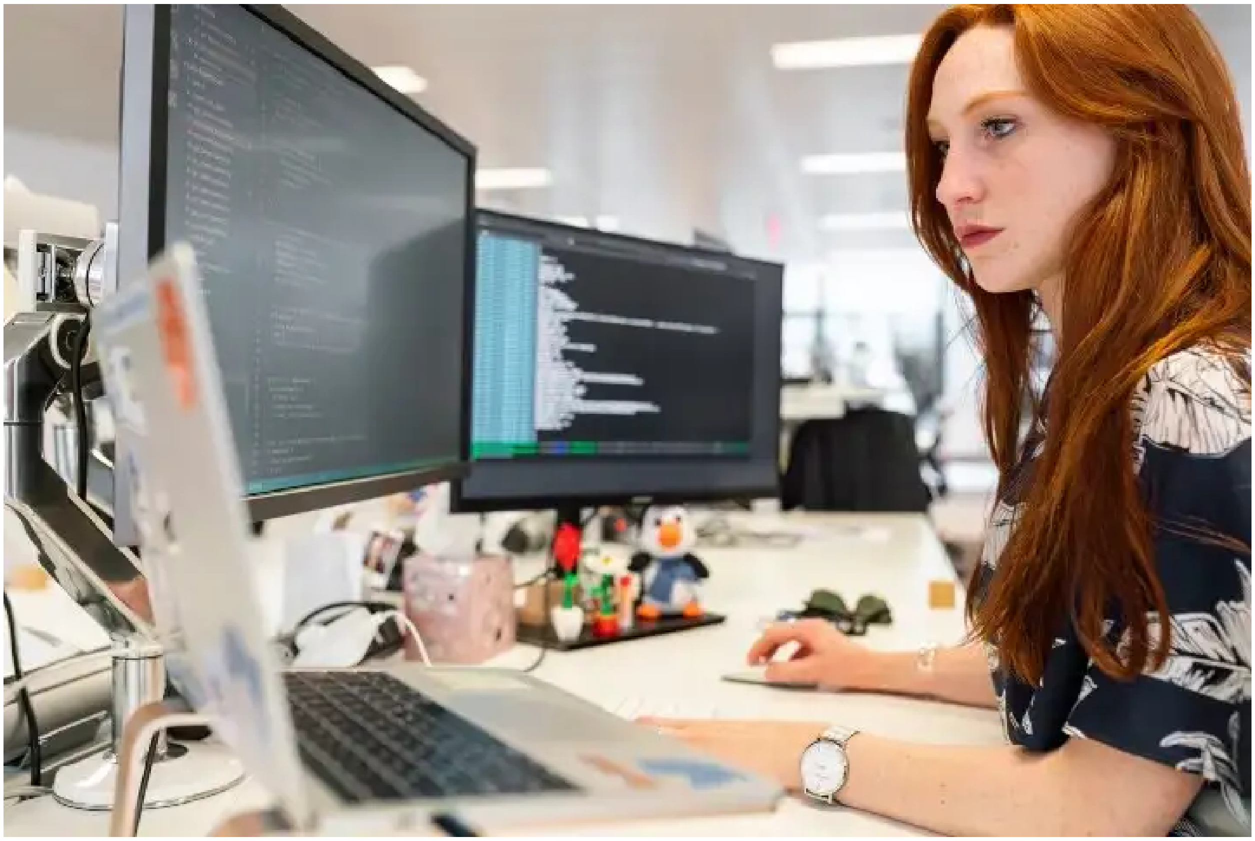
DS6103 - Intro to Data Mining
Fall 2022

Salary Prediction for Data Science related jobs

Authors: Tanmay Kshirsagar, Sunisha Harish, Kunal Inglunkar, Alejandra Mejia



Introduction

Having been called one of the “sexiest” jobs of the 21st Century, Data Science has become more and more relevant in today's world. Bringing together the domain expertise from programming, mathematics, and statistics, data scientists have the ability to transform and obtain insights from existing data that is not necessarily useful on its own, and use that information to reduce uncertainty. Organizations are becoming more aware that the value of data is soaring heights, and how through it, it is possible to unveil solutions and intelligent decisions across many industry verticals *(Figure 1. Ways data scientists add value to organizations*). According to Deloitte (2021), businesses are expected to increase their spending in data analytics by around 76%. Moreover, the 2020 LinkedIn U.S. Emerging Jobs Report highlights how data science has experienced continued growth on a tremendous scale in recent years. According to Fortune Education (2022), it is a great time to have a degree in data science due to a hot job market for these roles. As prospective data scientists, it becomes relevant for us to better understand the actual conditions of the labor market that we are going to face when we seek employment. For this particular reason, we decided to use the power of data and explore some market trends as well as try to predict the salaries in our future field of work.

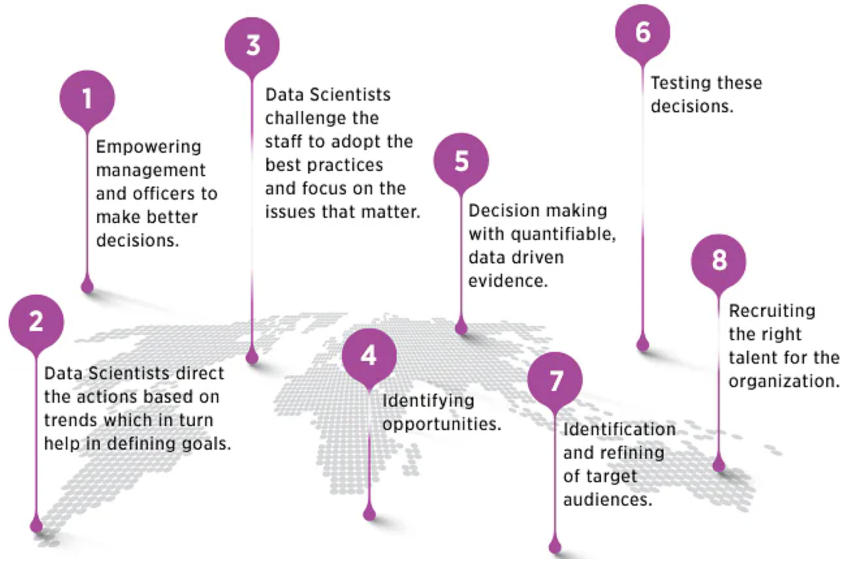

1
Empowering management and officers to make better decisions.
2
Data Scientists direct the actions based on trends which in turn help in defining goals.
3
Data Scientists challenge the staff to adopt the best practices and focus on the issues that matter.
4
Identifying opportunities.
5
Decision making with quantifiable, data driven evidence.
6
Testing these decisions.
7
Identification and refining of target audiences.
8
Recruiting the right talent for the organization.

Figure 1. Eight ways a data scientist can add value to any business
Source: Simplilearn

The following research will be structured into five sections that will include information about the dataset, exploratory data analysis, modelling, limitations and other considerations and conclusions.

About the Dataset

Glassdoor is an American website that started in 2008. The website provides a platform in which current and former employees anonymously review companies, where open roles can be found along with relevant information about companies names, rating, revenue, location and employees, as well as salary data. Glassdoor collects millions of job postings from a wide variety of online sources each month. The site does not allow bulk download of their data and checking each job manually could take an extremely long time. One of the best solutions to be able to access the data in a more efficient way is through web scraping. In the United States of America (USA) it is completely legal to scrape publicly available data, however, there are some caveats regarding copyrighted data and personal information. Given that Glassdoor mentions in their Terms of Use that they do not like to be scraped, we decided to use an existing dataset that was scrapped by someone else. For this particular project, we used the data created by *pickesueat* published in GitHub, which contained scrapped Glassdoor data from June 2020. The dataset was scrapped having in consideration four different job categories: Data Scientist, Data Analyst, Data Engineer and Business Analyst. Overall, the dataset had a total of 26,980 observations and 15 different features, which are detailed in the table below.

Table 1. Dataset features

No.	Feature	Description	Data type
1	Job Title	Job Title of the job opening	object
2	Salary Estimate	Minimum and maximum salaries for the job opening	object
3	Job Description	Details of the job requirements	object
4	Rating	Company ratings determined by recent employee feedback	float64
5	Company Name	Name of the company	object
6	Location	Location of the company offering the job position	object
7	Headquarters	Headquarter location of the company offering the job position	object

8	Size	Approximate number of employees in the company	object
9	Founded	Year the company was founded	int64
10	Type_Ownership	If the company is public, private or others	object
11	Industry	Industry in which the job is open	object
12	Sector	Sector in which the job is open	object
13	Revenue	Range of company revenue	object
14	Competitors	Name of the company’s main competitors	object
15	Easy_apply	If the job opening offers easy apply option	object

Exploratory Data Analysis (EDA)

The EDA was an important first step in our project that helped us identify general patterns in the data, like outliers and unexpected features, identify trends, determine whether a predictive model was a feasible analytical tool, gain an understanding of the data set beyond the formal modeling or hypothesis testing task, and comprehend the structure of the dataset, which allowed us to expand by leveraging the relationship between the variables and made the data modeling more streamlined. Based on our hypothesis that salaries were not static among places or industries, we wanted to be able to answer through the EDA three key SMART questions :

1. Does salary estimate vary between the states in the USA?
2. Is the salary estimate in the USA correlated with revenue of the company?
3. Does the salary estimate in the USA depend on the industry and the sector?

In order to do so, it was important for us to clean our data for better insights. The first section explains the details of the data manipulations that we performed.

1. Preparing The Data

Although the scraped data was fairly cleaned, it still required some manipulation which led to the creation and dropping of columns, and transformation of information from string to float. As an initial step in the data cleaning process we reviewed the amount of missing values in the dataset and we were able to find that :

- 0.00% or 0 values are Missing in job_title Column
- 0.11% or 29 values are Missing in Salary_Estimate Column
- 0.00% or 0 values are Missing in Job_Description Column
- 8.76% or 2347 values are Missing in Rating Column
- 0.02% or 5 values are Missing in Company_Name Column
- 0.00% or 0 values are Missing in Location Column
- 5.17% or 1386 values are Missing in Headquarters Column
- 4.90% or 1313 values are Missing in Size Column
- 24.77% or 6636 values are Missing in Founded Column
- 4.90% or 1313 values are Missing in Type_ownership Column
- 12.87% or 3448 values are Missing in Industry Column
- 12.86% or 3445 values are Missing in Sector Column
- 4.90% or 1313 values are Missing in Revenue Column
- 72.02% or 19294 values are Missing in Competitors Column
- 96.26% or 25788 values are Missing in Easy_apply Column

Given that the features “Competitors” and “Easy_apply” had more than 70% of missing values, we decided to remove them to avoid obtaining inaccurate conclusions from our data. Also we removed data where job locations were outside the USA, like Canada.

Secondly, we parsed the Job Descriptions given that they contain extremely long paragraphs of the description of the job position, that as a whole, were not easy to manipulate when analyzing the data. With this feature, we were also able to group the skills into reportable categories like Statistics, Machine Learning, Data Visualization, Data Engineering, Software Engineer, SQL, Trait Skill, Social Skills and Business. We also cleaned the company's name given that they came with rating information in the same column.

The job title was divided into two additional columns detailing job domain and role. The location and headquarters were transformed into two columns each, to be able to obtain city and state in different columns. For the size and revenue of the company we considered the maximum value, and also the ranges for the latter. Additionally, rather than using the year founded we obtain the years of existence of the companies by subtracting 2022 to the foundation year. Furthermore, we separated salaries into a minimum (min) and maximum (max) salary column and obtained the estimated mean salary by dividing min and max salaries by two. As a final step, we replaced the null values with mean, median or mode in accordance with the type of data available in the columns.

2. Exploring the Data

Salary expectations are among one of the main factors that are relevant when searching for new job opportunities. Given its importance, we wanted to identify from our data how much a data science job role can fetch a person on average. We also wanted to know the minimum and maximum salary ranges they could expect *(See Figure 2).*

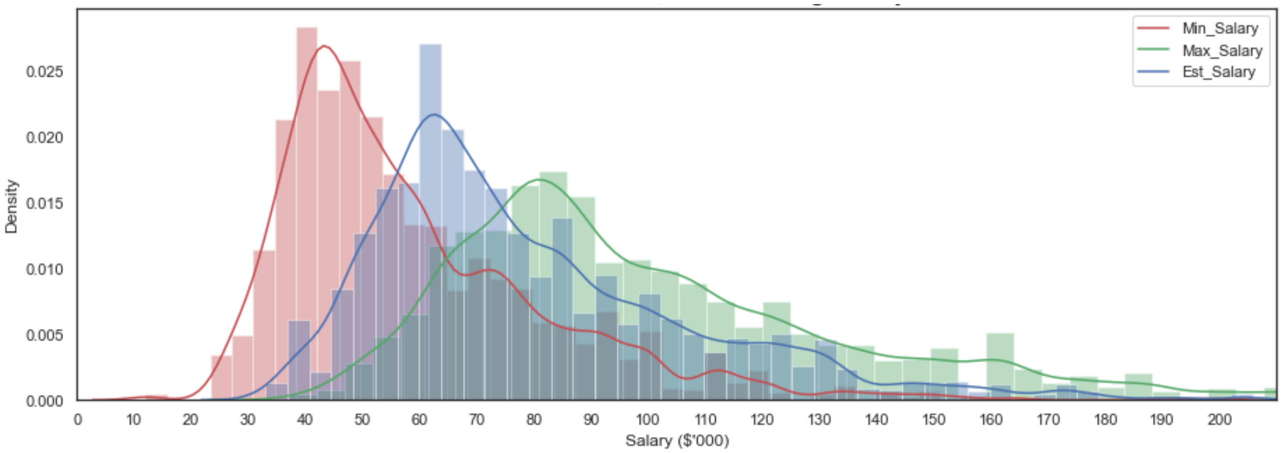


Figure 2. Distribution of data science salaries

We were able to unmask that the salaries for data science domain ranges between 20 to 200 thousand USD on average. The mean of the minimum salary is close to ~59,000 while the mean of the maximum salary is ~100,000. Moreover, when we analyze the average salaries, which are around USD 80,000, we see that they follow a unimodal distribution, making it a useful indicator for any subsequent statistical analysis.

An additional factor to consider while on a job hunt are the companies that are popular in the industry. In order to understand if the more popular companies usually offer higher salaries, we analyzed the salaries of the top 20 companies in terms of job openings offered. Regarding higher salaries, Apple and Amazon stand out, with average salaries above 100,000 US$. However, they both show a high level of variance in the offered salaries, which is a relevant finding given that it can be pinpointing at salary ranges based on skills, experience, or seniority. On the other hand, Staffigo, which stands out as the company with the highest number of job openings, shows lower salaries (US$75,000) but with low variance. A possible reason for this behavior is that this company specializes in permanent, contract and contract-hire IT jobs, which could be linked to standard skill sets that have a similar pay.

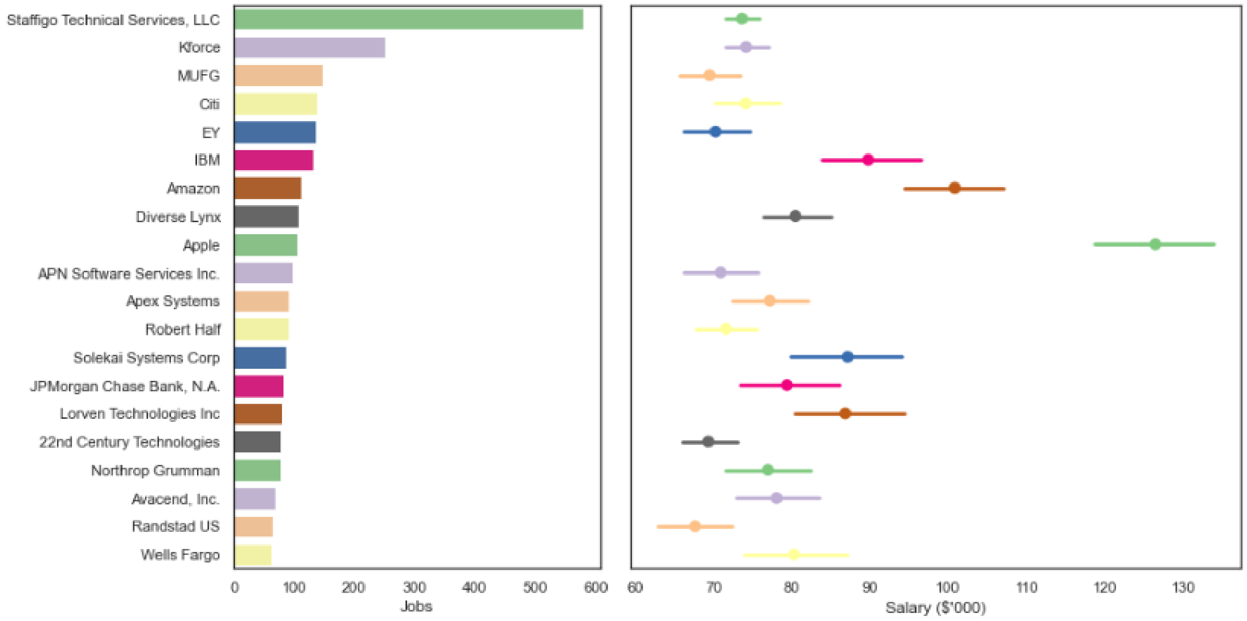


Figure 3. Job openings and estimated salaries by companies

As we mentioned before, skill sets could also be a key source of salary discrepancies. Data science is a rapidly evolving field, which requires training to be able to catch up with innovation. Therefore, it became relevant for us to understand what kind of skills were expected by different companies. Business and Social skills stand out among the main traits companies request in job descriptions, which includes words like reports, dashboards, business intelligence, teamwork, team, communication, leadership, and interpersonal, among others. Statistics is the third most requested skill which includes words like statistical, forecasting, R, pandas on the job descriptions.

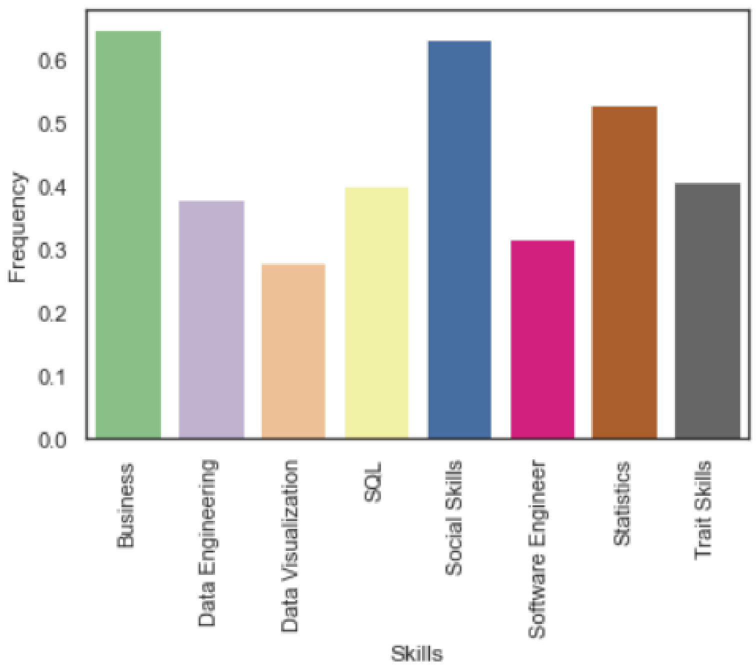


Figure 4. Main skills expected by hiring companies

Given that our alma mater is located in Washington, DC and considering networking, closeness to home as key factors when looking for job opportunities we decided to analyze the job opportunities in Virginia, DC and Maryland, from now on called the region, and compared it to the national average. The average salaries at the national level seems to be slightly lower than the regional average. The regional salaries seem to be more concentrated around USD 50-90K, having more outliers.

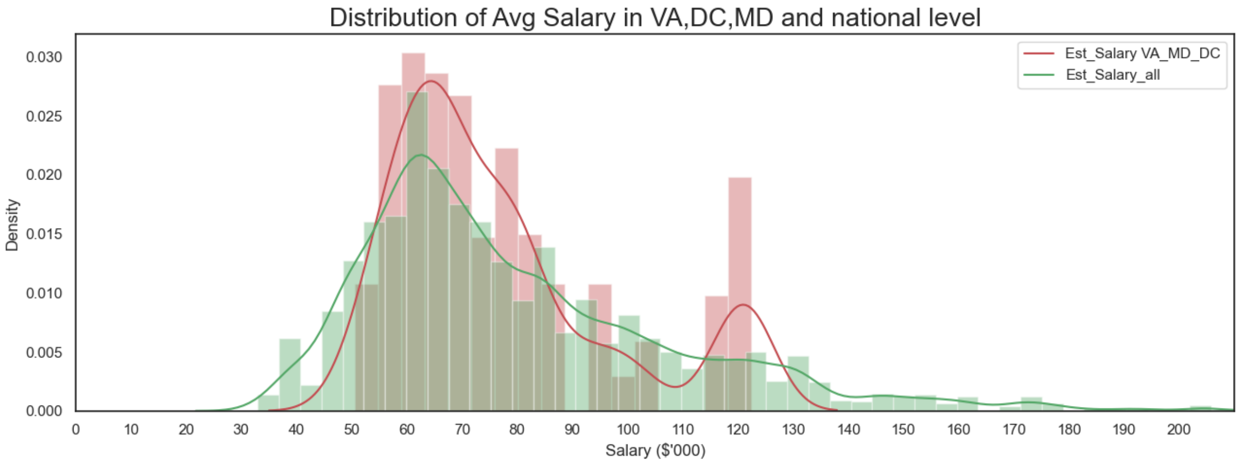


Figure 5. Distribution of data science salaries in Virginia, DC and Maryland and national level

Even though the previous visualization was useful to understand the average difference between salaries, it does not provide specific information regarding what could be the reason for these differences. In order to understand the region further, we developed heatmaps that were able to take into consideration revenue, number of employees and salaries offered by companies *(See Figure 6)*.

At a national level, a higher portion of jobs are offered by big companies that have 10 thousand or more employees and 10 billion or more in revenues. However, contrary to what researches like O'Reilly 2016 Report pinpoint that bigger companies pay higher salaries, it doesn't seem to be a conclusion that is supported by our data. At the regional level, the highest proportion of hiring is focused on small/medium businesses which usually pay more.

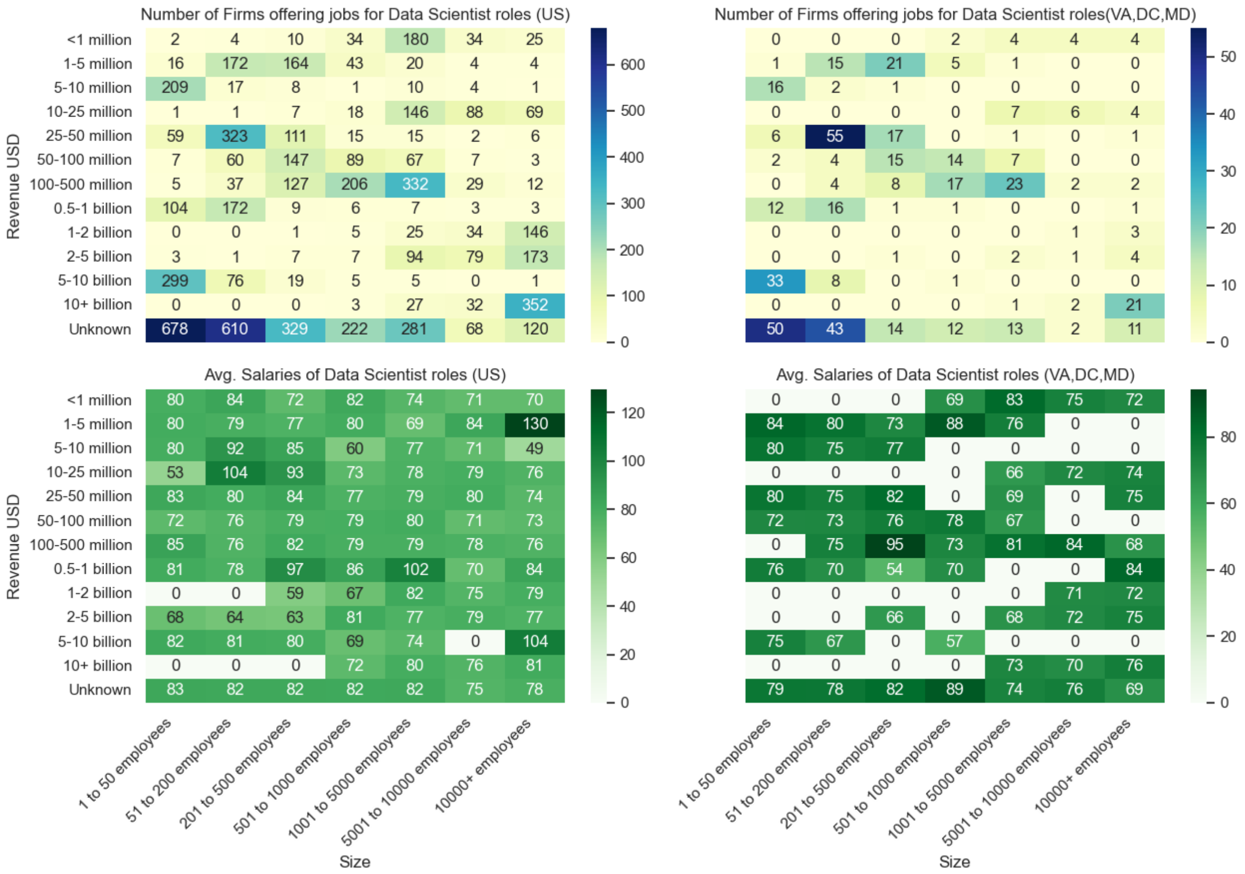


Figure 6. National and regional salaries and job openings comparison

Even though we would have preferred to do additional analysis in our region, our dataset only provided us with less than a thousand observations in our area. For this particular reason, we answered our SMART questions by focusing on the national data. The next section presents the analysis.

3. SMART Questions

Smart questions become a key element in helping us keep our research organized and reach a key conclusion about our data. As mentioned before, we identified three key relevant questions. In this section, we used our dataset to answer them.

1.Does the salary estimate vary between the states in the USA?

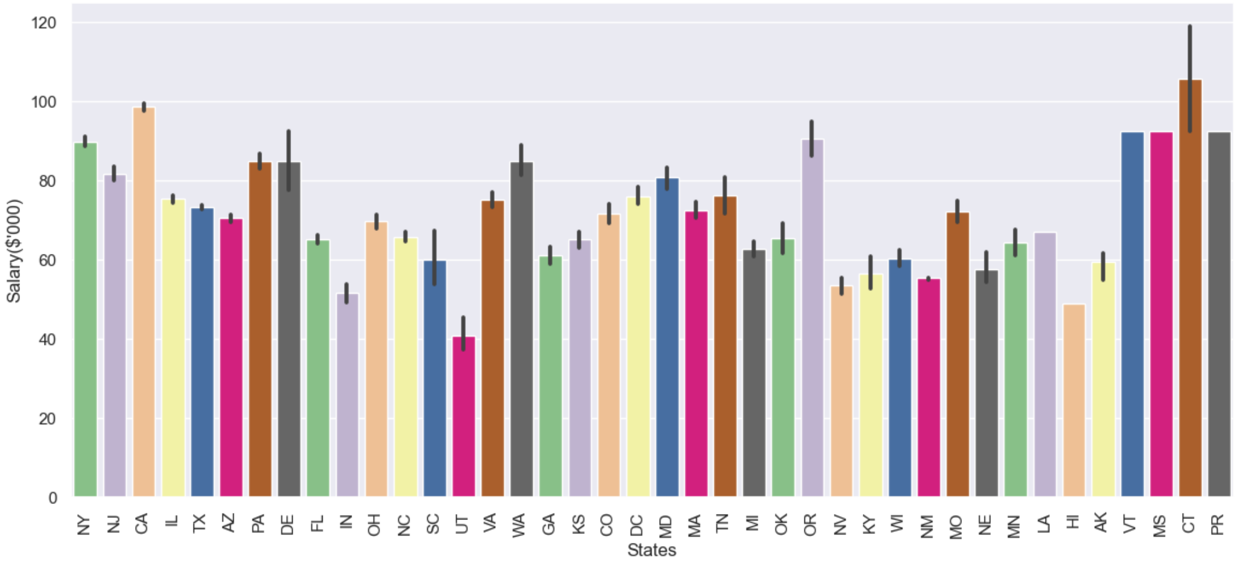


Figure 7. Estimated salary average for states in the US

Our first SMART question was about understanding the relationship between salaries and states. From the data we were able to see that salaries offered are strongly dependent on location. One interesting insight from the graph above is that the majority of the states that provide an average wage of more than 80,000 dollars are on either the west coast or the east coast.

California and Connecticut are the top two states with higher salary averages. Connecticut was a surprising discovery, whereas California was expected to have higher salary averages given that San Jose, the capital of Silicon Valley, is located there. Utah and Hawaii are among the states that have lower salary averages.

We did a deep dive into this question as it was also important for us to understand how the number of job openings played a role in the salary offered. For this, we considered and compared the top 20 states in terms of the number of job offerings.

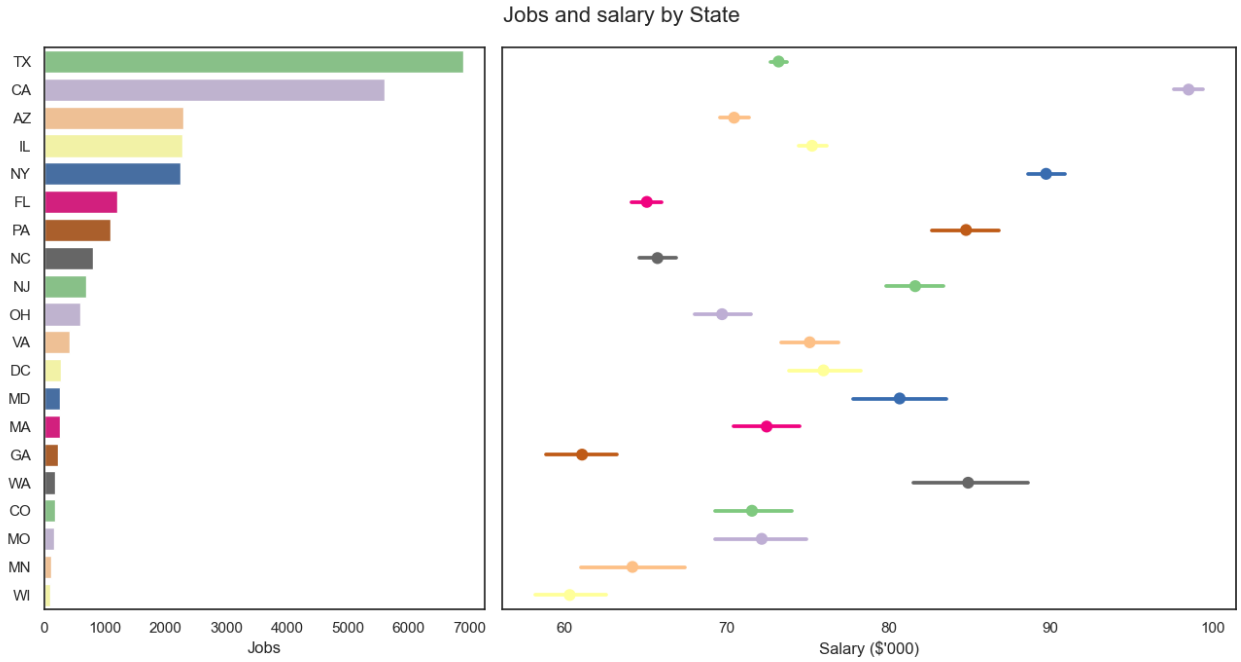
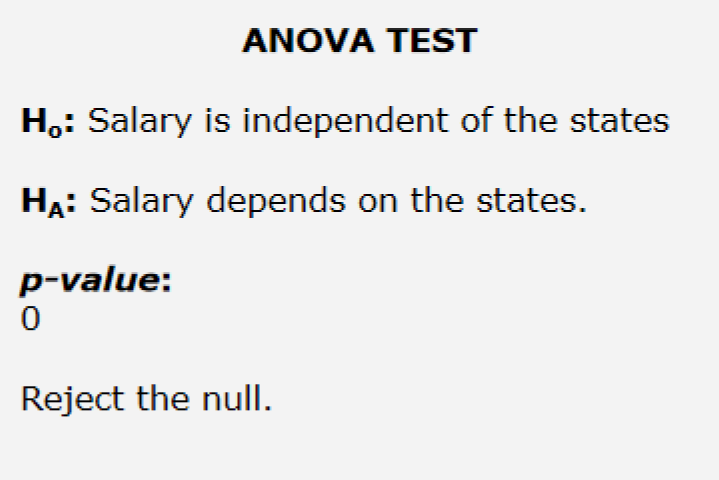


Figure 8. Job openings and estimated salaries by states

Texas and California have the highest number of job offerings, however the average pay is substantially lower in Texas when compared to California which has the highest average. New York has the highest salary average after California. Washington comes next in terms of higher salary but there is high variance in salary offered.

We also wanted to understand whether there were any statistically significant differences between the means of independent (unrelated) groups. Therefore we ran a one-way anova test.

ANOVA test results:



ANOVA TEST

H_o: Salary is independent of the states

H_A: Salary depends on the states.

***p-value*:**
0

Reject the null.

For this particular SMART question, our ANOVA test had the null hypothesis that defined that salaries are independent of the states, while the alternate hypothesis defined that salaries depend on the states. After running the test, the p value we obtained was 0, which given that it is less than 0.05, pointed out that there is some dependency between state and estimated salary.

2.Is the salary estimate in the USA correlated with revenue of the company?

Our second smart question focused on understanding further the dynamics between company revenues and salaries.

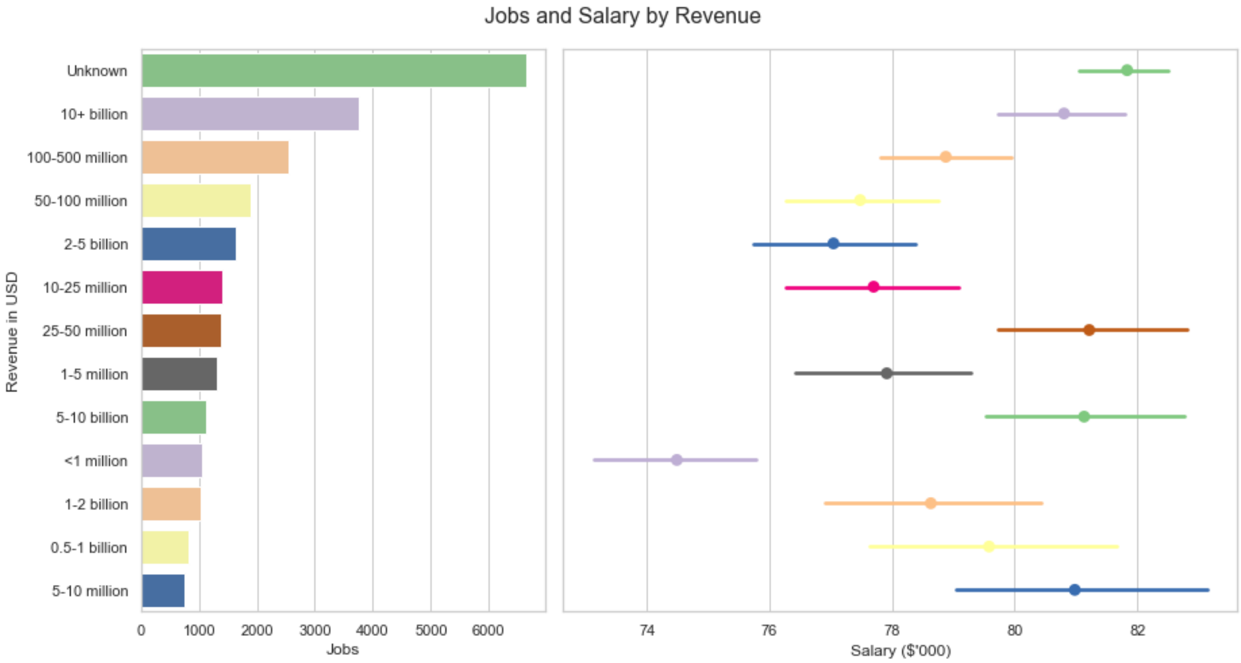
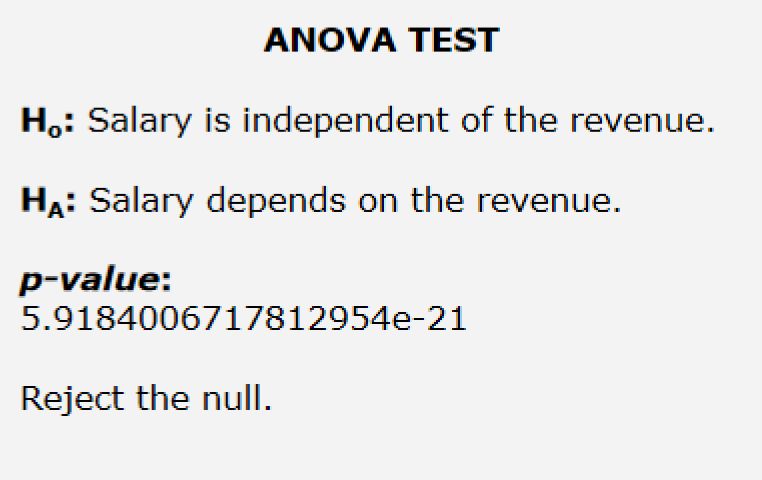


Figure 9. Job openings and estimated salaries by company revenue

Our graph showcases that company's revenues are affecting the average salary estimate. From the graph, we can notice that companies with less than one million revenue are offering a lower salary. But if we look at the remaining revenue categories there is no particular pattern of increase in salary with increase in revenue.

Further, we also did an ANOVA test to check for the dependency between these two variables, where the null hypothesis defined that salaries are independent of the company revenue, while the alternate hypothesis defined that salaries depend on it.

ANOVA test results:



ANOVA TEST

H_o: Salary is independent of the revenue.

H_A: Salary depends on the revenue.

***p-value*:**
5.9184006717812954e-21

Reject the null.

Given that the p value obtained was less than 0.05, we concluded that there is some dependency between revenue and estimated salary.

3a.Does the salary estimate in the US depend on the sector?

Our third and final question had two parts, the first one focused on salary and sectors and the second one on salaries and industry.

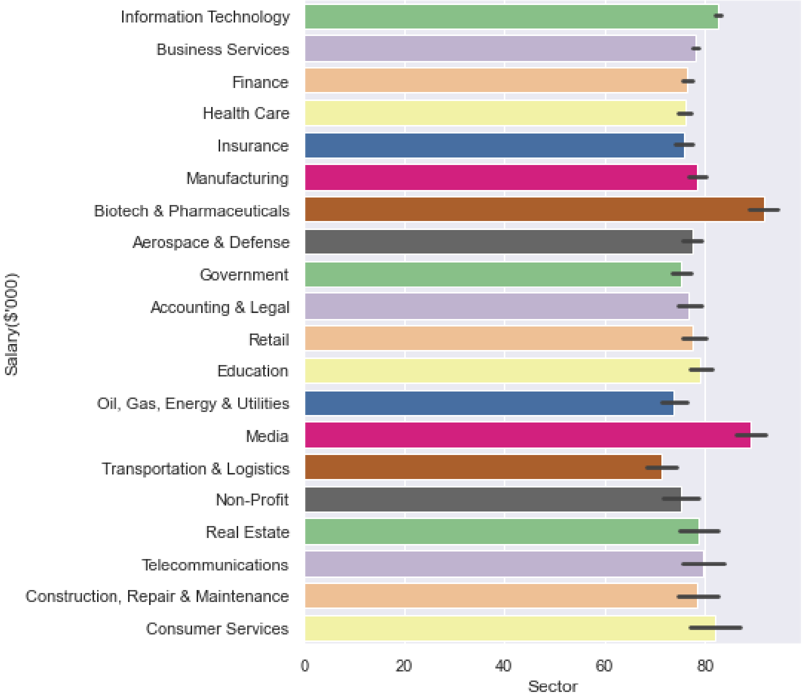
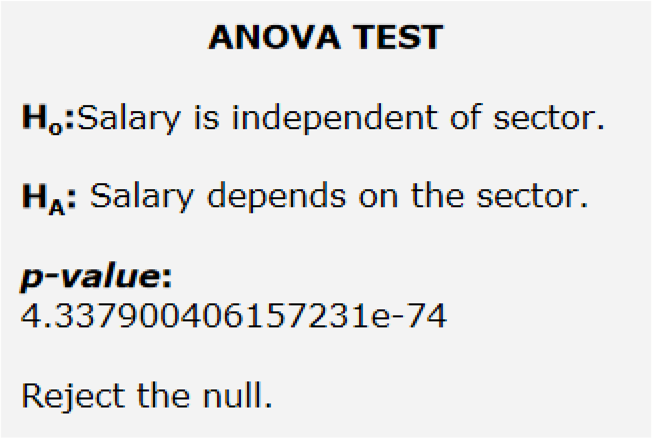


Figure 9. Estimated salaries by sector

Regarding what kind of sectors are offering a competitive salary average, the above bar graph shows the top 20 sectors in terms of job openings. ‘Biotech & Pharmaceuticals’ and Media sectors have the highest salary average followed by Consumer Services and Information technology whereas the Transportation & Logistics sector has the lowest salary average.

ANOVA Test results:



ANOVA TEST

H_o: Salary is independent of sector.

H_A: Salary depends on the sector.

***p-value*:**
4.337900406157231e-74

Reject the null.

Our ANOVA test null hypothesis defined that salaries are independent of the sector and the alternate hypothesis defined that salaries depend on it. The p value obtained was less than 0.05 and hence we concluded that there is some dependency between sector and estimated salary.

3b.Does the salary estimate in the US depend on the industry?

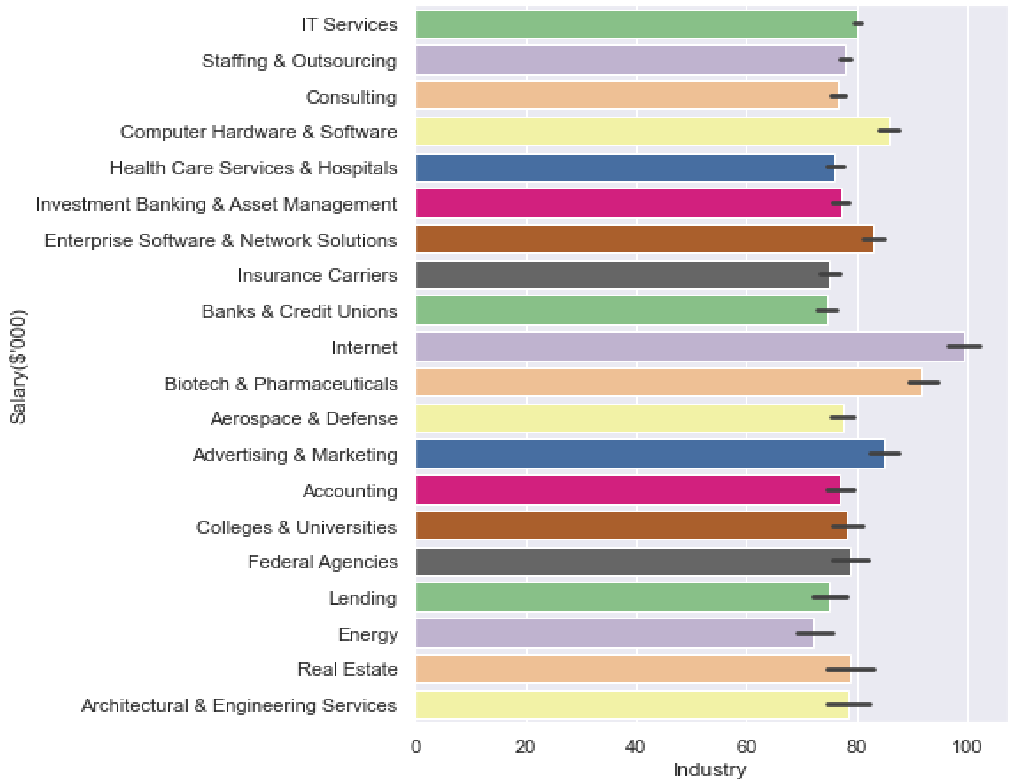
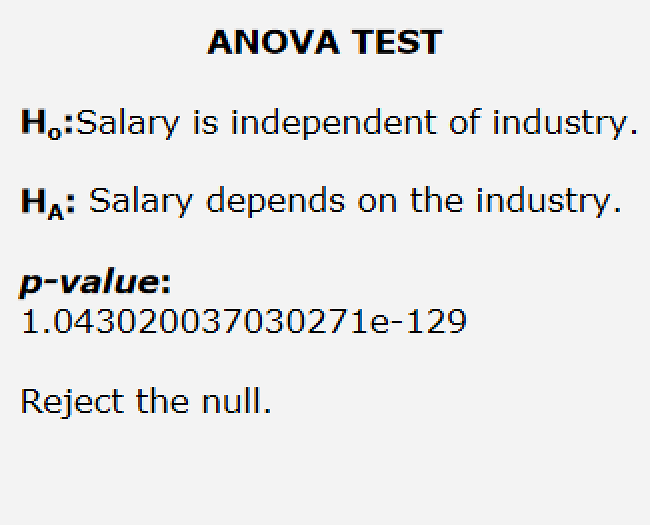


Figure 10. Estimated salaries by industry

Focusing on the relationship between salaries and industries, we were able to observe that Data Science jobs are a part of nearly every industry but some are more eager to pay well based on the needs. The above bar graph shows the average salary for the top 20 industries in terms of job openings. The Internet, Biotech & Pharmaceuticals and Computer Hardware & Software are among the top three industries that have the highest salary average whereas the Energy industry has the lowest salary average. Industries like Advertising & Marketing and Enterprise software are also offering a competitive salary.

ANOVA test results:



ANOVA TEST

H_o: Salary is independent of industry.

H_A: Salary depends on the industry.

***p-value*:**
1.043020037030271e-129

Reject the null.

The final ANOVA test that checked for the dependency between salaries and industries, had a p-value smaller than 0.05. Hence, we concluded that there is some dependency between industry and estimated salary.

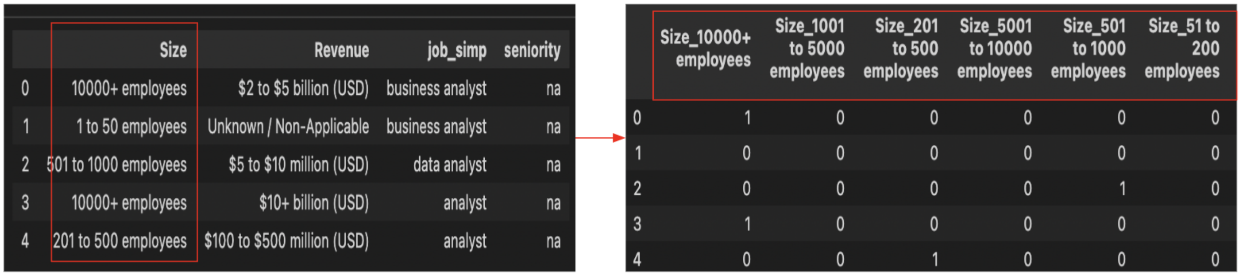
Given that our research also wanted us to identify if we could predict data science salaries, we undertook some modeling techniques which are present in the next section.

Modeling

For modeling, the dataset was divided into a 80:20 split. 80% was for the train data and the remaining 20% was for the test data.

As there were many categorical variables, we did one hot encoding on them to represent categorical variables as binary vectors. We did this as modeling techniques only accept numeric values.

For example, in size columns there were categories like 1000+ employees, 1 to 50 employees which were converted to binary vectors.



	Size	Revenue	job_simp	seniority
0	10000+ employees	$2 to $5 billion (USD)	business analyst	na
1	1 to 50 employees	Unknown / Non-Applicable	business analyst	na
2	501 to 1000 employees	$5 to $10 million (USD)	data analyst	na
3	10000+ employees	$10+ billion (USD)	analyst	na
4	201 to 500 employees	$100 to $500 million (USD)	analyst	na

	Size_10000+ employees	Size_1001 to 5000 employees	Size_201 to 500 employees	Size_5001 to 10000 employees	Size_501 to 1000 employees	Size_51 to 200 employees
0	1	0	0	0	0	0
1	0	0	0	0	0	0
2	0	0	0	0	1	0
3	1	0	0	0	0	0
4	0	0	1	0	0	0

As our problem statement is a regression problem hence we used the following six different modeling techniques to estimate the salary average:

1. Linear regression
2. Decision Tree
3. Random Forest
4. Ensemble Techniques (Bagging Meta Estimator)
5. Ensemble Techniques (ADA Boost)
6. XGBoost

Variables used for our modeling:

- Size
- Revenue
- job_simp (Job Title)
- seniority
- Rating
- Type_ownership
- Sector
- State_Location
- Est_Salary
- Years_Founded
- python
- spark
- aws

- excel
- sql
- sas
- Hadoop

We created skill variables like python, sql etc by extracting them from the job description column.

1. Linear Regression

The first model we used to predict the estimated salary was the basic linear regression model. Linear regression attempts to model the relationship between two or more variables by fitting a linear equation to observed data. Here the estimated salary was the dependent variable and other mentioned variables as independent.

By evaluating the model we got the following result

RMSE	R-Squared	Adj. R-Squared
25.876329	0.220101	0.218488

We got the RMSE value as 25.876329, R-Squared value as 0.22101 and Adjusted R-Square value as 0.218488 (which is too low). Hence we concluded that this model was not the best fit for our dataset.

2. Decision Tree

The second model we decided to use to predict the estimated salary was the decision tree. To make a prediction, the decision tree uses the value of the independent variable to traverse the tree and arrive at a leaf node which represents the predicted value of the dependent variable. To predict the estimated salary the tree would use variables such as state, skill, sector etc to determine the estimated salary.

After doing model evaluation of decision tree we got the following results

RMSE	R-Squared	Adj. R-Squared
26.771088	0.180409	0.178675

We can see Adj. R-Squared value is 0.178675 which is much lower when compared to the previous model. Hence we concluded that this model is also not a good model for our dataset.

3. Random Forest

The third modeling technique we used was Random Forest which is an ensemble decision tree. Random forest uses the value of the independent variable to pass the data point through each decision tree in the forest. The final prediction is determined by aggregating the predictions of all the individual trees such as taking average or the mode.

Below is the model evaluation of Random Forest on our dataset

RMSE	R-Squared	Adj. R-Squared
22.168408	0.438003	0.436814

We got the RMSE value as 22.168408, R-Squared value as 0.438003 and Adj.R-Squared value as 0.436814 which is much better than our previous two models. Hence this model is a good fit for our dataset.

4. Bagging Regressor

The next modeling technique we used is Bagging Regressor which is an Ensemble Technique. A Bagging regressor is an ensemble meta-estimator that fits base regressors on random subsets of the original dataset and then aggregates their individual predictions to form a final prediction. We used the decision Tree Regressor as the base estimator. By evaluating the model we got the following result:

RMSE	R-Squared	Adj. R-Squared
22.472628	0.386786	0.385489

We got the RMSE value as 22.472, R-squared value as 0.386 and the Adjusted R-squared as 0.385 which is better than the normal decision tree score.

5. ADA Boost

The next modeling technique we used is ADA Boost. In short, an AdaBoost regressor is a meta-estimator that begins by fitting a regressor on the original dataset and then fits additional copies of the regressor on the same dataset but where the weights of instances are adjusted according to the error of the current prediction. By evaluating the model we got the following result:

RMSE	R-Squared	Adj. R-Squared
25.528108	0.386786	0.385489

Coming to model evaluation on our dataset, the RMSE is 25.528, R-squared value is 0.386 and the Adjusted R-squared is 0.385 which is same as the Bagging Regressor but this model has a higher RMSE, thus we can say that our earlier model was better.

6. XGBoost

The last modeling technique we used was XGBoost. XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable. By evaluating the model we got the following result:

RMSE	R-Squared	Adj. R-Squared
22.339057	0.429317	0.428110

We got the RMSE value as 22.339, R-squared value as 0.429 and the Adjusted R-squared as 0.428 which is on par with our Random Forest model.

Model Evaluation

Model	RMSE	R-Squared	Adj. R-Squared
Random Forest	22.168408	0.438003	0.436814
XGBoost	22.339057	0.429317	0.428110
Bagging Meta Estimator	22.472628	0.386786	0.385489
ADA Boost	25.528108	0.386786	0.385489
Linear Regression	25.876329	0.220101	0.218488
Decision Tree	26.771088	0.180409	0.178675

Out of all the models we created for this dataset, the Random Forest and XGBoost model performed the best with Adj. R-Squared values of 0.436 and 0.428 respectively. They also had lower RMSE values compared to the other models. The ensemble techniques also performed well.

Limitations and other considerations

Our dataset is from the real world and there are factors that affect the Estimated Salary which are not present in the dataset. For example, the presence of any data regarding the cost of living of the particular State in our dataset would have improved our analysis. Knowing the cost of living would have enabled us to determine whether higher earnings were a result of the high cost of living.

Another way our analyses could have benefitted was by having data regarding the number of years of experience expected by the company for a certain job role. This might have made it clearer to us whether the pay offered rose with years of expertise.

There are some data cleaning and feature engineering techniques that we were unable to perform due to our limited knowledge and time constraints. Additionally, given sufficient amounts of data for each role, we could have performed our analysis on each role rather than all the roles together. This might have increased the amount of insights that we could have generated for each role and we would have been able to create better prediction models.

Conclusions

Based on our analysis, we were able to draw the following conclusions:

- The average salary between all the states in the USA varies from around 40k to 120K.
- California and Texas are the states with the highest number of job offerings for data science professionals with California offering the highest salary average.
- The variables in our SMART questions: state, revenue, industry and sector are all statistically significant.
- The best fitting models are the Random Forest model and XGBoost with Adj R-Squared values of 0.436 and 0.429 respectively.
- Further data is required to obtain more robust conclusions from the relationship of data science salaries.

References

Bleach.W. (2022). Web Scraping Laws. Retrieved from: https://www.termsfeed.com/blog/web-scraping-laws/#:~:text=Even%20though%20it's%20completely%20legal,Personal%20information

Deloitte Access Economics (2018). The future of work: Occupational and education trends in data science in Australia. Retrieved from: https://online.jcu.edu.au/sites/default/files/The%20future%20of%20work%20occupational%20and%20education%20trends%20in%20Australia_Data%20Science2.pdf

Zhen Tee. (July 2022). Salary Prediction in Data Science Field Using Specialized Skills and Job Benefits -A Literature Review. Retrieved from: https://www.researchgate.net/publication/362280362_Salary_Prediction_in_Data_Science_Field_Using_Specialized_Skills_and_Job_Benefits_-A_Literature_Review

Fortune|Education(2022). A hot market for data scientists means starting salaries of $125K and up this year. Retrieved from: https://fortune.com/education/articles/a-hot-market-for-data-scientists-means-starting-salaries-of-125k-and-up-this-year/

Glassdoor (2019). Methodology: Glassdoor Job Market Report. Retrieved from: https://www.glassdoor.com/research/app/uploads/sites/2/2019/04/Methodology-Glassdoor-Job-Market-Report-2-2.pdf

Picklesueat (2022). Data science jobs data. Retrieved from: https://github.com/picklesueat/data_jobs_data

O'Reilly (2016). 2016 Data Science Salary Survey. Retrieved from: https://www.oreilly.com/radar/2016-data-science-salary-survey-results/