Data Science Salaries

Group 1/ Wannabe Inc.
Clayton Chan, Hannah Aguirre, Karen Hong,
Juyi Yang, Elbert Liu

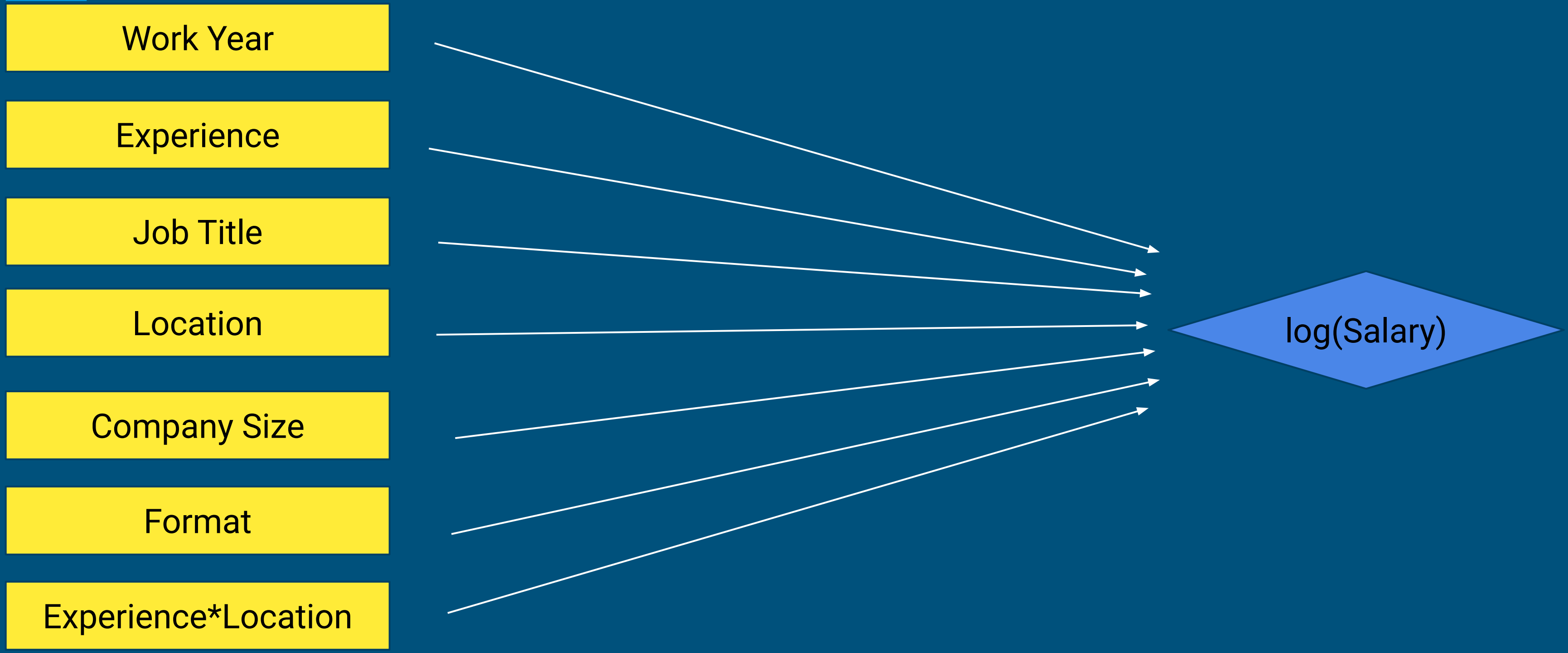
Abstract

We're given a dataset for data science salaries and we want to see if we can build something to help us predict the salary for jobs not in the dataset. We had questions where we wanted to see how well the model predicts the salaries and which variables are the most important predictors. We were also focused on whether the factors like location, job title, format, and year had an effect on the salary. We fitted both a regression and anova model. A summary of the results found that all the variables (including the ones of interest) were significant predictors and had an effect with experience*location being the most important interaction effect. The answer to our main question regarding predictive power was that the regression model had a fit that wasn't good/bad at R^2=.4 .

Research Questions

-How well can we predict salaries from our dataset?
-Which variables are the most important predictors for salary?
-Is there a difference between US and Non-US Markets?
-Is there a difference between job titles?
-Is there a difference between remote, hybrid, and in person?
-Has salary increased with year?

Schematic/ Roadmap



Codebook

1. **work_year:** The year of the data related to the job salary. 5 levels 2020,2021,2022,2023,2024
2. **experience_level:** The level of experience of the employee 4 levels entry-level, mid-level, senior-level, executive
3. **employment_type:** The type of employment 4 levels full-time, part-time, contract, freelance
4. **job_title:** The title or role of the employee within the data science field. 100+ levels
5. **salary:** The salary of the employee.
6. **salary_currency:** The currency in which the salary is denoted.
7. **salary_in_usd:** The salary converted to US dollars for standardization.
8. **employee_residence:** The residence location of the employee. 70+ levels
9. **remote_ratio:** The ratio of remote work allowed for the position. 3 levels Remote, Hybrid, In-Person
10. **company_location:** The location of the company. 70+ levels
11. **company_size:** The size of the company based on employee count or revenue. 3 levels Small, Medium, Large

Data Cleaning

-Duplicates

-Dropping Redundant Variables

-Feature Engineering

Employment Type

Full-Time	Part-Time	Contract	Freelance
9061	27	26	13

-too imbalanced!

-focus on full-time

Work Year

2020	2021	2022	2023	2024
69	206	1099	4616	3071

-Also imbalanced

-Remedy by combining 2020-2022 into one level called “Pandemic”

Company Location

-70+ countries

-Combine the levels based on economy

-Leave US as its own level for reference

-3 levels US, First World, Developing

Job Title

-Over 100+ different job titles

-Many are redundant/similar ex. Data Analyst and Applied Data Analyst

-Remedy by combining similar titles

-6 total levels Data Analyst, Data Scientist, BI, ML, Data Engineer, Other

Updated Codebook

1. **work_year:** The year of the data related to the job salary. 3 levels Pandemic,2023,2024
2. **experience_level:** The level of experience of the employee 4 levels entry-level, mid-level, senior-level, executive
3. ~~**employment_type:** The type of employment ALL full-time~~
4. **job_title:** The title or role of the employee within the data science field. 6 levels DA, DS, DE, ML, BI, Other
5. ~~**salary:** The salary of the employee.~~
6. ~~**salary_currency:** The currency in which the salary is denoted.~~
7. **salary_in_usd:** The salary converted to US dollars for standardization.
8. ~~**employee_residence:** The residence location of the employee.~~
9. **remote_ratio:** The ratio of remote work allowed for the position. 3 levels In-Person, Hybrid, Remote
10. **company_location:** The location of the company. 3 levels US, First World, Developing
11. **company_size:** The size of the company based on employee count or revenue. 3 levels Small, Medium, Large

Exploratory Data Analysis: Salary by Work Year

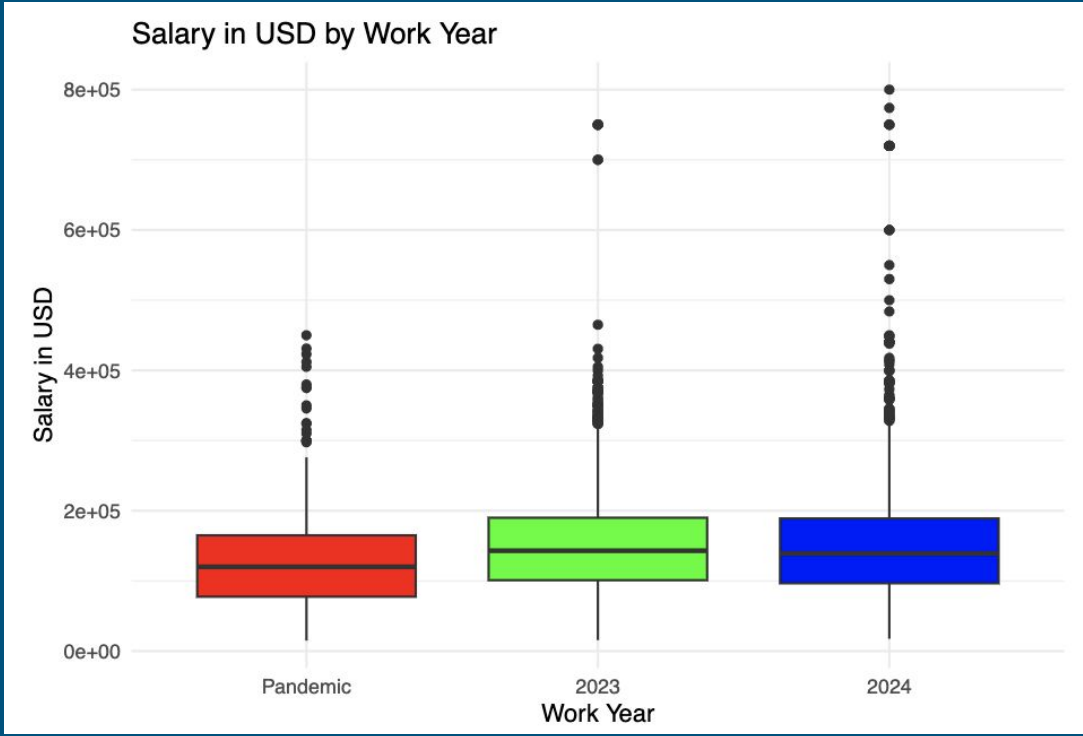
Median Salary is highest in

2023-2024.

All years are skewed,

with 2024 having the

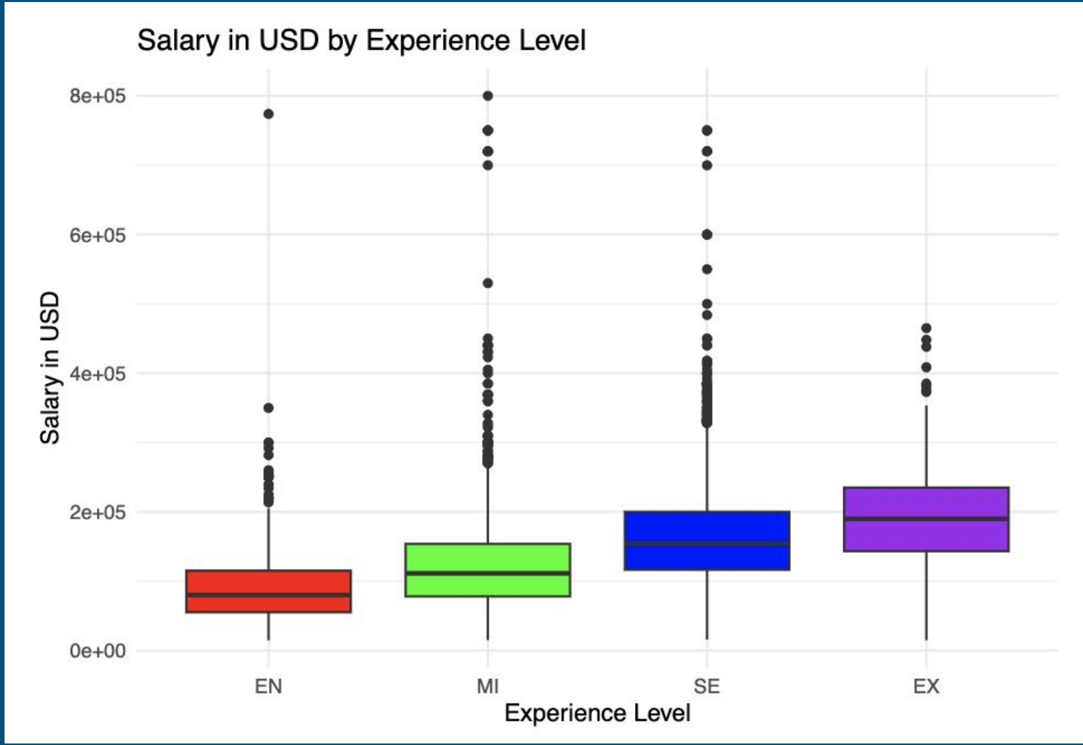
highest variability.



EDA: Salary by Experience Level

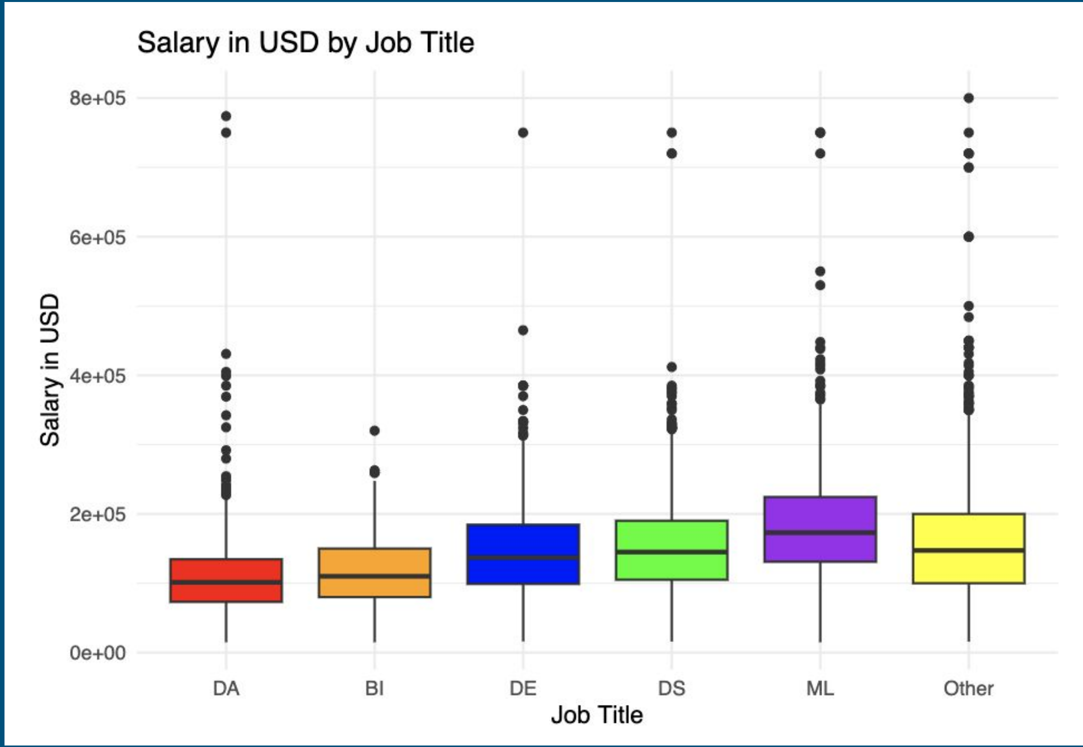
Executive make the highest median salary with the least amount of variability.

While Entry Level has the lowest median salary.



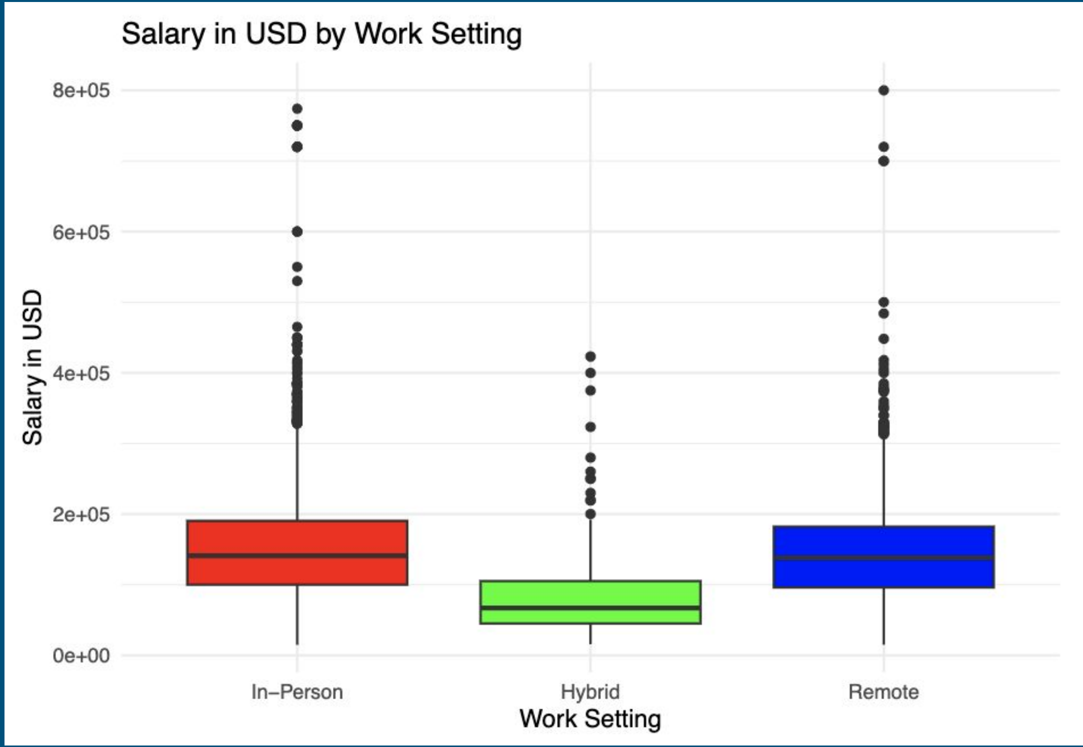
EDA: Salary by Job Title

Those who work in Machine Learning make the highest median salary, followed by other, Data Scientists, and Data Engineers. While Data Analysts make the lowest median salary.



EDA: Salary by Work Setting

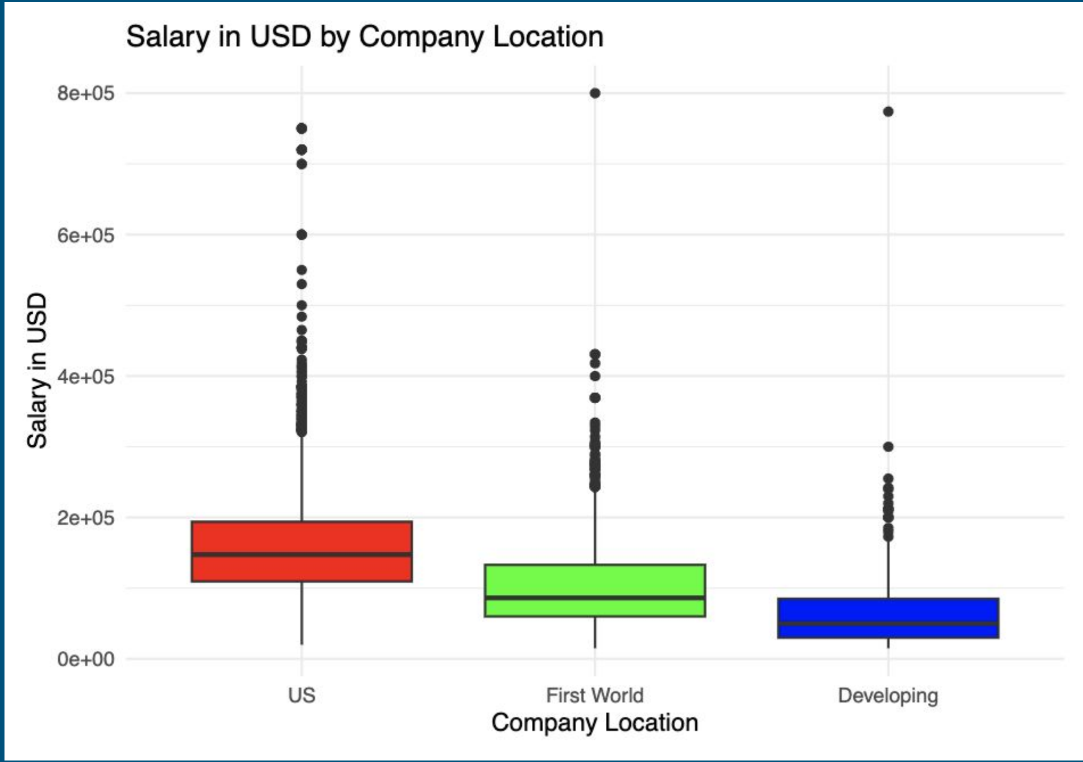
In-person and Remote workers make very similar median salaries, with those who work hybrid making a significant lower median salary.



EDA: Salary by Company Location

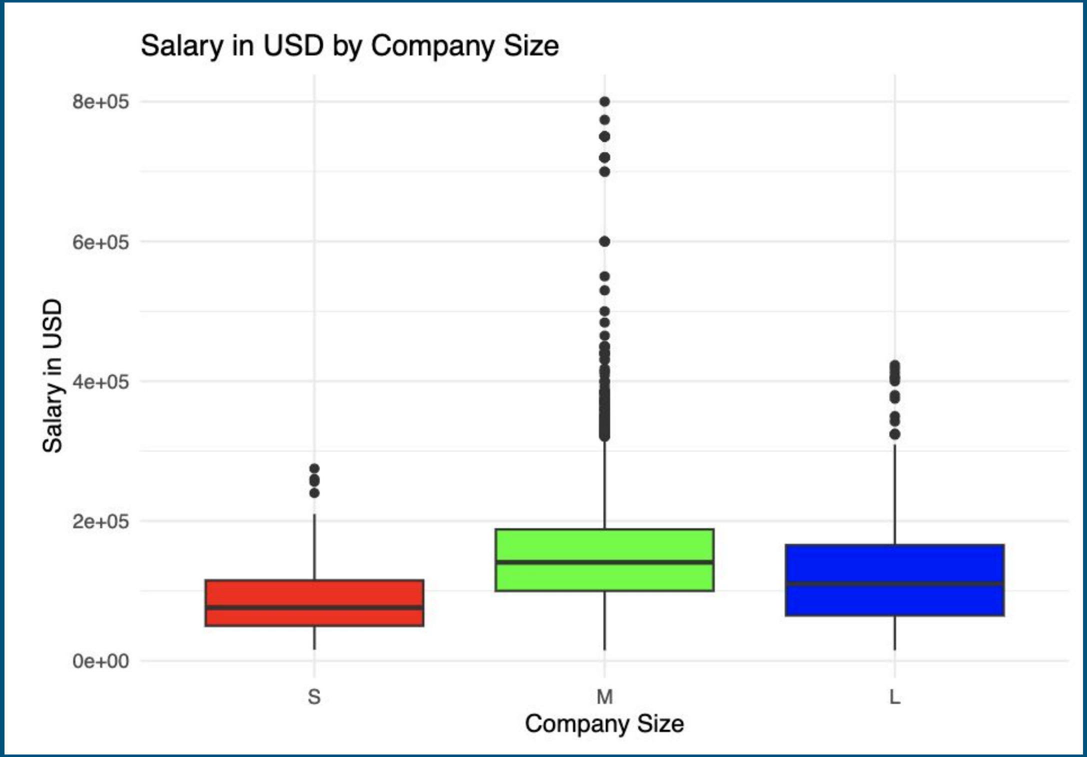
The US leads in median salary followed by First world, then developing countries.

All have large outliers.



EDA: Salary by Company Size

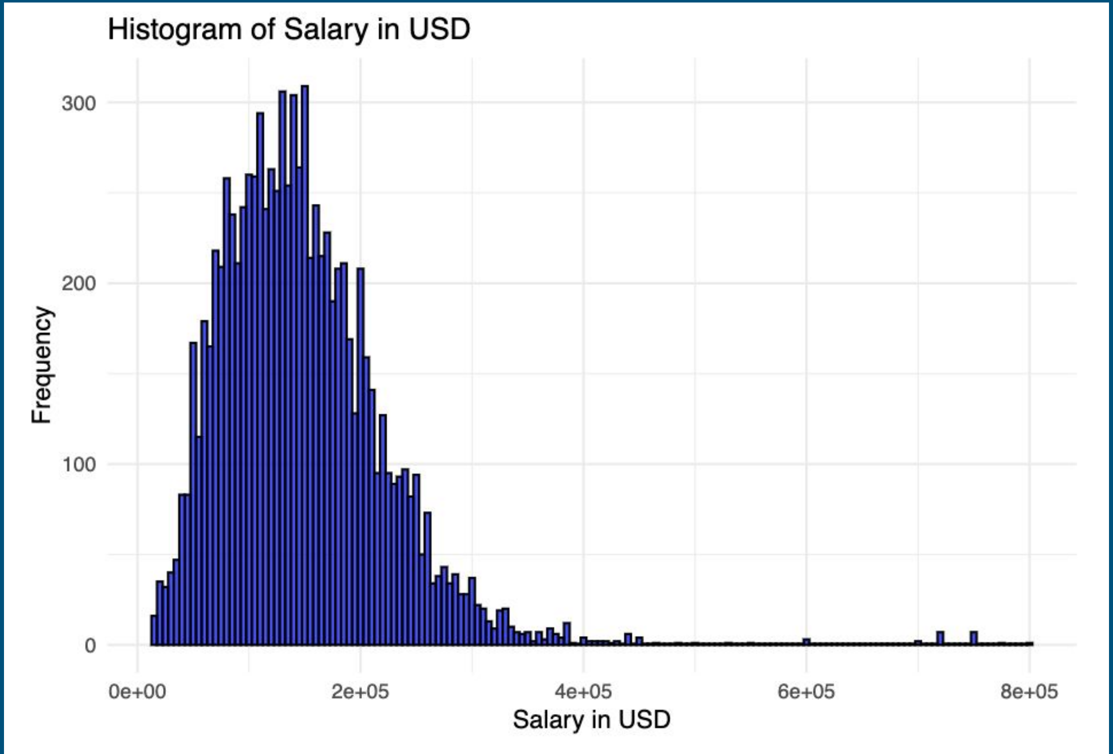
Medium size companies make the highest median salary , followed by Large, then Small size companies.



EDA: Response Variable

As we saw in all boxplots the data for salary in USD has a large amount of high outliers and is skewed right.

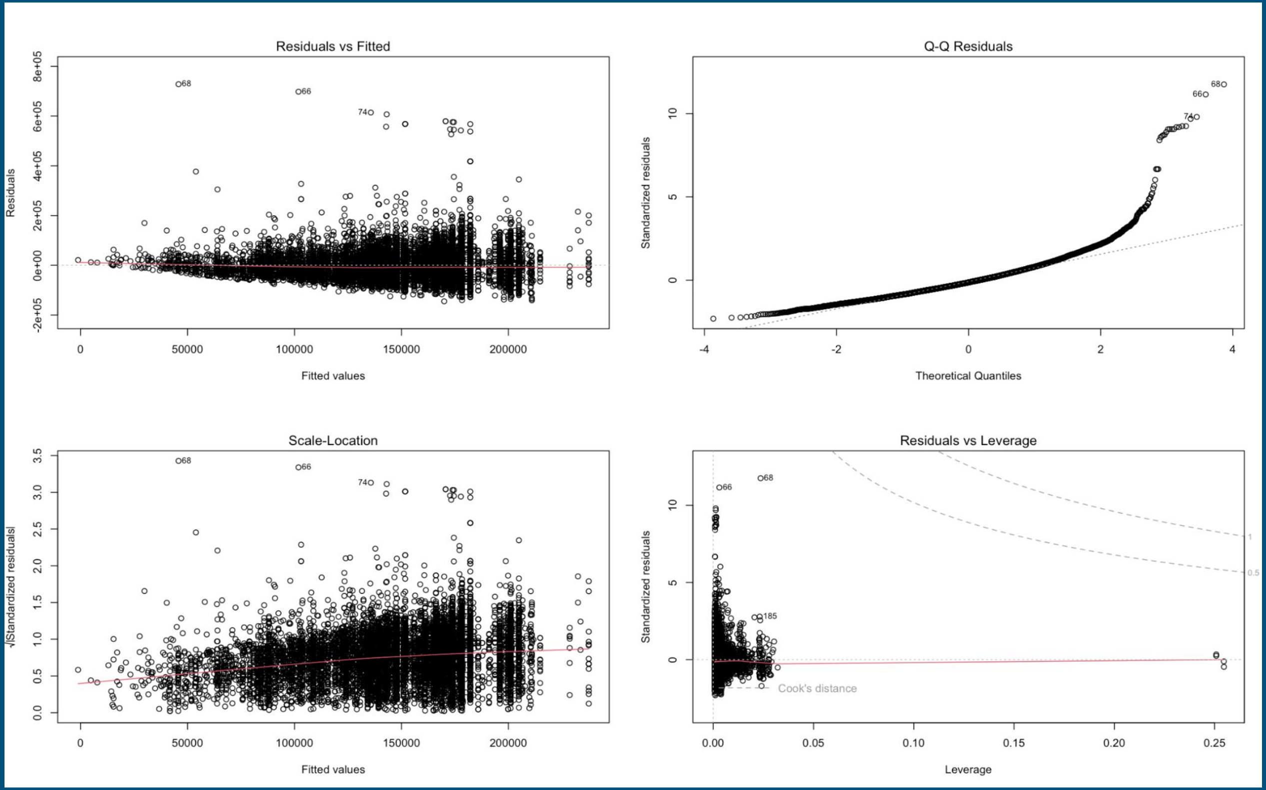
We want to consider transforming the data.



Residual Plot shows funnel shape.

= Heteroscedasticity

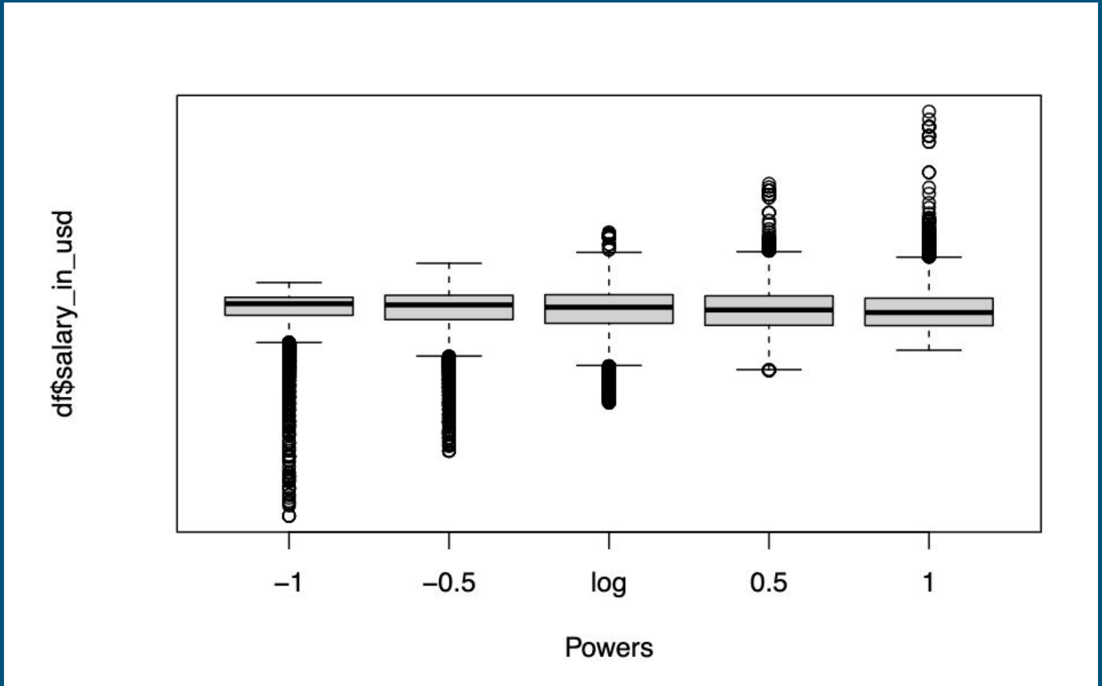
Q-Q Residual plot deviates from Normal lines in the 3rd quartile.



Different Linear Transformations using Salary in USD

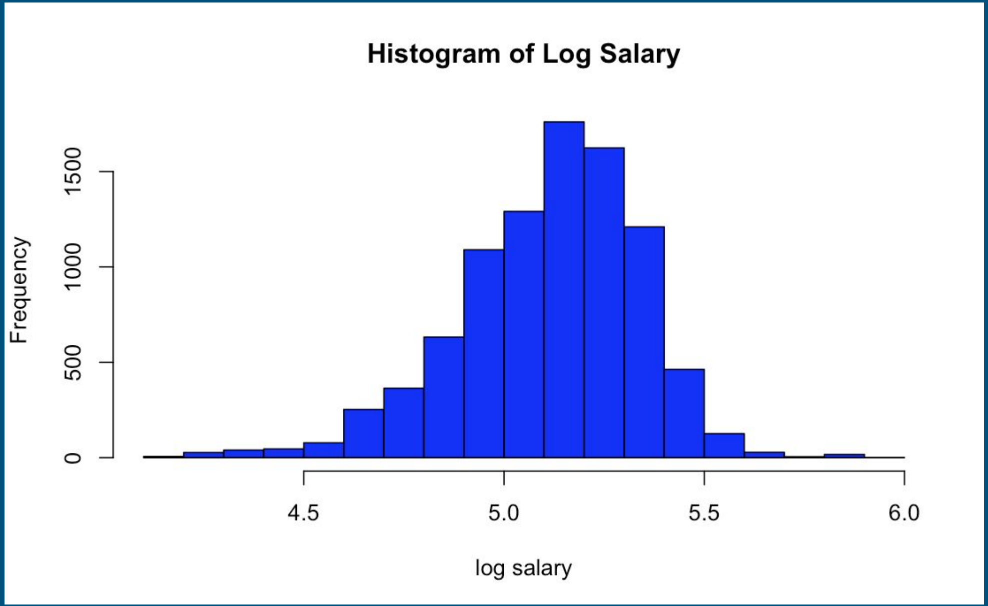
Points to log transformation to normalize our data.

We chose to use the log10 Transformation of Salary_in_USD.



EDA: Response Variable (Transformed)

Transformed data makes Histogram appear more symmetrical.



Exploratory Data Analysis: Response Variable

Improvement in homoscedasticity.

Q-Q Residual Plot more closely

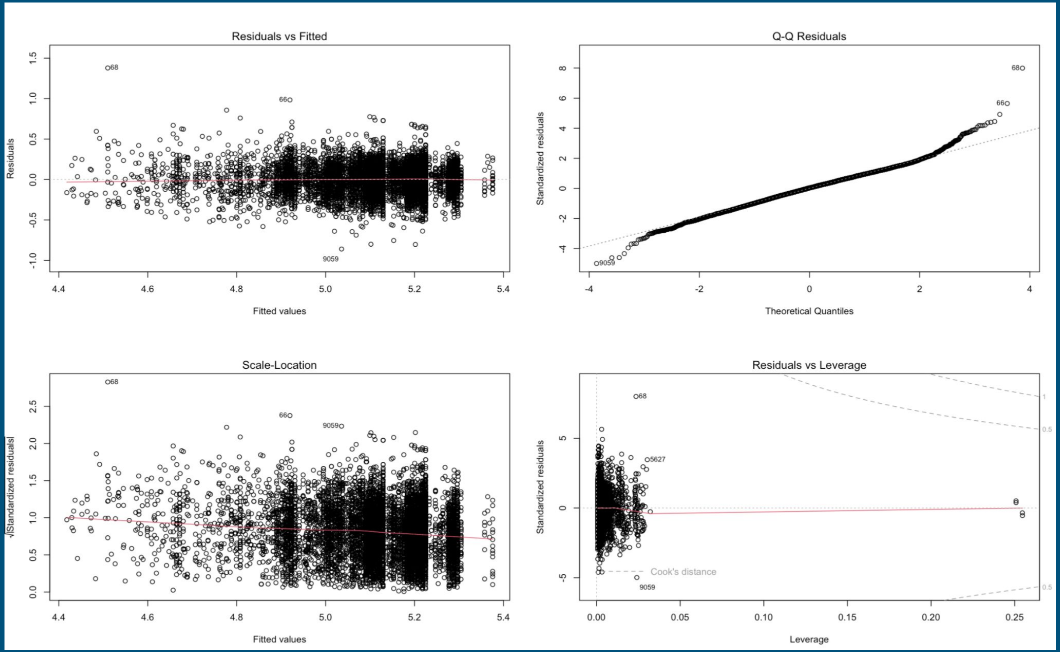
follows normal line.

Still has high outliers and leverage

points, but all are within

Cook's distance and will not

disproportionately impact the model.



Testing Assumptions:

ncvTest result: statistically significant

- Heteroscedasticity
- violates assumption of equality of variance
- Large Sample size: proceed with ANOVA

Non-constant Variance Score Test	
Chi-square	486.1625
Df	1
p-value	< 2.22e-16

Test for Multicollinearity

All predictors have low multicollinearity.

We will explore the Interaction effects between company_location and experience _level

Predictor	GVIF	Df	$GVIF^{\frac{1}{2 \times Df}}$	Interacts With
work_year	1.255330	2	1.058497	--
experience_level	1.303991	11	1.012138	company_location
job_title	1.099184	5	1.009502	--
remote_ratio	1.358309	2	1.079567	--
company_location	1.303991	11	1.012138	experience_level
company_size	1.387064	2	1.085236	--

Model Selection

- ANOVA on all individual factors as well as interaction effects to identify statistical significance.
- Compute and sort each variable by R^2.
- Multiple regression model with significant and meaningful independent variables.
- Examine interaction effects.

ANOVA

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
work_year	2	10.10	5.050	169.920	< 2e-16	***
experience_level	3	84.62	28.207	949.172	< 2e-16	***
job_title	5	23.38	4.676	157.355	< 2e-16	***
remote_ratio	2	9.07	4.533	152.543	< 2e-16	***
company_location	2	56.43	28.216	949.461	< 2e-16	***
company_size	2	0.82	0.412	13.851	9.86e-07	***
work_year:experience_level	6	0.58	0.097	3.255	0.003381	**
work_year:job_title	10	0.62	0.062	2.100	0.021189	*
work_year:remote_ratio	4	0.12	0.030	1.012	0.399650	
work_year:company_location	4	1.26	0.315	10.586	1.48e-08	***
work_year:company_size	4	0.42	0.104	3.504	0.007270	**
experience_level:job_title	15	1.23	0.082	2.765	0.000277	***
experience_level:remote_ratio	6	1.12	0.187	6.280	1.34e-06	***
experience_level:company_location	6	1.81	0.302	10.170	3.06e-11	***
experience_level:company_size	6	0.54	0.090	3.034	0.005778	**
job_title:remote_ratio	10	0.50	0.050	1.670	0.081409	.
job_title:company_location	10	0.42	0.042	1.417	0.165781	
job_title:company_size	10	1.09	0.109	3.664	6.69e-05	***
remote_ratio:company_location	4	0.49	0.121	4.088	0.002597	**
remote_ratio:company_size	4	0.51	0.126	4.249	0.001947	**
company_location:company_size	4	1.19	0.297	10.000	4.51e-08	***
Residuals	8941	265.71	0.030			

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

- All individual variables are statistically significant.
- Quite a few statistically significant interaction effects.

Important Predictors R^2

	R2
Residuals	0.5750943025
experience_level	0.1831551124
company_location	0.1221406205
job_title	0.0506062279
work_year	0.0218588891
remote_ratio	0.0196234498
experience_level:company_location	0.0039248805
work_year:company_location	0.0027237114
experience_level:job_title	0.0026676584
company_location:company_size	0.0025727473
experience_level:remote_ratio	0.0024236697
job_title:company_size	0.0023566482
company_size	0.0017818422
work_year:job_title	0.0013508417
work_year:experience_level	0.0012561126
experience_level:company_size	0.0011708209
remote_ratio:company_size	0.0010933180
job_title:remote_ratio	0.0010743141
remote_ratio:company_location	0.0010517302
job_title:company_location	0.0009111540
work_year:company_size	0.0009015959
work_year:remote_ratio	0.0002603527

- Listing out R^2 of each individual variable as well as all interaction effects gives us a better understanding of their predicting power for data science salary.
- Everything from work_year * company_location down contributes less than .3% to R^2 so we are not using them in the multiple regression model due to their lack of practical significance.

Multiple Regression

- Final multiple regression model consists of all individual variables and the interaction effect between experience level and company location.
- This interaction effect is interesting because it gives us a comparison of career trajectory between different job markets.

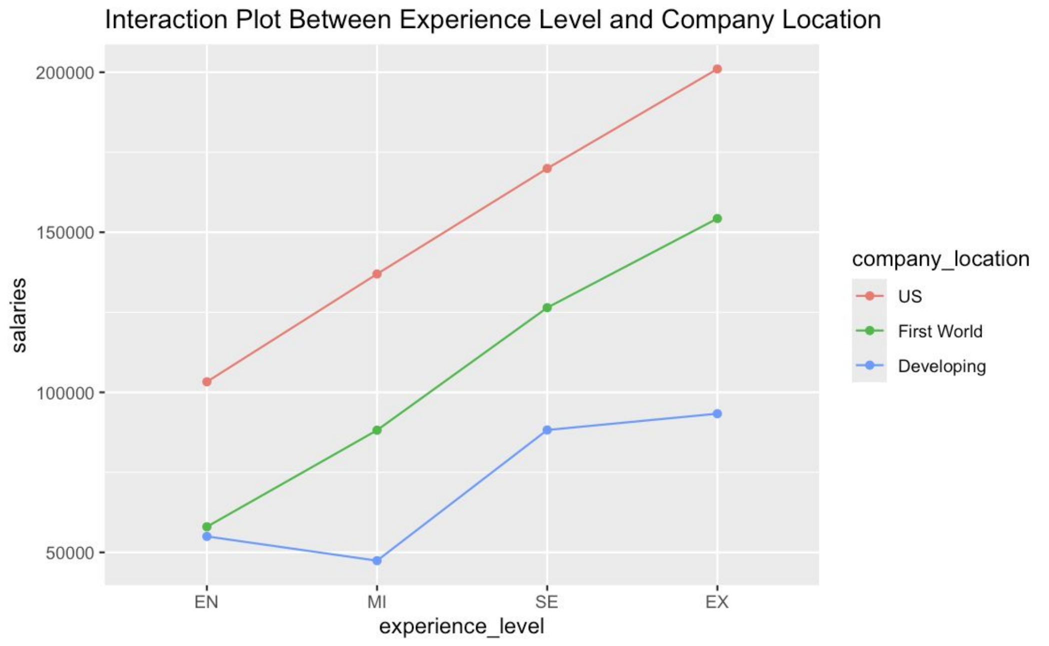
Table: Multiple Regression Model
Dependent Variable: **Log 10 of Salary in USD**

	Estimates
Intercept	4.8107***
	(.016)
Work Year (Base = Pandemic)	
2023	0.0307***
	(.0058)
2024	.0375***
	(.0063)
Experience Level (Base = Entry Level)	
Mid Level	.0909***
	(.0082)
Senior Level	.1866***
	(.0076)
Executive	.2664***
	(.0124)
Job Title (Base = Data Analyst)	
Business Intelligence	,0186*
	(.0093)
Data Engineer	.0933***
	(.0064)
Data Scientist	.1160***
	(.0061)
Machine Learning Engineer	.1873***
	(.0069)
Other	.1163***
	(.0062)
Remote Ratio (Base = Onsite)	
Hybrid	-.0612***
	(.0132)
Remote	-.0113**
	(.0040)
Company Location (Base = U.S)	
Non-U.S First World Economy	-.2504***
	(.0148)
Developing Economy	-.4120***
	(.0272)

Table: Multiple Regression Model
Dependent Variable: **Log 10 of Salary in USD**

	Estimates
Company Size (Base = Small)	
Medium	.0738***
	(.0147)
Large	.0647***
	(.0156)
Interaction Effect: Experience Level & Company Location (Base = Entry Level & U.S)	
Mid Level & Non-U.S First World	.0522**
	(.0170)
Senior Level & Non-U.S First World	.1093***
	(.0165)
Executive & Non-U.S First World	.1352***
	(.0315)
Mid Level & Developing	-.0598
	(.0342)
Senior Level & Developing	.0837*
	(.0334)
Executive & Developing	.1091
	(.0919)
Observations	9061
Multiple R-Squared	.4045
Adjusted R-Squared	.403

Interaction Effect



- U.S is still the best place for a career, since at all experience level, the average salaries are the highest.
- Salary trajectory is steeper in non-U.S first world and US economies.
  - Going from Entry level to Senior level in the U.S saw around 60% gain in nominal salary, while it's more than 100% in non-U.S first world economies.
  - Experience seems to matter less for developing countries since entry/mid and senior/executive don't have a huge disparity in salary

Tukey HSD

- Tukey's Honestly Significant Difference (HSD) test is a statistical tool that determines if the difference between sample means is statistically significant.

- By using Tukey's HSD, we gain a more precise understanding of exactly where these salary differences exist, enabling a more targeted interpretation of the factors influencing salary.

Tukey HSD Work Year Table Summary

Work Year	Difference (diff)	Lower CI (lwr)	Upper CI (upr)	p-value (p adj)
2023 - Pandemic	0.0082	0.0072	0.0093	0.0000
2024 - Pandemic	0.0080	0.0069	0.0091	0.0000
2024 - 2023	-0.0002	-0.0011	0.0006	0.7661

- Shows salaries significantly increased from the pandemic period to 2023 and remained high in 2024.
- Not much difference from 2023-2024 because the p-value is too large.

Tukey HSD Experience Level Table Summary

Experience Level	Difference (diff)	Lower CI (lwr)	Upper CI (upr)	p-value (p adj)
MI - EN	0.0124	0.0108	0.0139	0.0000
SE - EN	0.0244	0.0230	0.0258	0.0000
EX - EN	0.0310	0.0286	0.0334	0.0000
SE - MI	0.0120	0.0111	0.0130	0.0000
EX - MI	0.0186	0.0164	0.0208	0.0000
EX - SE	0.0066	0.0044	0.0087	0.0000

- Salaries increase significantly with experience level, with each comparison (mid-level, senior, executive) showing a positive and significant difference
- Indicates that more experience is consistently associated with higher pay.

Tukey HSD Job Title Table Summary

Job Title	Difference (diff)	Lower CI (lwr)	Upper CI (upr)	p-value (p adj)
BI - DA	0.0003	-0.0020	0.0026	0.9990
DE - DA	0.0068	0.0053	0.0084	0.0000
DS - DA	0.0086	0.0071	0.0101	0.0000
ML - DA	0.0139	0.0123	0.0156	0.0000
Other - DA	0.0089	0.0074	0.0104	0.0000
DE - BI	0.0065	0.0043	0.0087	0.0000
DS - BI	0.0083	0.0061	0.0105	0.0000
ML - BI	0.0136	0.0113	0.0159	0.0000
Other - BI	0.0086	0.0065	0.0108	0.0000
DS - DE	0.0018	0.0004	0.0031	0.0033
ML - DE	0.0071	0.0055	0.0087	0.0000
Other - DE	0.0021	0.0007	0.0035	0.0002
ML - DS	0.0053	0.0038	0.0068	0.0000
Other - DS	0.0003	-0.0010	0.0017	0.9796
Other - ML	-0.0050	-0.0065	-0.0034	0.0000

- ***Machine Learning Engineers (ML)*** earn significantly more compared to Data Analysts (DA), Business Intelligence (BI), Data Engineers (DE), and even Data Scientists (DS)
- Data Engineers (DE) and Data Scientists (DS) also show significantly higher salaries than Data Analysts (DA), with p-values all close to zero.
- There is no significant salary difference between salaries for roles labeled "Other" and Data Scientists (DS) & Business Intelligence(BI) and Data Analysts (DA)

Tukey HSD Remote Ratio Table Summary

Remote Ratio	Difference (diff)	Lower CI (lwr)	Upper CI (upr)	p-value (p adj)
Hybrid - In-Person	-0.0169	-0.0192	-0.0145	0.0000
Remote - In-Person	-0.0015	-0.0023	-0.0007	0.0000
Remote - Hybrid	0.0154	0.0130	0.0178	0.0000

- Hybrid workers earn significantly less than both fully in-person and fully remote workers, with a disadvantage in salary.
- Fully remote workers earn slightly more than hybrid workers, highlighting a small but significant salary benefit for remote work.

Tukey HSD Company Location Table Summary

Location	Difference (diff)	Lower CI (lwr)	Upper CI (upr)	p-value (p adj)
First World - US	-0.0132	-0.0143	-0.0122	0.0000
Developing - US	-0.0315	-0.0338	-0.0292	0.0000
Developing - First World	-0.0183	-0.0208	-0.0158	0.0000

- Salaries in the US are significantly higher compared to both "First World" and developing countries.
- Developing countries have significantly lower salaries compared to both the US (-0.0315) and "First World" countries (-0.0183), indicating a clear disparity in compensation across regions.

Tukey HSD Company Size Table Summary

Company Size	Difference (diff)	Lower CI (lwr)	Upper CI (upr)	p-value (p adj)
Medium - Small (M-S)	0.0057	0.0029	0.0085	0.0001
Large - Small (L-S)	0.0055	0.0024	0.0087	0.0001
Large - Medium (L-M)	-0.0002	-0.0017	0.0013	0.9553

- Medium and large companies both offer significantly higher salaries compared to small companies
- There is no significant difference in salaries between large and medium-sized companies (p-value = 0.9553)

Cross-Validation

We will evaluate two regression models using **K-fold Cross-Validation** and **Leave-One-Out Cross-Validation (LOOCV)** to compare their predictive performance. These methods ensure a robust assessment of model accuracy and generalizability.

Subsequently, we will compare the **Akaike Information Criterion (AIC)** of both models to determine which strikes a better balance between goodness of fit and complexity, with a lower AIC indicating a more optimal model.

Important Concepts

RMSE:

- the square root of the average of the squared differences between the actual and predicted values.
- A measure of how well the model's predictions match the actual data.

R-squared:

- measures the proportion of the variance in the dependent variable that is predictable from the independent variables.
- It shows how well the data fits the model.

MAE:

- the average of the absolute differences between actual and predicted values.
- It measures the average magnitude of errors in the predictions, without considering their direction.

AIC:

- A metric used for model selection that takes into account the goodness of fit and the complexity of the model.
- A lower AIC value indicates a better model, balancing fit and complexity.

Comparison (10-fold)

Model contains individual variables and interaction effect experience level*company_location:

- **RMSE: 0.4024323**
- **R-squared: 0.4019016**
- **MAE: 0.3140735**

Model contains individual variables only:

- **RMSE: 0.4037524**
- **R-squared: 0.3973163**
- **MAE: 0.3153437**

The **first model** has lower RMSE and MAE values, suggesting better predictive accuracy. The **R-squared** value is also higher in the first model, indicating that it explains more variance in the dependent variable.

Comparison (LOOCV)

Model contains individual variables and interaction effect experience level*company_location:

- **RMSE: 0.4025759**
- **R-squared: 0.400528**
- **MAE: 0.3140832**

Model contains individual variables only:

- **RMSE: 0.4039293**
- **R-squared: 0.3964852**
- **MAE: 0.3153773**

The first model has a lower RMSE and MAE, indicating better predictive accuracy and lower average error compared to the first model. The R-squared for the first model is also higher than that of the second model, suggesting that the second model explains a greater proportion of the variance in the dependent variable.

AIC

Model contains individual variables and interaction effect:

- **AIC: 9213.094**

Model contains individual variables only:

- **AIC: 9281.726**

We believe the model that contains individual variables and interaction effect is the better one between these two

Recommendations/Shortcomings

-No numerical variables. It might be more helpful to have experience as a numerical variable in years

-Missing lots of important data like education level that could improve the fit

Conclusion

- **How well can we predict salaries from our dataset?** The regression model has moderate predictive power, with an R^2 value of 0.4, therefore it wasn't good/bad.

- **Most Important Variable:** Experience * Location are the most important interaction variable that affecting salary.

- **US vs Non-US Markets**: Salaries are higher in the US compared to first-world and developing countries. The US also shows stronger salary growth with experience.

Conclusion (cont’)

- **Job titles**: Machine Learning Engineers earn the most, while Data Analysts earn the least. There are some similar salaries among other roles like Data Scientists and Business Intelligence analysts.

- **Remote vs Hybrid vs In-Person**: Hybrid workers earn less compared to fully remote or in-person employees, who have similar earnings.

- **Salary trends**: Salaries have increased from the pandemic period to 2023 and remained high in 2024, possibly reflecting adjustments for inflation.

Conclusion - Revisit Research Question

- How well can we predict salaries from our dataset?
  - Our model's R^2 is 0.4. It has moderate predictive power and is neither particularly good or bad.
- Which variables are the most important predictors for salary?
  - The individual R^2 chart shows us the top 3 ranking goes: Experience Level > Company Location > Job Title
- Is there a difference between US and Non-US Markets?
  - Yes. The U.S is the best place to progress your data science career, followed by non-U.S first world economies, and then developing economies.

Conclusion - Revisit Research Question

- Is there a difference between job titles?
  - Yes, as we can see from the regression estimates and Tukey test, Machine Learning engineers make the most.
- Is there a difference between remote, hybrid, and in person?
  - Yes, in person positions has the highest average salary, followed by fully remote, and then hybrid.
- Has salary increased with year?
  - While the year 23 & 24 salaries are significantly higher than pandemic years, our Tukey test shows that the difference between 23 & 24 are not significant. So salary went up and plateaued.