Section C - Team 37 - M/Th 10:30 am - 12:45 pm

Alex Gille, Ana Paula Gonzalez, Arushi Dheer, Monil Soni, Xuchen Tan

Data Science Job Salaries

Business Understanding

Using this data, we want to better understand, as managers, what it may cost to hire data professionals of specific titles and skill levels in various locations. This has potential cost-saving benefits as we will explore in the next section.

Data Understanding

- The data itself is from salaries.ai-jobs.net as of 9.26.2022
- The dataset has 11 Columns, with 949 total observations. Further information can be found in Appendix A.

Through exploring the data, we observed that we had a clean dataset that did not contain any null values. We then performed the following information analysis:

- We did notice that there were 120 duplicate observations, but we chose to keep those observations in the dataset, as we treated those as jobs with similar compositions as opposed to accidental double entries by an individual.
- For this project, we will focus on the “salary_in_usd” field for our salary-related questions, as this allows us to be consistent in our exploration.
- Because this data source relies on user-submitted data, we have the inherent risk of not having 100% accurate data.

Limitations

One limitation of the analysis of this model is the sample size. Low sample size may impact the standard errors and consequently p-values. Thus, we might end up concluding that a change is significant

when it is actually not, or vice versa. Another potential limitation is that there are many external factors that may have an impact on employee salary that are not included in the data set.

Question 1

As a company local to the US, how much can we change the salary of an employee if they live outside of the US? Furthermore, how does the salary of remote, non-US workers compare to those located in the US?

Data Preparation

Our treatment variable is a dummy variable called EmployeeResidency which will be a 1 if the employee lives in the US and 0 otherwise. To use company size as a controlling factor, we categorize the variable into two: large and small. We are assuming that both "M" and "L" entries are considered large. Therefore, J = 0 if the company size is small, and J = 1 if the company size is large.

To use experience level as a control variable, we again categorize our data into two: inexperienced and experienced. For 'SE' and 'EX', we assume that the employee is experienced. For 'EN' and 'MI', we assume the employee is inexperienced. Therefore, K = 0 for an inexperienced employee, and K = 1 for an experienced employee.

Modeling

Data	Information on employees that work for a US company (X1), data on remote employees (X2), Salary (Y)
Uncertainty	One thing we cannot measure is the number of people that do 100% remote work outside of their company location which is not accounted for. We also do not know the cost of living of the different places that the employees live in, and if that plays a role in the salaries from the data.
Decision/Action/Deployment	Salary to offer

Our model would then be looking to answer this question:

$$E[salary\_in\_usd \mid employee\_residency = US] - E[salary\_in\_USD \mid employee\_residency != US]$$

We account for control variables such as level of experience because the amount of people who work remotely changes with experience level as well as company size that could have an impact on salaries.

We will start with a linear model to see if there is a significant linear relationship between the EmployeeResidence variable and the salary in USD.

$$salary(USD) = \beta_0 + \beta_1 * EmployeeResidency + Control\,Variables$$

Evaluation

Modeling with Linear Regression:

Our first linear model was:

$salary(USD) = \beta_0 + \beta_1 * D$, where D = employee residency

The expected average treatment effect that living and working in the same location has on an employee's salary is $67,916 holding everything else constant. This shows a significant impact of the variable on salary. We now take into account the following control variables: experience level and the interaction between the experience level and the treatment effect. However, this new model did not result in statistically significant values.

$$salary(USD) = \beta_0 + \beta_1 * D + B_2 * J + \beta_3 * K$$

When trying out a model, controlling only for experience level, the resulting model was more statistically significant and therefore, we decided to go forward with that. If we control for experience level, the point estimate decreases to $60,506 at a 95% confidence level. This might be because of selection bias since employees that live in the US as well also have different characteristics in general. D is not randomly assigned, as it depends on characteristics, such as experience or job title.

Based on these models and the analysis previously done, we can see that there is a difference between the salary that an employee residing in the US gets and an employee that lives outside of the US but still works for an American company.

Deployment

Leaving out some factors that may be important that are out of the scope of this project, such as the productivity and attrition rates of not going into the office, and potentially managing their own time, you can expect to pay a different salary base rate depending on the residential locations. If the employee wants to work somewhere else, a company can take advantage of this to decrease its costs.

Question 2

Is it cheaper to hire a part-time employee for different job lengths and titles?

Data Preparation

At the beginning of EDA, we check the sample size for each group found in employment_type. From the bar graph in Appendix B, we can see that sample sizes are extremely unbalanced, with the majority being full-time.

Since we are exploring the idea of hiring part-time employees as opposed to full-time to lower the salary cost, we will consider employment_type as a treatment variable, hiring part-time employees as a treatment group, and hiring full-time employees as a control group. The sample sizes of our treatment group (PT) and our control group (FT) are extremely unbalanced, with the majority being full-time. Therefore, we can not be very confident about the treatment effect of hiring part-time, but one possible solution is to do random sampling across two groups to minimize the noise in the control group to increase our confidence, another approach is downsampling, and the last approach is propensity score matching.

We performed downsampling by excluding those full-time samples which have characteristics that won't be available to part-time samples. For example, some job titles such as Director of Data Science are not available for part-time employment. Downsampling eliminates job titles that are not available for part-time employment, restricting countries, and Experience Levels. After Downsampling, we decreased the original 926 samples of full-time to 160 samples. This increased the ratio of part-time samples over full-time samples from the original 1.3% to 7.5%. It is important to acknowledge, however, that this smaller sample size may impact the standard errors and consequently p-values of further analysis.

Modeling

The mathematical model for dealing with the causal of hiring part-time employees is below:

$$Y_{Salary} = \beta_0 + \beta_t X_{part-time\ or\ not} + \beta_c X_{control\ variables}$$

- Treatment Effect of part-time:

$$E[salary \mid employment\_type = PT] - E[salary \mid employment\_type = FT]$$

- Treatment Variable: part_time - Dummy variable of part-time or not
- Control Variables: entry_level - Dummy variable EN = 1, MI = 0
- List of 7 job title dummies: After the downsampling, we have the following titles with “Computer Vision Researcher” serving as the baseline: "Computer Vision Software Engineer", "Data Scientist", "Data Engineer", "Data Analyst", "AI Scientist", "3D Computer Vision Researcher", "ML Engineer", "Computer Vision Engineer"
- List of 8 company location dummies: We have these country codes after downsampling "DK", "FR", "DE", "DZ", "ES", "US", "IN", "NL", "IT" with “DE” serving as the baseline.

Evaluation

After the first result of linear regression, We found out that variables in job titles have extremely low significance in the regression. We decided to drop the list of job_title dummies, and we also dropped some high p-value company location dummies. We then ran a second linear regression, and our model improved, but the p-value of 0.11976 for the part-time variable is still high. This may be in part due to the unbalanced sample. We derive the below model:

$$Y_{salary} = -18181 * X_{part\ time\ or\ not} + \beta_{controls} X_{controls}$$

Deployment

Our model shows that part-time is not statistically significant to a change in salary. This could be due to the unbalanced sample sizes, even after we attempted to eliminate the noise surrounding it. We may, however, be able to lower our cost of salary by hiring an employee outside of the US. It may, in turn, be beneficial for global firms to open more branches around the world to seek low-cost talent markets.

Question 3

We want to understand that, for a given workload, how much does an increase in remote ratio change the salary expected by the employee? This will allow us to make sure we at least offer the baseline expectation of the employee if we want to retain them.

Data Preparation

1. Create dummy variables for remote ratios at 0%, 50%, and 100%.
2. Create an ordinal factor for experience levels and company size.
3. Create a factor for employment type.
4. Create a dummy variable that dictates whether the job is managerial based on the job title.

Modeling

Data	Information on jobs that were accepted (X), Salary (Y)
Uncertainty	Candidates accepting the offer
Decision/Action/Deployment	Salary to offer

Since we only have 0, 50, or 100 as remote ratios, we will model these as dummy variables. 0% will be the baseline as our question asks about switching jobs to remote.

$Salary = \beta_0 + \beta_1 D_{50} + \beta_2 D_{100} + \beta' Control$, where β' is a coefficient vector for controls.

Adding control variables will ensure that we're able to isolate the effect of remote from the workload. We take the following variables into account: (refer to the graphs in Appendix B)

1. Experience Level
2. Employment Type
3. Job Title (by splitting it into a binary for individual contributor vs manager)
4. Company Size,

In addition, we think the following interaction terms should help us capture more nuanced considerations:

1. Job Title x Experience Level (Leadership): A manager's workload is usually a compounding effect of the work of the people they are leading. As such it might also have a non-linear effect on Salary.
2. Experience Level x Company Size (Seniority at scale): A senior at a large-scale company will have more products to look after compared to a smaller company.
3. Job Title x Company Size (Leadership at scale): Similar to the previous point, a manager at a large-scale company will have more employees reporting to them.

Evaluation

We worked with a linear model to elicit the salary difference when the remote ratio is changed. We found that our model objective then becomes to fit the distribution of the target variable. For our case, since our treatment variable is ternary, the model essentially tries to fit the available distribution. We obtain a better fit when we try to introduce non-linearity that makes sense to the model but not economically.

To understand the next steps, we looked back to our goal - to compare salaries for similar workloads. This means that conditional on the workload, the difference we observe in our salaries would essentially be the treatment effect. Using this conditional independence assumption, we used Propensity Score Matching to assign propensity scores based on how likely an employee is to be part of a job with a particular remote ratio and then compare the effect of switching to a job with a different remote ratio.

Our results show that the change in expected salary is 35672 USD (p-value 0.0023) when offering a 100% remote ratio compared to a similar job with 50%. This result is shown to be statistically significant in our model.

We also observed that the change in expected salary is a decrease of $13,124 USD (p-value 0.14) when I offer a 50% remote ratio compared to a similar job onsite mode and that the change in expected salary is 5062.6 USD (p-value 0.3272) when offered a 100% remote ratio compared to a similar job with 0%. However, these are not statistically significant given the p-values meaning that 50% and 100% remote ratios may not have a material effect on expected salaries.

Deployment

In addition to the type of work, salaries must be an important tool in job design. For example, with the dawn of remote collaboration tools like Slack and Zoom, remote work has become possible and desired by employees. This question is meant to exhibit how companies can adjust the remote ratio to reduce salary expenses paid to an employee.

Companies can deploy this model with their internal and external hiring database to update the resulting parameters or change in expected salary so that the hiring team is aware of the flexibility they have when offering packages to a new hire or promoting someone internally.

Another opportunity to better this model would be to have a numeric measure of both job requirements and employee performance. With these two inputs, we can reduce our dependence on Propensity Score Matching or even possibly eliminate it.

We can also extend this question to a larger question of how we can maximize the value we can derive from an employee, modeling for which is performed in Appendix C.

Question 4

We want to observe how job titles affect the salary of roles with similar responsibilities, and if this information may be utilized when it comes to creating job listings and attracting talent. For example, might we be able to offer a lower salary in exchange for a more prestigious title, or would this more prestigious title lead to the expectation of a higher salary?

Data Preparation

For this, we'll be comparing similar roles, and we use the assumption that jobs with the title containing the phrase “Data Science / Data Scientist” at similar experience levels follow this assumption. To test this, we will take a look at when a job has a managerial title, controlling for experience level, how does the salary of that position compare to one with a baseline title of ‘Data Scientist”. We first looked at titles that contained ‘Data Scien’ to get data science roles, and from this, the titles that seemed to be more managerial were picked.

We use the experience level field as a proxy to determine whether jobs would have similar responsibilities based on their experience. For this, we decided to filter down further to senior-level jobs only, as the base data science title becomes less common in executive-level jobs, while the managerial titles are not found in entry-level jobs.

- Base Data Scientist Title = Data Scientist: 128 observations
- Managerial Titles = Data Science Manager, Director of Data Science, Principal Data Scientist, Head of Data Science, and Lead Data Scientist: 29 observations

Finally, we created a dummy variable based on whether the job had a managerial title and assigned the value 1 if it did and a value of 0 otherwise.

Modeling

Data	Data Science-related job title information, Salary in USD
Uncertainty	How much the experience_level field relates to in-role responsibilities, Candidates accepting the offer
Decision/Action/Deployment	Salary to offer

We designed a linear model to see how managerial titles affect salary in USD:

$Salary\ in\ USD\ = \beta_0 + \beta_1 * D_0$, where

- β_0 = The base expected salary of a Data Scientist
- β_1 = The salary change associated with having a managerial job title
- D_0 = Dummy variable representing whether a role has a managerial-title.

Evaluation

Using the subset of senior-level roles only, we observed that for similar responsibility, managerial-title commanded $ 8,018 (p-value 0.449) more than the baseline.

However, due to the low p-value, managerial job titles in our model were not statistically significant, and the variability in salaries seems to be largely unattributed to this type of title change.

In the case that there was a business case developed to evaluate improvements, we'd likely want employment data within the firm regarding the salaries of new hires for open roles. We'd hope to see trends of our senior roles being filled and performed up to company standards with employee costs lowering over time as we feel more comfortable with offers and acceptances surrounding specific titles and salaries.

Deployment

This model, as with other models considered, will need to be deployed with both internal and external databases if we're able to verify that our results are statistically significant with larger data. There would also be inherent risks in making major salary-offering decisions based on a change in the title if all else is kept constant, as an employee might come into an interview with specific salary expectations if the job title contains specific keywords, such as senior or director.

If we had access to more data on the subject, it would have been interesting to explore how specific job titles at similar experience levels change expected salary. It would have also been helpful to be able to look at the specific years of experience for these roles, as opposed to just general experience levels, as we may have had a better idea of what titles are truly analogous in this space.

Appendix

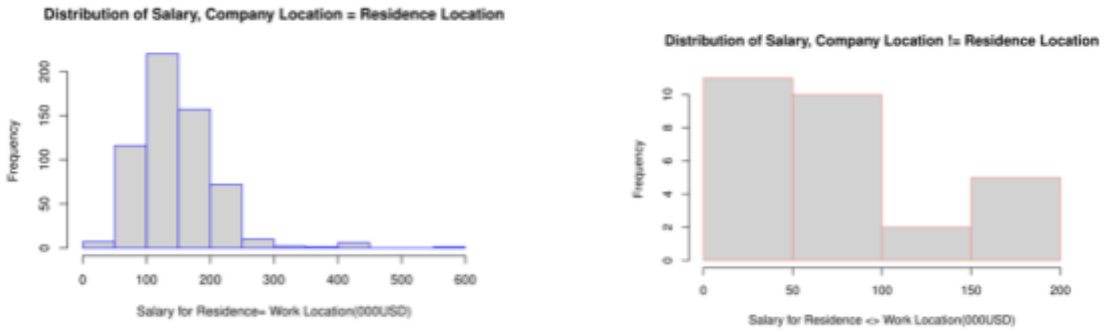
Appendix A: Fields in Dataset

The following fields are found in our dataset, along with the number of unique values in each field:

- **Work_year -** The year the salary was paid (3 unique values: 2020, 2021, 2022)
- **Experience_level -** The experience level in the job during the year (4 unique values):
    - **EN** - Entry-level / Junior
    - **MI -** Mid-level / Intermediate
    - **SE -** Senior-level / Expert
    - **EX -** Executive-level / Director
- **Employment_type -** The type of employment for the role (4 unique values):
    - **PT -** Part-time
    - **FT -** Full-time
    - **CT -** Contract
    - **FL -** Freelance
- **Job_title -** The role worked in during the year (57 unique values)
- **Salary -** The total gross salary amount paid.
- **Salary_currency -** The currency of the salary paid (17 unique values)
- **Salary_in_usd -** The salary in USD (Foreign exchange rate via fxdata.foorilla.com as of 9.26.2022).
- **Employee_residence -** Employee's primary country of residence during the work year (62 unique values)
- **Remote_ratio -** The overall amount of work done remotely (3 unique values):
    - **0 -** No remote work (less than 20%)
    - **50 -** Partially remote (between 20% and 80%)
    - **100 -** Fully remote (more than 80%)

- **Company_location -** The country of the employer's main office or contracting branch (52 unique values)
- **Company_size -** The average number of people that worked for the company during the year (3 unique values):
    - **S -** less than 50 employees (small)
    - **M -** 50 to 250 employees (medium)
    - **L -** more than 250 employees (large)

Appendix B: Supplementary Visualizations



Figures B.1: These graphs show the histograms of salaries for when the employee residence is same as the work location and when it is not.

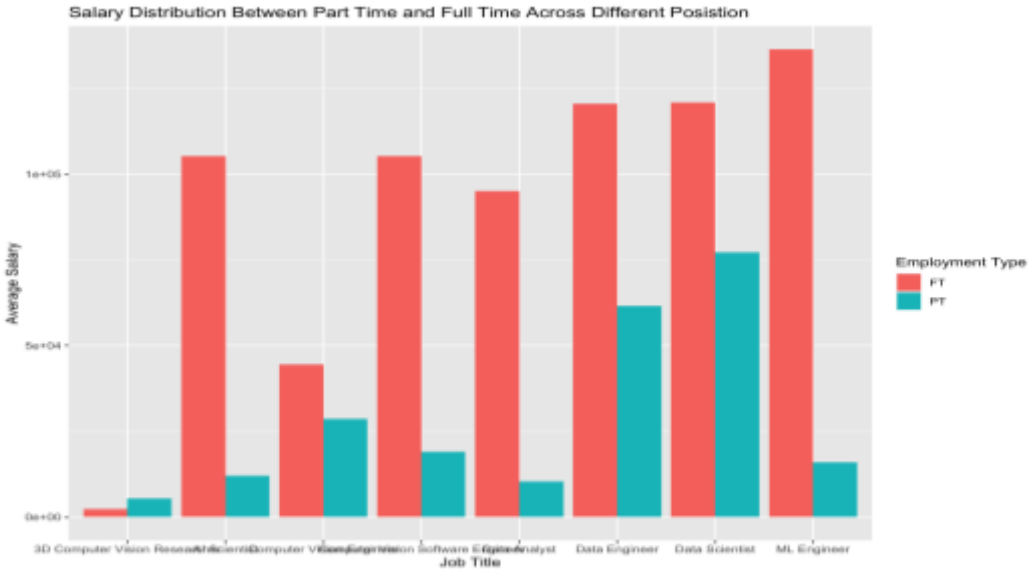


Figure B.2: Above graph shows the average salary comparing part-time and full-time employment across different positions

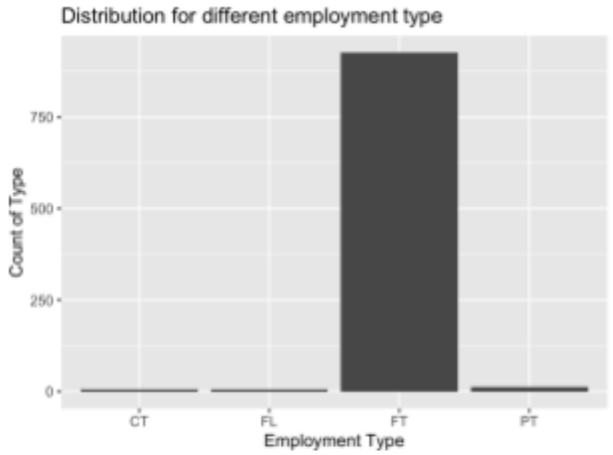


Figure B.3: above graph shows the unbalanced sample size before downsampling

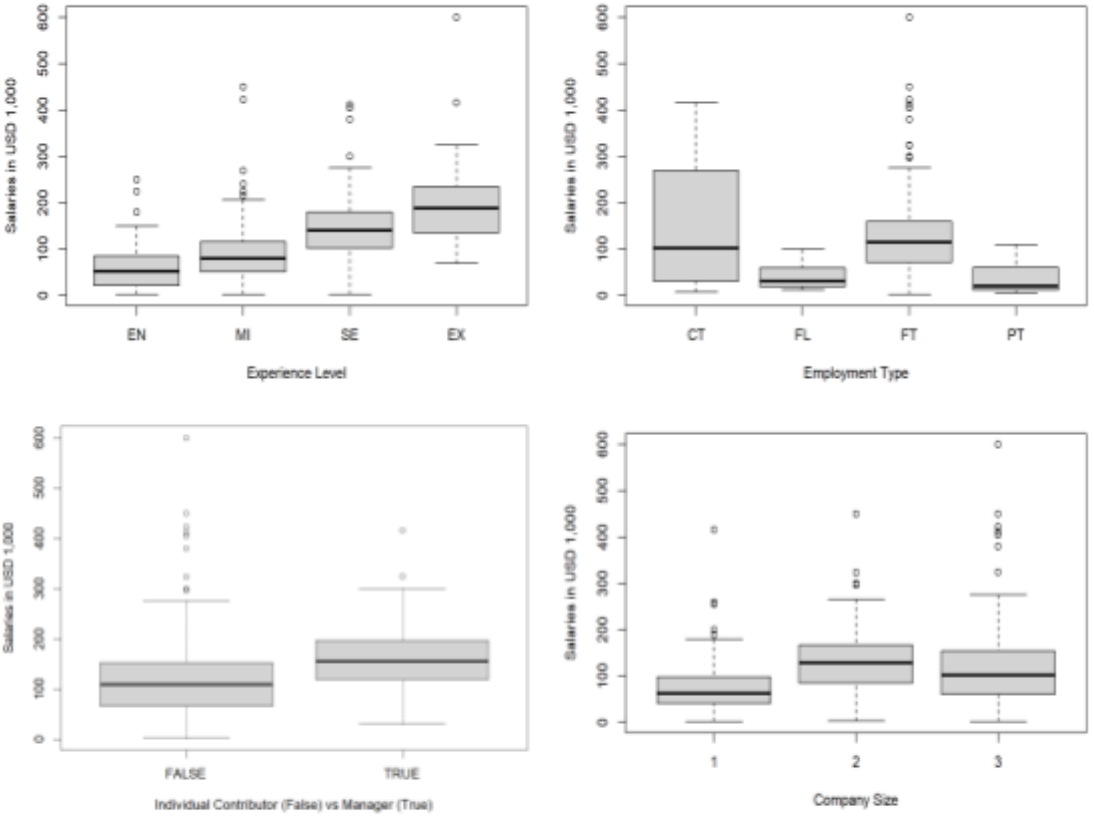


Figure B.4: Above plots show the distribution of salaries across different categories.

Appendix C

The larger question can be to understand how we can maximize the value we can derive from an employee. The value that a company obtains for its expenses in a year, given that the employee stay is:

$VC(X) = V(X) - S(X)$, where

- $VC(X)$ is the value that the company obtains for its expenses in a year
- $V(X)$ is the value that the employee brings to the firm in a year,
- $S(X)$ is the salary that we have to pay the employee in a year

Since there is a chance that an employee can leave the firm before the year ends, we can account for the probability of the employee leaving. This can be modeled as $P(Stay \mid X, C(X))$ and so our $VC(X)$ now becomes:

$VC(X) = P(Stay \mid X, C(X)) \cdot \{V(X) - S(X)\}$. The hiring data that we will have integrated will help us model this equation and optimize over the value.

We can also use Quantile Regression to find the 0th Quantile or 25th Quantile. Given additional data about when candidates reject the job, we can build a cost-benefit matrix and get the quantile that maximizes our gain.

Appendix D

This section notes the contribution of each team member. In addition to working on the final report, individual contributions of the members are included below:

- Alex Gille - Worked on question 4 to understand how job titles affect the salary of roles with similar responsibilities
- Ana Gonzalez and Arushi Dheer - Worked on question 1 to see if there was a difference in salaries between people that lived in the US and those that lived elsewhere given that they worked for a US company
- Monil Soni - Worked on question 3 to understand the salary impact of making the job remote
- Xuchen Tan - Worked on question 2 to understand if cheaper to hire a part-time employee for different job lengths and titles