Final Project

Exploratory Data Analysis of

Data Scientist Salary

DS 23B

Irma Rahma Suci

www.linkedin.com/in/irma-rahma-suci

Data Scientist Salary

Data Science menjadi bidang pekerjaan yang berkembang pesat saat ini. Data Scientist memiliki peran yang sangat penting dalam menganalisa, mengeksplorasi data sampai mendapatkan insight yang diperlukan dalam pengambilan sebuah keputusan dalam industrial pekerjaan. Hampir semua bidang industry kerja membutuhkan peran Data Scientist saat ini. Untuk itu, saya ingin mengetahui factor-factor yang mempengaruhi gaji sebagai Data Scientist.

Data set diambil melalui Kaggle.

https://www.kaggle.com/datasets/henryshan/2023-data-scientists-salary

Table of Content

About Data Science

Pendahuluan

Perkembangan ilmu data science dapat ditelusuri kembali ke tahun 1960-an ketika para ilmuwan komputer mulai mempertimbangkan pentingnya data dalam pengambilan keputusan. Namun, pada saat itu, teknologi komputer masih terbatas dan tidak ada alat yang cukup canggih untuk memproses dan menganalisis data dengan cepat.

Definisi

Data science adalah ilmu yang mencakup pengumpulan, pengorganisasian, pemrosesan, dan analisis data untuk menghasilkan informasi yang berguna. Ilmu ini melibatkan penerapan metode ilmiah, algoritma, dan teknologi komputer untuk mengeksplorasi dan memecahkan masalah yang melibatkan data

Peran

Data scientist adalah profesi yang bertanggung jawab untuk menerapkan metode data science dalam konteks bisnis atau ilmu pengetahuan. Mereka mengumpulkan dan menganalisis data, mengembangkan model prediksi, dan menyajikan hasilnya kepada pemangku kepentingan.

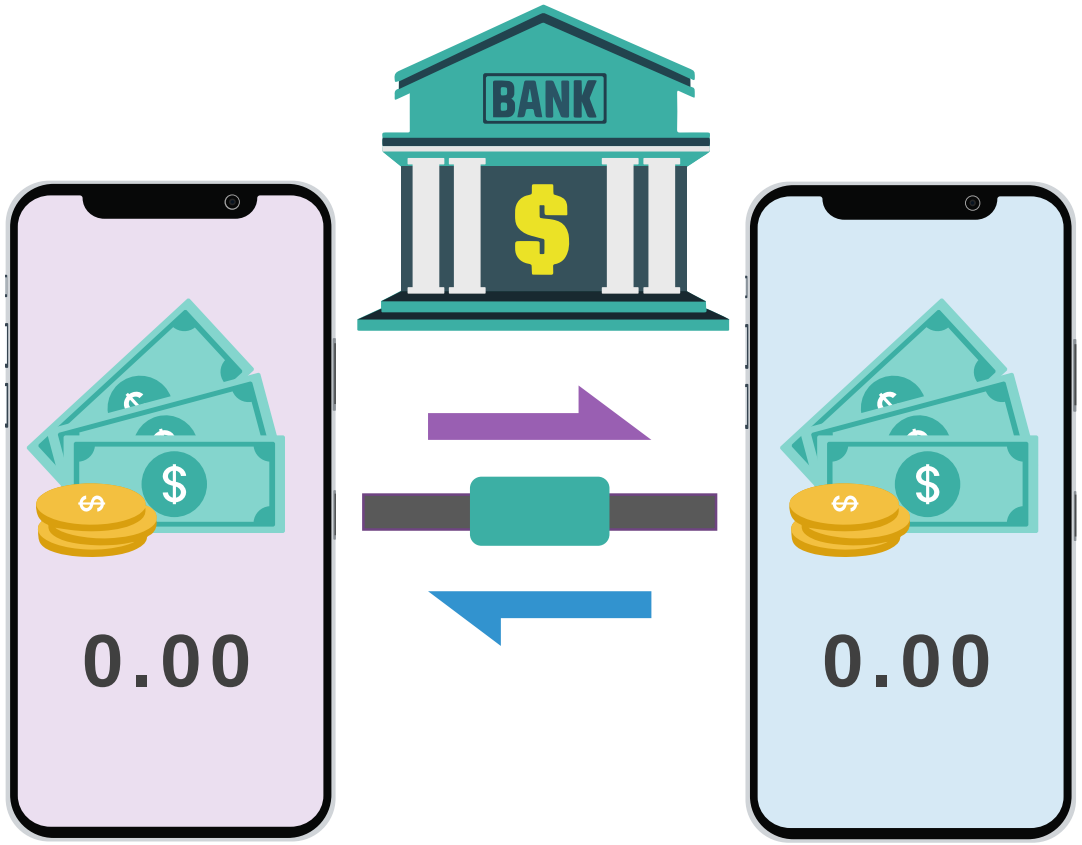
Era Big Data

Perkembangan yang signifikan dalam ilmu data science terjadi pada tahun 2000-an dengan munculnya fenomena "big data". Big data merujuk pada jumlah data yang sangat besar dan kompleks yang tidak dapat diproses dengan menggunakan alat tradisional. Pada saat ini, para ilmuwan data mulai mengembangkan metode dan algoritma baru untuk mengolah data dalam skala

Kesimpulan

Dengan peningkatan kapasitas penyimpanan data, pengembangan algoritma dan teknik analisis data yang canggih, dan perkembangan teknologi komputasi, ilmu data science dapat mengolah dan menganalisis data dengan lebih efisien. Peran penting big data juga tidak bisa diabaikan dalam perkembangan ilmu data science ini. Selain itu, implementasi data science di berbagai bidang juga menunjukkan betapa pentingnya ilmu ini dalam mengambil keputusan yang lebih baik dan efisien. Dengan perkembangan yang terus berlanjut, ilmu data science akan terus menjadi bidang yang sangat menarik dan berpotensi besar di masa depan.

Data Understanding

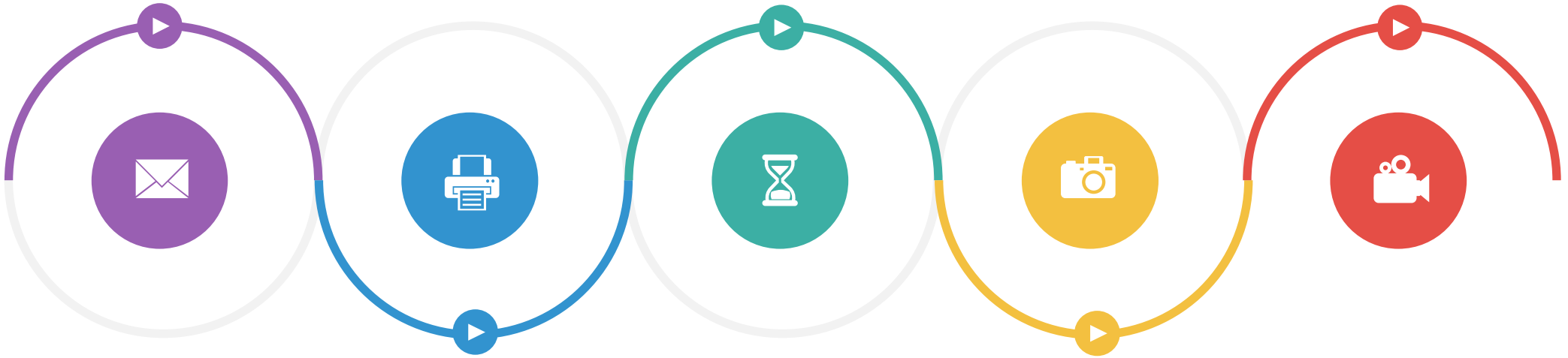


Statistical Summary

Univariate Analysis

Multivariate Analysis

Predictor Variables



Experience Level

Employment Type

Job Title

Salary in USD

Company Location

Dataset

What are Top 10 job_title with experience_level?

```
# general info
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3755 entries, 0 to 3754
Data columns (total 11 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   work_year           3755 non-null   int64
 1   experience_level    3755 non-null   object
 2   employment_type     3755 non-null   object
 3   job_title           3755 non-null   object
 4   salary              3755 non-null   int64
 5   salary_currency     3755 non-null   object
 6   salary_in_usd       3755 non-null   int64
 7   employee_residence  3755 non-null   object
 8   remote_ratio        3755 non-null   int64
 9   company_location    3755 non-null   object
 10  company_size        3755 non-null   object
dtypes: int64(4), object(7)
memory usage: 322.8+ KB
```

Cleaning Data

```
[6]  df.isnull().sum()
```

```
work_year             0
experience_level      0
employment_type       0
job_title             0
salary                0
salary_currency       0
salary_in_usd         0
employee_residence    0
remote_ratio          0
company_location      0
company_size          0
dtype: int64
```

Drop the columns that are not needed

```
[7]  df = df.drop(['salary_currency', 'salary'], axis = 1)
```

Job Position

Top 10 Job Positions Analysis

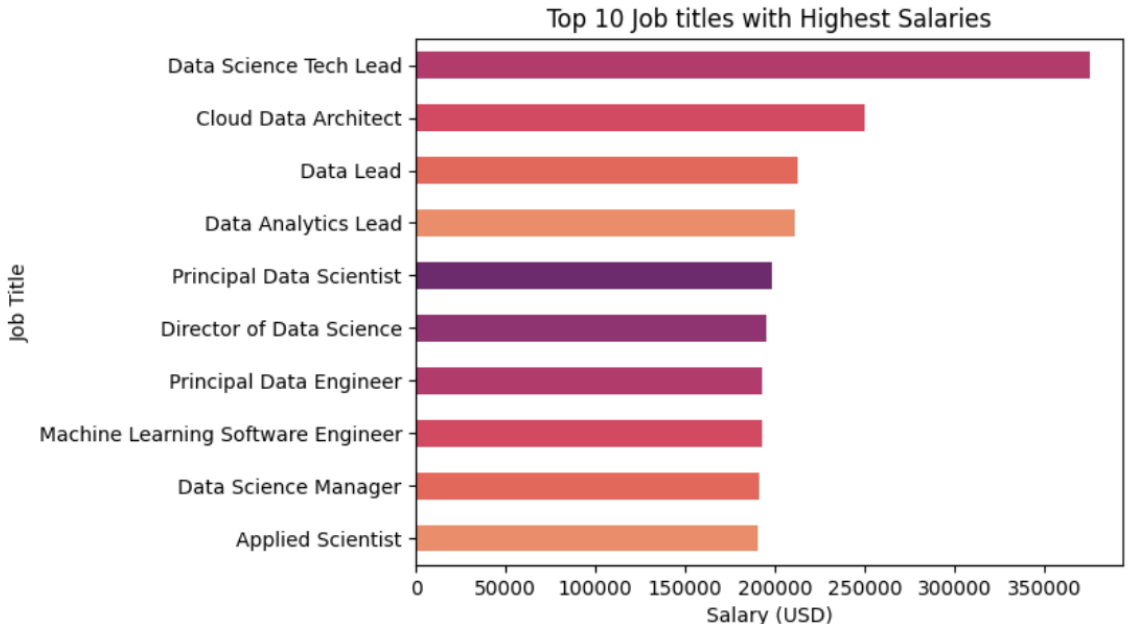
```
df['job_title'].value_counts().head(10)
```

```
Data Engineer                1040
Data Scientist                840
Data Analyst                  612
Machine Learning Engineer     289
Analytics Engineer            103
Data Architect                101
Research Scientist             82
Data Science Manager           58
Applied Scientist              58
Research Engineer              37
Name: job_title, dtype: int64
```

Top 10 Job Positions Analysis

```
job_title_salary = df.groupby('job_title')['salary_in_usd'].mean()
job_title_salary = job_title_salary.sort_values().tail(10)
job_title_salary
```

```
job_title
Applied Scientist                    190264.482759
Data Science Manager                 191278.775862
Machine Learning Software Engineer   192420.000000
Principal Data Engineer              192500.000000
Director of Data Science             195140.727273
Principal Data Scientist             198171.125000
Data Analytics Lead                  211254.500000
Data Lead                            212500.000000
Cloud Data Architect                 250000.000000
Data Science Tech Lead               375000.000000
Name: salary_in_usd, dtype: float64
```



Posisi Data Science Tech Lead memiliki gaji tahunan tertinggi USD 375.000

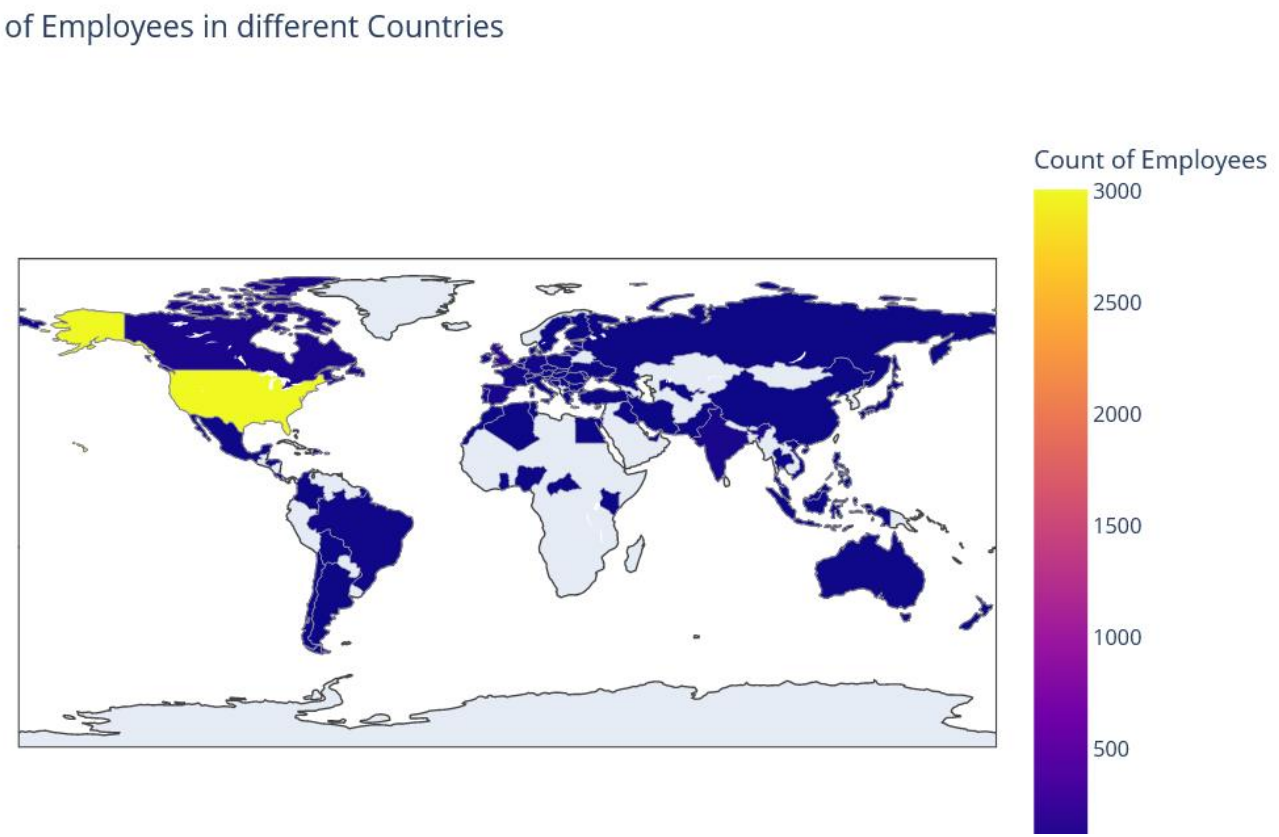
Location

Top 10 Location

```
employee_residence = df.groupby("ISO-3").size().reset_index(name="No. of Employees")
employee_residence.sort_values(by="No. of Employees", ascending=False, inplace=True)
employee_residence
```

	ISO-3	No. of Employees
75	USA	3004
29	GBR	167
12	CAN	85
25	ESP	80
37	IND	71
...	...	...
43	JEY	1
41	ISR	1
40	IRQ	1
33	HND	1
39	IRN	1

78 rows × 2 columns

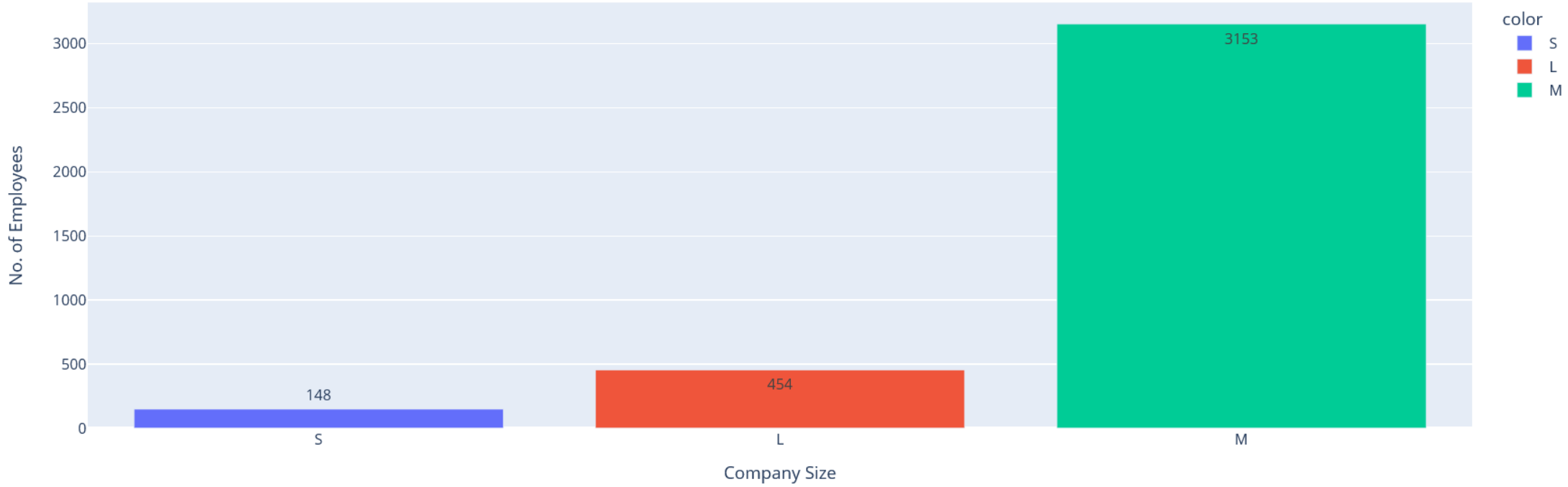


Dari data, jumlah terbanyak terdapat di negara USA, dengan jumlah 3.004 orang. Diikuti dengan Inggris 167 orang, Kanada 85 orang, Spanyol 80 orang dan India 71 orang.

	ISO-3	No. of Employees	Country
75	USA	3004	United States
29	GBR	167	United Kingdom
12	CAN	85	Canada
25	ESP	80	Spain
37	IND	71	India
20	DEU	48	Germany
28	FRA	38	France
10	BRA	18	Brazil
63	PRT	18	Portugal
31	GRC	16	Greece
57	NLD	15	Netherlands
4	AUS	11	Australia
52	MEX	10	Mexico
42	ITA	8	Italy
59	PAK	8	Pakistan
56	NGA	7	Nigeria
38	IRL	7	Ireland
44	JPN	7	Japan
61	POL	6	Poland
5	AUT	6	Austria
1	ARG	6	Argentina
6	BEL	5	Belgium
62	PRI	5	Puerto Rico
73	TUR	5	Türkiye
66	SGP	5	Singapore
74	UKR	4	Ukraine
69	SVN	4	Slovenia
49	LVA	4	Latvia
16	COL	4	Colombia
65	RUS	4	Russian Federation
13	CHE	4	Switzerland
71	THA	3	Thailand
35	HUN	3	Hungary
64	ROU	3	Romania
0	ARE	3	United Arab Emirates
77	VNM	3	Viet Nam
9	BOL	3	Bolivia, Plurinational State of
21	DNK	3	Denmark
34	HRV	3	Croatia
60	PHL	2	Philippines
14	CHL	2	Chile
70	SWE	2	Sweden
19	CZE	2	Czechia
47	LTU	2	Lithuania
45	KEN	2	Kenya
27	FIN	2	Finland
30	GHA	2	Ghana
3	ASM	2	American Samoa
32	HKG	2	Hong Kong
76	UZB	2	Uzbekistan
11	CAF	2	Central African Republic
18	CYP	1	Cyprus
17	CRI	1	Costa Rica
15	CHN	1	China
36	IDN	1	Indonesia
67	SRB	1	Serbia
68	SVK	1	Slovakia
72	TUN	1	Tunisia
7	BGR	1	Bulgaria
2	ARM	1	Armenia
8	BIH	1	Bosnia and Herzegovina
22	DOM	1	Dominican Republic
58	NZL	1	New Zealand
23	DZA	1	Algeria
55	MYS	1	Malaysia
54	MLT	1	Malta
53	MKD	1	North Macedonia
24	EGY	1	Egypt
51	MDA	1	Moldova, Republic of
50	MAR	1	Morocco
26	EST	1	Estonia
48	LUX	1	Luxembourg
46	KWT	1	Kuwait
43	JEY	1	Jersey
41	ISR	1	Israel
40	IRQ	1	Iraq
33	HND	1	Honduras
39	IRN	1	Iran, Islamic Republic of

Company Size

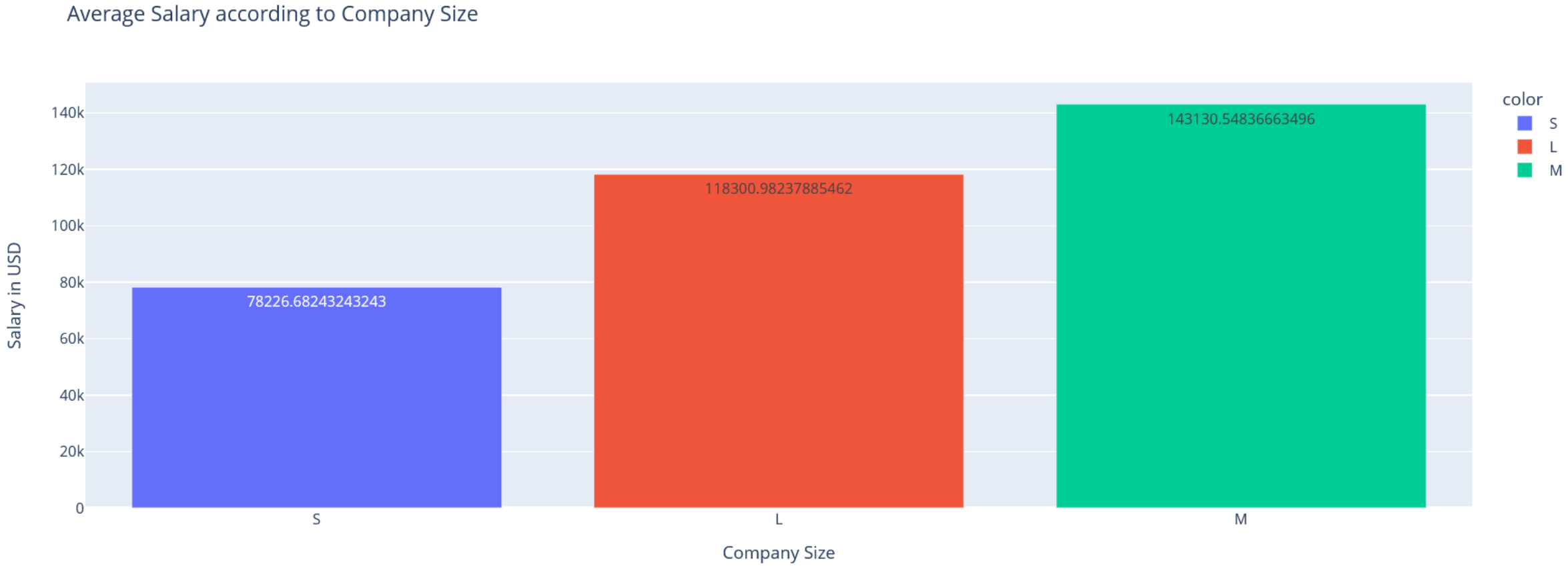
the population of employees per company size



Company Size

average salary range per company size

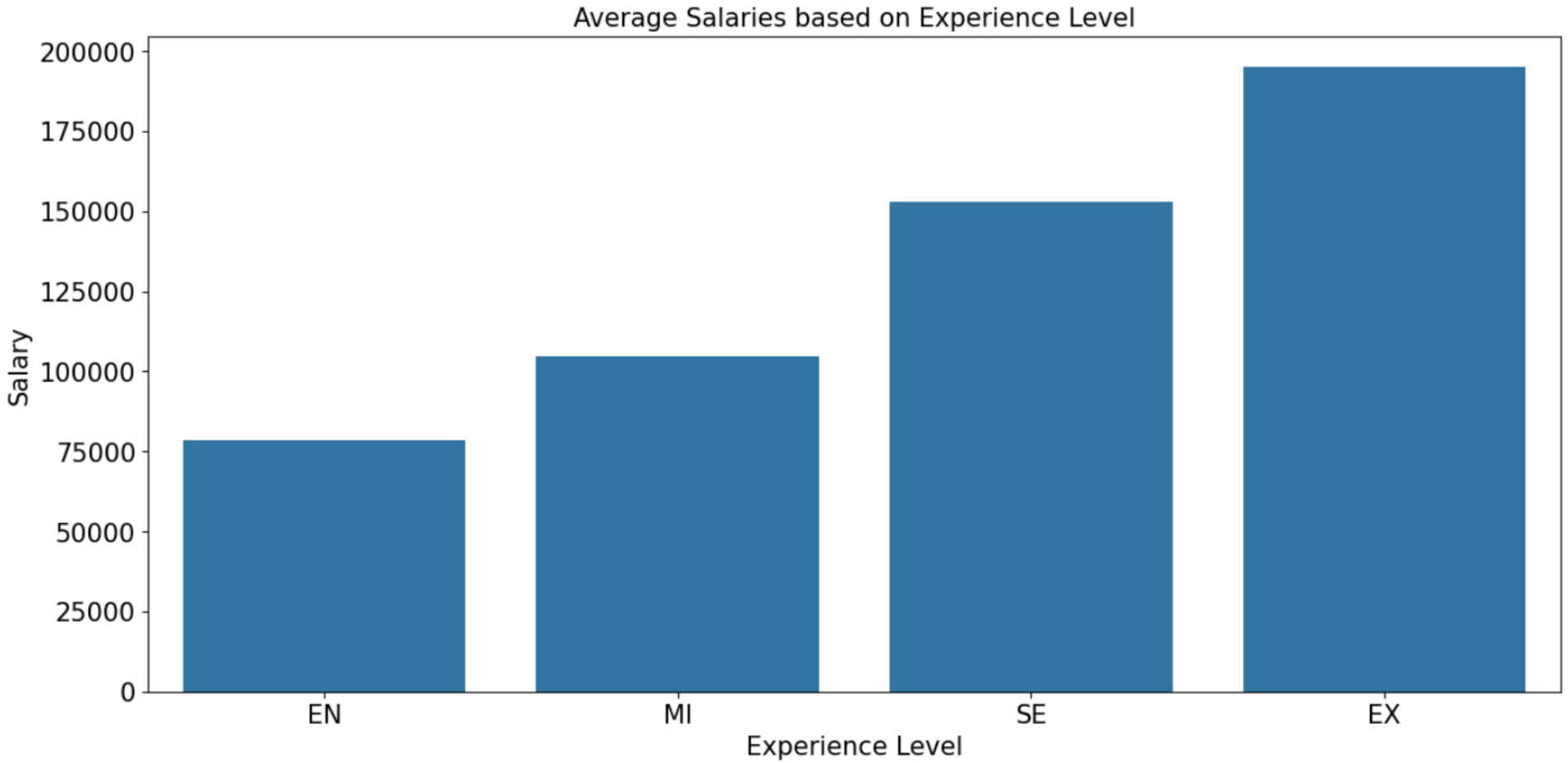
Gaji rata-rata yang lebih tinggi, berada di perusahaan M yaitu USD 143.131 sedangkan perusahaan Besar USD 118.301 dan perusahaan kecil USD 78.227.



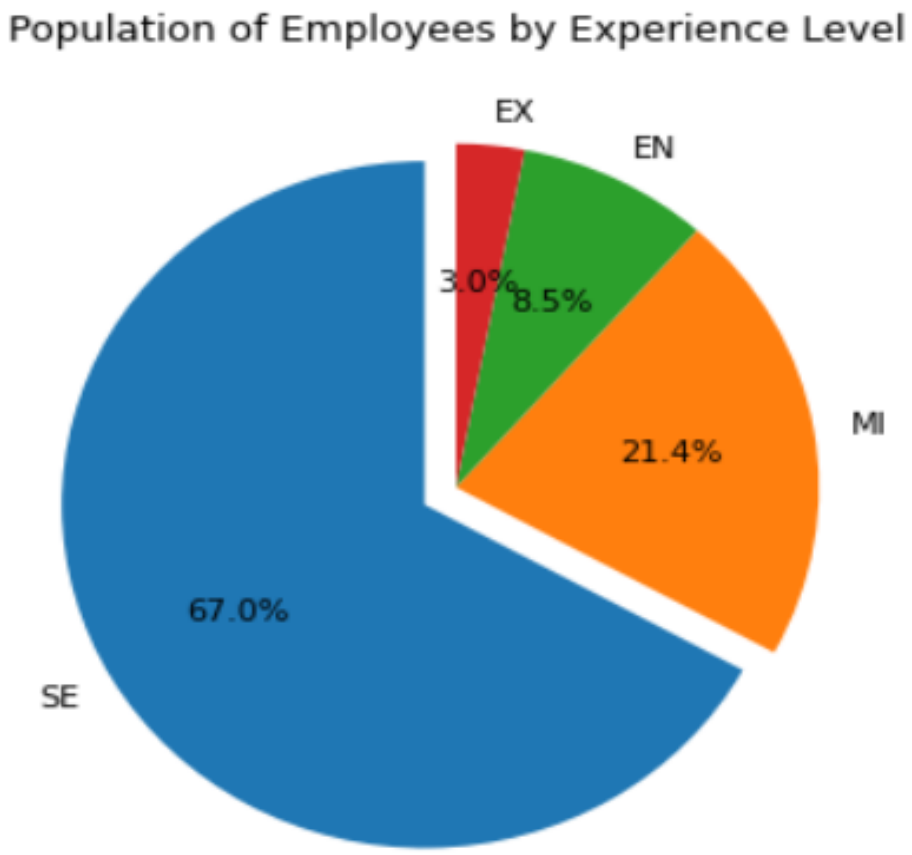
Salary Analysis

the average salaries per experience level

Rata-rata gaji tertinggi seiring dengan bertambahnya pengalaman kerja.



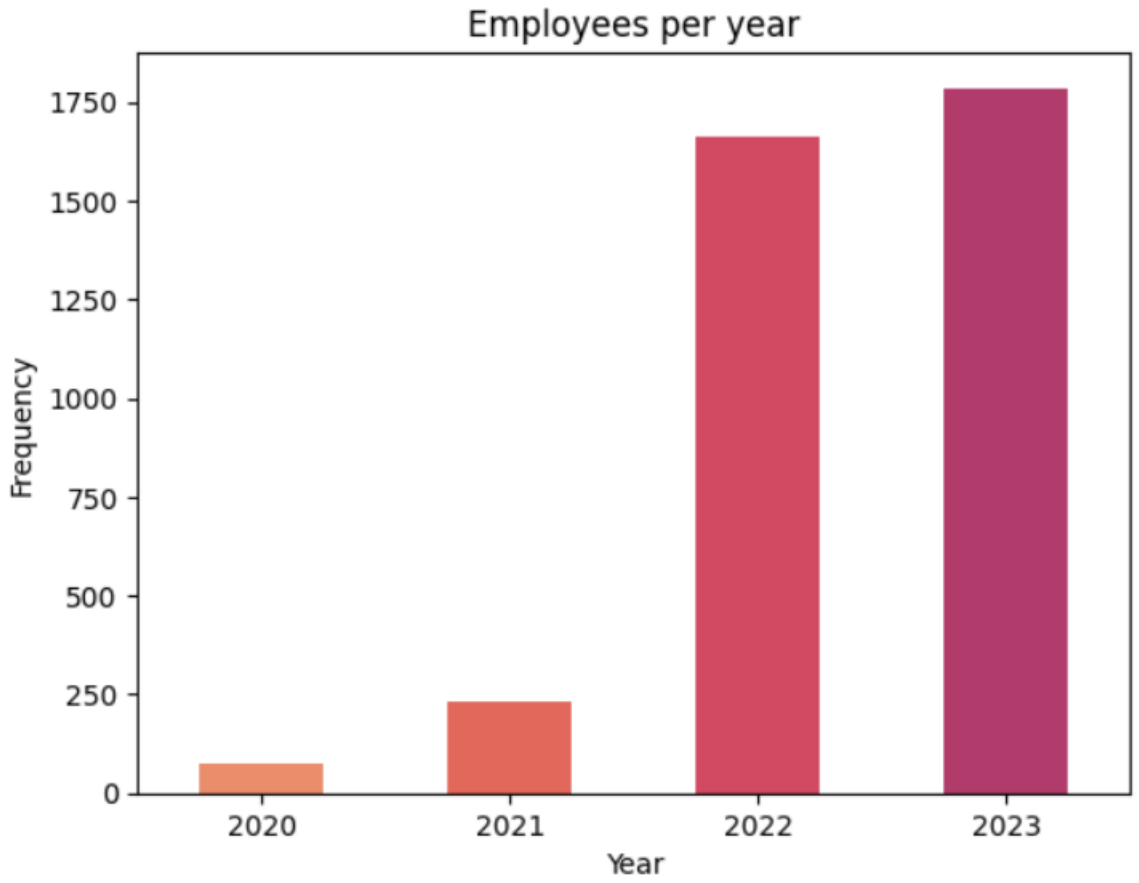
Experience Level

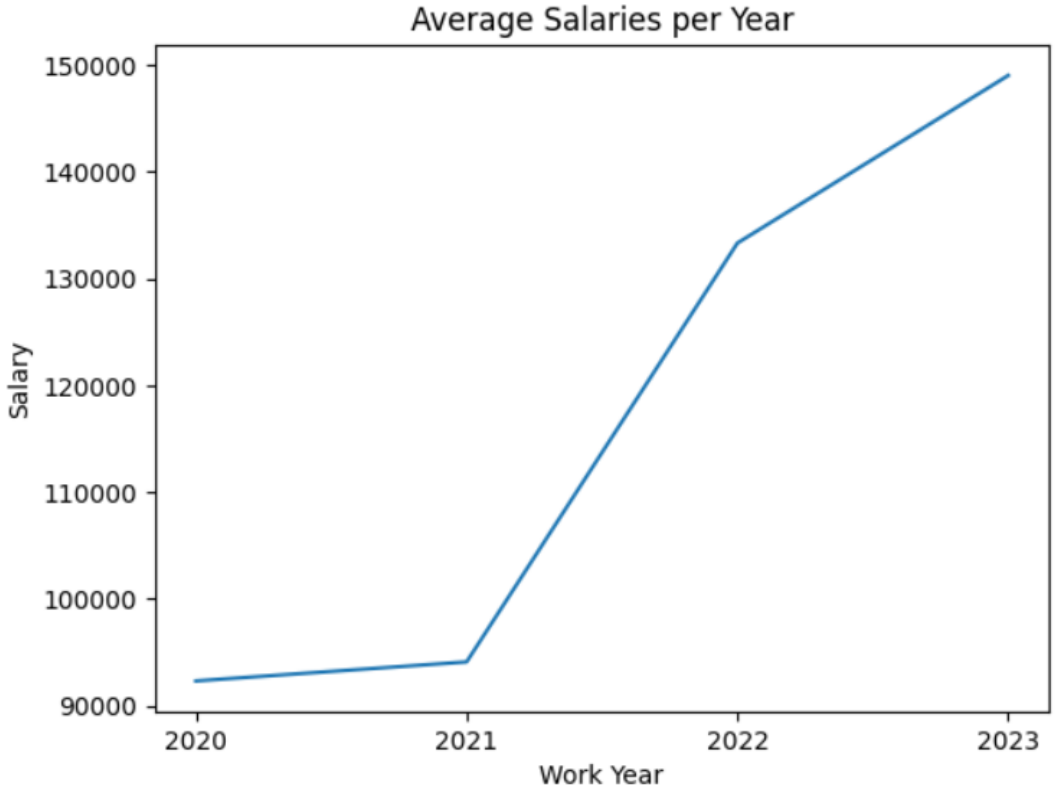


Pengalaman kerja SE dibidang Data Scientist memiliki jumlah terbanyak, sementara EN dan MI masih sedikit.

Yearly Progress

the number of hired employees per year



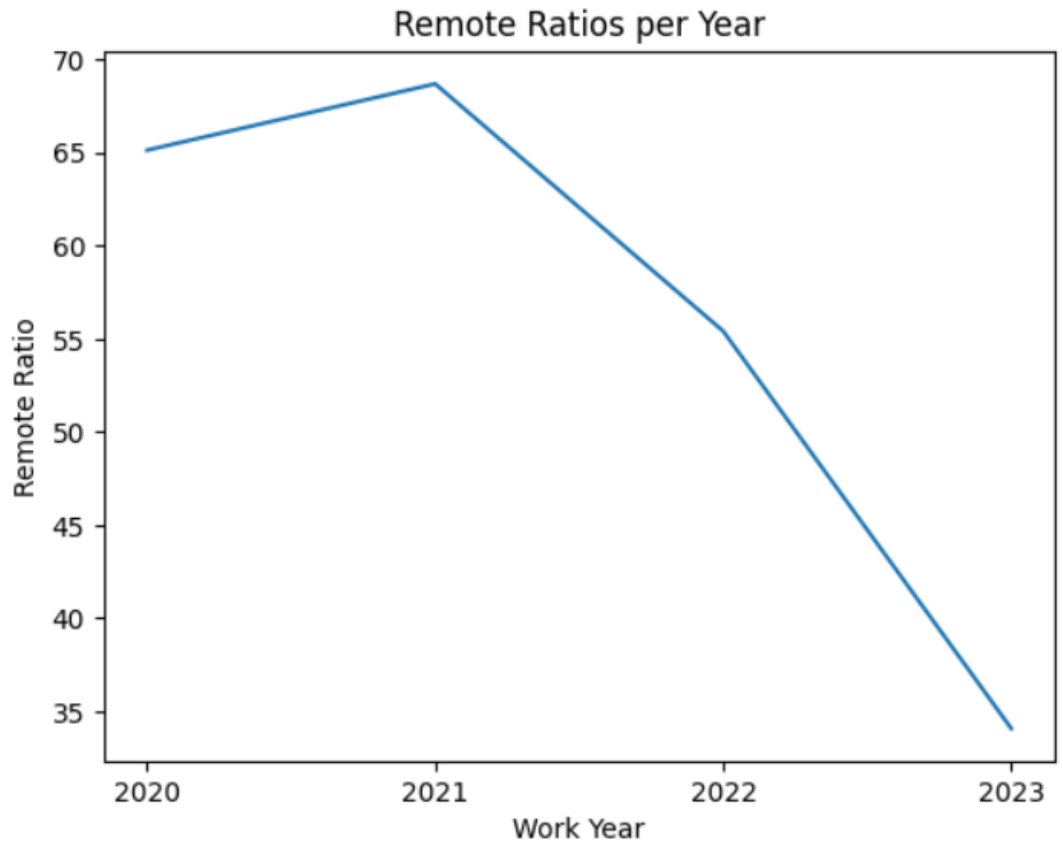


Periode dari tahun 2020 sampai sekarang, kebutuhan data scientist terus meningkat diiringi dengan bisnis focus pada Big data. Diikuti dengan peningkatan rat-rata gaji data scientist.

Yearly Progress

the average salaries per year and remote ratio per year

work_year	salary_in_usd	remote_ratio
2020	92302.631579	65.131579
2021	94087.208696	68.695652
2022	133338.620793	55.408654
2023	149045.541176	34.061625

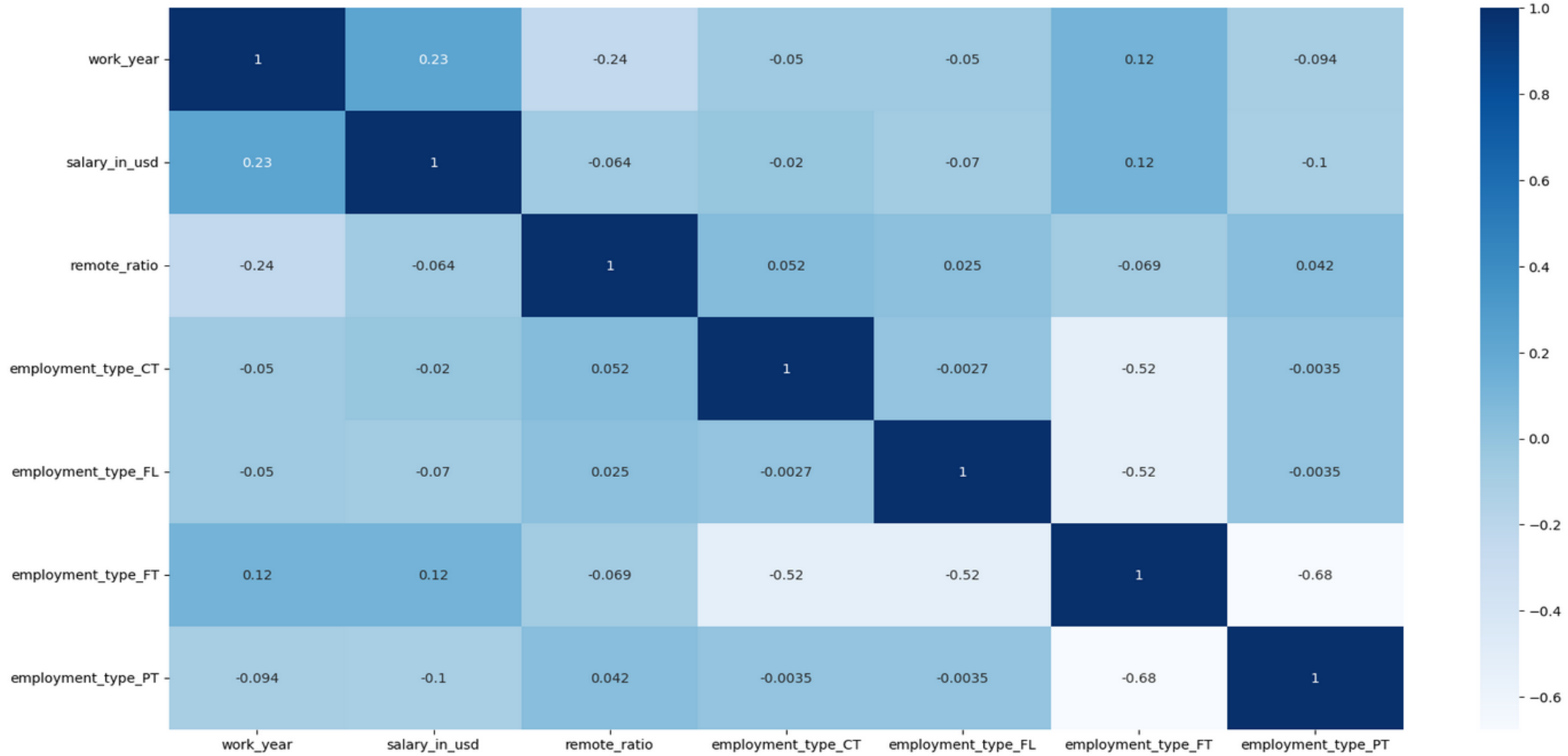


Periode sebelum tahun 2021 saat terjadi pandemic, pekerjaan WFO meningkat pesat. Setelah tahun 2021, seiringnya meredanya pandemic, pekerjaan WFO semakin berkurang.

Korelasi data

```
plt.figure(figsize = (20,10))
sns.heatmap(filtered_data.corr() , annot = True , cmap = "Blues")
```

```
<Axes: >
```



Statistical Summary

```
US_data[numericals].describe()
```

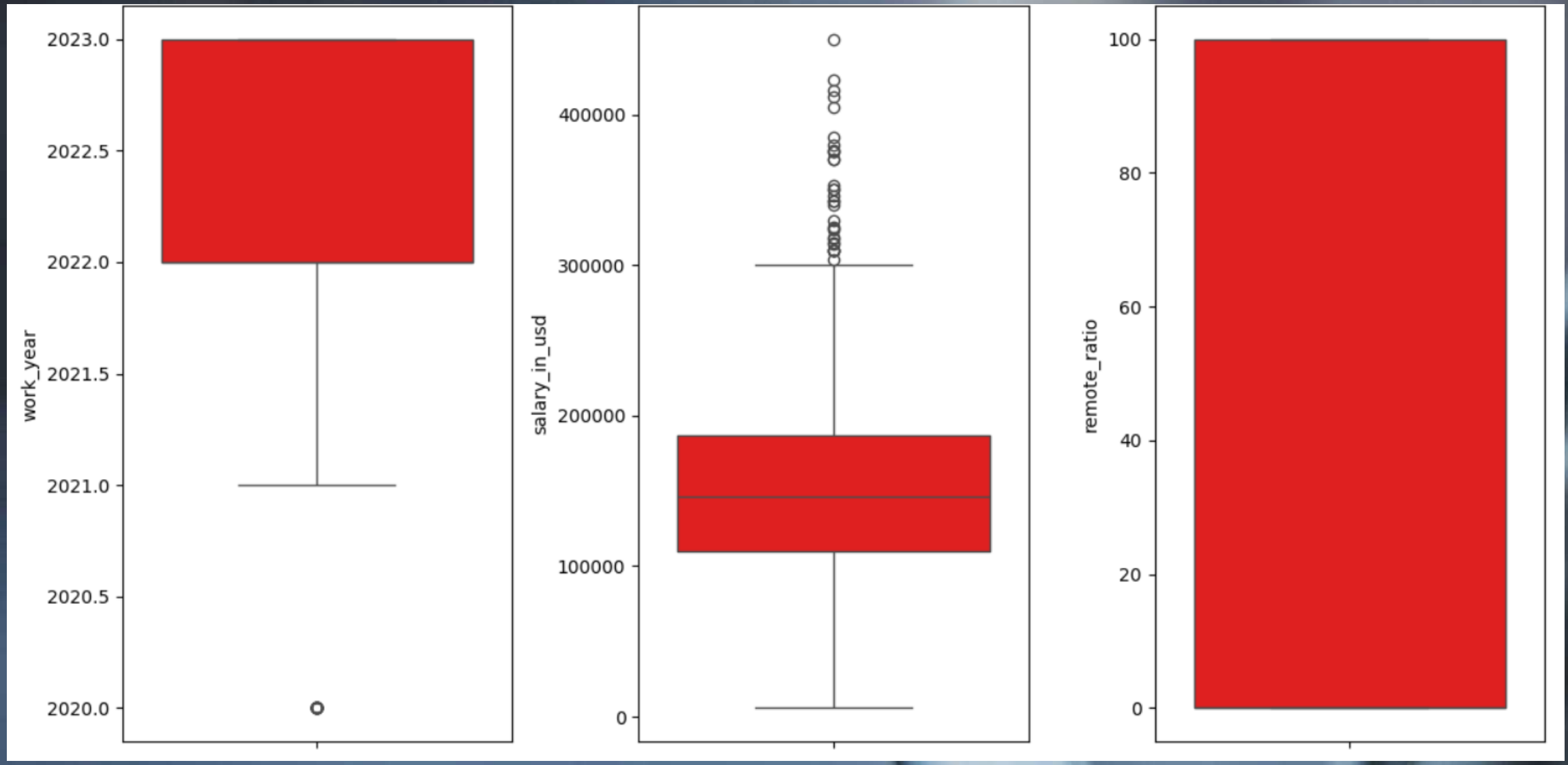
	work_year	salary_in_usd	remote_ratio
count	1929.000000	1929.000000	1929.000000
mean	2022.412131	152374.791602	48.548471
std	0.664511	59786.145995	49.483514
min	2020.000000	5679.000000	0.000000
25%	2022.000000	110000.000000	0.000000
50%	2022.000000	145885.000000	0.000000
75%	2023.000000	187200.000000	100.000000
max	2023.000000	450000.000000	100.000000

```
Value counts of work_year column
2023    1156
2022    1125
2021     228
2020      75
Name: work_year, dtype: int64

Value counts of salary_in_usd column
100000    58
150000    56
120000    51
200000    47
130000    39
          ..
314100     1
195800     1
262500     1
209450     1
94665      1
Name: salary_in_usd, Length: 1035, dtype: int64

Value counts of remote_ratio column
100    1211
0      1186
50      187
Name: remote_ratio, dtype: int64
```

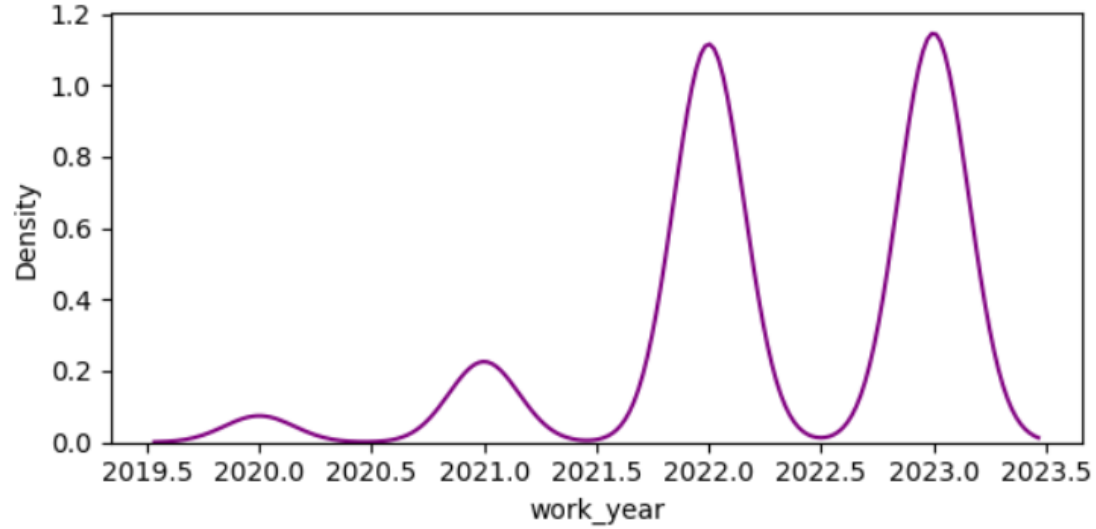
```
US_data[categoricals].describe()
```

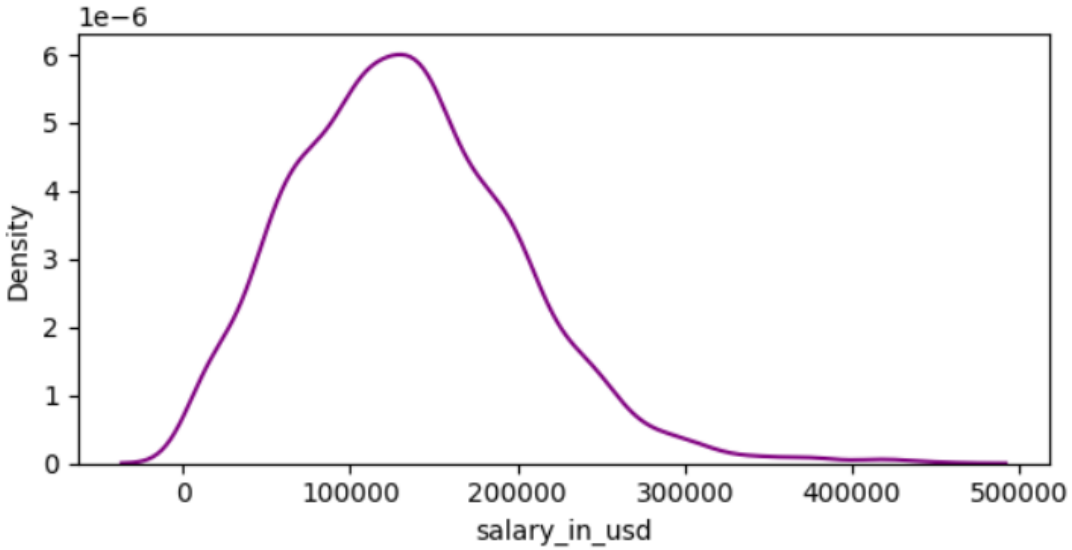
	experience_level	employment_type	job_title	company_location	employee_residence	company_size
count	1929	1929	1929	1929	1929	1929
unique	4	4	70	1	28	3
top	SE	FT	Data Engineer	US	US	M
freq	1335	1911	487	1929	1888	1656

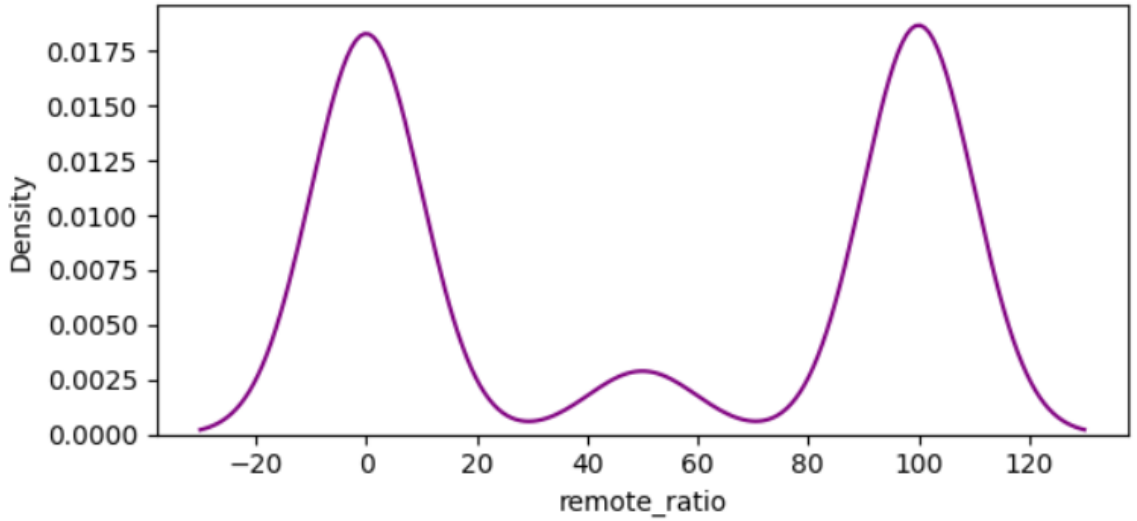
Univariate Analysis



KDE Plot for knowing the distribution form







Multivariate Analysis

Remove 2 variable with high correlation (>0,8)

1. Salary
2. Salary_in_usd


FEDERAL RESERVE NOTE
THE UNITED STATES OF AMERICA
THIS NOTE IS LEGAL TENDER
FOR ALL DEBTS, PUBLIC AND PRIVATE
L 89823401 J
WASHINGTON, D.C.
L
L 89823401 J
ONE DOLLAR

Summary

Job Positions

Posisi Data Science Tech Lead memiliki gaji tahunan tertinggi USD 375.000

Location

Dari data, jumlah terbanyak terdapat di negara USA, dengan jumlah 3.004 orang. Diikuti dengan Inggris 167 orang, Kanada 85 orang, Spanyol 80 orang dan India 71 orang.

Company Size

Gaji rata-rata yang lebih tinggi, berada di perusahaan M yaitu USD 143.131 sedangkan perusahaan Besar USD 118.301 dan perusahaan kecil USD 78.227.

.

Salary in USD

Rata-rata gaji tertinggi seiring dengan bertambahnya pengalaman kerja. Yaitu EX dengan rata-rata gaji USD 194.931

Experience Level

Pengalaman kerja SE dibidang Data Scientist memiliki jumlah terbanyak, sementara EN dan MI masih sedikit.

.

Work Year

Periode dari tahun 2020 sampai sekarang, kebutuhan data scientist terus meningkat diiringi dengan bisnis focus pada Big data. Diikuti dengan peningkatan rat-rata gaji data scientist.

.

THANK YOU

Do You Have any Questions?

+62 81381483860

iramasu3101@gmail.com