# Salaries in the Fields of Data Science, Artificial Intelligence, and Machine Learning

Abhishek Anand, Bolun Lu, Siming Xu, Rijin Lnu

2022-10-17

**Abstract**

[[Add summary of our report when analyses is complete.]]

**Index Terms / Keywords**

[[Complete list of keywords/acronyms that we reference.]]

AI: Artificial Intelligence DS: Data Science ML: Machine Learning USD: United States Dollar

**Introduction**

A shift towards data-centered decision-making is taking hold and at the center of this shift are those working in the fields of data science ("DS"), artificial intelligence ("AI"), and machine learning ("ML"). As businesses seek to optimize operations and information using vastly increased computing power, the shape and size of data analytics has increased exponentially over the last two decades, and even more prominently in the last decade with the rise of machine learning, cloud computing, and advances in the field of artificial intelligence.

Exploring the composition and trends of salaries in this sector is equally as intriguing, both from a research perspective and from the perspective of what can be waiting for students of Data Analytics (DS/AI/ML) in academic programs globally.

The data we've chosen for this project comes from https://salaries.ai-jobs.net/, a website collecting salary data from around the world.

This data set has salary (in USD) for a variety of technical roles, and a variety of information on the characteristics of the employee, employer, the job itself, and the market.

The data set contains the following variables:

**Citations and Related Work (> Siming)**

[[Add citations we consulted, and a summary of related work!]]

**Data**

**Data Cleaning (> Abhishek)** Follow all cleaning steps outlined in class _clean.csv

```
##   work_year experience_level employment_type                 job_title salary
## 1      2022               SE              FT              Data Analyst 144000
## 2      2022               SE              FT              Data Analyst 113000
## 3      2022               EN              FT              AI Scientist  30000
## 4      2022               SE              FT            Data Architect 195400
## 5      2022               SE              FT            Data Architect 131300
## 6      2022               SE              FT Machine Learning Engineer 195400
```

| | |
|---|---|
| **work_year** | The year the salary was paid. |
| **experience_level** | The experience level in the job during the year with the following possible values:<br>**EN** Entry-level / Junior<br>**MI** Mid-level / Intermediate<br>**SE** Senior-level / Expert<br>**EX** Executive-level / Director |
| **employment_type** | The type of employement for the role:<br>**PT** Part-time<br>**FT** Full-time<br>**CT** Contract<br>**FL** Freelance |
| **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. |
| **salary_in_usd** | The salary in USD (FX rate divided by avg. USD rate for the respective year via **fxdata.foorilla.com**). |
| **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, possible values are as follows:<br>**0** No remote work (less than 20%)<br>**50** Partially remote<br>**100** Fully remote (more than 80%) |
| **company_location** | The country of the employer's main office or contracting branch as an ISO 3166 country code. |
| **company_size** | The average number of people that worked for the company during the year:<br>**S** less than 50 employees (small)<br>**M** 50 to 250 employees (medium)<br>**L** more than 250 employees (large) |

Figure 1: A caption

```
##   salary_currency salary_in_usd employee_residence remote_ratio
## 1             USD        144000                 US          100
## 2             USD        113000                 US          100
## 3             EUR         31981                 PT          100
## 4             USD        195400                 US          100
## 5             USD        131300                 US          100
## 6             USD        195400                 US          100
##   company_location company_size
## 1               US            M
## 2               US            M
## 3               ES            M
## 4               US            L
## 5               US            L
## 6               US            L
```

**Recode**

```
Salaries <- Salaries[,c(1:4,7,9:11)]
```

**Remove Redundant Data**

```
#Salaries$remote_ratio <- as.factor(Salaries$remote_ratio)
#Salaries$experience_level <- as.factor(Salaries$experience_level)
#Salaries$employment_type <- as.factor(Salaries$employment_type)
#Salaries$company_size <- as.factor(Salaries$company_size)
#Salaries$remote_ratio <- as.factor(Salaries$remote_ratio)
```

**Convert to Factor (may not be needed, but leaving it here for now)**

```
summary(Salaries[c(1:3,5:6,8)])
```

**Missing Data**

```
##    work_year    experience_level employment_type salary_in_usd
##  Min.   :2020   Min.   :1.000    Min.   :1.000   Min.   :  2324
##  1st Qu.:2021   1st Qu.:2.000    1st Qu.:2.000   1st Qu.: 69440
##  Median :2022   Median :3.000    Median :2.000   Median :113465
##  Mean   :2022   Mean   :2.487    Mean   :2.005   Mean   :119224
##  3rd Qu.:2022   3rd Qu.:3.000    3rd Qu.:2.000   3rd Qu.:160000
##  Max.   :2022   Max.   :4.000    Max.   :4.000   Max.   :600000
##   remote_ratio    company_size
##  Min.   :1.000   Min.   :1.000
##  1st Qu.:1.000   1st Qu.:2.000
##  Median :3.000   Median :2.000
##  Mean   :2.356   Mean   :2.178
##  3rd Qu.:3.000   3rd Qu.:3.000
##  Max.   :3.000   Max.   :3.000
```

```
sum(is.na(Salaries))
```

```
## [1] 0
```

There are no missing data in our data set. Also, as we see from the summaries above, all the categorical factors and the continuous variables are correctly assigned.

```
round(apply(Salaries[,-c(4,7)],2,mean),1)
```

**Means**

```
##        work_year experience_level  employment_type    salary_in_usd
##           2021.6              2.5              2.0         119224.4
##     remote_ratio     company_size
##              2.4              2.2
```

```
round(apply(Salaries[,-c(4,7)],2,sd),1)
```

**Standard Deviations**

```
##        work_year experience_level  employment_type    salary_in_usd
##              0.6              0.8              0.2          68259.2
##     remote_ratio     company_size
##              0.9              0.6
```

```
mahal <- mahalanobis(Salaries[,-c(4,7)],
                     colMeans(Salaries[,-c(4,7)]),
                     cov(Salaries[,-c(4,7)],use = "pairwise.complete.obs"))

cutoff <- qchisq(1-0.001,ncol(Salaries[,-c(4,7)]))
```

```
ncol(Salaries[,-c(4,7)])
```

**Outliers**

```
## [1] 6
```

```
summary(mahal < cutoff)
```

```
##    Mode   FALSE    TRUE
## logical      31     947
```

From the above summary, we see that of the 978 rows of data, 31 contain outliers. The other 947 do not. We will remove the 31 rows to create a data set without outliers.

```
Salaries_Outliers <- subset(Salaries, mahal >= cutoff)
Salaries_NoOutliers <- subset(Salaries, mahal < cutoff)
write.csv(Salaries_NoOutliers, "salaries_clean.csv")
Salaries_clean <- read.csv("salaries_clean.csv")
```

**Summary of Data (> Bolun)**

- median, mean, sd, range, IQR
- anything else discussed in class
- maybe include boxplots

```
## AA note, please use "Salaries" rather than "Salaries_clean" for summary of data
```

[[Add language describing data.]]

**Descriptive Statistics (> Bolun)** [[Add language summarizing descriptive statistics.]]

**Descriptive Plots**

**Scatterplots (> Bolun)** [[Add language summarizing the relationships shown in descriptive plots.]]

**Barplots / Histograms (> Abhishek)** [[Add language summarizing the relationships shown in descriptive plots.]]

**Multivariate plots (> Abhishek)** [[Add language summarizing the relationships shown in descriptive plots.]]

[[Overarching summary of plots.]]

**Assumptions Tests**

**Additivity (Rijin)**

**Linearity (Rijin)**

**Homogeneity (Rijin)**

**Normality (Rijin)** [[Summary of the tests.]]

**Correlation Check / Comments** [[Describe any correlations that were observed. How did we deal with this?]]

**Technical Approach (> Siming)**

[[Language on our overall approach to the analysis.]]

**Modeling (All)**

**Linear models**

**t-tests**

**ANOVA tests**

**Findings (Siming + All)**

**Conclusion (Siming + All)**

**Future Studies (All)**