Report

Visualisations based on the "Data Related Jobs in US" dataset

Yozlem Ramadan y.ramadan@student.vu.nl

Malgorzata Zdych m.j.zdych@student.vu.nl

VU VRIJE UNIVERSITEIT AMSTERDAM

1. Introduction

The dataset used in this report was obtained from Kaggle, a widely known platform among data science enthusiasts used to share, explore and analyse datasets (Sayed, 2023). This dataset in particular was created by Mohamed Sayed, a Kaggle contributor. It contains a comprehensive collection of key factors relevant to job evaluation across different career positions in the data-related field in the US.

The size of the dataset is relatively substantial, consisting of 2084 occupations, each one categorised within 23 different categories: company name, job title, location (in US state), job description, salary estimate, company size, company type, company sector, company industry, company founded, company revenue, hourly payment or not, company rating on glassdoor, required skills (such as python, spark, azure, aws, excel, machine_learning), job title simple, seniority, job description length and company age. In the following section, ten chosen features from this broader set will be further specified.

While it may not be a widely known or popular dataset, it offers valuable insights into data-related jobs in the US. In 2023, the US was characterised as the top-paid nation for the data scientists (Zaveria, 2023). The constant growth in the data-related job market is observed, as there is an increase in technological advancements (Carlos Orellana Fantoni, Mero and Vaca, 2020). Therefore, it may serve as a representative sample for similar trends in the data-related field worldwide. It is worth noting that the dataset was updated, approximately eight months ago. This ensures to some extent that the data provides relevant insights into the current job market for data-related professions. The dataset’s quantity and diversity make it suitable for research purposes, allowing comprehensive analysis of the data. People who can particularly benefit from this evaluation include job seekers, such as Data Science students, professionals looking to switch to technology-related roles, and aspiring entrepreneurs interested in the data science field. This data can provide them with valuable insights into the field's potential and dynamics, enabling them to make well-informed career decisions.

2. Dataset overview

The ten clearly different features were chosen from the dataset and divided into two categories numerical data and non-numerical data. For the numerical data minimum, maximum, average and missing percentage was counted. For the non-numerical data, such as job titles or locations, missing percentage, most and least common values were calculated to show the distribution of the data. It's worth noting that there are several values in the dataset that are considered least common, as those are usually the ones that appear only once in the dataset. Thus, not all the least common values will be listed, just a few examples.

In the dataset, any missing data in non-numerical columns, represented by values like "Unknown" or "Unknown / Non-Applicable," was replaced with NaN values to make sure that the sample does not contain quality problems. Moreover, the “salary estimate ($)” signifies yearly salary estimate, but it will be referred to as just salary estimate.

2.1 Numerical values

	Minimum	Maximum	Average	Missing percentage
Salary estimate ($)	3760.0	297000.0	108768.64	0.00
Company age	1.0	397.0	57.26	33.62
Rating	1.0	5.0	4.06	17.89

Figure 1.1 The distribution of the numerical values in the dataset

2.2 Non-numerical values

	Common Values (Mode)	Least Common Values	Missing percentage
Company	I28 Technologies	Swyftt, Liberty Mutual Insurance, Meta, (...).	0.00
Location	Remote	Exton, PA, Farmers Branch, TX, (...)	0.00

Company type	Company - Private	Self-employed	8.35
Company industry	Internet & Web Services	Stock exchanges, Metal and Mineral Manufacturing, (...)	24.75
Company founded	2016	1919, 1852, 1886	33.62
Python yn	1	0	0.00
Seniority	Senior	junior	0.00

Figure 1.2 The distribution of the non-numerical values in the dataset

3. Charts

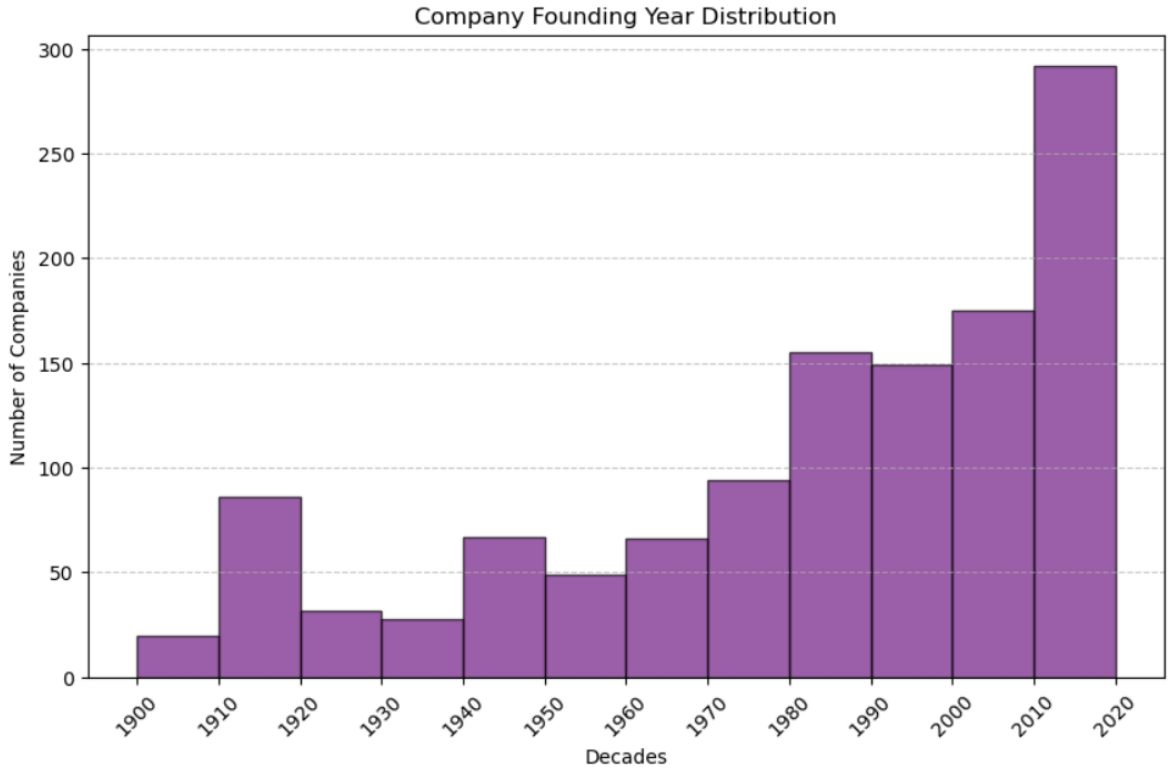

Figure 2.1 The distribution of salary estimate by average rating

The graphical representation provided in Figure 2.1 illustrates a timeline, grouped by decades and the number of companies founded in this period. The histogram captures the slice of the dataset that takes into account the 20th and 21st century, so most historically

recent times. The distribution of founding years ranges from 1900 to 2020. Overall, it reveals a gradual growth in the number of companies being founded for a given decade.

There are two notable exceptions in the graph, which occur right after the decades of 1910-1920 and 1940-1950. Interestingly, even though there is an increase in the number of companies founded during these decades, the subsequent two decades experienced a significant drop in the number of new companies. This occurrence might likely be attributed to the disruptive impact of World War I and World War II during these periods.

However, starting from 1950 to 2020 there is almost gradual increase in the newly established companies. Most significant and impressive growth can be observed in the last decade marked between the years of 2010-2020, where the number of companies nearly doubled that of the previous decade (Zaveria, 2023).

As this graph focuses on companies within the realm of data science employment, the last decade before 2020 stands as an excellent example of the outstanding expansion of this field. This observation aligns with the initial expectation of the group that well-established companies have existed for many years, but the majority of companies in this domain were founded in the last decade. Lastly, it suggests rising interest and popularity of these roles and companies in recent years, with potential for continued growth in the future. Particularly, it gives an insight for the job seekers that the field is worth exploring and should be taken into consideration.

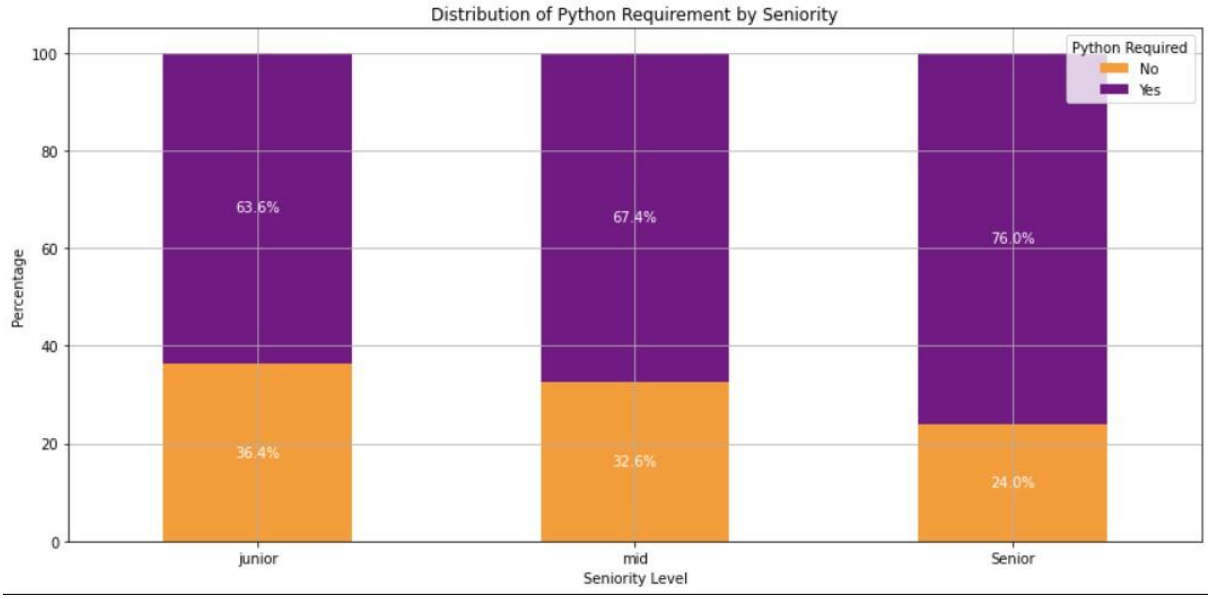

Figure 2.2 The distribution of Python requirements based on job seniority levels.

Figure 2.4 illustrates the relationship between seniority levels in job postings and the demand for Python skills. The chart provides an overview of the percentages of job listings requiring Python proficiency at different seniority levels. The bar chart shows that for junior-level positions, approximately 63.6% require Python, while the remaining 36.4% do not. In mid-level postings Python is required with a similar score, but slightly higher - 67.4%. Eventually, among the senior-level positions the demand for Python is the highest and 76% of the offers require Python proficiency.

A relatively high demand for Python among the junior positions could be related to its fundamentality in the technological industry, particularly nowadays (Raschka, Patterson and Nolet, 2020). Data analysis or scripting may be examples of the tasks that can be done in Python and are usually assigned to junior roles. Such a finding can give an idea to the job seekers at the junior level about how Python proficiency may increase their chance of finding suitable positions.

The responsibilities of the mid-level jobs are often more complex, so they can require skills like Python more often, therefore a slight increase from 63.6% to 67.4%. Mid-level professionals with Python skills may be at a more advantageous position in the job market, as the significant majority of mid-level positions demand such a proficiency.

Senior-level positions typically demand not only high-level professional skills but also strong leadership and strategic abilities. Consequently, individuals applying for such positions need to be proficient in managing and supervising employees in lower-level positions. This often entails having advanced knowledge of languages like Python, ideally at a higher level of proficiency or at least a similar one. Furthermore, the data-driven insights that can be derived from Python are crucial for analytics in senior and management roles. The senior level professionals with Python skills are highly demanded, it may be expected that soon almost every data-related higher level job will require this skill. This can indicate a growing acknowledgment of its use in the decision-making processes and automation (Raschka, Patterson and Nolet, 2020).

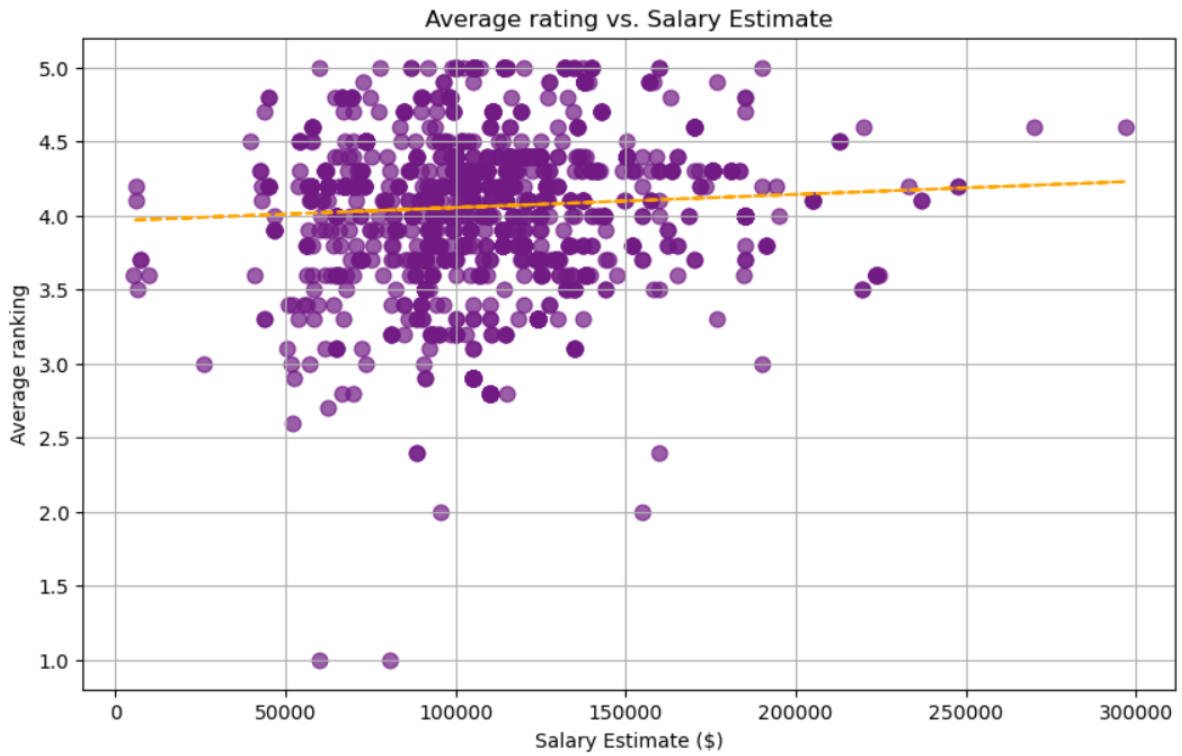

Figure 2.3 The distribution of salary estimate by average rating

Figure 2.2 presents a scatterplot illustrating the relationship between two key variables: the average ranking of companies, which is measured on a scale from 0 to 5, and salary estimates of a position in a specific company, corresponding to the ranking data.

On first inspection it is notable that in this scatterplot the concentration of data points fall within the salary range of $50,000 to $150,000, with the centre of this cluster occurring around the $100,000 region. On the average ranking axis, data points are primarily distributed between 3.5 and 4.5, with the central clusters centred around an average ranking of approximately 4. As a result, the scatterplot suggests that there is a common tendency among the companies in the dataset to offer job positions with salary estimates falling within the $50,000 to $150,000 range, while their average rankings hover around 4.

It is worth mentioning that in this dataset, the companies with the highest ratings do not always offer the highest-paying job positions. This observation suggests that correlation between a company's rating and their salaries is unlikely. It is further supported by the fact that two jobs with the lowest company ratings provide comparable salaries to those offered by companies with the highest ratings. Additionally it is interesting to mention that the companies with top ranking still fall within the salary range that is most common in the

dataset. The highest-paying job offers tend to come from companies with an average of around 4.6 points. The trend line in the scatterplot supports the observation that higher ranking of the company does not necessarily result in significantly higher salary estimates for the jobs they offer. This might be the case, because the ratings mostly reflect a safe working environment, which often is not strictly associated with the high salary payments. On the contrary, usually the companies with higher ratings are small and still developing companies, which can provide employees with collaborative and rewarding workplaces, but smaller salaries (Hasle et al., 2012). This suggests that other factors potentially play a more prominent role in salary estimation than the ranking of the companies.

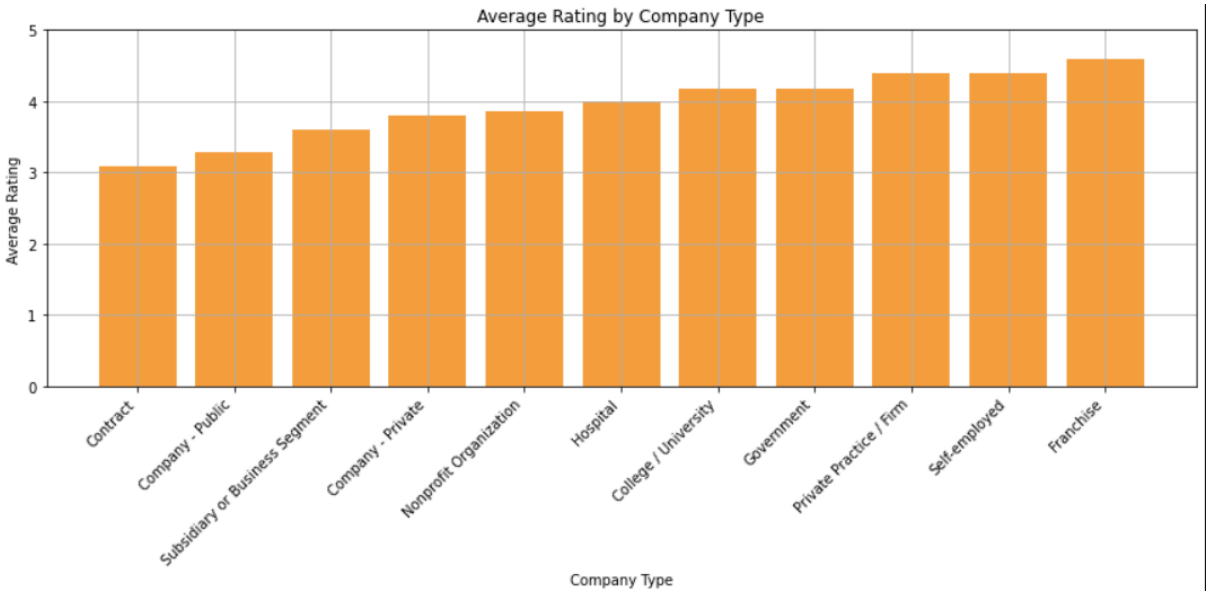

Figure 2.4 The distribution of average rating by the company type

In Figure 2.3 bar chart illustration in ascending order can be observed. The aim is to observe if there is a relation between the types of companies and their average rates. Therefore on its x-coordinate are presented types of companies in the dataset and on the y-coordinate average rating of these companies, for example the same company type. The rating ranges from minimum of 3 points to a maximum or over 4.5 points, as they gradually increase from “contract” type to “franchise”. This can be the case, because the franchises tend to be rated more favourably compared to any other company types. This could be due to factors such as customer expectations or loyalty to a given franchise. The franchises usually are an established brand on the market and they care about their public image, therefore they may put a lot of effort into the brand promotion and ensuring that the ethics and environment in

their company is consistent with their reputation. In the franchises they are more likely to engage in the brand supporting behaviour (Nyadzayo, Matanda and Ewing, 2015). Similar situation can apply to the self-employed company type, where usually a small group of people take care of the company, so they have a personal insight and influence in their image. They also have a direct and personal gain from how they operate. Therefore, they put more effort into ensuring safe and collaborative work environments, which contributes to positive employee’s reviews. Contrarily, the contract type of company is usually temporary, more of a freelance agreement, so thus there is a smaller emphasis on the work environment. The workers are usually hired for a given task and they usually have a limited time to be integrated into the company's culture and become familiar with their values.

Such an overview can give an insight for the job seekers about how the type of the company influences work culture, environment and integrity. It is clearly seen, where employees feel at home and where the employer puts emphasis on the company's branding and value in everyday life of their workers.

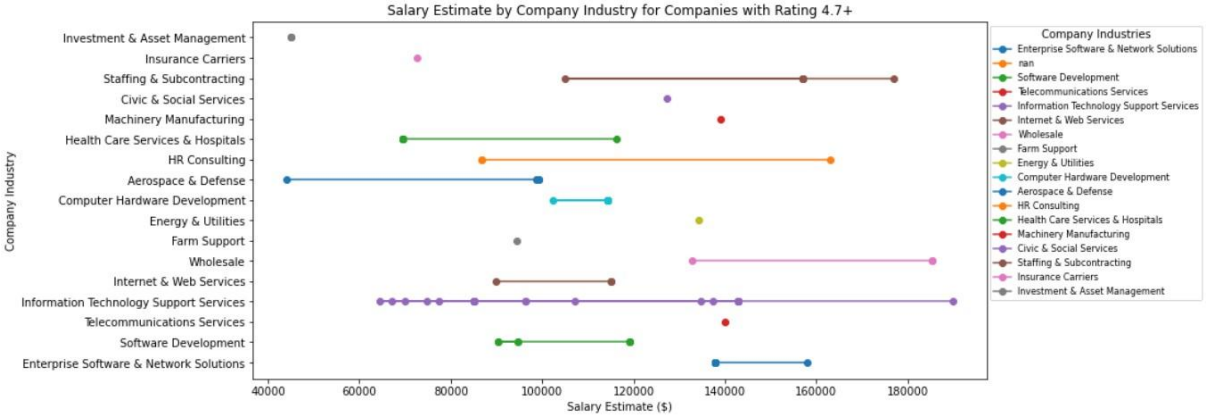

Figure 2.5 The distribution of average salaries among industries for companies with a rating of 4.7+

As it can be seen on Figure 2.5, the relationship between company industry and the salary estimate in that field is presented. Moreover, only the companies with rating 4.7 or higher are taken into account, since the interest was in the firms, where employees are “very satisfied” about their workplace. However, it is important to notice that some of the companies do not have ratings, so therefore they are not taken into consideration. In general, it is worth noticing that the ratings can be a very subjective and inaccurate measure, but according to their

proprietary algorithms try to prevent that by giving the more recent reviews a higher weight towards the overall rating. It ultimately gives users of Glassdoor.com the most up-to-date perspective on the workplace and its environment. Moreover, the Glassdoor significantly contributed to exposing malpractice in the workplace and it had an enormous impact on the improvement in workplace disclosures based on the employees reviews (DUBE and ZHU, 2021). Therefore, it can be considered a reliable source of information about the company’s work ethic.

It can be seen in the graph that there are many industries that have only one job offering, therefore the range of their salaries is not visible. They can be very niche industries, but because of the applied filter of the rating 4.7+ it could be a potentially plausible industry for the job-seekers. Usually the smaller companies, freshly developing, have more to offer to their employees and due to their size their work environment is more safe and friendly (Hasle et al., 2012). Additionally, it can be noted that the salaries of those singular companies do not exceed $140,000, which can also be influenced by their smaller budgets. It can be easily noticed that Information Technology Support Services is the most diversified when it comes to the salary estimate, as it can range from approximately $60,000 to $190,000. It is a very broad and large industry, according to (www.factmr.com, 2023) it was worth $66.3 billion only in 2023 and it was experiencing a steady growth throughout the years. It is expected to reach $111 billion by the end of 2023. Thus, there are many companies operating within this field. Another interesting case can be the Aerospace and Defense industry, which is more of a niche field, however their salaries also range heavily between $40,000 and $100,000. It represents the lowest salary on the graph, so therefore it might be a small company offering the job, which would explain their high rating. In general, the salary estimates can heavily differ within the industry, which is the case for HR Consulting, Wholesale, Information Technology Support Services and Staffing and Subcontracting. It may be due to many factors like the company size or the popularity of the industry. However, still the highest salaries remain very high regardless.

The potential job seekers can derive valuable insights into the work environments with a good and proven record of employee satisfaction. A healthy and safe workplace is a very good attribute that facilitates people’s job search. Additionally, the chart's depiction of industries with only one job offering, despite potentially niche markets, could signify sectors with potential that are worth exploring. Moreover, the prevalence of small and developing firms can imply that the work experience there is better than in big companies, so it can

strongly influence decision-making. Finally, the chart highlights the diversity of salary estimates within different industries, helping job seekers find the industry that would meet their financial expectations and give an idea, in which direction they should go. To sum up, this chart can give job seekers insights to make more informed decisions and find unique opportunities.

4. Problem description

One problem that can be solved using the selected dataset “Data related jobs in US” is predicting the salary estimate for job positions in the field of data science and analytics. The target value that can be predicted could be the salary estimate, and the columns that can be used for prediction include location, company industry, job seniority level, company type, and their average rating. The problem could be faced by the HR professionals and managers that often struggle with determining competitive salaries for their employees. Whereas, it is one of the most significant aspects that job seekers look at and moreover, it shows the current employees that they are an asset to a company that is fairly valued. Motivating salary is one of the factors that can lead to employee’s enthusiasm about the job (Lee and Lin, 2014).

As of now, the professionals who deal with making salary estimates often base their insights on market research from comparable companies and institutions, past salary values, and current trends in the nation, which, for instance, can depend on inflation or the economic situation (Rees, 1993). Such a process can be very time-consuming and may not take into consideration factors that are listed in this dataset, such as company industry or company’s rating based on the employee’s experience. Therefore, a data science model could significantly improve estimating salaries in the data-related fields, especially considering the fact that it is a rapidly growing and changing field.

The given dataset is very recent and up-to-date, thus it can provide accurate information and estimates. A well-trained model can also take into account many features, for instance rating, industry, or company type and make precise predictions based on that. The data-science models can also increase the speed of the estimating process, in comparison to HR professional’s market search. The model could also provide insights into its decision-making process and estimations, which would enhance its understanding for HR professionals and managers. What is more, the model can provide consistent, impartial and objective salary estimates, reducing for example gender or race bias, in contrast to human’s assessment. This could be achieved by only considering input features that are relevant for the job and not

directly focusing on the candidate’s personal attributes. The model can also find which features have a crucial impact on the salary estimations, so the companies can make more data-driven decisions in the future.

5. References

5.1 Scientific Papers and main results of the references:

Carlos Orellana Fantoni, Mero, A. and Vaca, C. (2020). Tech for Hire: Data Science-Related Jobs Signal Economic Growth. doi:https://doi.org/10.1109/icedeg48599.2020.9096765.

There is a strong increase in the technological advancements and data science and analytics jobs rise in popularity.

DUBE, S. and ZHU, C. (2021). The Disciplinary Effect of Social Media: Evidence from Firms’ Responses to Glassdoor Reviews. *Journal of Accounting Research*. doi:https://doi.org/10.1111/1475-679x.12393.

The Glassdoor contributed to detecting malpractice in the work environments with their transparency of reviews and made an impact in changing work dynamics in many companies.

Hasle, P., Limborg, H.J., Kallehave, T., Klitgaard, C. and Andersen, T.R. (2012). The working environment in small firms: Responses from owner-managers. *International Small Business Journal: Researching Entrepreneurship*, 30(6), pp.622–639. doi:https://doi.org/10.1177/0266242610391323.

The work environment patterns in small firms are usually very positive.

Lee, H.-W. and Lin, M.-C. (2014). A study of salary satisfaction and job enthusiasm – mediating effects of psychological contract. *Applied Financial Economics*, 24(24), pp.1577–1583. doi:https://doi.org/10.1080/09603107.2013.829197.

The employees that earn fair income are more motivated and excited about their job.

Nyadzayo, M.W., Matanda, M.J. and Ewing, M.T. (2015). The impact of franchisor support, brand commitment, brand citizenship behavior, and franchisee experience on

franchisee-perceived brand image. *Journal of Business Research*, 68(9), pp.1886–1894. doi:https://doi.org/10.1016/j.jbusres.2014.12.008.

The franchises are more likely to support their brand image and engage in the brand supporting behaviours.

Raschka, S., Patterson, J. and Nolet, C. (2020). Machine Learning in Python: Main Developments and Technology Trends in Data Science, Machine Learning, and Artificial Intelligence. *Information*, 11(4), p.193. doi:https://doi.org/10.3390/info11040193.

Python has seen exponential growth in popularity in recent years.

Rees, A. (1993). The Role of Fairness in Wage Determination. *Journal of Labor Economics*, 11(1, Part 1), pp.243–252. doi:https://doi.org/10.1086/298325.

The salaries are estimated based on the surveys from the competitive industries and companies.

5.2 Other references

Zaveria (2023). *Top 10 Highest-Paying Countries in Need of Data Scientists in 2023*. [online] Analytics Insight. Available at: https://www.analyticsinsight.net/top-10-highest-paying-countries-in-need-of-data-scientists-in-2023/.

Sayed, M. (2023). *Data Related Jobs in US*. [online] www.kaggle.com. Available at: https://www.kaggle.com/datasets/mohamedsiika/data-related-jobs-in-us [Accessed 26 Sep. 2023].

www.factmr.com. (2023). *Fact.MR – Tech Support Services Market Analysis By Services (Customer Acquisition, Call Center Services, Email & Live Chat Support), By End-Use Industry (BFSI, Education, Healthcare, Manufacturing) & By Region (North America, Latin America, Europe) - Global Market Insights 2023 to 2033*. [online] Available at: https://www.factmr.com/report/tech-support-services-market#:~:text=The%20global%20tech%20support%20services.

5.3 Alternative Datasets

1. Data Science Jobs in India

This dataset was retrieved from Ambition Box’s website. It has variables such as company name, job titles, minimum experience, average salary, minimum salary, maximum salary and number of salaries. Even though it is a nice aggregation of data science jobs it does not meet the standards of the assignment to have at least 10 variables. Moreover, the websites from which the data was collected is not available in every country. The size of the dataset is smaller (1601 instances) in comparison to the one chosen for this assignment. Lastly, it is twice as popular as the chosen dataset, which decreases the chances of having a unique dataset.

2. Data Science Jobs Salaries Dataset

This dataset is in a similar field, created mainly for salary comparison between the data science jobs. It includes the variables: work year, experience level, employment type, job title, salary, salary currency, salary in USD, employee residence, remote ratio and company location. The aim in this dataset is to compare mainly the salary changes between the years. It was a relatively famous dataset with a high number of downloads, which means the probability of this dataset being used in this course is higher. Moreover, the last time it was updated was 2 years ago, which may affect the current accuracy of the data in the market.

Pant, M. (2022). *Data Science Jobs in India*. [online] www.kaggle.com. Available at: https://www.kaggle.com/datasets/madhurpant/data-science-jobs-in-india [Accessed 27 Sep. 2023].

Shahane, S. (2021). *Data Science Jobs Salaries Dataset*. [online] www.kaggle.com. Available at: https://www.kaggle.com/datasets/saurabhshahane/data-science-jobs-salaries.

Front page picture:

https://studyonline.unsw.edu.au/blog/data-science-degree-jobs