PROJECT

Presented by vidhi sharma

Overview

- Introduction
- Domain Knowledge
- EDA
- Graphical EDA
- Statistical EDA
- Hypothesis testing
- Insights

Introduction about data

- work_year:

The year the salary was paid.

- experience_level:

The experience level in the job during the year with the following possible values: EN Entry-level / Junior MI Mid-level / Intermediate SE Senior-level / Expert EX Executive-level / Director

- employment_type:

The type of employement for the role: PT Part-time FT Full-time CT Contract FL Freelance

- job_title:

The role worked in during the year.

- salary:

The total gross salary amount paid.

- salary_currency:

The currency of the salary paid as an ISO 4217 currency code.

- salary_in_usd:

The salary in USD (FX rate divided by avg. USD rate for the respective year via fxdata.foorilla.com).

- employee_residence:

Employee's primary country of residence in during the work year as an ISO 3166 country code.

- remote_ratio:

The overall amount of work done remotely, possible values are as follows: 0 No remote work (less than 20%) 50 Partially remote 100 Fully remote (more than 80%)

- company_location:

The country of the employer's main office or contracting branch as an ISO 3166 country code.

- company_size:

The average number of people that worked for the company during the year: S less than 50 employees (small) M 50 to 250 employees (medium) L more than 250 employees (large)

Introduction about EDA

Exploratory Data Analysis (EDA) is a crucial initial step in any data science project. It involves examining and visualizing data to understand its characteristics, uncover patterns, detect anomalies, and formulate hypotheses. Through EDA, analysts gain insights into the underlying structure of the data, which informs subsequent modeling and decision-making processes

Two Types of EDA:

1) Graphical EDA
2) Statistical EDA

DESCRIPTIVE STATISTICS

Descriptive Statistics

Descriptive Statistics

	work_year	salary	salary_in_usd	remote_ratio
Valid	607	607	607	607
Missing	0	0	0	0
Mode	2022.000[a]	80000.000[a]	100000.000[a]	100.000[a]
Median	2022.000	115000.000	101570.000	100.000
Mean	2021.405	324000.063	112297.870	70.923
Std. Deviation	0.692	$1.544\times10^{+6}$	70957.259	40.709
IQR	1.000	95000.000	87274.000	50.000
Variance	0.479	$2.385\times10^{+12}$	$5.035\times10^{+9}$	1657.233
Skewness	−0.736	14.053	1.668	−0.904
Std. Error of Skewness	0.099	0.099	0.099	0.099
Kurtosis	−0.644	247.426	6.354	−0.888
Std. Error of Kurtosis	0.198	0.198	0.198	0.198
Minimum	2020.000	4000.000	2859.000	0.000
Maximum	2022.000	$3.040\times10^{+7}$	600000.000	100.000
25th percentile	2021.000	70000.000	62726.000	50.000
50th percentile	2022.000	115000.000	101570.000	100.000
75th percentile	2022.000	165000.000	150000.000	100.000
10th percentile	2020.000	42720.000	33689.200	0.000

[a] The mode is computed assuming that variables are discreet.

- Dataset contains 607 entries with complete data.
- No missing values are observed.
- Average salary is $324,000.63, with a maximum salary of $30,000,000.
- Average remote work ratio is 70%.
- Minimum recorded salary is $4,000.
- The salary column demonstrates positive skewness.
- The Interquartile Range (IQR) for 'salary_in_usd' is 87,274, while for 'salary' it is 95,000, representing the range where the majority of the data is concentrated.

DESCRIPTIVE STATISTICS FOR CATEGORICAL

Descriptive Statistics

Descriptive Statistics

	experience_level	job_title	company_location	company_size	employee_residence
Valid	607	607	607	607	607
Missing	0	0	0	0	0
Mode	4.000	23.000	49.000	2.000	56.000
Range					
Minimum					
Maximum					

Note. Not all values are available for *Nominal Text* variables

- There are five categorical columns: 'experience_level', 'job title', 'company location', 'company size', and 'employee residence'.
- The modes for these categorical columns are as follows:
  - 'experience_level': 4
  - 'job title': 23
  - 'company location': 49
  - 'company size': 2
  - 'employee residence': 56

FREQUENCY TABLE FOR CATEGORICAL

Frequency Tables

Frequencies for experience_level

experience_level	Frequency	Percent	Valid Percent	Cumulative Percent
EN	88	14.498	14.498	14.498
EX	26	4.283	4.283	18.781
MI	213	35.091	35.091	53.871
SE	280	46.129	46.129	100.000
Missing	0	0.000		
Total	607	100.000		

- Experience Level Distribution:
- Entry-level / Junior (EN): Approximately 14.4%
- Mid-level / Intermediate (MI): Approximately 4.23%
- Senior-level / Expert (SE): Approximately 35%
- Executive-level / Director (EX): Approximately 45%
- The Senior-level / Expert (SE) category constitutes the largest proportion of experience levels.

Frequencies for company_size

company_size	Frequency	Percent	Valid Percent	Cumulative Percent
L	198	32.619	32.619	32.619
M	326	53.707	53.707	86.326
S	83	13.674	13.674	100.000
Missing	0	0.000		
Total	607	100.000		

- Large companies: 298
- Medium companies: 326 (the most prevalent category)

Frequencies for job_title ▼

job_title	Frequency	Percent	Valid Percent	Cumulative Percent
3D Computer Vision Researcher	1	0.165	0.165	0.165
AI Scientist	7	1.153	1.153	1.318
Analytics Engineer	4	0.659	0.659	1.977
Applied Data Scientist	5	0.824	0.824	2.801
Applied Machine Learning Scientist	4	0.659	0.659	3.460
BI Data Analyst	6	0.988	0.988	4.448
Big Data Architect	1	0.165	0.165	4.613
Big Data Engineer	8	1.318	1.318	5.931
Business Data Analyst	5	0.824	0.824	6.755
Cloud Data Engineer	2	0.329	0.329	7.084
Computer Vision Engineer	6	0.988	0.988	8.072
Computer Vision Software Engineer	3	0.494	0.494	8.567
Data Analyst	97	15.980	15.980	24.547
Data Analytics Engineer	4	0.659	0.659	25.206
Data Analytics Lead	1	0.165	0.165	25.371
Data Analytics Manager	7	1.153	1.153	26.524
Data Architect	11	1.812	1.812	28.336
Data Engineer	132	21.746	21.746	50.082
Data Engineering Manager	5	0.824	0.824	50.906
Data Science Consultant	7	1.153	1.153	52.059
Data Science Engineer	3	0.494	0.494	52.554
Data Science Manager	12	1.977	1.977	54.530
Data Scientist	143	23.558	23.558	78.089
Data Specialist	1	0.165	0.165	78.254
Director of Data Engineering	2	0.329	0.329	78.583
Director of Data Science	7	1.153	1.153	79.736
ETL Developer	2	0.329	0.329	80.066
Finance Data Analyst	1	0.165	0.165	80.231
Financial Data Analyst	2	0.329	0.329	80.560
Head of Data	5	0.824	0.824	81.384
Head of Data Science	4	0.659	0.659	82.043
Head of Machine Learning	1	0.165	0.165	82.208
Lead Data Analyst	3	0.494	0.494	82.702
Lead Data Engineer	6	0.988	0.988	83.690

- Top 5 categories in job title:

1. Data Scientist
2. Data Engineer
3. Data Analyst
4. Machine Learning Engineer
5. Research Scientist

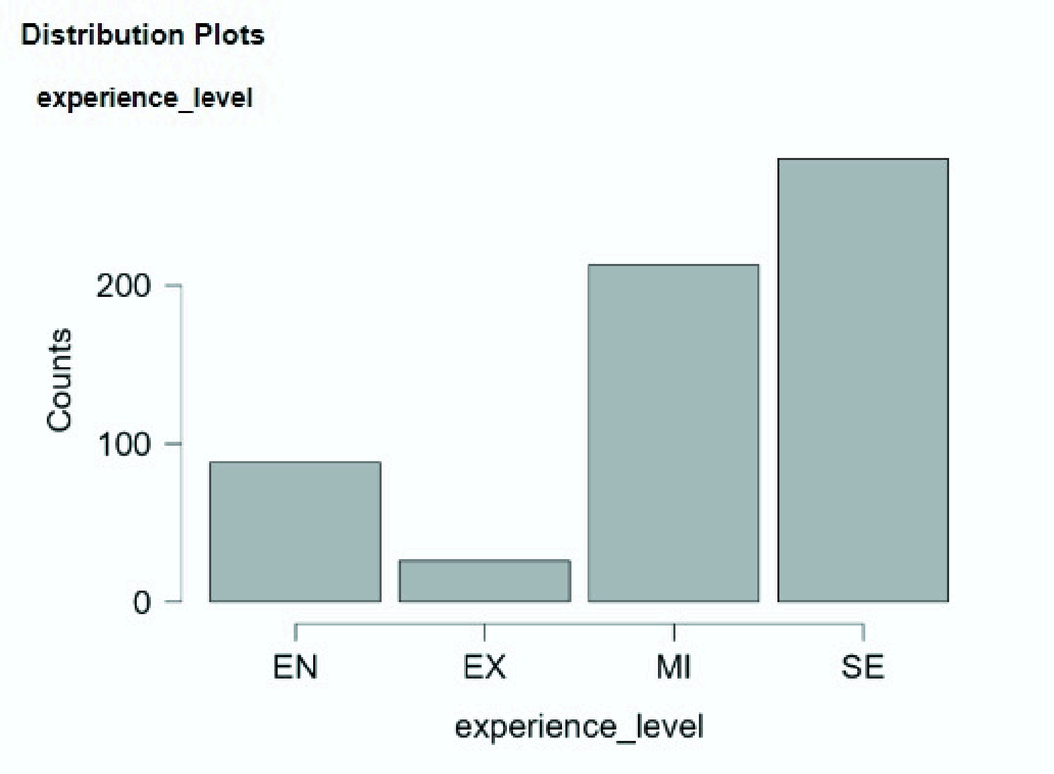
- Top 5 categories in company location:

1. US
2. GB
3. CA
4. DE
5. IN

Frequencies for company_location

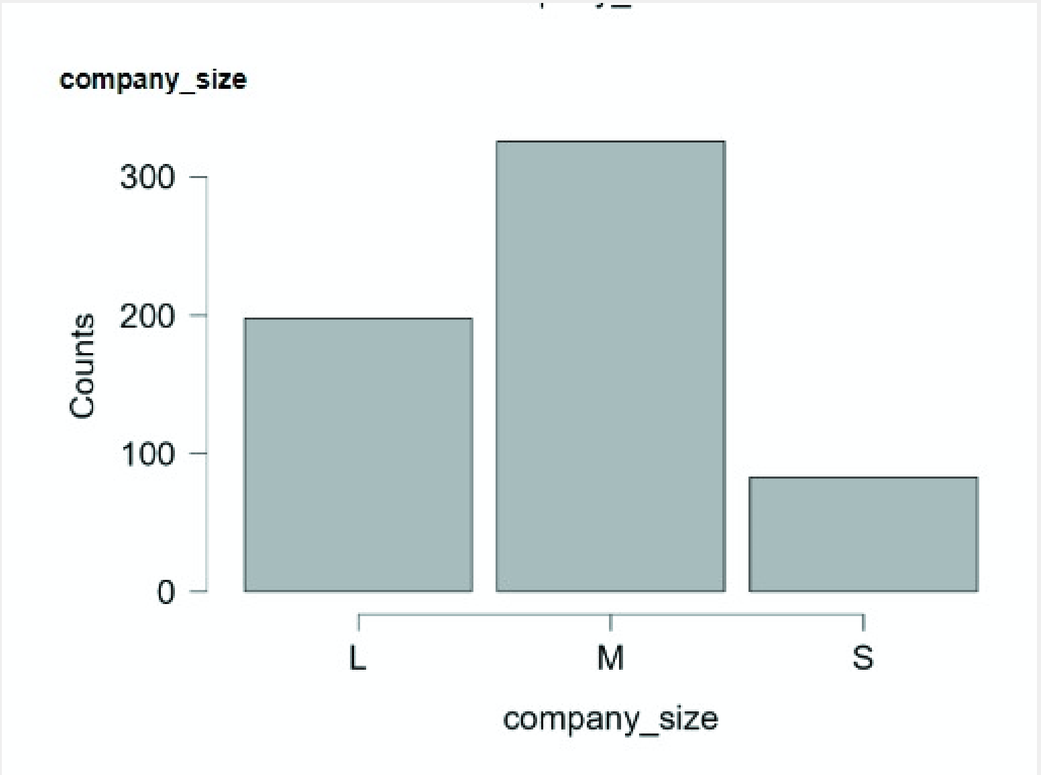
company_location	Frequency	Percent	Valid Percent	Cumulative Percent
AE	3	0.494	0.494	0.494
AS	1	0.165	0.165	0.659
AT	4	0.659	0.659	1.318
AU	3	0.494	0.494	1.812
BE	2	0.329	0.329	2.142
BR	3	0.494	0.494	2.636
CA	30	4.942	4.942	7.578
CH	2	0.329	0.329	7.908
CL	1	0.165	0.165	8.072
CN	2	0.329	0.329	8.402
CO	1	0.165	0.165	8.567
CZ	2	0.329	0.329	8.896
DE	28	4.613	4.613	13.509
DK	3	0.494	0.494	14.003
DZ	1	0.165	0.165	14.168
EE	1	0.165	0.165	14.333
ES	14	2.306	2.306	16.639
FR	15	2.471	2.471	19.110
GB	47	7.743	7.743	26.853
GR	11	1.812	1.812	28.666
HN	1	0.165	0.165	28.830
HR	1	0.165	0.165	28.995
HU	1	0.165	0.165	29.160
IE	1	0.165	0.165	29.325
IL	1	0.165	0.165	29.489
IN	24	3.954	3.954	33.443
IQ	1	0.165	0.165	33.608
IR	1	0.165	0.165	33.773
IT	2	0.329	0.329	34.102
JP	6	0.988	0.988	35.091
KE	1	0.165	0.165	35.255
LU	3	0.494	0.494	35.750
MD	1	0.165	0.165	35.914

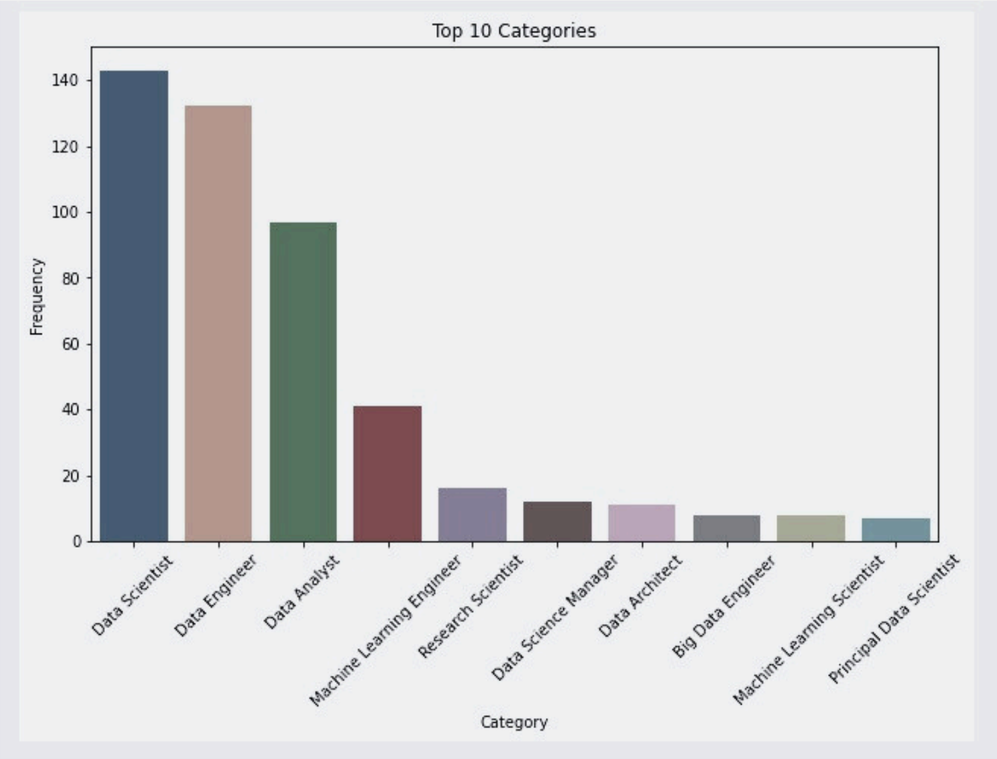
DISTRIBUTION PLOTS



- A significant portion of organizations falls within the medium size category, boasting over 300 employees, potentially indicating a trajectory towards unicorn status in the coming years.

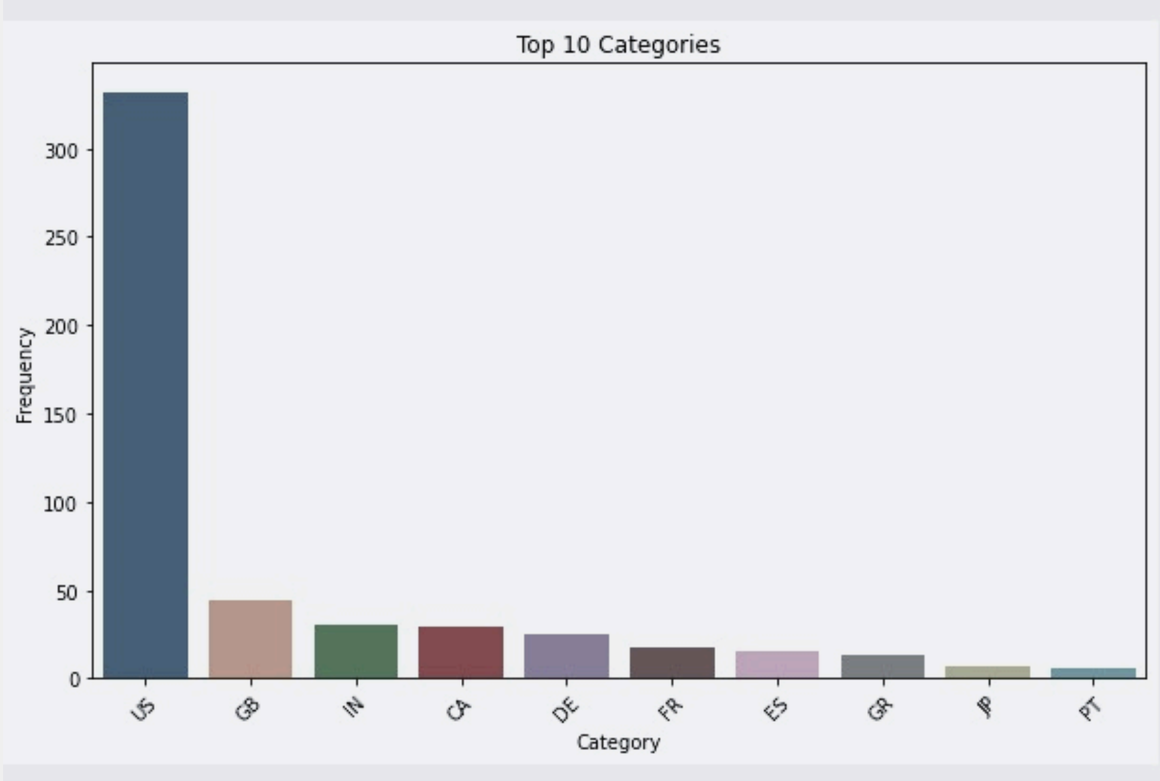
- Majority of individuals hold positions as Senior Executives within the data science industry, with a notable scarcity in Executive roles.

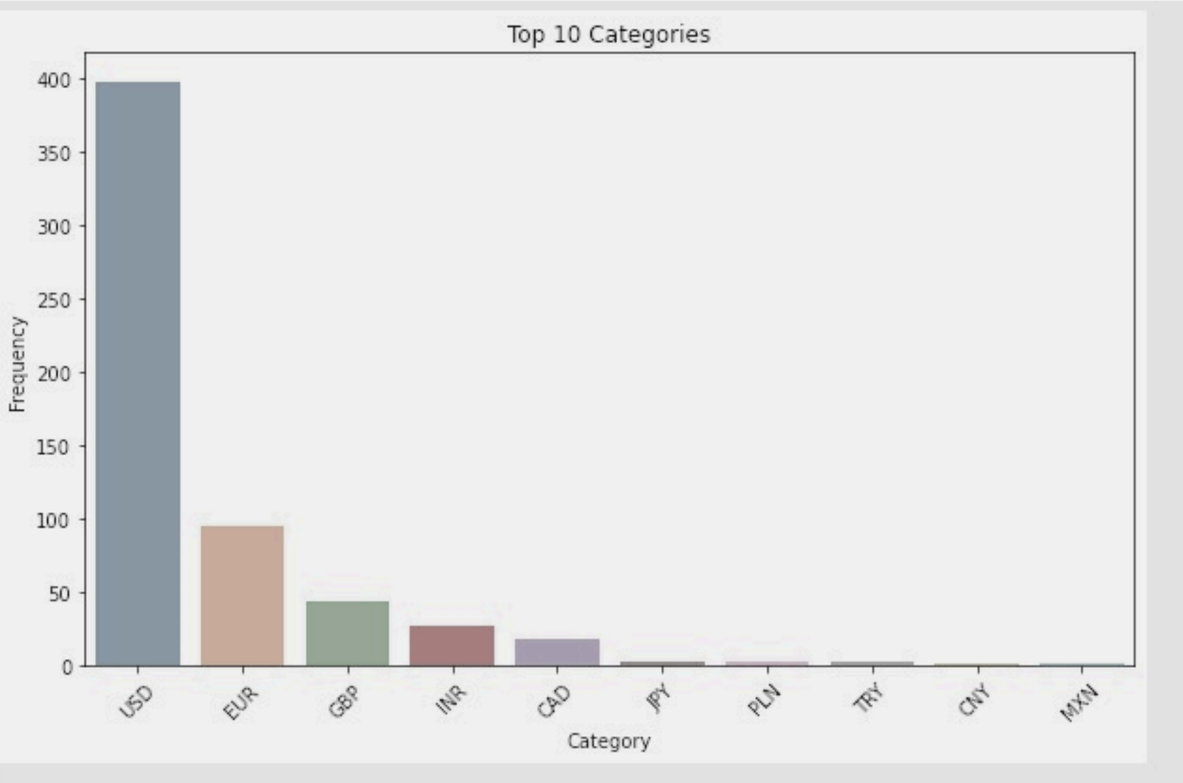




The graph illustrates the top 10 job profiles within the data science field, suggesting that these roles may be associated with higher salary brackets.

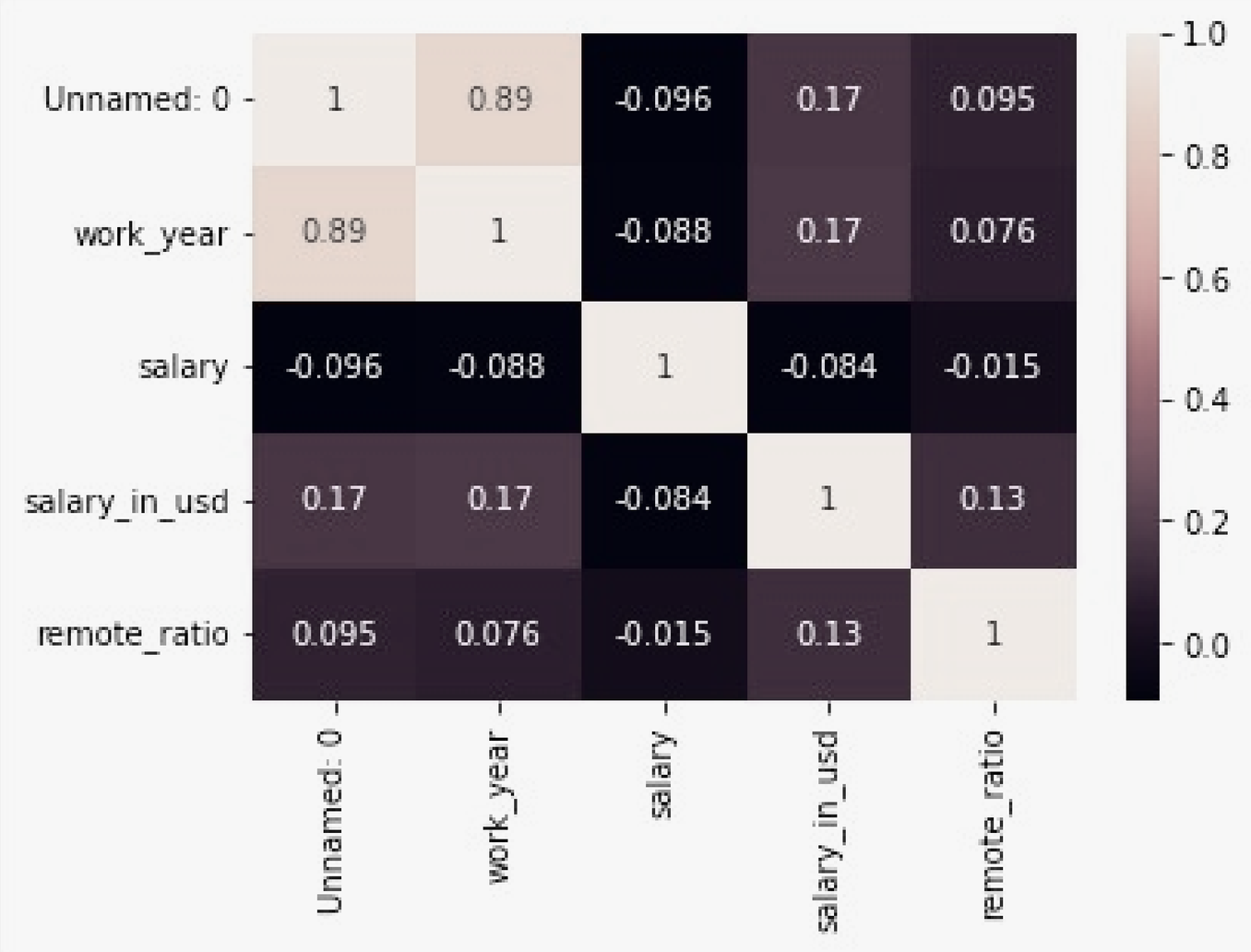
The concentration of professionals predominantly in the United States, Great Britain, Canada, Germany, and India could potentially correlate with higher pay scales. Additionally, these countries, being developed, actively engage in AI and deep learning initiatives, further contributing to competitive salaries within the field.





The high concentration of professionals in roles like Data Scientist, Data Engineer, Data Analyst, Machine Learning Engineer, and Research Scientist in the USA aligns with the predominant use of USD currency, indicating a potential correlation between job roles and currency usage.

CORELATION THROUGH HEATMAPS



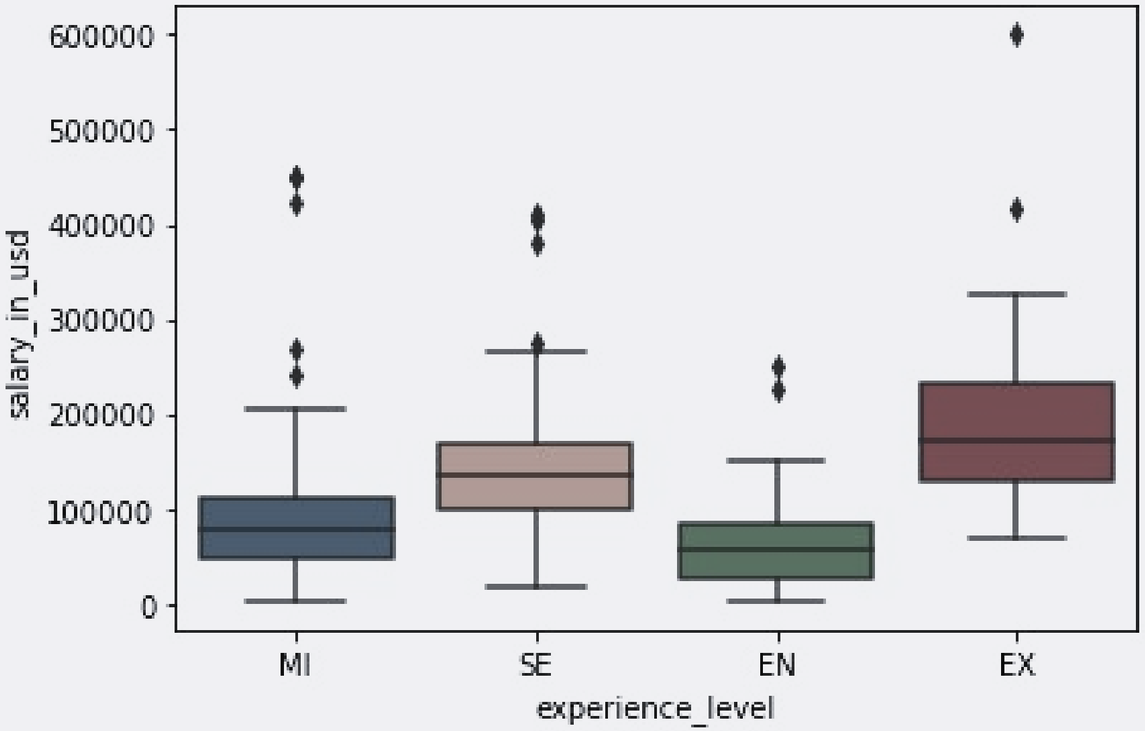
The correlation plot indicates the relationships between variables. In this instance, it suggests that there's minimal correlation among numerical columns, implying limited linear relationships between them.

Correlation ▼

Correlation Table ▼

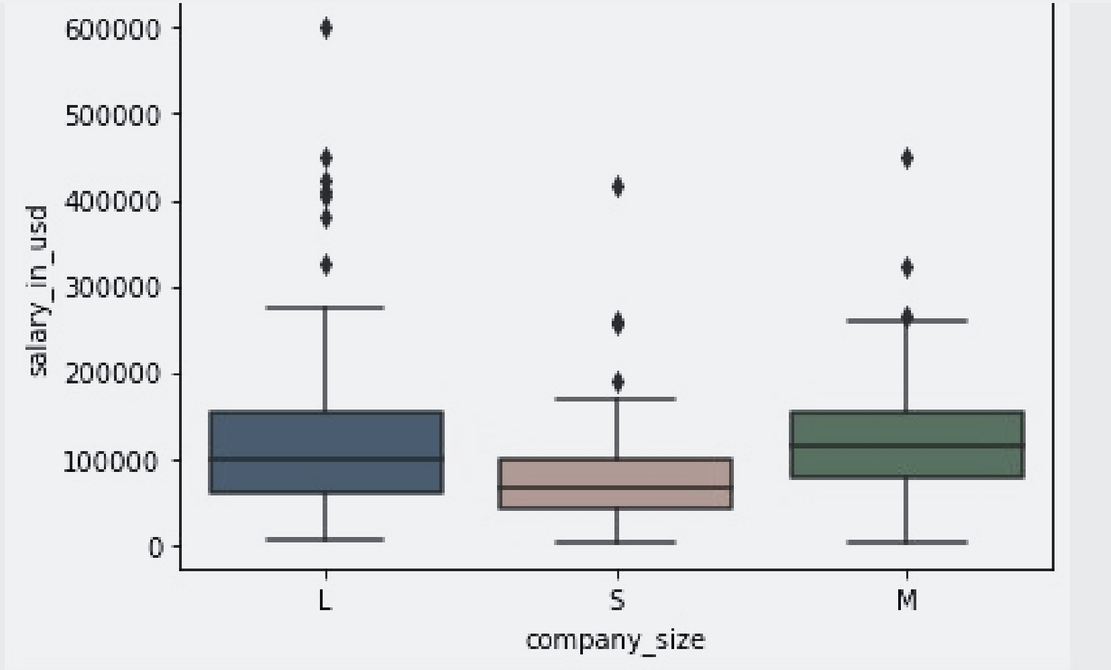
Variable		work_year	salary	salary_in_usd	remote_ratio
1. work_year	Pearson's r	—			
	p-value	—			
	Spearman's rho	—			
	p-value	—			
	Kendall's Tau B	—			
	p-value	—			
2. salary	Pearson's r	−0.088	—		
	p-value	0.031	—		
	Spearman's rho	0.100	—		
	p-value	0.014	—		
	Kendall's Tau B	0.078	—		
	p-value	0.015	—		
3. salary_in_usd	Pearson's r	0.170	−0.084	—	
	p-value	< .001	0.039	—	
	Spearman's rho	0.275	0.662	—	
	p-value	< .001	< .001	—	
	Kendall's Tau B	0.215	0.696	—	
	p-value	< .001	< .001	—	
4. remote_ratio	Pearson's r	0.076	−0.015	0.132	—
	p-value	0.060	0.719	0.001	—
	Spearman's rho	0.125	0.114	0.181	—
	p-value	0.002	0.005	< .001	—
	Kendall's Tau B	0.116	0.088	0.142	—
	p-value	0.002	0.006	< .001	—

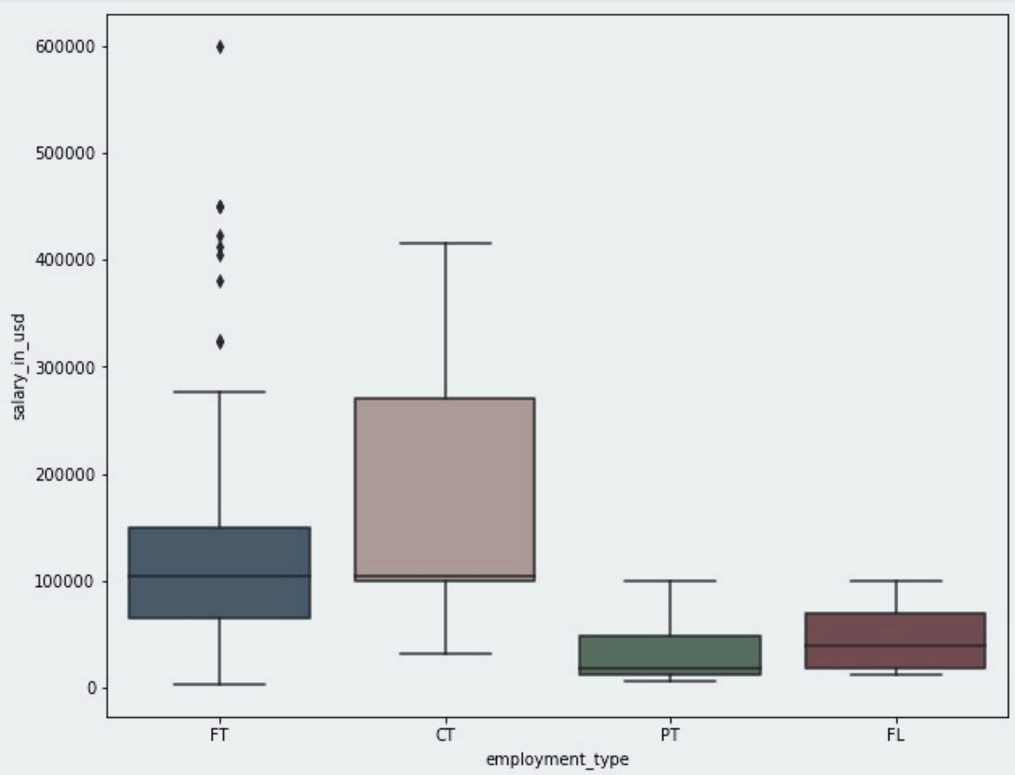
BOXPLOTS



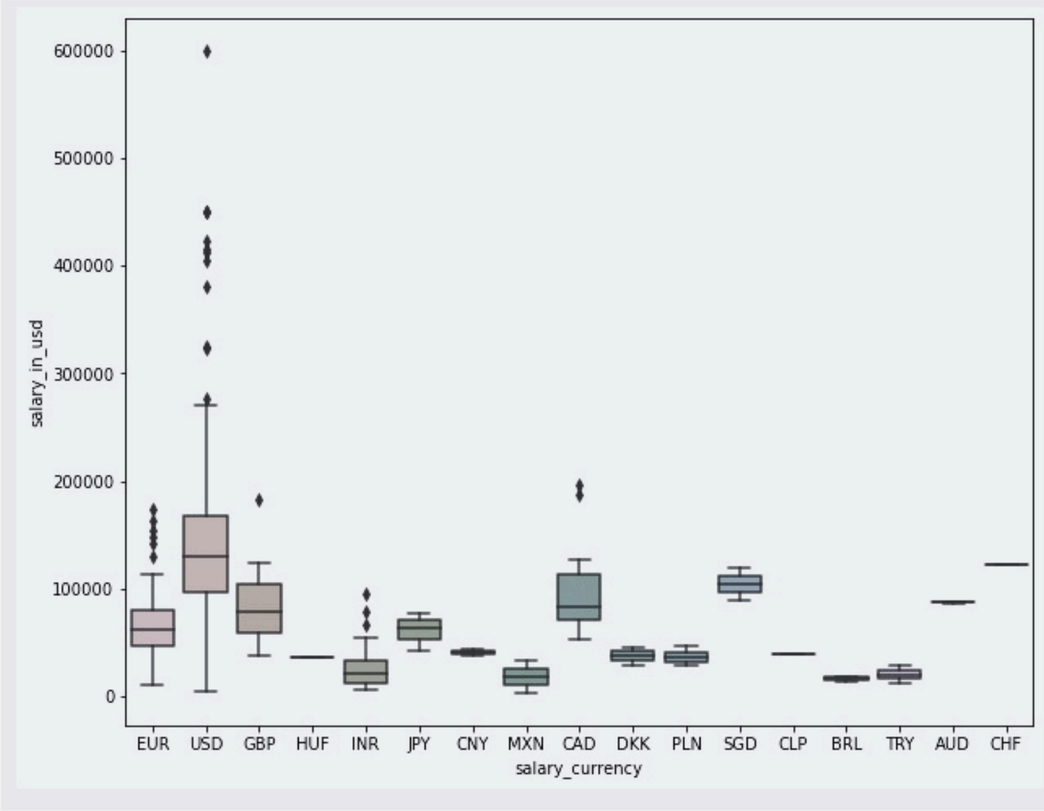
A boxplot is a graphical representation that displays the distribution of a dataset ,identifies outliers in the relationship between experience level and salary in USD. It's evident that some executives receive significantly higher salaries, which could be attributed to their extensive experience in the field.

This observation suggests a potential correlation between company size and salary levels, with larger companies being more likely to provide higher compensation packages.

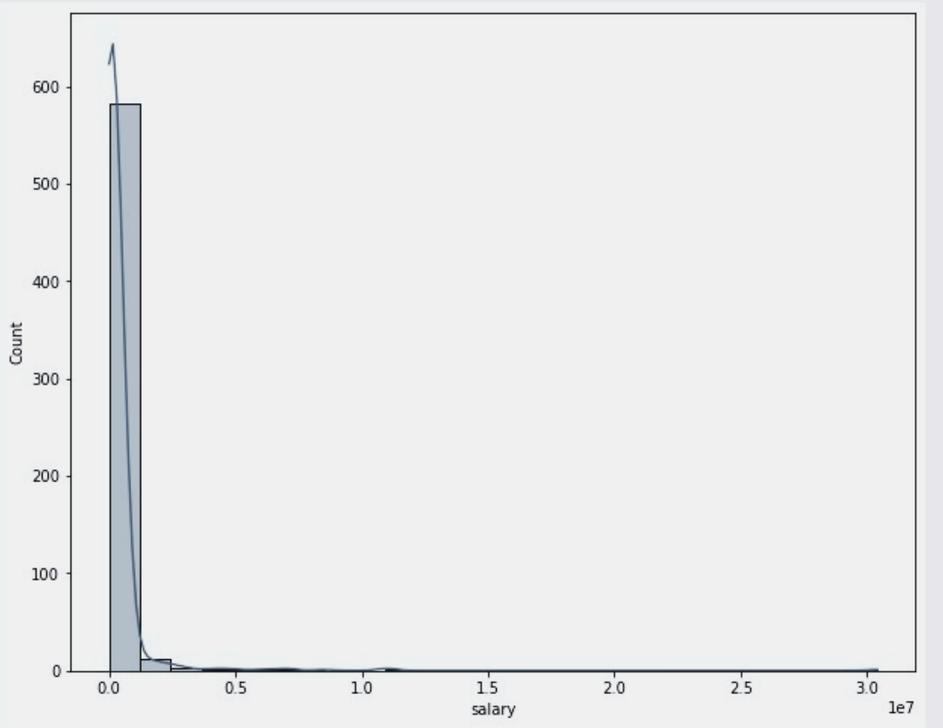




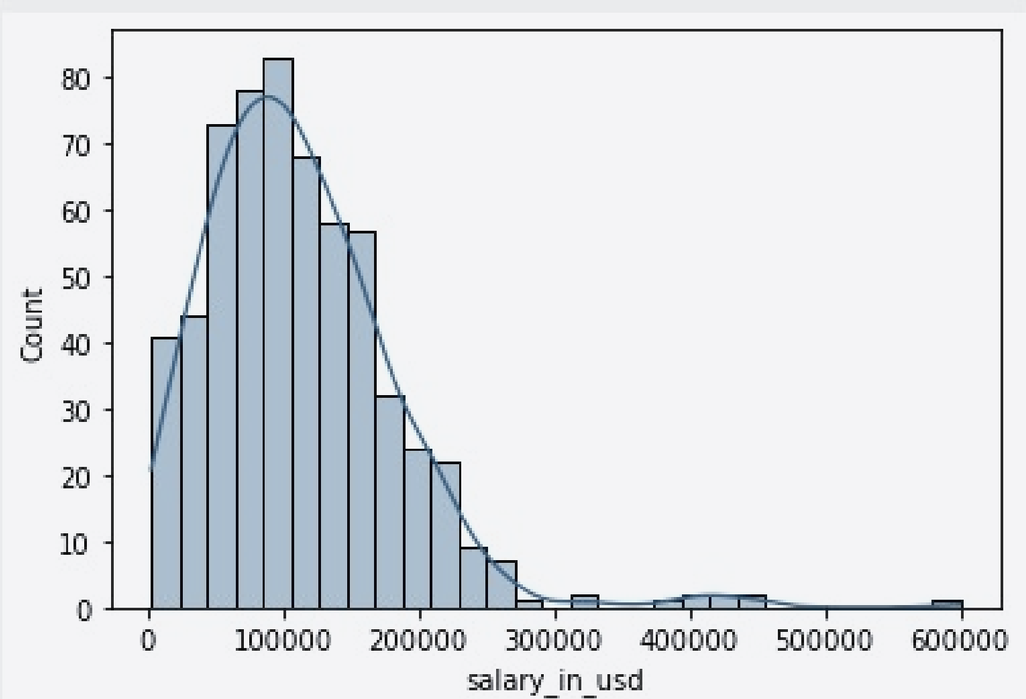
Contract workers typically receive higher salaries compared to freelancers, with full-time employees earning the highest salaries overall.



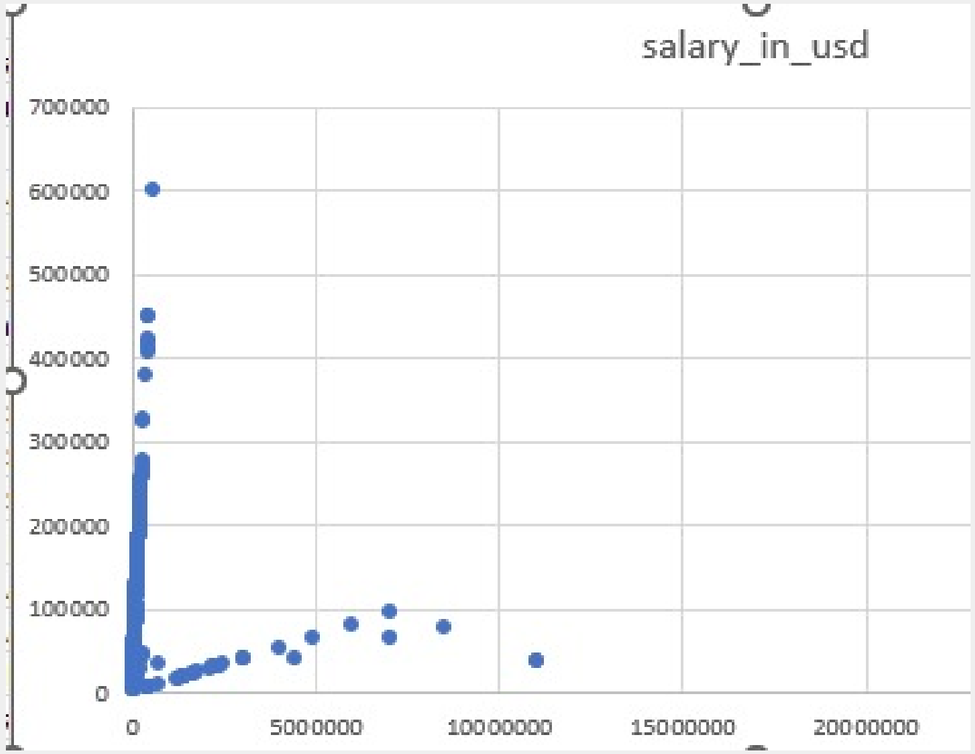
HISTOGRAM



The salary data exhibits right skewness, indicating that a smaller portion of individuals possess significantly higher salaries compared to the majority.

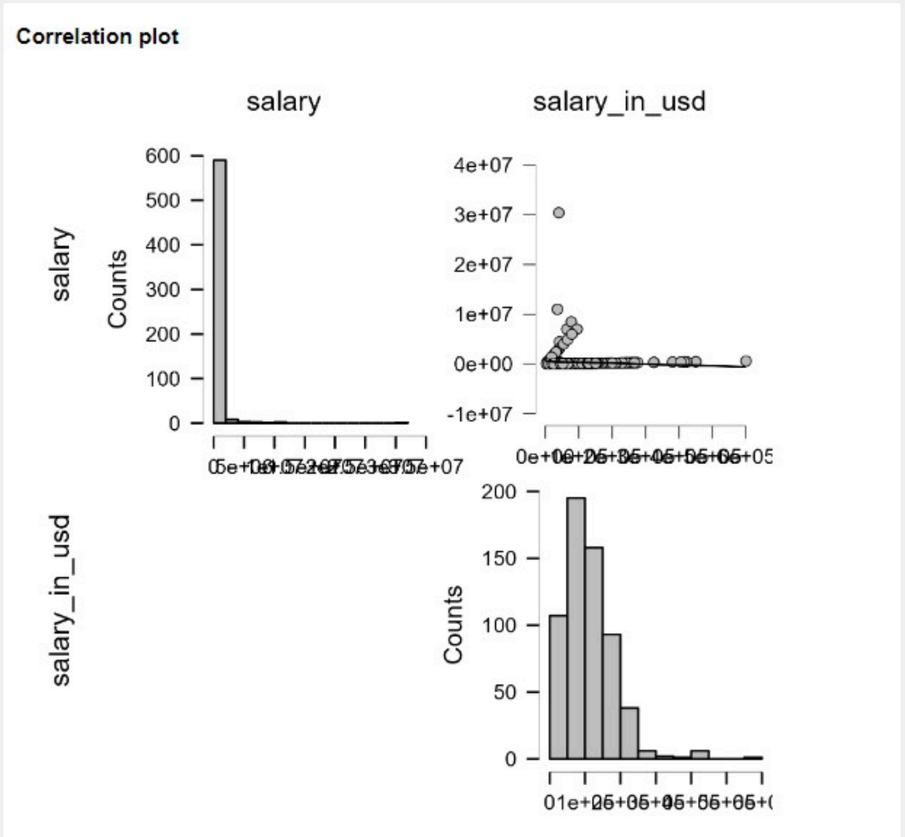


SCATTER PLOT

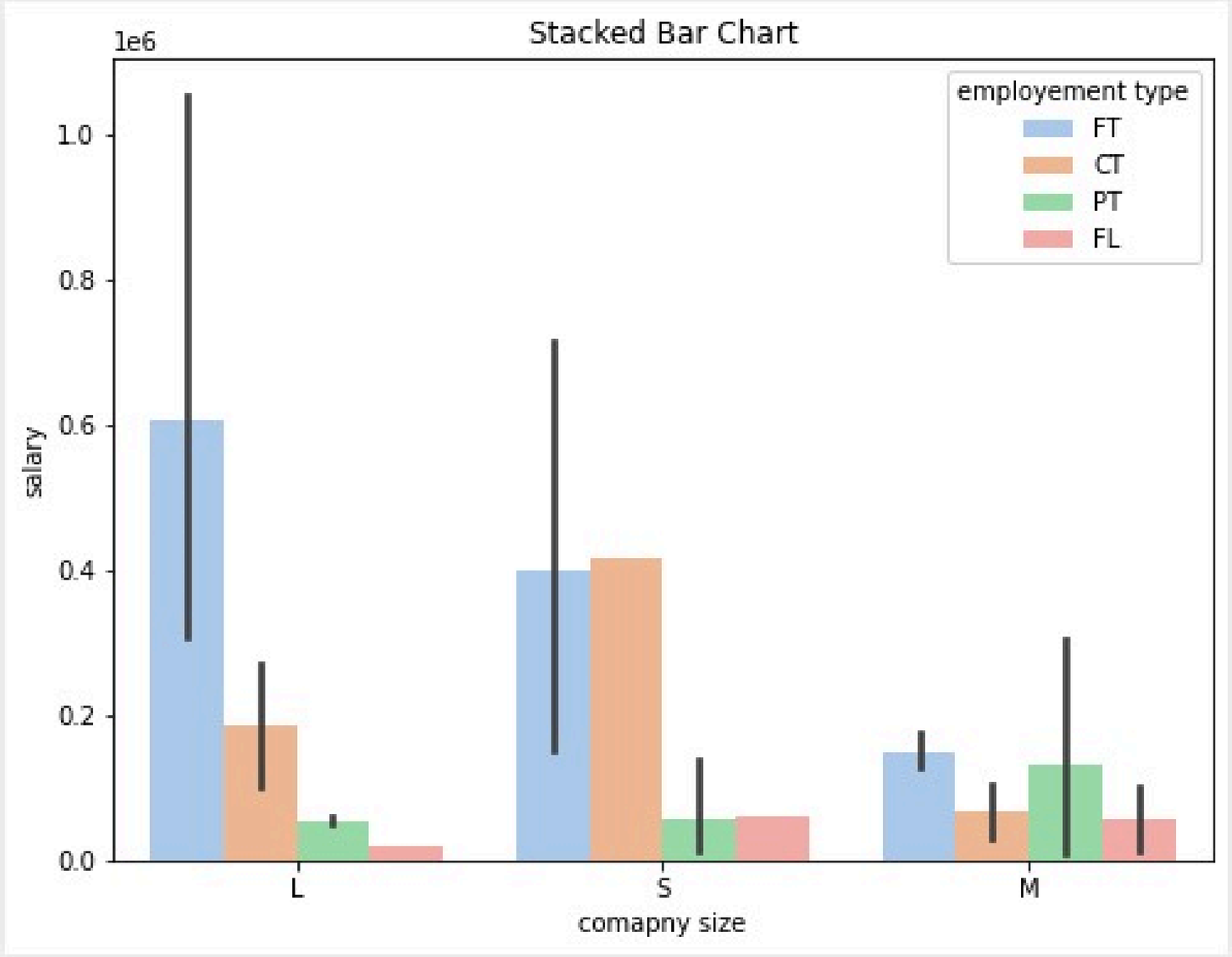


No significant correlation is observed between the columns.

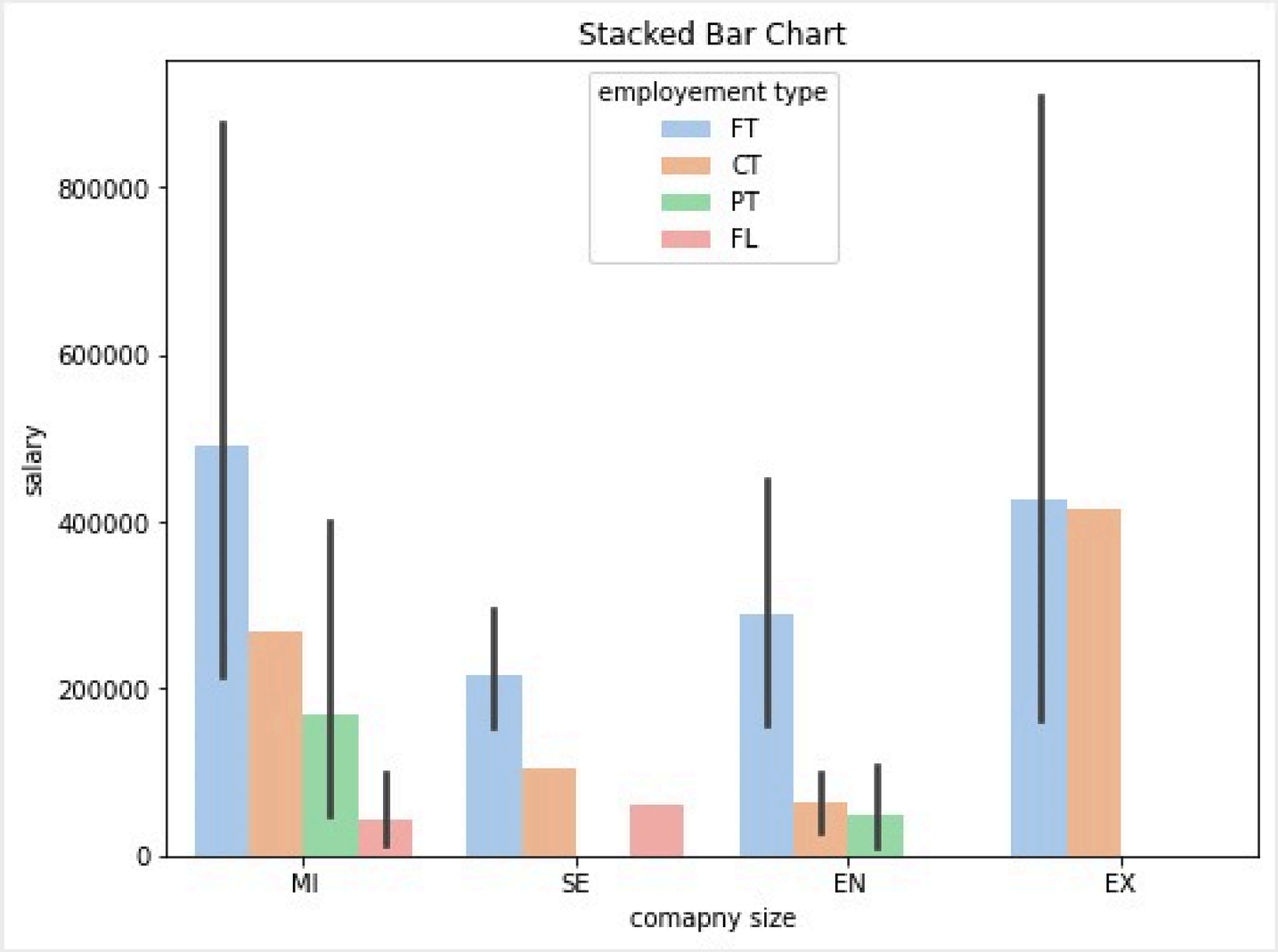
The salary data exhibits right skewness, indicating that a smaller portion of individuals possess significantly higher salaries compared to the majority.



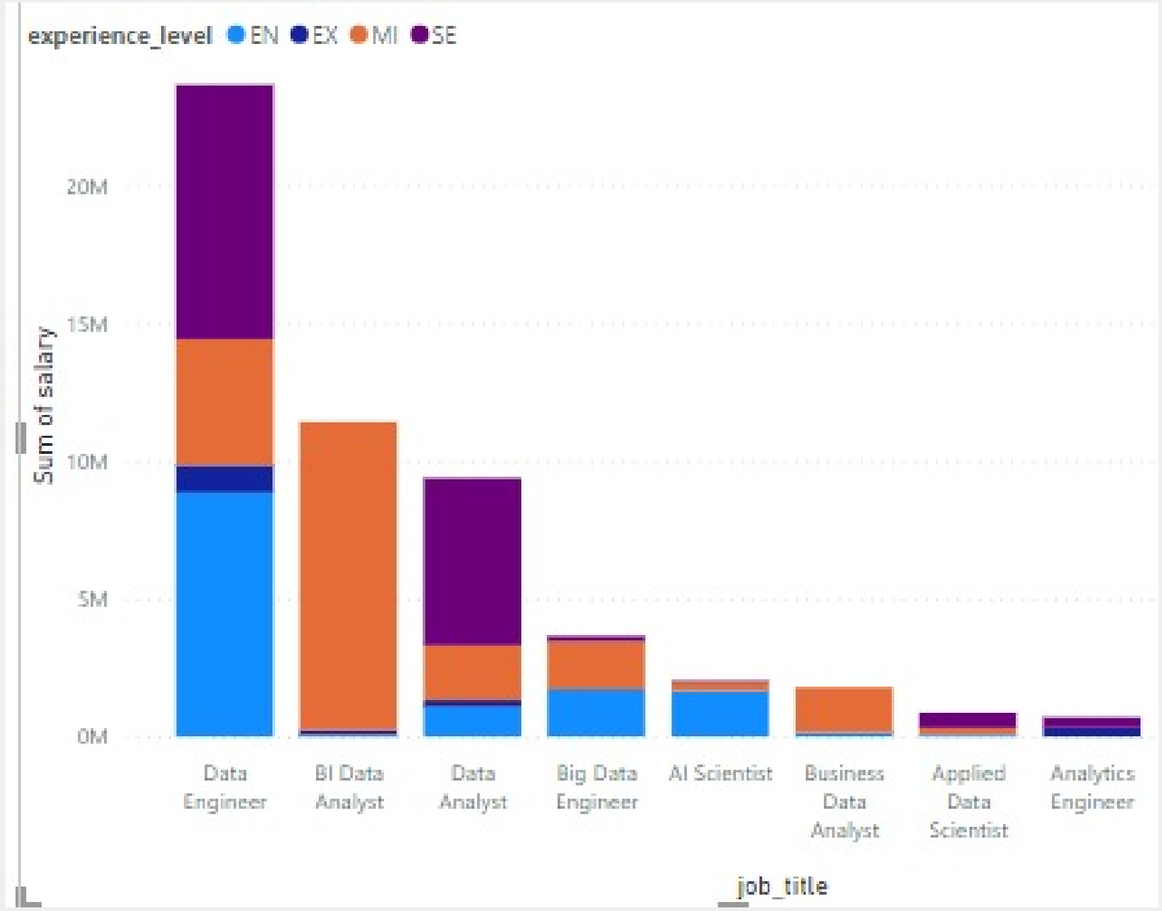
STACKED BAR CHART



In large companies, full-time roles predominate; small companies favor contractors, while medium-sized ones utilize both full-time and part-time positions.



In executive roles, both full-time and contractor employment arrangements are common. Entry-level positions typically lack full-time opportunities, while individuals with mid-level experience may encounter various employment options.



In the field of data engineering, professionals exhibit a diverse range of experience levels. BI analysts primarily possess mid-level expertise, whereas data analysts tend to be predominantly senior-level practitioners. Overall, the collective experience profile across these roles tends toward the mid-level range.

CONTINGENCY TABLES

Contingency Tables

Contingency Tables

work_year	company_size L	M	S	Total
2020	33	14	25	72
2021	119	53	45	217
2022	46	259	13	318
Total	198	326	83	607

Chi-Squared Tests

	Value	df	p
X^2	217.387	4	< .001
N	607		

In 2021, there were 119 large companies, whereas in 2022, the count of medium-sized companies surged to 250.

Contingency Tables ▼

Contingency Tables ▼

experience_level	company_size L	M	S	Total
EN	29	30	29	88
EX	11	12	3	26
MI	86	98	29	213
SE	72	186	22	280
Total	198	326	83	607

Chi-Squared Tests

	Value	df	p
X^2	57.081	6	< .001
N	607		

In large companies, the majority of employees hold mid-level positions, while in medium-sized companies, senior-level roles are more prevalent.

Contingency Tables ▼

Contingency Tables

company_size	remote_ratio 0	50	100	Total
L	32	60	106	198
M	79	21	226	326
S	16	18	49	83
Total	127	99	381	607

Chi-Squared Tests ▼

	Value	df	p
X^2	53.773	4	< .001
N	607		

Remote work preferences are becoming omnipresent across all types of companies, irrespective of their size—be it large, medium, or small.

The number of remote workers has been steadily increasing over the years: in 2020, there were 36 remote workers; this number rose to 117 in 2021, and further increased to 228 in 2022.

Contingency Tables ▼

Contingency Tables

work_year	remote_ratio 0	50	100	Total
2020	15	21	36	72
2021	34	66	117	217
2022	78	12	228	318
Total	127	99	381	607

Chi-Squared Tests

	Value	df	p
X^2	77.867	4	< .001
N	607		

HYPOTHESIS TESTING

One Sample T-Test

One Sample T-Test

	t	df	p	Mean Difference
salary	5.169	606	< .001	323996.063

Note. For the Student t-test, location difference estimate is given by the sample mean difference *d*.
Note. For the Student t-test, the alternative hypothesis specifies that the mean is different from 4.
Note. Student's t-test.

Descriptives

Descriptives

	N	Mean	SD	SE	Coefficient of variation
salary	607	324000.063	$1.544 \times 10^{+6}$	62683.537	4.767

H0: there is no significant difference in their mean

H1: there is significant difference in their men

One Sample T-Test ▼

One Sample T-Test

	t	df	p	Mean Difference
salary_in_usd	38.990	606	< .001	112293.870

Note. For the Student t-test, location difference estimate is given by the sample mean difference *d*.
Note. For the Student t-test, the alternative hypothesis specifies that the mean is different from 4.
Note. Student's t-test.

Descriptives ▼

Descriptives ▼

	N	Mean	SD	SE	Coefficient of variation
salary_in_usd	607	112297.870	70957.259	2880.066	0.632

Insights:

Mean is different

Paired Samples T-Test ▼

Paired Samples T-Test ▼

							95% CI for Mean Difference	
Measure 1		Measure 2	t	df	p	Mean Difference	SE Difference	Lower	Upper
salary	-	salary_in_usd	3.361	606	< .001	211702.193	62990.603	87995.810	335408.575

Note. Student's t-test.

Assumption Checks

Test of Normality (Shapiro-Wilk)

			W	p
salary	-	salary_in_usd	0.121	< .001

Note. Significant results suggest a deviation from normality.

A paired sample t-test was conducted on salary data, measured in USD, revealing a significant difference in their means. Both sets of data exhibited deviations from normality.

The obtained p-value, being less than 0.05, indicates no statistically significant difference in the variance between the two columns.

ANOVA

ANOVA - salary

Cases	Sum of Squares	df	Mean Square	F	p
remote_ratio	$1.761\times10^{+13}$	2	$8.804\times10^{+12}$	3.724	0.025
Residuals	$1.428\times10^{+15}$	604	$2.364\times10^{+12}$		

Note. Type III Sum of Squares

ANOVA

ANOVA

ANOVA - salary

Cases	Sum of Squares	df	Mean Square	F	p
experience_level	$9.202\times10^{+12}$	3	$3.067\times10^{+12}$	1.288	0.278
Residuals	$1.436\times10^{+15}$	603	$2.382\times10^{+12}$		

Note. Type III Sum of Squares

H0: There is no difference in the variance

H1: there is a significant difference

The p-value exceeding 0.05 indicates a statistically significant difference in the variance between the two columns.(salary and experience level)

One Sample T-Test

	Z	p
salary	$7.982\times10^{+6}$	< .001

Note. For the Z-test, the alternative hypothesis specifies that the mean is different from 5.

Note. Z test.

One Sample T-Test

	Z	p
salary	$7.982 \times 10^{+6}$	< .001

Note. For the Z-test, the alternative hypothesis specifies that the mean is different from 5.
Note. Z test.

Assumption Checks

Test of Normality (Shapiro-Wilk)

	W	p
salary	0.139	< .001

Note. Significant results suggest a deviation from normality.

The Shapiro-Wilk test is a statistical test used to assess the normality of a data sample. It evaluates whether the data follows a normal distribution based on the sample's observed values and their expected frequencies under the assumption of normality.

Insights:

Data (salary) is slightly different from normal

CONCLUSIONS:

- Senior executives typically earn higher salaries due to their extensive experience and leadership roles
- Remote work is increasingly favored, reflecting evolving work preferences and technological advancements.
- Fresh graduates often start their careers in medium-sized companies.
- Data-related roles such as data scientists, data engineers, and machine learning engineers offer lucrative salary packages and are highly sought after.
- Countries like the USA, Britain, and Canada have substantial numbers of data scientists, indicating the importance of data expertise in these economies.


thank
you.