June 28th 2024
EDA for ML Honors Project – Data Science Salaries 2022-2023

Description of the data set / summary of its attributes:

`work_year` : The year the salary was paid.

`experience_level` : The experience level in the job during the year

`employment_type` : The type of employment for the role

`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.

`salaryinusd` : The salary in USD

`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

`company_location` : The country of the employer's main office or contracting branch

`company_size` : The median number of people that worked for the company during the year

Plan for Data Exploration:

- Look for the information of the DataSet and the datatypes of its features.
- Determine whether the DataSet has NA values, and if so fill them with the necessary values (mean, median, 0, previous value, etc).
- Filter/Delete the columns that don't seem necessary for the analysis.
- Normalize/Standardize Data.
- Look for outliers with visualization tools and statistic methods.
- Create visualizations.

Actions taken for data cleaning and feature engineering:

- Explored the dataset for missing values.
- Some column names renamed to be more informative.
- Some records' data replaced with more informative categorical values.
- Dropped records where the year wasn't 2022 or 2023.
- Encoded some columns with get_dummies() to be later fitted with the Standard Scaler method of scikit-learn.
- Checked the skewness level for the continuous variable (salary).
- Left without doing the log transformation (skew = 0.5080).
- Did PCA() on the predictor features of the dataset already encoded with a 95% of n_components.

June 28th 2024

EDA for ML Honors Project – Data Science Salaries 2022-2023

Key Findings and Insights:

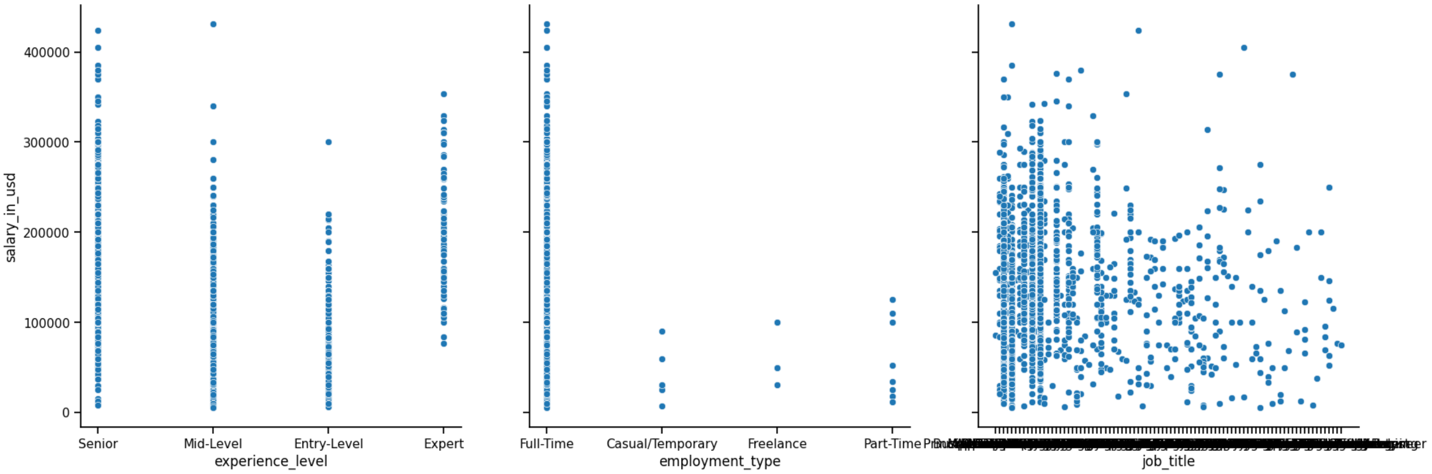


Fig 1. Salary in USD vs ExpLevel, EmploymentType & JobTitle.

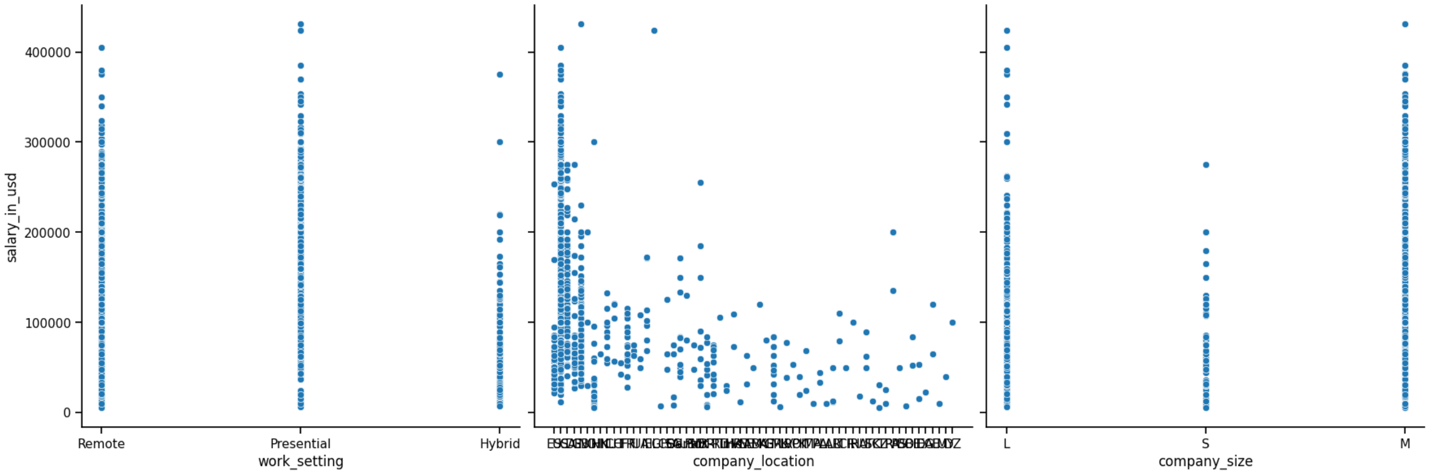


Fig 2. Salary in USD vs WorkSetting, CompLocation, CompSize.

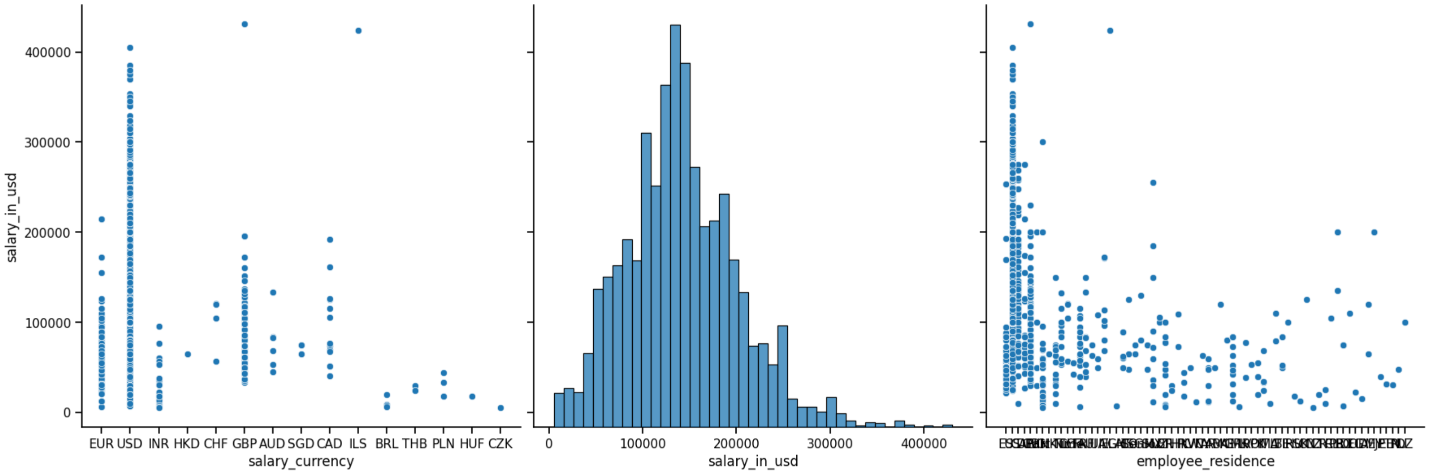
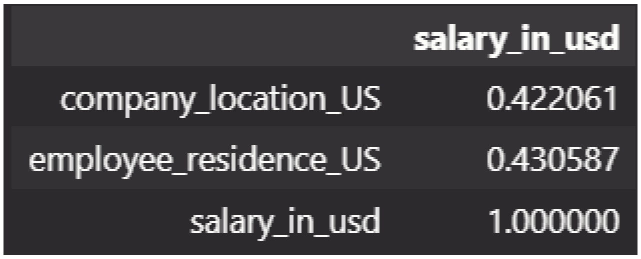


Fig 3. Salary in USD vs Currency, EmployeeResidence & Salary in USD distribution.

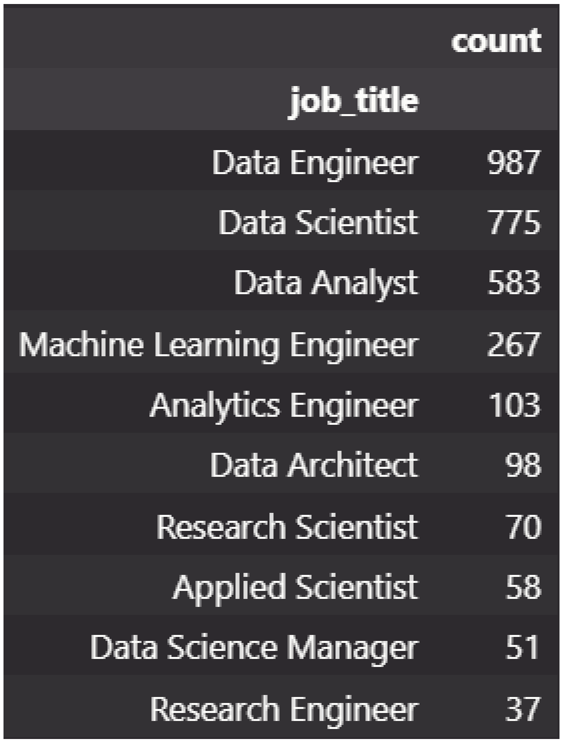
Table 1. Average USD Salary through 2022 & 2023, by Work Setting.

	Hybrid	Presential	Remote
2022	84.56087k USD	134.7192k USD	135.6311k USD
2023	72.0535k USD	152.3935k USD	146.3801k USD



	salary_in_usd
company_location_US	0.422061
employee_residence_US	0.430587
salary_in_usd	1.000000

Fig 4. Highest corr factors with salary.



	count
job_title	
Data Engineer	987
Data Scientist	775
Data Analyst	583
Machine Learning Engineer	267
Analytics Engineer	103
Data Architect	98
Research Scientist	70
Applied Scientist	58
Data Science Manager	51
Research Engineer	37

Fig 5. Most common job titles.

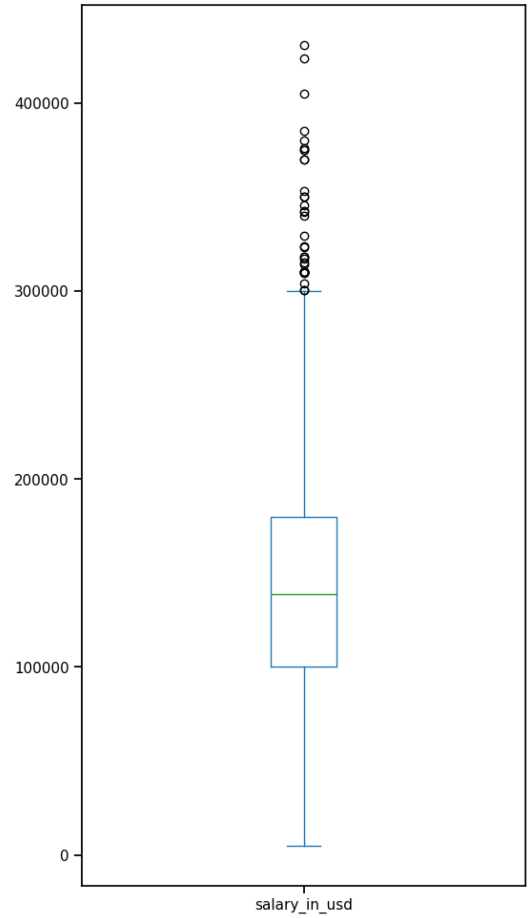


Fig 6. Salary in USD boxplot.

19 outliers were found following the z-score analysis. All of which had a z-score higher than 3 which in turn represented a salary higher than 324,000 usd.

3 Hypothesis about this Data:

Impact of Experience on Salary: Null Hypothesis (H0): There is no significant difference in salary across different experience levels. Alternative Hypothesis (H1): There is a significant difference in salary in at least one of the experience levels..

Remote Work and Salary: H0: There is no correlation between remote_ratio and salary_in_usd. H1: There is a significant correlation between remote_ratio and salary_in_usd.

Company Size and Salary: H0: The median salary does not differ significantly among different company sizes. H1: The median salary differs significantly among different company sizes.

Employee Residence vs Company Location: H0: There is no significant difference in salary between employees working in their country of residence and those working for companies in different countries. H1: There is a significant difference in salary between employees working in their country of residence and those working for companies in different countries.

Employment Type and Salary: H0: The type of employment (employment_type) has no significant effect on salary_in_usd. H1: The type of employment (employment_type) has a significant effect on salary_in_usd.

Significance test for one of the Hypotheses:

Impact of Experience on Salary: H0: There is no significant difference in salary across different experience levels (Mid-Level, Senior, Expert). H1: There is a significant difference in salary in at least one of the experience levels.

H0: salary average is the same across the 3 different experience levels

H1: salary average is significantly different in at least one of the experience levels.

Decision Criteria: significance level 5% ~ 0.05.

Our data is 2 tailed, we'll divide our Alpha = 0.05 by 2. Alpha = 0.025.

If our p-value is les than 0.05 we'll reject the null hypothesis and accept that ther is a significant difference in salary in at least one experience level.

We'll use ANOVA and f-score statistic.

The DataFrame is filtered by excluding the Entry-Level experience_level records.

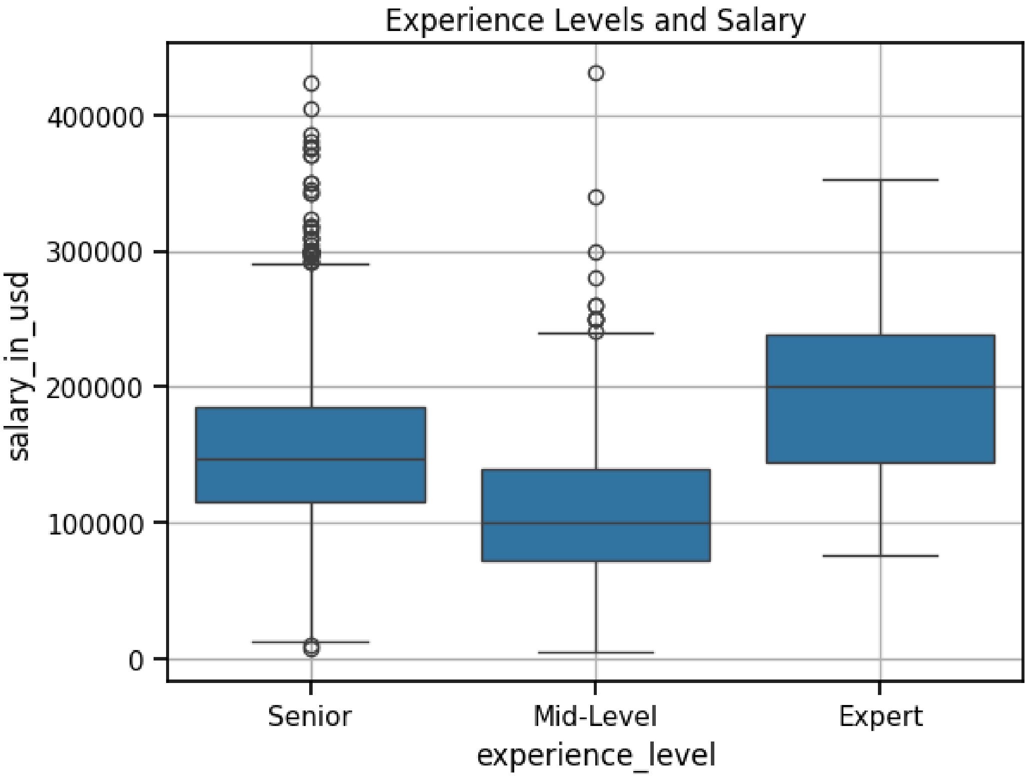
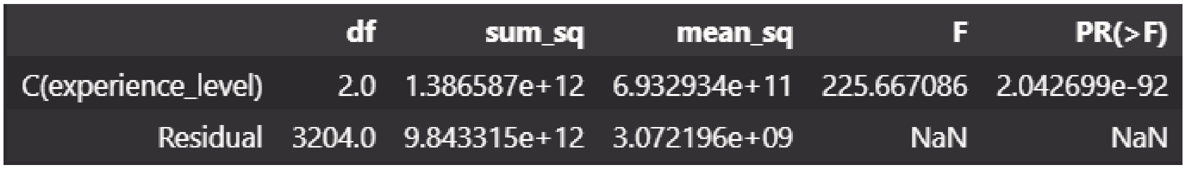


Fig 7. Experience Levels and Salaries boxplot.



	df	sum_sq	mean_sq	F	PR(>F)
C(experience_level)	2.0	1.386587e+12	6.932934e+11	225.667086	2.042699e-92
Residual	3204.0	9.843315e+12	3.072196e+09	NaN	NaN

Fig 8. Analysis of Variance result table p-value (PR(>F)) 2.04e-92.

The P-Value is way smaller than the Alpha of 0.05 therefore we reject the null hypothesis and conclude that there is a significant difference in salary in at least one of the experience levels.

Next steps in analyzing this data:

One way that we could further analyze the data is by making even more comparissons within the features and for a greater understanding of the employee distribution worldwide it could be valuable to créate a choropleth map.

The data quality was really prime, I chose this dataset because I'd found it interesting, but I had no idea that the Wrangling and Cleaning was already done, I had a blast diving into the hidden meaning of the data and I love how much learned from this course.