# Analysis of Self-reported Remote Work Salaries

**Stacey N**
University of Tübingen

## Abstract

This project will analyse a dataset of anonymously collected yearly salaries of employees in the IT sector to estimate a standard range of expected salaries per company location, experience level and company size. Using the cleaned data a regression model is trained to identify the most important indicators for making an accurate salary prediction.

## 1 Datasets

This report analyses two databanks of self-reported salaries of people working remotely in different parts of the world. Both databanks are collected in the same format and thus contain the same columns[1]. The first databank contains 1546 salaries from many different IT based professions whereas the second one contains 275 salaries of people in professions related to artificial intelligence and machine learning specifically and is hence a subset of the first databank. Both datasets were combined and the duplicates removed. An extra column was added to indicate job descriptions from the AI/ML dataset as being AI/ML-related jobs.

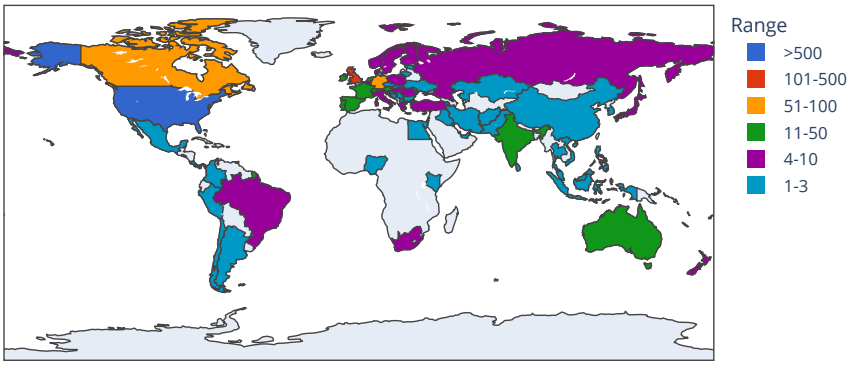


Figure 1: Chloropleth showing the company locations represented in the data collected and how many salaries were reported per country. The USA with 830 reported salaries is the dominating company location followed by Great Britain with 113.

Figure 1 shows the company locations the entries in the dataset that were collected from and includes all reported entries. For the purpose of this analysis, the combined dataset without duplicates and with the filter that only salaries reported as full-time work (96% of the total data) available are used further.

[1]For column descriptions, please see `https://salaries.ai-jobs.net/download/`.

## 2 Data analysis

### 2.1 Exploratory data analysis

The histogram in figure 2 helps to gather better understanding of the various subsets in the dataset. Most of the salaries in the dataset are from the work year "2021e". The label "2021e" means that the salary information was collected during 2021 and is an estimate of how much the employee expects to earn at the end of 2021. The values without an "e" signify a concrete salary submission after the year has ended. Based on the histogram it is clear that most of the data in the dataset is salary information pertaining to 2020 and 2021. The data is sorted in reported experience levels. "EN" is short for an entry-level and "MI" for mid-level position, whereas "SE" refers to senior and "EX" to executive positions. The histograms shows an imbalance of data per experience level with the bulk of the entries being from the categories "MI" and "SE" and only few salaries reported by executive level employees. Since the dataset is sparse when splitted into countries and experience level the following

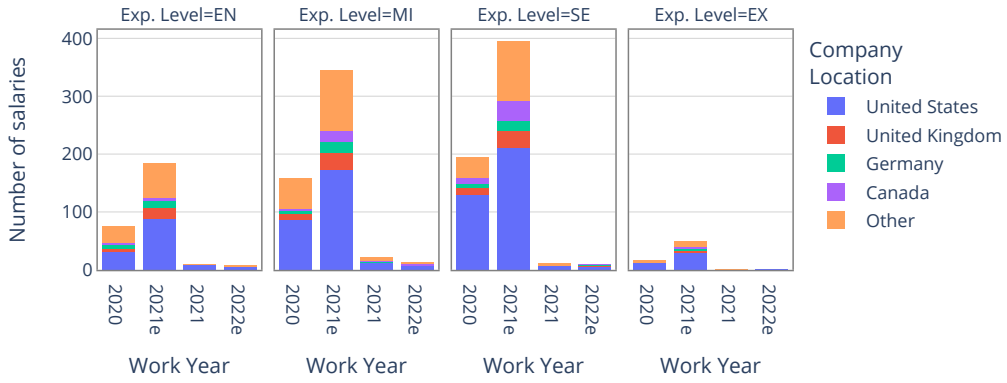


Figure 2: Histograms showing data available per experience level, company location and work year. The company locations have here been grouped into the four countries with the largest number of salary entries in the dataset and "Other" containing all other company locations as visible in figure 1.

chloropleth map in figure 3 was made with the median salary in USD. The median salary was chosen in figure 3 because it is more robust to outliers than the mean and the currency in USD was chosen instead of the local currency in order to make sensible comparisons between the countries.

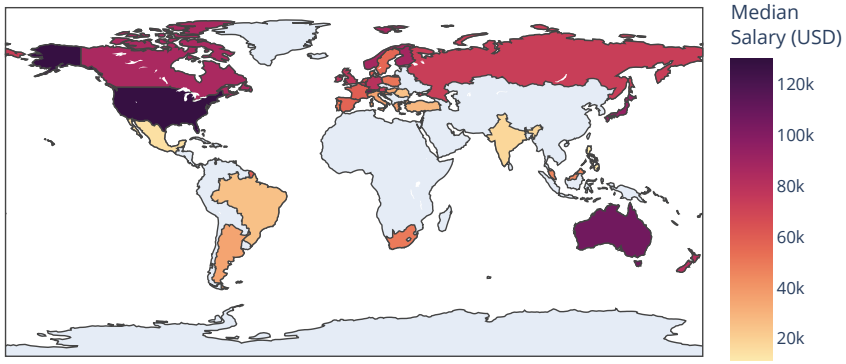


Figure 3: Median reported salary in USD for all employees in 2021 by country of company location. Here, median values are only displayed for countries which had more than three entries, see figure 1.

### 2.2 Determining confidence intervals

To better understand the trends in the data, three countries with the highest number of entries in the dataset plus all other countries grouped into one category "Other" were examined as shown in figure 4.

The upper and lower limits in figure 4 were calculated using the interquartile range ($IQR$) defined as the difference between the 25th and 75th percentiles of the data, denoted as the lower quartile ($Q_1$) and the upper quartile ($Q_2$) respectively. The upper and lower limits were thus defined as

$$\text{Upper limit} = Q_2 + 1.5 \cdot IQR \quad , \quad \text{Lower limit} = Q_1 - 1.5 \cdot IQR \tag{1}$$

and marked in figure 4 with red triangles based on the fences of a box plot [1].

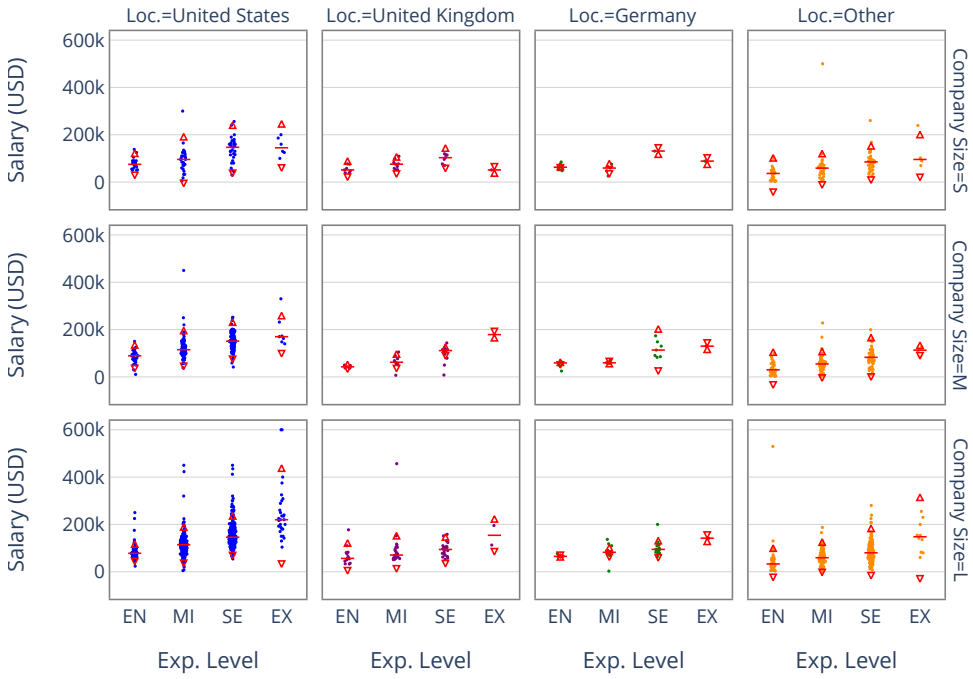


Figure 4: The figure shows salaries reported in USD for different experience levels for four different company location groups and three different company size groups. The raw data is presented as individual datapoints. The red horizontal line markers show the median of each individual distribution and the red triangles indicate the upper and lower limits.

### 2.3 Regression analysis

The values outside the upper and lower limits defined in section 2.2 were removed from the dataset for this section of the analysis, in order to gather meaningful conclusions about more general trends in the data. Some factors already identified as influencing the magnitude of the reported salary are the company location and employee's experience level. This section attempts to use this data to optimally fit the reported salary. For this purpose a generalized linear model using the Poisson distribution was used, as the reported salaries are independent of each other and most importantly must be predicted strictly positive. The Poisson regression models the logarithm of the target variable as a linear combination of the input variables and calculates the coefficients using maximum likelihood estimation under the assumption of a poisson distribution. The predicted mean of the Poisson distribution is thus [2]:

$$\log \mathbb{E}(\mathbf{y} \mid \mathbf{x}') = \beta_0 + \beta \mathbf{x}' \quad \Leftrightarrow \quad \mathbb{E}(\mathbf{y} \mid \mathbf{x}) = e^{\theta \mathbf{x}} \tag{2}$$

Where $\mathbf{x}' \in \mathbb{R}^{\mathbf{n}}$ is a vector of $n$ independent variables and $\beta_0 \in \mathbb{R}$ and $\beta_n \in \mathbb{R}^n$ are the y-intercept and coefficients respecitively. In this case $n$ is 21 as that many features were used as input variables and they can be seen on the right with their calculated coefficients in figure 5. $\theta$ is $\beta$ with $\beta_0$ concatenated to it and $\mathbf{x}$ has a 1 concatenated to it. The $\theta$ can be estimated by maximum likelihood

estimation. The likelihood to be maximized is the Poisson distribution's probability mass function given by

$$p(y \mid x; \theta) = \frac{\lambda^y}{y!} e^{-\lambda} = \frac{e^{y\theta x} e^{-e^{\theta x}}}{y!} \tag{3}$$

for a single $(x, y)$. For $x_i, y_i \in \mathbb{R}^{n+1}$, $i = 1, \ldots, m$, the probability and thus the likelihood function to be maximized is defined as:

$$L(\theta \mid X, Y) := p(y_1, \ldots, y_m \mid x_1, \ldots, x_m; \theta) = \prod_{i=1}^{m} \frac{e^{y_i \theta x_i} e^{-e^{\theta x_i}}}{y_i!} \tag{4}$$

The negative logarithm of equation 4 is a convex function and hence can be maximized using standard convex optimization techniques like gradient descent. In this analysis the Poisson regression was

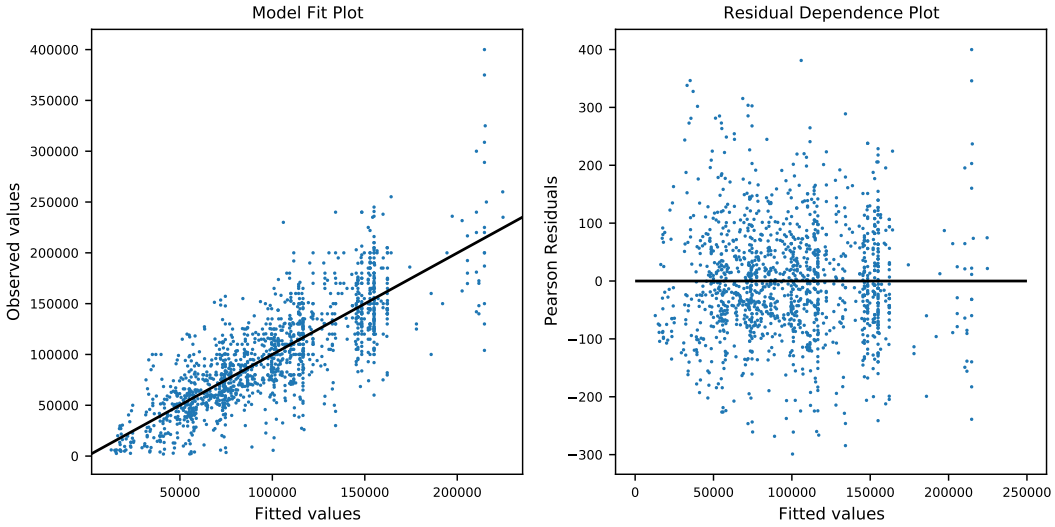


Figure 5: Left: Performance of Poisson regression on the cleaned data (training and test set) and outliers removed from the dataset according to limits set in section 2.2. The $y = x$ line is drawn for comparison purposes. Right: features that were used in the regression and their coefficients after fitting the regression on the training set. Note that all features contain binary values (0 for False or 1 for True) except Remote Ratio which contains the values (0,0.5,1).

used without any regularization and the coefficients were fitted on 80% of the cleaned dataset and then tested on the remaining 20% as shown in figure 5. The figure shows the outliers defined in the previous section in red and the performance of the fitted regression model on them. As expected the large majority of them are out of the scope of the model as indicated by the negative $R^2$ score.

## 3 Conclusion

The analysis presented introduces some methods to remove outliers from anonymously reported salary data using robust measures of scale like the interquartile range. Fitting a generalized linear model without these outliers and evaluating the coefficients provides a way to find the most influential factors on the salary prediction. In this dataset, the company location and the experience level were found to be the most important factors on salary prediction. Finally, another visual way to remove outliers was presented which uses the GLM model fitted on the cleaned dataset on new data entries and plots the discrepancy between the reported salary and the predicted salary. It should be noted that for executive-level and salaries above 200k USD the model is not usually capable of providing good estimates, due to the sparsity of data in this range in the training set.

## References

[1] F. Dekking, C. Kraaikamp, H. Lopuhaä, and L. Meester, "A Modern Introduction to Probability and Statistics: Understanding Why and How". Springer Texts in Statistics. Springer, 2005.

[2] P. McCullagh and J. Nelder, “Generalized Linear Models, Second Edition”. Chapman and Hall/CRC Monographs on Statistics and Applied Probability Series. Chapman & Hall, 1989.