Evaluation of Different Means on AI/ML Job Salaries

Enes Duran, 6036777
enes.duran@student.uni-tuebingen.de

Jakob Laing, 3868294
jakob.laing@student.uni-tuebingen.de

Abstract

In this work, we observe the AI/ML industry positions. We form our dataset by augmenting the AI Jobs Net Salaries dataset by using the cost of living indexes provided by the Cost of Living dataset. We perform statistical tools and concepts covered in the lectures such as correlation analysis, likelihood fitting, and hypothesis testing to observe the relationship between salary and employee traits. We see that company location and experience level have a strong relation whereas the impact of job title appears to be negligible. Our code can be seen on here.

1 Data Overview and Preprocessing

1.1 Cost of Living Dataset

The Cost of Living Dataset (4) provides different indexes associated with cost of living per country as of 2022. We used the parameter **Cost of Living Plus Rent Index** which is a combination of the *Cost Of Living Index*, a indicator of consumer good prices, including groceries, restaurants, transportation and utilities, and the *Rent Index*, an estimation of the price of renting apartments.

All indexes are relative to New York City, i.e. a *Cost of Living Index* of 80 indicates that the estimated cost of living in that country is 80% of that in New York City. The indexes for the seven countries with the most entries in the *AI Jobs Net Salaries Dataset* are shown in the bottom right plot in Fig. 1. We added a column containing the country codes using pycountry to match the data with the salary dataset.

1.1.1 The Source

Numbeo is a private crowd-sourced global database collecting self reported data to generate indexes about quality of life and cost of living (5). Numbeo has no external control instance and all values are self reported, but the indexes have been shown to be comparable to other databases and have been used by major organizations (5).

1.1.2 Method

While the dataset is freely available on their website, no file is provided. For our analysis we collect the data in phyton from the html file of the website.

1.2 AI Jobs Net Salaries Dataset

The AI Jobs Net Salaries Dataset(1) is a dataset containing currently 278 entries of self reported information about the employment type and salary from people working in AI worldwide.

In our analysis we used the parameters *Job Title, Experience Level, Employment Type, Remote Ratio, Company Size, Employee Residence, Company Location and Salary in USD* and created the new variables *Lead* and *Buying Power*.

Experience Level	Employment Type	Remote Ratio	Company Size	Lead*
• EN: Entry-level	• FT: Full Time	• 0 %	• S: Small	• 0
• MI: Mid-level	• PT: Part Time	• 50 %	• M: Medium	• 1
• SE: Senior-level	• CT: Contractor	• 100 %	• L: Large	
• EX: Executive	• FL: Freelancer			

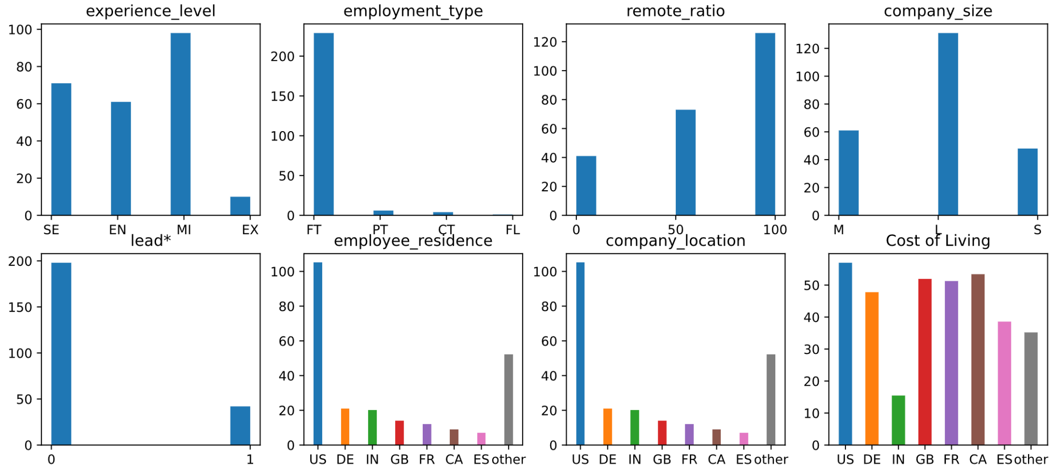


Figure 1: 1-7: Categorical variables of the *AI Jobs Net Salaries Dataset*. Variables marked with * were modified or created out of the other variables by us. 8: Cost of living in percent relative to New York City for the seven countries most represented in the *AI Jobs Net Salaries Dataset*.

Small Categorical Variables

Salary and Buying Power

Using the cost of living index described above we computed the *Adjusted Buying Power* for each entry. An *Adjusted Buying Power* of e.g. 200.000 USD would mean that the buying power of the given salary is equivalent to 200.000 USD in New York City. Salary and buying power are displayed in figure 2.

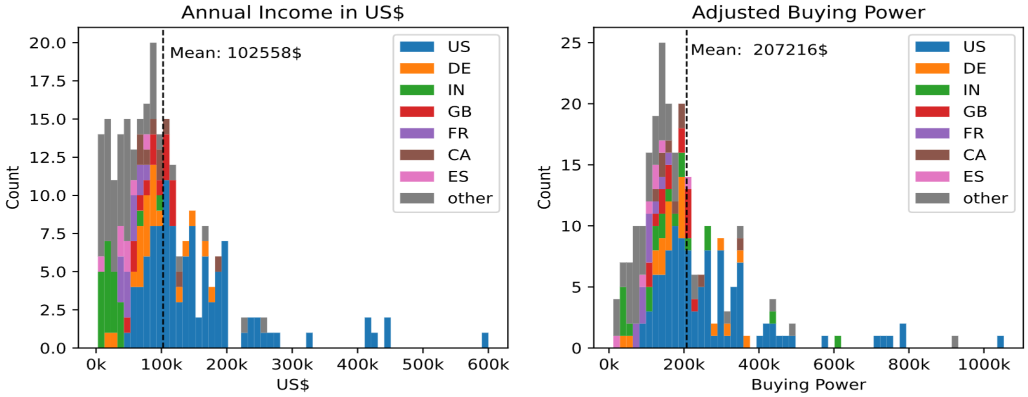


Figure 2: Salary and Adjusted Buying Power per Country

1.2.1 Preprocessing

In preprocessing we added the *Lead* variable out of the provided *Job Title* variable. *Lead* is 1 if *Job Title* contains one of the following keywords: *Manager, Director, Head, Head, Principal* and 0 otherwise. We also excluded cases where the employee residence differs from company location to have meaningful results for the analysis of the adjusted buying power. Besides that, we parsed the job titles into different Job title categories, to make them comparable. The title names are 'Data

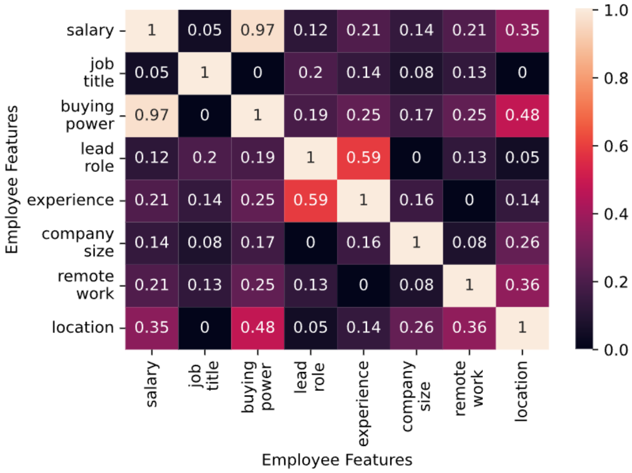
Scientist', 'Data Engineer', 'Data Analyst', 'Data Architect', 'Data Science Consultant', 'Machine Learning Engineer', 'Machine Learning Scientist'.

1.2.2 The Source

The data gets collected on https://ai-jobs.net/ a job portal for jobs in AI. Since all data was voluntarily admitted there is no ethical concern about the data.

2 Analysis of Data: Correlation Analysis

The variable types we want to find the correlation between are not at in the same space, meaning that some variables are categorical, e.g. job title, company size, experience level, some of them continuous, like salary in USD and buying power. To cope up with this complication we searched the statistics literature and found a technique named *Cramer's V* (2). This method allows us to calculate correlation value between two categorical values or categorical and continuous variables. Table 3 shows the results.



	salary	job title	buying power	lead role	experience	company size	remote work	location
salary	1	0.05	0.97	0.12	0.21	0.14	0.21	0.35
job title	0.05	1	0	0.2	0.14	0.08	0.13	0
buying power	0.97	0	1	0.19	0.25	0.17	0.25	0.48
lead role	0.12	0.2	0.19	1	0.59	0	0.13	0.05
experience	0.21	0.14	0.25	0.59	1	0.16	0	0.14
company size	0.14	0.08	0.17	0	0.16	1	0.08	0.26
remote work	0.21	0.13	0.25	0.13	0	0.08	1	0.36
location	0.35	0	0.48	0.05	0.14	0.26	0.36	1

Figure 3: Correlation Values obtained via Cramer's V method.

As seen from the table 3 the salary in dollars is strongly correlated with the buying power. Thus, there is no need to observe the buying power related hypothesis. Despite the company size does not correlate with salary, it correlates with the company location. This may be the reflecting the accumulation big AI/ML countries in certain countries, e.g. US. Apart from that, the high correlation between experience level and lead roles implies that more experienced people are generally employed in leading positions.

Surprisingly, the experience level statistically plays more significant relation with salary than lead role. This may be mean that companies value educated and experienced figures more. Another counter-intuitive implication is the correlation between the job title and salary. This may be the result of the dominating effect of company location on salary. For example a research scientist in India may earn the same as a Data Engineer in the US. We also see that the work location considerably correlates with the salary and buying power. This may be the reason for educated individuals to emigrate.

3 Hypothesis Testing: Company location does not affect the salary

We intend to observe the effects of employee features on salary. Hence, our analysis is mainly oriented towards salary values. We assume that the salary values in the dataset is derived from a continuous distribution. For that reason, we cannot apply the hypothesis testing techniques in the lectures which was the Fischer's exact test for the binomial distributions. We find that we need to know the likelihood of the data distribution. Beta distribution is the best explaining distribution

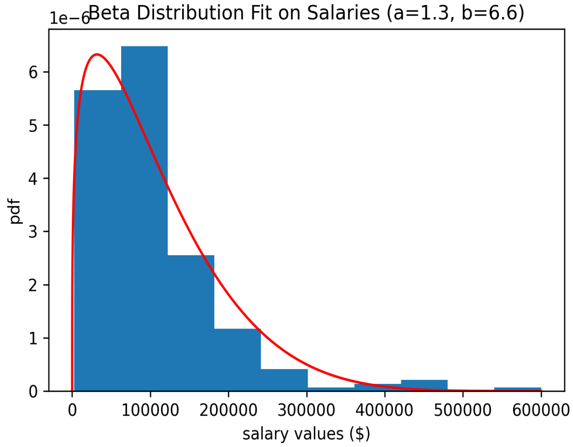


Figure 4: Salary Distibution

among all other tested distributions like Uniform, Gaussian, Poisson, Gamma 4. In addition to that we have multiple classes in our categorical feature, meaning that company location can take several values. Considering all these, Kruskal test which makes no assumptions on the likelihood function provides a good framework for our testing (3).

$$p(salary|H_0) = p(salary|H_0, location{=}x) \quad \forall x \in \{US, UK, CA, IN, DE, ...\}.$$

We hypothesise the company location does not affect the salary. This hypothesise from our personal curiosity. As people studying this field we want to know if there exists a place where working is better in. The Kruskal test takes flexible number of distributions and compare their originating distribution.

$$p(salary|H_0) = Kruskal(D(i) \quad \forall i \in \{US, UK, IN, DE, ...\}) = 6.3 * 10^{-18}.$$

The $D(i)$ stands for the salary values of class i. Based on the p-value, we reject the null hypothesis.

4 Conclusion

In this work we analyse the AI/ML jobs dataset (1) and perform data preprocessing, correlation analysis, hypothesis testing. Although our main motivation was to perform the concepts covered in the lectures we went beyond the concepts covered due to the required framework. From our analyses we see that there is a considerable relation between working place and experience level of the employee on the salary. Also we do not find any strong correlation of salary with job title. It is also worth noticing that our assumptions in the preprocessing and data augmentation parts may induce error or bias to our statistical analysis.

References

[1] foorilla. Ai jobs net salaries github, 2022. [Online; accessed 5-February-2022]. URL: https://github.com/foorilla/ai-jobs-net-salaries/blob/main/salaries.csv.

[2] Michael Kearney. *Cramér's V*. 12 2017. doi:10.4135/9781483381411.n107.

[3] Thomas MacFarland and Jan Yates. *Kruskal–Wallis H-Test for Oneway Analysis of Variance (ANOVA) by Ranks*, pages 177–211. 07 2016. doi:10.1007/978-3-319-30634-6_6.

[4] Numbeo. Nuembo - cost of living index by country, 2022. [Online; accessed 5-February-2022]. URL: https://www.numbeo.com/cost-of-living/rankings_by_country.jsp.

[5] Wikipedia contributors. Numbeo — Wikipedia, the free encyclopedia, 2021. [Online; accessed 5-February-2022]. URL: https://en.wikipedia.org/w/index.php?title=Numbeo&oldid=1047384398.