Project Phase 3:
Salary Distribution Across Different Experience Levels and Job Categories

Computer and Technology 1

Gursimran Sujlana

Table of Contents:

Problem/Challenge

What we want to know: What influences salary more, Experience level or Job Title

The dataset classifies the **experience level** of employees ranging from “**Entry-Level**” to “**Executive**”.

The dataset also gives specific **job titles** within the data field such as ‘**Data Scientist**’ or ‘**Data Engineer**’.

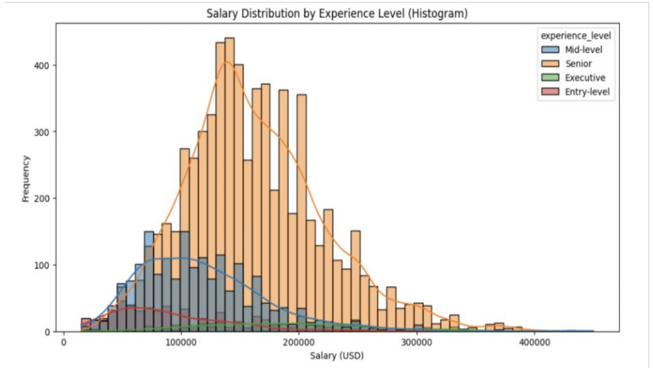
Our objective is to determine whether experience level or job title has a greater impact on salary.

Now we have leveraged **predictive modeling** to see which feature (experience level or job title) influences the salary the most.

Summary Metrics: Experience Level

Summary by Experience Level:

experience_level	Count	Mean	Std	Min	25%	Median	75%	Max	Skewness	Kurtosis
Entry-level	496	88534.8	49102.1	15000	51726	80000	120000	281700	1.03048	1.0189
Executive	281	189463	68793	15000	140000	185000	235000	416000	0.367679	-0.102954
Mid-level	1869	117524	55453.6	15000	75000	110000	149600	450000	1.27149	3.25047
Senior	6709	162356	59523	18381	122600	155000	198800	412000	0.629989	0.607562



The **Mid-level experience** category exhibits the **highest skewness (1.27149) and kurtosis (3.25047)** among the groups, indicating that **salaries are largely concentrated on the lower end with a right-skewed distribution**, and there's a pronounced **presence of outliers** with more extreme salary values.

Summary Metrics: Job Title

Summary by Job Title (Top 10):

job_title	Count	Mean	Std	Min	25%	Median	75%	Max	Skewness	Kurtosis
Analytics Engineer	256	155239	55607.5	37573	116920	149400	185175	430640	0.8867	2.27501
Applied Scientist	272	190172	50196.3	20000	136000	192000	222200	350000	0.155353	0.0788192
Business Intelligence Engineer	144	151405	52944.3	43064	104300	156400	185225	259000	-0.0899781	-1.06411
Data Analyst	1388	109911	42994.1	15000	80000	105320	135000	430967	1.10128	4.15142
Data Architect	213	164061	56105.9	52500	120000	159500	192564	376080	0.90724	1.37404
Data Engineer	2195	146620	56643.6	18000	106800	140000	180000	385000	0.581873	0.324287
Data Scientist	1989	156681	59914.4	16000	120000	154800	190000	412000	0.390596	0.601277
Machine Learning Engineer	991	184786	61760.6	20000	142200	182200	220000	392000	0.21831	0.0371832
Research Engineer	144	182840	68469.4	16455	139750	169056	226250	385000	0.686569	0.618615
Research Scientist	269	184376	68479	23000	144000	175000	220000	450000	0.686175	0.917388

The **Data Scientist** role exhibits a **moderate kurtosis** (0.601277), which is **higher than some other roles like Machine Learning Engineer and Applied Scientist,** but **Lower than Research Scientist and Engineer** indicating a **slightly more peaked distribution** with the potential for more outliers compared to a normal distribution, but less so than the Data Analyst role.

Baye's Theorem

Bayes' Theorem Results:

	Description	Value
P(A)	Probability of earning above the high-salary threshold	0.250027
P(B1)	Probability of being a Senior	0.717157
P(B2)	Probability of being a Data Scientist	0.212614
P(A|B1)	Probability of earning above the threshold given being a Senior	0.297511
P(A|B2)	Probability of earning above the threshold given being a Data Scientist	0.262946

- **Earning Above Threshold:** There's a 25% chance overall of earning above the high-salary threshold.
- **Senior Likelihood:** Seniors are more likely to earn above this threshold, with a probability of approximately 30%.
- **Data Scientist Potential:** Data Scientists have around a 26% probability of earning above the threshold.

Null Hypothesis

H0: There is no significant difference in salaries (USD) across different experience levels
H1: There is a significant difference in salaries (USD) across different experience levels

H0: There is no significant difference in salary across different job categories.

H1: There is a significant difference in salary across different job categories.

Experience Level vs Salary

Entry-Level	Mid-Level	Senior	Executive
95000	95012	186000	210000
75000	224400	81800	168000
72000	138700	212000	219650
64000	43064	93300	136000
100000	36912	130000	170000
75000	140000	100000	145000
49216	120000	224400	250000
36912	204500	138700	210000
105000	142200	300000	212000
133000	155000	234000	190000
58300	110000	266500	220000
43187	222200	152000	120000
31310	136000	273400	185000
92280	185000	182200	125000
67672	79600	167500	212000
92280	133000	106500	190000
67672	58400	185900	125000
85000	90000	129300	87500
65000	70000	122000	135000
32974	170884	94500	100000
32974	113923	247300	230000
133000	184000	139700	180000
58400	123000	176000	247500
163800	165000	100000	172200
88200	118800	204500	220000

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Column 1	281	25823418	91898.2847	2512662834
Column 2	281	34900004	124199.3025	2596971326
Column 3	281	46016943	163761.363	3279772591
Column 4	281	53239079	189462.9146	4732473682

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	1.56036E+12	3	5.20121E+11	158.5508031	1.20663E-85	2.612848859
Within Groups	3.67413E+12	1120	3280470108			
Total	5.23449E+12	1123				

With a p-value of 1.21e−85, which is lower than our alpha of 0.05 **we reject our null hypothesis** therefore we know there is a significant difference of salaries across different experience levels.

Job Title on Salary

Anova: Single Factor

SUMMARY

Groups	Count	Sum	Average	Variance
Data Enginee	2260	330406703	146197.656	3262261521
Data Archite	259	40404611	156002.359	3252368739
Data Science	3014	493568348	163758.576	4007867059
Machine Lea	1428	255506110	178925.847	4726396791
Data Analysis	1457	158092836	108505.721	1924846713
Leadership a	503	73174438	145476.02	3609265750
BI and Visuali	313	42283828	135092.102	2428585987
Data Quality	55	5548371	100879.473	2834288206
Data Manage	61	6291536	103139.934	1937547759
Cloud and Da	5	775000	155000	825000000

ANOVA

Source of Variati	SS	df	MS	F	P-value	F crit
Between Gro	4.6618E+12	9	5.1798E+11	148.146914	9.327E-263	1.88088427
Within Group	3.2674E+13	9345	3496370579			
Total	3.7335E+13	9354				

The p-value is very small, less than the significance level of 0.05 meaning **we reject the null hypothesis (H0)** and conclude that there is a significant difference in salaries across job categories

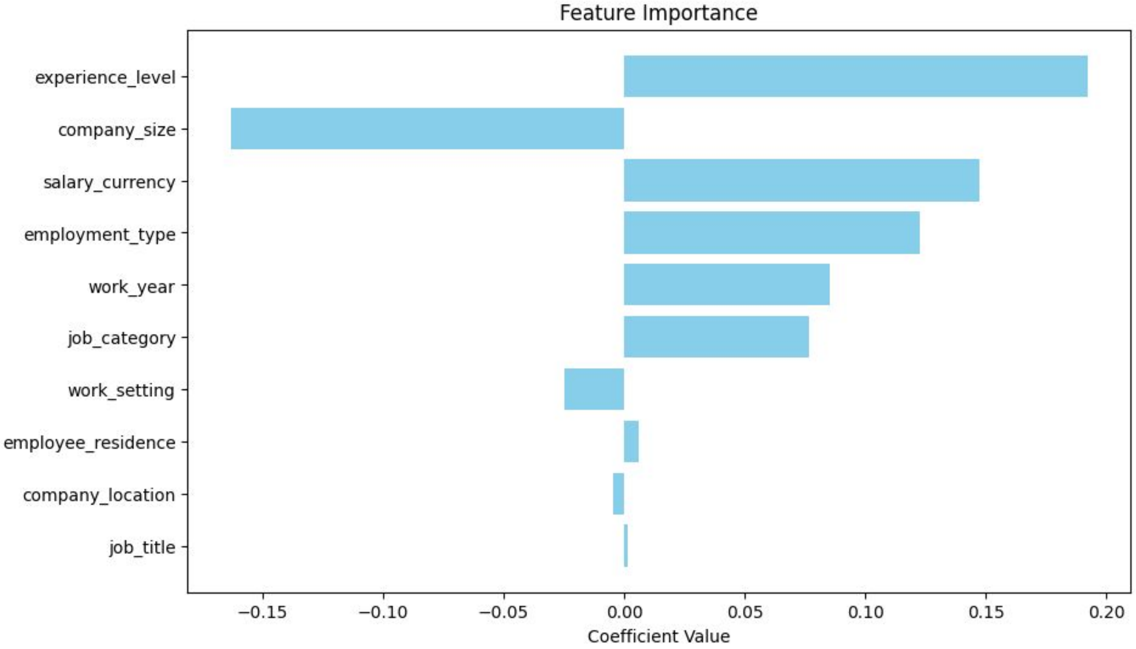
Linear Regression: All Features on Salary

	Train Metrics		**Test Metrics**
R-squared	0.23867	R-squared	0.18809
MAE	0.52434	MAE	0.53150
MSE	0.46425	MSE	0.47965
RMSE	0.68136	RMSE	0.69257

Key Insights

- Signs of overfitting with higher training R-squared
- More difficult time generalizing on new data
- General difference between predicted and actual is 0.69
- Possibility of Irrelevant Features and nonlinear relationships

Linear Regression: All Features on Salary



Top Coefficients:

Feature	Absolute Coefficient
experience_level	0.192399
company_size	(-) 0.163101
salary_currency	0.147401
employment_type	0.122811
work_year	0.085313
job_category	0.076912
work_setting	(-) 0.024810
employee_residence	0.006011
company_location	(-) 0.004534
job_title	0.001631

Linear Regression: All Features on Salary

Feature Analysis

- Experience level has the most significant impact on salary
- Larger companies tend to have lower salaries
- Salaries range differently in employment types
- Positive significance in work_year indicate salaries may increase overtime
- Job category, work settings, employee residence, and company location have the lesser importance in affecting salary

Conclusion

The analysis highlights significant salary differences across experience levels and job categories. While predictive accuracy was limited, the feature analysis underscores the importance of experience level in salary determination.

Future Work:

Refining the model for better prediction accuracy and exploring additional features and outlier detection methods would be essential steps for further enhancing insights and decision-making capabilities.

Thank You,

Questions?