah : 117
      4. 501 - 1000 jumlah : 134
      5. 1001 - 5000 jumlah : 150
      6. 5001 - 10000 jumlah : 76
      7. 10000+ jumlah : 130
      8. Unknown jumlah : 10

dengan data kosong.

df["Size"].value\_counts()

# output



1. **Categorical-Binary**

Hourly, Employer provided, Python, spark, aws, excel, sql, sas, keras, pytorch, scikit, tensor, hadoop, tableau, bi, flink, mongo, google\_an

* 1. (1 digunakan, 0 tidak digunakan) : Python, spark, aws, excel, sql, sas, keras, pytorch, scikit, tensor, hadoop, tableau, bi, flink, mongo, google\_an
  2. (1 ada, 0 tidak ada) : Hourly, Employer provided

1. **Quantitative-Discrete**

Rating

Rating Range : [-1, – 5]

-1 berarti data tidak diisi

Data 98.51% terisi dan 1.49% tidak terisi

count = 0

empty = 0

for i in df["Rating"]:

if(i == -1):

empty += 1

count += 1

(1- empty/count)\*100

# output

98.51752021563343

1. **Quantitative-Continues**

Founded, Lower Salary, Upper Salary, Avg Salary, Age, Salary Estimate

1. Founded Range : [-1, - 2021]

-1 berarti data tidak diisi

Data 93.26% terisi dan 6.74% tidak terisi

count = 0

empty = 0

for i in df["Founded"]:

if(i == -1):

empty += 1

count += 1

(1- empty/count)\*100

# output

93.26145552560648

1. Lower salary Range : 15 - 202
2. Upper salary Range : 16 - 306
3. Avg salary Range :  15.5 - 254

# Template kode untuk mencari nilai minimum dan maksimum data kuantitatif

df[<nama kolom>].min()

# output

{nilai minimum}

df[<nama kolom>].max()

# output

{nilai maksimum}

1. Salary Estimate :  dalam format {$<angka1>K-$<angka2>K}  
   Namun tidak semua isi pasti berformat seperti itu maka perlu memakai regular expression (regEx) untuk mengeluarkan angkanya saja

import re

salary\_est = []

for est in df["Salary Estimate"]:

    est = est.split("-")

    rnge = [0,0]

    count = 0

    for i in est:

        num = re.findall("\d",i)

        n = ""

        for a in num:

            n += a

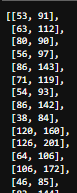
        rnge[count] = int(n)

        count += 1

    salary\_est.append(rnge)

Salary\_est

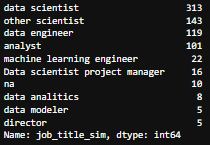
# output



1. Job\_title\_sim :
   1. data scientist jumlah : 313
   2. other scientist jumlah : 143
   3. data engineer                    jumlah : 119
   4. analyst                           jumlah : 101
   5. machine learning engineer          jumlah : 22
   6. Data scientist project manager     jumlah : 16
   7. data analitics                      jumlah : 8
   8. data modeler                        jumlah : 5
   9. director                            jumlah : 5
   10. na                                jumlah : 10

df["job\_title\_sim"].value\_counts()

# output



1. Sampel Data
2. 5 pekerjaan dengan rating tertinggi dan terendah

|  |
| --- |
| col = list(pd.read\_csv("D:/data\_cleaned\_2021.csv",nrows = 1).columns)  df = pd.read\_csv("D:/data\_cleaned\_2021.csv", usecols=[ i for i in col if i != "index"])  dfr = df.sort\_values(["Rating"],ascending=[0])  # Data 5 baris pertama berdasarkan rating  dfr.loc[dfr["Rating"] > 0] |

|  |
| --- |
|  |

1. 5 pekerjaan dengan rata rata gaji tertinggi dan terendah (dalam ribu USD)

|  |
| --- |
| df1 = df.sort\_values(["Avg Salary(K)"],ascending=[0])  df2 = df1[["Job Title", "Avg Salary(K)"]]  df2.head(5)                                                         df2.tail(5) |

1. 5 pekerjaan dengan gaji atasan tertinggi dan terendah (dalam ribu USD)

|  |
| --- |
| df1 = df.sort\_values(["Upper Salary"],ascending=[0])  df2 = df1[["Job Title", "Upper Salary"]]  df2.head(5)                                                         df2.tail(5) |

1. 5 pekerjaan dengan gaji bawahan tertinggi dan terendah (dalam ribu USD)

|  |
| --- |
| df1 = df.sort\_values(["Lower Salary"],ascending=[0])  df2 = df1[["Job Title", "Lower Salary"]]  df2.head(5)                                                         df2.tail(5) |

Berdasarkan data diatas, pekerjaan di bidang teknologi informatika memiliki range yang cukup luas. Untuk mengetahui lebih lanjut mengenai pemasukan dari pekerjaan tersebut, kita dapat menganalisis datanya

Rata-rata data:

Rating            3.618868

Lower Salary     74.754717

Upper Salary     128.214286

Avg Salary(K)    101.484501

Standar deviasi data

Ukuran sebaran data-data

Rating            0.801210

Lower Salary      30.945892

Upper Salary 45.128650

Avg Salary(K) 37.482449

Persentil data

|  |
| --- |
| df3 = df[["Rating","Lower Salary","Upper Salary","Avg Salary(K)"]] |
| df3.quantile(0.1) #10%  #output  Rating            2.9  Lower Salary     40.0  Upper Salary     72.0  Avg Salary(K)    56.5 |
| df3.quantile(0.25) #20%  #output  Rating            3.3  Lower Salary     52.0  Upper Salary     96.0  Avg Salary(K)    73.5 |
| df3.quantile(0.5) #50%  #output  Rating             3.7  Lower Salary      69.5  Upper Salary     124.0  Avg Salary(K)     97.5 |
| df3.quantile(0.75) #75%  #output  Rating             4.0  Lower Salary      91.0  Upper Salary     155.0  Avg Salary(K)    122.5 |
| df3.quantile(0.9) 90%  #output  Rating             4.4  Lower Salary     114.0  Upper Salary     189.0  Avg Salary(K)    151.4 |

Berdasarkan data di atas, dapat disimpulkan bahwa kita cukup berada pada persentil top 80% untuk memenuhi biaya rata-rata hidup di amerika yang sebesar 66,928 USD\* per tahunnya jika kita menggunakan gaji rata-rata. Bahkan jika seseorang menjadi pekerja atas dari pekerjaan tersebut, pekerja cukup masuk pada persentil top 90%. Statistik diatas sangat menjelaskan bahwa pekerjaan di bidang informatika memiliki peluang pendapatan yang sangat besar (\*Statista, 2021)

Distribusi frekuensi nilai pada data

#Jumlah frekuensi berdasarkan sektor pekerjaan

df["Sector"].value\_counts()

#output

Information Technology                180

Biotech & Pharmaceuticals             112

Business Services                   97

Insurance                              69

Health Care                           49

Finance                                42

Manufacturing                          34

Aerospace & Defense                    25

Education                              23

Retail                                 15

Oil, Gas, Energy & Utilities           14

Government                             11

-1                                     10

Non-Profit                              9

Transportation & Logistics              8

Real Estate                             8

Travel & Tourism                        8

Telecommunications                      6

Media                                   6

Arts, Entertainment & Recreation        4

Consumer Services                   4

Mining & Metals                       3

Construction, Repair & Maintenance     3

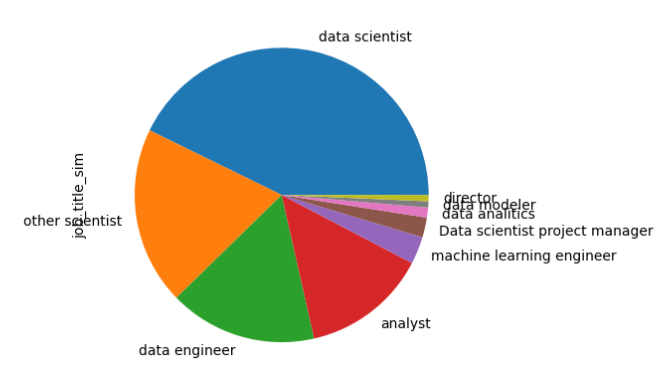
Agriculture & Forestry             1

Accounting & Legal                      1

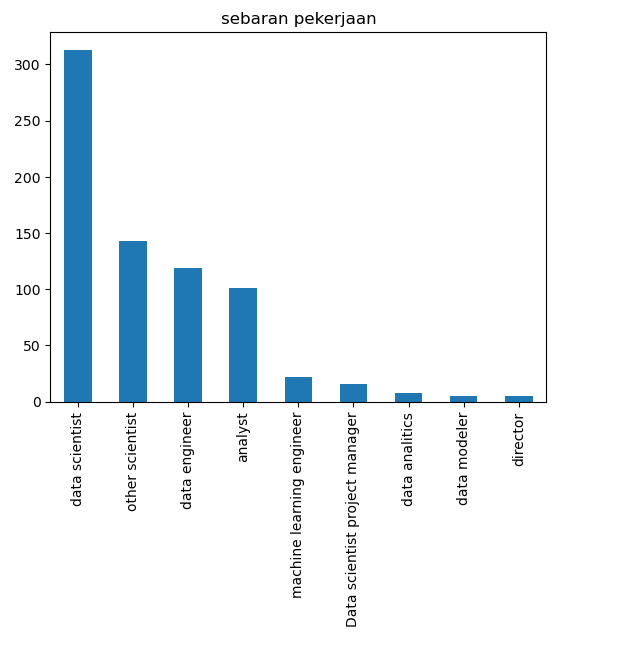
|  |
| --- |
| #Jumlah frekuensi berdasarkan sektor pekerjaan  df["Degree"].value\_counts()  #Output  na    383  M     252  P     107 |

1. Visualisasi

Sebaran Pekerjaan Menggunakan Pie Chart

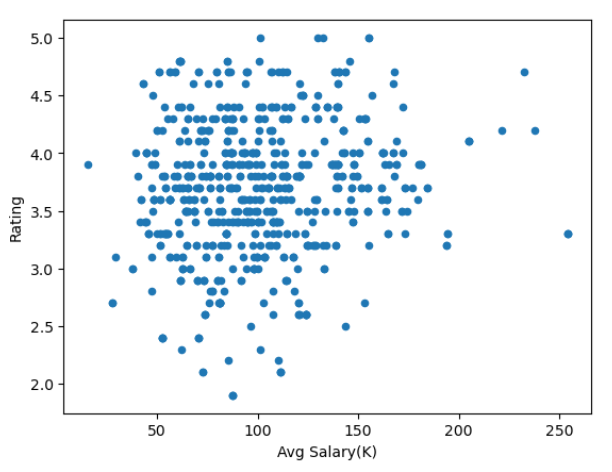


Sebaran Pekerjaan Menggunakan Bar Plot

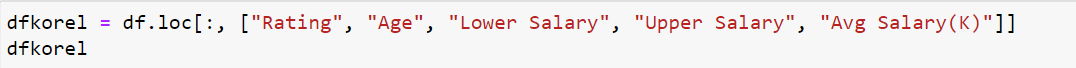


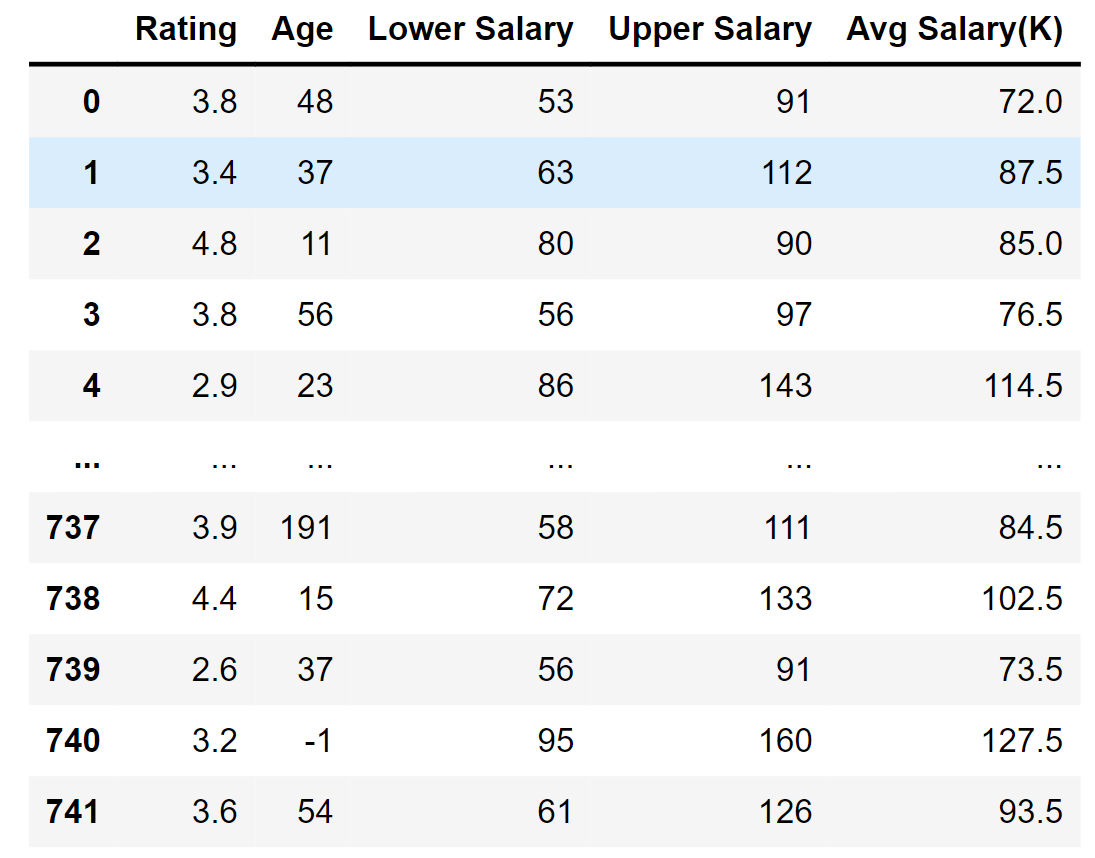
Sebagian besar pekerjaan di bidang informatika menjadi data scientist. Dua bagian besar lainnya yaitu data engineer dan analyst

Scatter Plot antara rata-rata gaji dengan rating pekerjaan

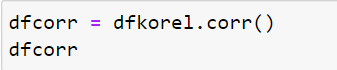
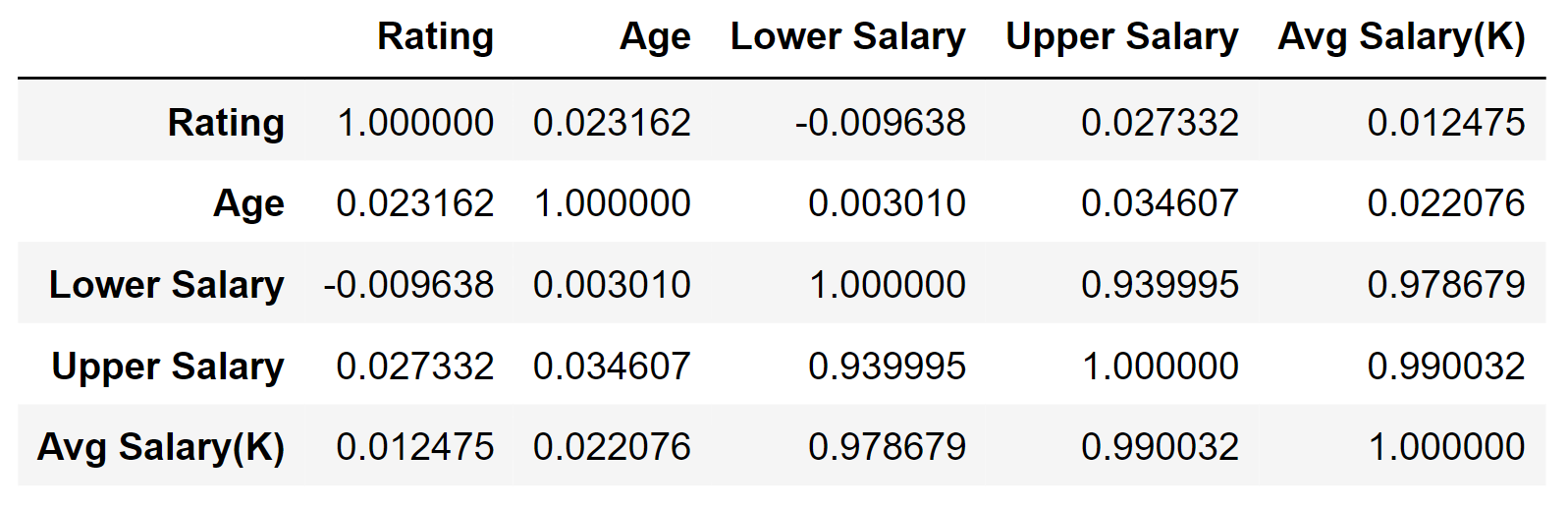


1. Analisis Korelasi

Korelasi yang kami analisis pada dataset yang digunakan hanyalah pada atribut kuantitatif yang kemudian difilter kembali agar korelasi yang dihasilkan tidak bias. Atribut yang dikorelasi ialah atribut “Rating”, “Age”, “Upper Salary”, “Lower Salary”, serta “Avg Salary(K)”, sedangkan atribut kuantitatif lain tidak kami korelasikan atribut tersebut bersifat binary categorical yang hanya berisi yes/no.

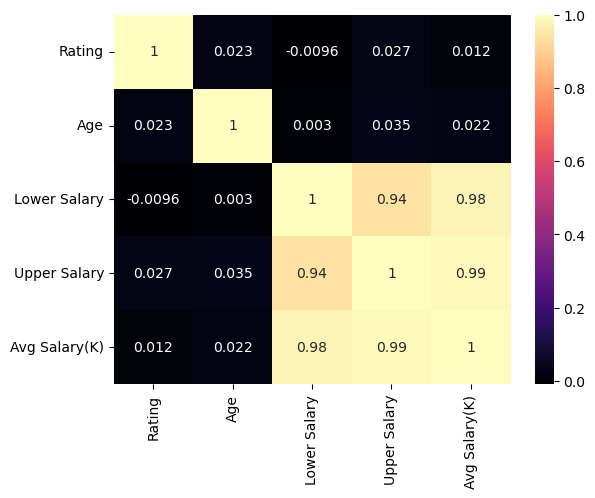


1. Tabel Korelasi antar Atribut



Pada tabel korelasi antar atribut ini, kami memfilter atribut mana saja yang akan dikorelasikan, kemudian digunakan fungsi matplotlib.pyplot.corr() untuk menampilkan tabel korelasinya seperti yang ada pada gambar di atas.

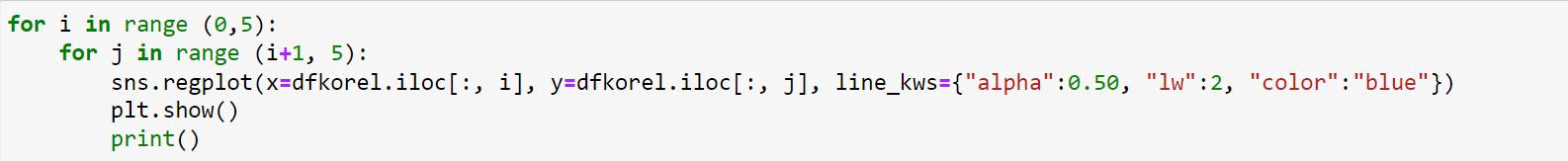
1. *Heatmap* Korelasi



Untuk menampilkan heatmap korelasi antar atribut berikut digunakan library seaborn. Warna ungu gelap menuju hitam menunjukkan koefisien korelasi mendekati nol yang berarti kurang/tidak berkorelasi, sedangkan warna jingga menuju krem muda menunjukkan koefisien korelasi mendekati nol yang berarti hampir sempurna berkorelasi lurus. Pada hubungan korelasi antar atribut tidak ditemukan adanya atribut yang saling berkorelasi berbanding terbalik (mendekati minus satu).

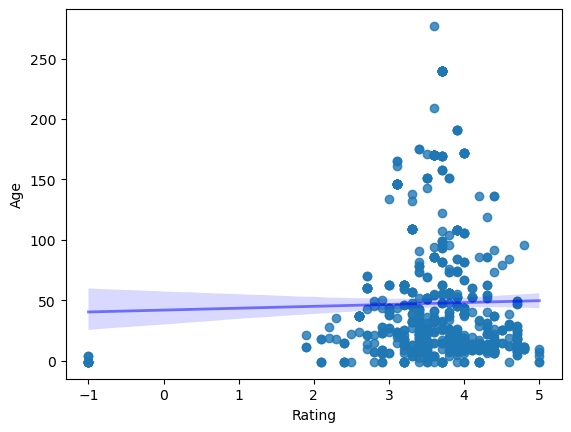
Dari *heatmap* di atas, dapat disimpulkan bahwa antar atribut di atas menunjukkan hubungan korelasi berbanding lurus dan tidak berkorelasi. Sementara itu, tidak ada atribut yang menunjukkan sifat korelasi berbanding terbalik. Hal itu diketahui dari nilai koefisien antar atribut yang kebanyakan mendekati 1 atau mendekati 0, namun tidak mendekati minus satu (-1).

Adapun keragaman korelasi yang ditunjukkan sedikit atau monoton (kurang/tidak tersebar). Nilai korelasi terkonsentrasi dekat 0 atau 1, yang mana dapat disimpulkan bahwa atribut yang ada pada data bersifat cenderung acak atau berpola sama (nilai suatu atribut semakin besar, nilai atribut lain juga semakin besar, dan kebalikannya).

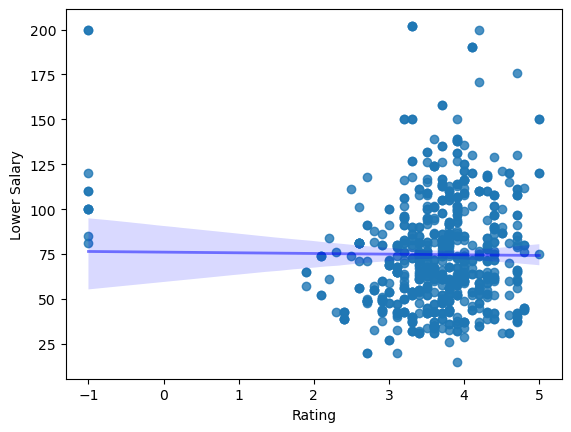
1. Grafik Korelasi Antar Atribut

Untuk menampilkan grafik korelasi, digunakan juga library seaborn agar dapat menunjukkan regresi linear pada grafik *scatter plot*. Untuk menampilkan masing-masing grafik, kami menggunakan for looping. Berikut masing-masing grafik korelasi antar atribut.

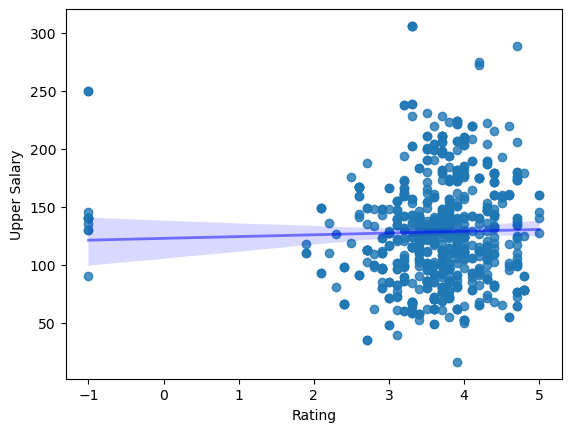
* Grafik Rating terhadap Age



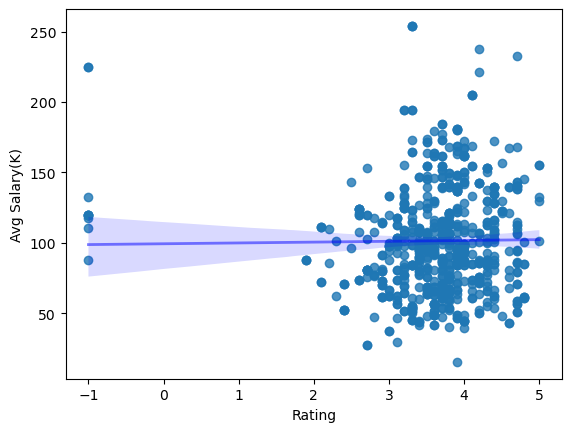
* Grafik Rating terhadap Lower Salary



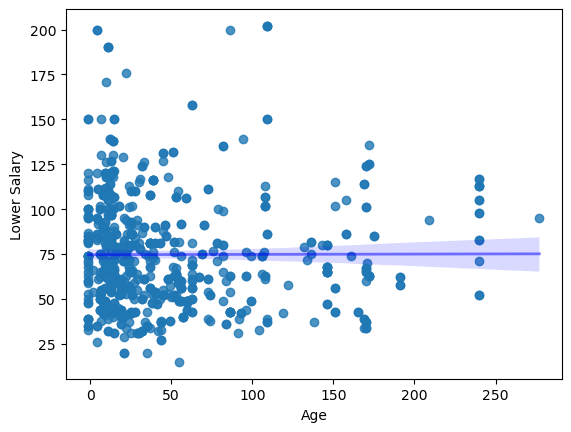
* Grafik Rating terhadap Upper Salary



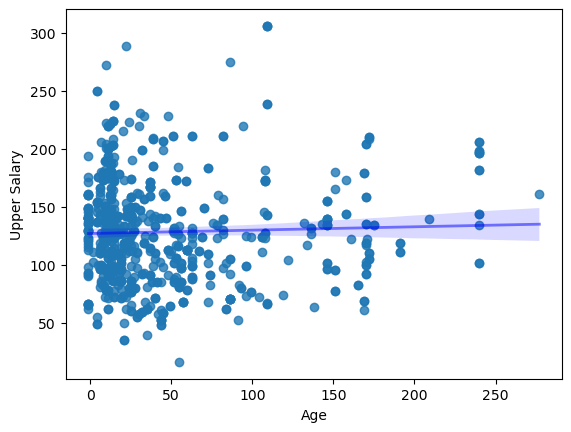
* Grafik Rating terhadap Avg Salary(K)



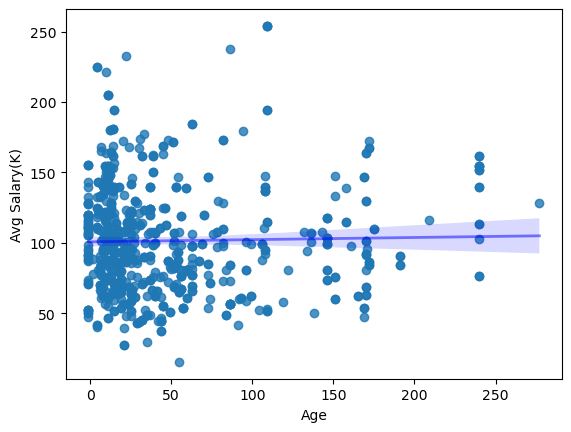
* Grafik Age terhadap Lower Salary



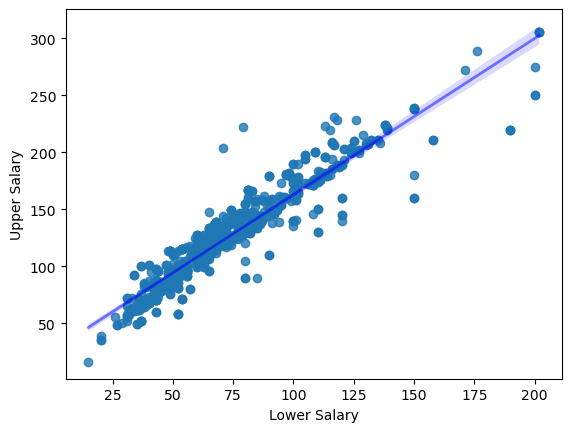
* Grafik Age terhadap Upper Salary



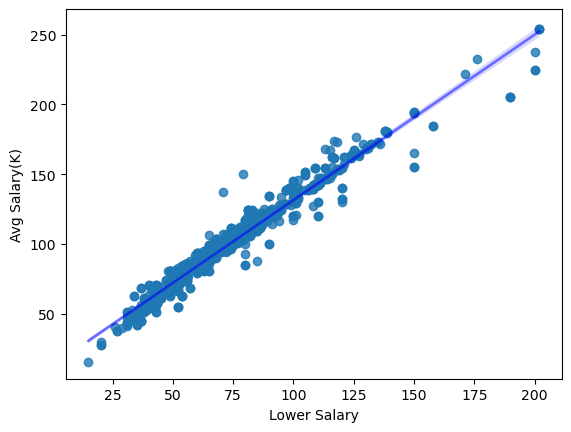
* Grafik Age terhadap Avg Salary(K)



* Grafik Lower Salary terhadap Upper Salary



* Grafik Lower Salary terhadap Avg Salary(K)



* Grafik Upper Salary terhadap Avg Salary(K)

