Name: Cheng Ho Wang
SID: 32353391

Introduction

Project Title: Data Scientists' Salary

The aim of this report is to give information for the student to make better decisions for their career after graduating. Based on the factors that affect data scientists' salaries, students can decide which country they are going to stay in and what skills they could improve in order to get a better salary.

The dataset is originating from the Glassdoor website. It contains information about the minimum salary, maximum salary, average salary, job description, age of the company in years, etc. With the given information, I have a few aspects on how to give precise and concise information to data science students to make their decision. First of all, I will investigate the topic from a macro view, in terms of countries. Then, I will dive deeper into finding the factors affecting the salary in the company and the employee aspect. After getting the data, I will focus on why these factors are important in determining data scientists' salaries and which one or a few is/are important factors. Therefore, data science students can focus on developing these skills and decide based on these factors. Finally, in terms of different countries, I will compare the salary of data scientists in different countries (after considering their daily expenses in that country). Based on this information, I aim to help students to decide their future career paths.

What are the factors that affect data scientists' salaries?
- How does data scientists' salary vary from:
the country
the company
-projects that the companies are doing
-rating
-size
-history
-industry of the company
the employee themselves
-education
-skills
-experience

Why are these factors important in determining data scientists' salaries? Which one is the most important factor?

Comparing different countries, which country has a higher salary? After compensation for living

Data Wrangling and Checking

Description of the data sources with links if available, the steps in data wrangling (including data cleaning and data transformations), and tools that you used.

Description of the data checking that you performed, errors that you found, your method and justification for how you corrected them, and the tools that you used. A comprehensive checking process is still expected, even though the data set is believed to be clean (i.e., to justify its correctness).

For data wrangling, I am using R to do the whole process. I have got two data sets, therefore, I will try to clean and summarize data and findings for both data.

Dataset1:

For the first data set, the data is mostly organized and cleaned, the thing I focus on here is selecting the column that I need for the data analysis. Another thing is to find ways for summarizing the column "Job description". There are a lot of words in this section, therefore, I may have to use text analysis. However, I am still not sure if the word analysis result is useful to the whole project.

The first thing we have to do is delete a few columns that are not too useful to the project. The first one is a Salary estimate since the average salary, maximum and minimum salary are provided and more useful. The company name is also not useful for us since it combines the company's name and rating in a single column. It is useful for us to remove the rating here. The original version has already separated it for us, therefore, we only have to delete the company name column and rename company_txt into the company name. There are also some other rows that may be not too useful for my analysis, such as competitor, hourly, and employer-provided. Therefore, I will just remove them all.

For the skills column, such as python, and TensorFlow looks a bit messy but this is easier for me to summarize the data later on. Therefore, I will leave it here at this point.

For the data checking of dataset 1, first of all, I have to find if there are any duplicated rows. I have used the unique function to remove all the duplicated rows and the number of rows is the same as the original one.

Besides, I have to check the spelling of words and make sure the category does not have any miss-spelling and cause the result to have a different group but in the same category. I used a unique function to list out all the unique values for each column.

The NA value in this dataset is not represented in a uniform symbol. Some columns use -2 to represent NA and some columns use unknow for NA value. This is not a big concern at this point but I will filter the NA data later on for our better analysis.

Dataset 2:

For dataset 2, there are not many things for me to clean since it is so tidy overall. Every column just contains a particular value I do not have to remove useless data and mutate the dataset. In addition, the dataset also changed all the currency into USD units so that we do not have to deal with the original salary. Therefore, I will start data checking right after.

First of all, I use unique to check if there are any duplicated rows for the whole dataset. It turns out there are 1 duplicated row in this dataset and the unique function has removed it. After that, I used the same process as before which is to find is there any value that has spelling mistakes and different symbols with the same meaning.

After all the data wrangling and checking, the next step is to do some data transformation in order to get the data organized for plotting the graph.

The aim of making these tables are to find the factors that affect the salary of the data scientist. I would like to find the correlation of two variables by using these tables. First of all, I investigate the data from a company aspect.

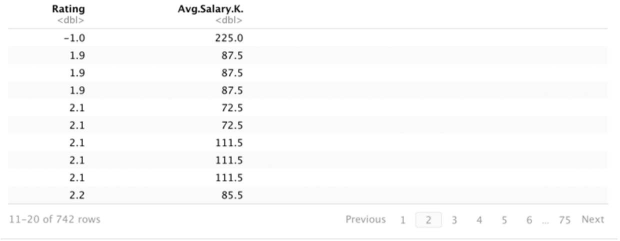
Here is a code for the first table as an example.

```
ratevsal <- deldata1 %>%
  group_by(Rating, Avg.Salary.K.) %>%
  summarize(Rating, Avg.Salary.K.) %>%
  filter(Rating != -1)
```

Here are the tables that I have made:

Rating vs salary

For this table, Rating = -1 means that the rating is unknown. Therefore, I have to use filter(Rating != -1) to remove all the unknown data since it is useless for us.



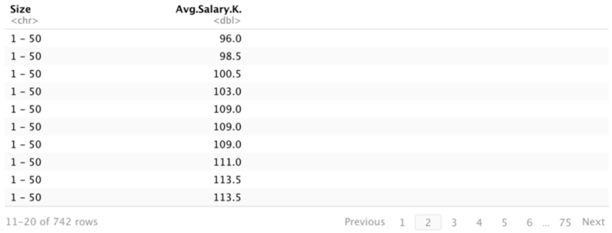
Rating	Avg.Salary.K.
<dbl>	<dbl>
-1.0	225.0
1.9	87.5
1.9	87.5
1.9	87.5
2.1	72.5
2.1	72.5
2.1	111.5
2.1	111.5
2.1	111.5
2.2	85.5

11–20 of 742 rows

Previous
1
2
3
4
5
6 ...
75
Next

Size vs salary

For this table, Rating = -1 means that the rating is unknown. Therefore, I have to use filter(Rating != -1) to remove all the unknown data since it is useless for us.

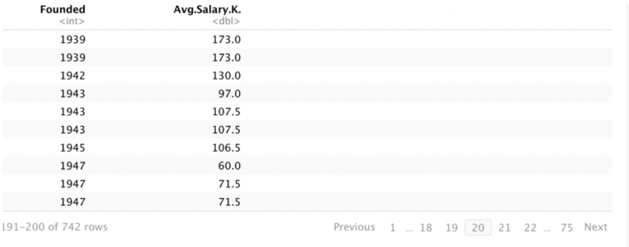


Size	Avg.Salary.K.
<chr>	<dbl>
1 - 50	96.0
1 - 50	98.5
1 - 50	100.5
1 - 50	103.0
1 - 50	109.0
1 - 50	109.0
1 - 50	109.0
1 - 50	111.0
1 - 50	113.5
1 - 50	113.5

11–20 of 742 rows

Previous
1
2
3
4
5
6 ...
75
Next

Founded year vs salary



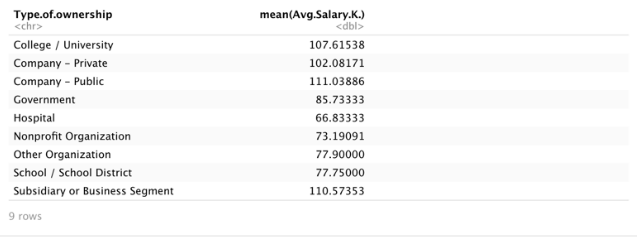
Founded	Avg.Salary.K.
<int>	<dbl>
1939	173.0
1939	173.0
1942	130.0
1943	97.0
1943	107.5
1943	107.5
1945	106.5
1947	60.0
1947	71.5
1947	71.5

191–200 of 742 rows

Previous
1 ...
18
19
20
21
22 ...
75
Next

For the factors below, it is hard to do the correlation graph. Therefore, I have summarized the average salary for each group and see is there any effect on the data scientists' salaries.

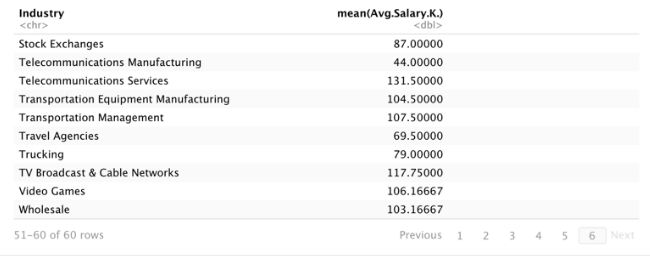
Type of Ownership



Type.of.ownership <chr>	mean(Avg.Salary.K.) <dbl>
College / University	107.61538
Company - Private	102.08171
Company - Public	111.03886
Government	85.73333
Hospital	66.83333
Nonprofit Organization	73.19091
Other Organization	77.90000
School / School District	77.75000
Subsidiary or Business Segment	110.57353

9 rows

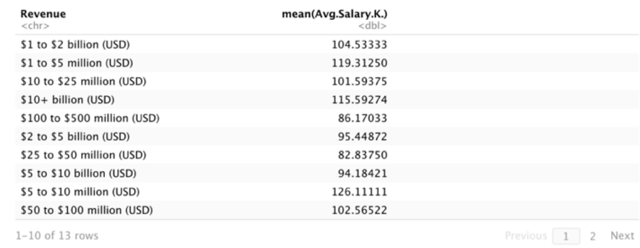
Industry vs salary



Industry <chr>	mean(Avg.Salary.K.) <dbl>
Stock Exchanges	87.00000
Telecommunications Manufacturing	44.00000
Telecommunications Services	131.50000
Transportation Equipment Manufacturing	104.50000
Transportation Management	107.50000
Travel Agencies	69.50000
Trucking	79.00000
TV Broadcast & Cable Networks	117.75000
Video Games	106.16667
Wholesale	103.16667

51-60 of 60 rows Previous 1 2 3 4 5 6 Next

Revenue vs salary

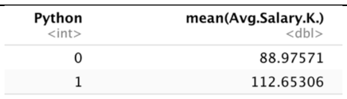
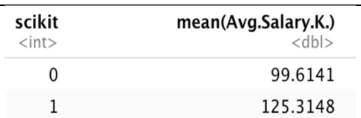
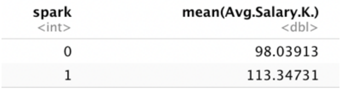
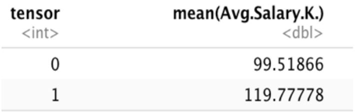


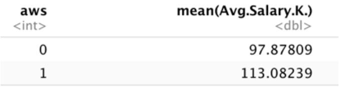
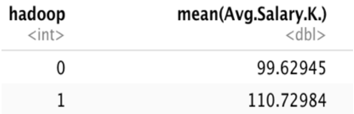
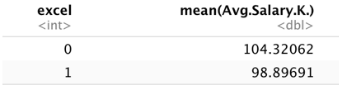
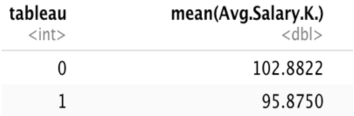
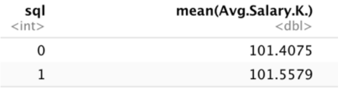
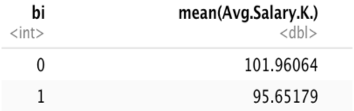
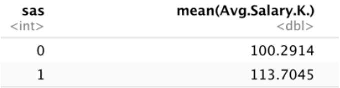
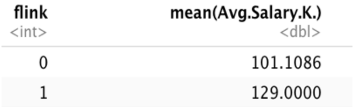
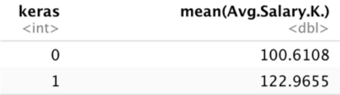
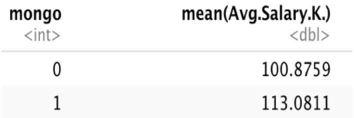
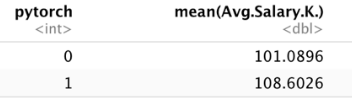
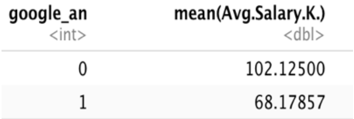
Revenue <chr>	mean(Avg.Salary.K.) <dbl>
$1 to $2 billion (USD)	104.53333
$1 to $5 million (USD)	119.31250
$10 to $25 million (USD)	101.59375
$10+ billion (USD)	115.59274
$100 to $500 million (USD)	86.17033
$2 to $5 billion (USD)	95.44872
$25 to $50 million (USD)	82.83750
$5 to $10 billion (USD)	94.18421
$5 to $10 million (USD)	126.11111
$50 to $100 million (USD)	102.56522

1-10 of 13 rows Previous 1 2 Next

Then, we will start to investigate a personal aspect. This may include the skillset and seniority of the data scientist.

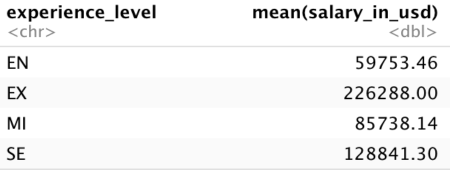
First of all, I grouped by whether the people know a particular programming language or not. Then, I calculated the mean of their salary. This is the method that gives us a big picture of generally whether learning a new language helps data scientists to get a higher salary. However, we have to keep in mind that the data is not fully presenting the fact because there are other factors affecting the result and some other outlier is affecting the mean. Therefore, it is only a reference and this method is not a perfect model for determining which skillset is affecting the salary.

Skillset	Data	Skillset	Data
Python	 Python <int> mean(Avg.Salary.K.) <dbl> 0 88.97571 1 112.65306	Scikit	 scikit <int> mean(Avg.Salary.K.) <dbl> 0 99.6141 1 125.3148
Spark	 spark <int> mean(Avg.Salary.K.) <dbl> 0 98.03913 1 113.34731	Tensor	 tensor <int> mean(Avg.Salary.K.) <dbl> 0 99.51866 1 119.77778

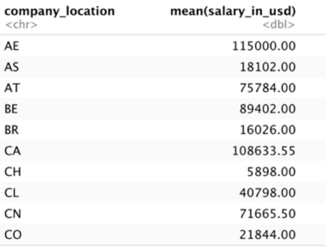
Aws	 aws mean(Avg.Salary.K.) <int> <dbl> 0 97.87809 1 113.08239	Hadoop	 hadoop mean(Avg.Salary.K.) <int> <dbl> 0 99.62945 1 110.72984
Excel	 excel mean(Avg.Salary.K.) <int> <dbl> 0 104.32062 1 98.89691	Tableau	 tableau mean(Avg.Salary.K.) <int> <dbl> 0 102.8822 1 95.8750
SQL	 sql mean(Avg.Salary.K.) <int> <dbl> 0 101.4075 1 101.5579	Bi	 bi mean(Avg.Salary.K.) <int> <dbl> 0 101.96064 1 95.65179
Sas	 sas mean(Avg.Salary.K.) <int> <dbl> 0 100.2914 1 113.7045	Flink	 flink mean(Avg.Salary.K.) <int> <dbl> 0 101.1086 1 129.0000
Keras	 keras mean(Avg.Salary.K.) <int> <dbl> 0 100.6108 1 122.9655	Mongo	 mongo mean(Avg.Salary.K.) <int> <dbl> 0 100.8759 1 113.0811
Pytorch	 pytorch mean(Avg.Salary.K.) <int> <dbl> 0 101.0896 1 108.6026	Google Analytics	 google_an mean(Avg.Salary.K.) <int> <dbl> 0 102.12500 1 68.17857

From the above data, we can see that some of the tools, such as python, Keras, and mongo make the salary higher. On the other hand, those skills such as excel and google analytics are the basic skills that most employees know. Then, these skills do not affect the salary so much. As I interpret it, those skills that are basic, such as excel, python, and SQL must be understood by the employee which is the basic requirement of a data scientist. Some advanced tools such as TensorFlow, Keras and Pytorch are some tools for machine learning. Knowing this kind of tool helps data scientist to boost their salaries.

Then, we started to determine how seniority affects the salary. The method of data transformation is the same as before. I have grouped by the seniority and calculated the average salary for each of the seniority. Also using the same method as the country.

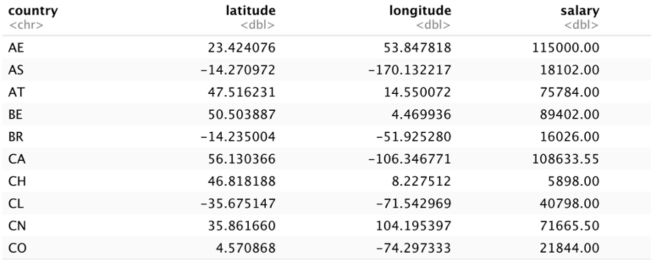


```
experience_level   mean(salary_in_usd)
<chr>                            <dbl>
EN                            59753.46
EX                           226288.00
MI                            85738.14
SE                           128841.30
```



```
company_location   mean(salary_in_usd)
<chr>                            <dbl>
AE                           115000.00
AS                            18102.00
AT                            75784.00
BE                            89402.00
BR                            16026.00
CA                           108633.55
CH                             5898.00
CL                            40798.00
CN                            71665.50
CO                            21844.00
```

For the country data, I need to plot the data on a map. I must get the latitude and longitude data for each country. The company location in data2 is the international standard symbol for representing the country. Therefore, we can just use the data without any modification. Then, I found a data set that includes symbols, latitude, and longitude of each country. Since the symbol in the country dataset is the same as those in company_location, therefore, I just rename the company_location column to country and then use left join to join two tables. By the way, data2 is the left table since the left join could eliminate those rows that are not useful in the country dataset automatically. After that, I group by country, latitude and longitude. Then, I used summarize for calculating the mean salary of each country. Here is the result:



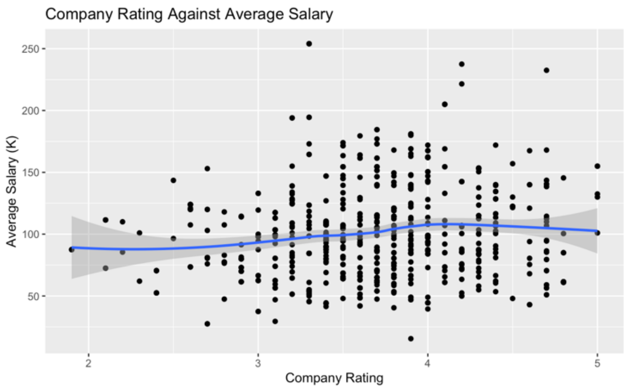
country	latitude	longitude	salary
<chr>	<dbl>	<dbl>	<dbl>
AE	23.424076	53.847818	115000.00
AS	-14.270972	-170.132217	18102.00
AT	47.516231	14.550072	75784.00
BE	50.503887	4.469936	89402.00
BR	-14.235004	-51.925280	16026.00
CA	56.130366	-106.346771	108633.55
CH	46.818188	8.227512	5898.00
CL	-35.675147	-71.542969	40798.00
CN	35.861660	104.195397	71665.50
CO	4.570868	-74.297333	21844.00

Data Exploration

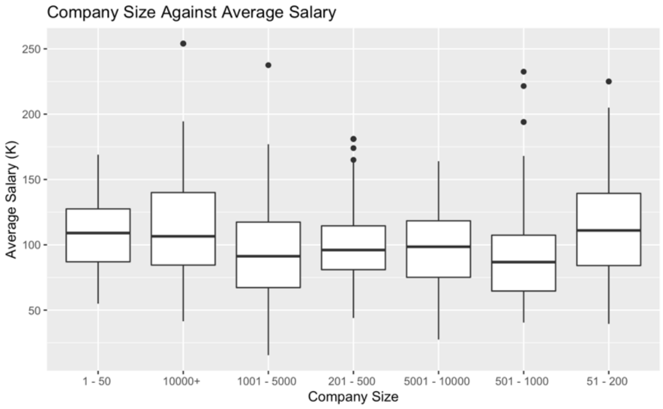
After processing the data wrangling and cleaning, we can start plotting graphs for visualization. First of all, I would like to find the correlation of different factors with the average salary. The tool for making graphs is by RStudio, ggplot2. Before making the graphs, I also have to adjust the data for better visualization. For example, to make a better grouping and eliminate all the NA values.

To start with, I have plotted a graph which is showing the relationship between the company's rating with the data scientists' salaries.

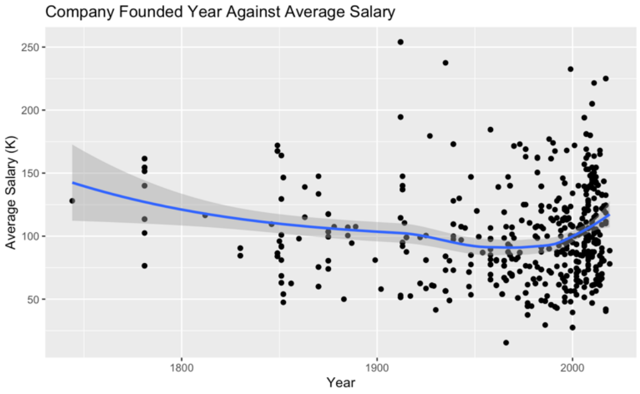
In the first graph, we can see the relationship between a company's rating and salary. The points are quite dispersed and we cannot see a clear relationship. However, there are more points clustered between ratings 3 to 4 and the highest salary appears at above 3 of the company's rating. Comparing those points ranging from 1 to 3, those data scientists' salary in the rating range of 3 to 4 is generally higher. Besides, the line is the regression line, which is plotted by geom_smooth, that tries to fit every data. At the point of rating 4, a slight peak exists. It means the average salary is slightly higher than others. Therefore, We can summarize that companies with an average rating generally provide a relatively stable salary or a bit higher than those companies having a lower rating.



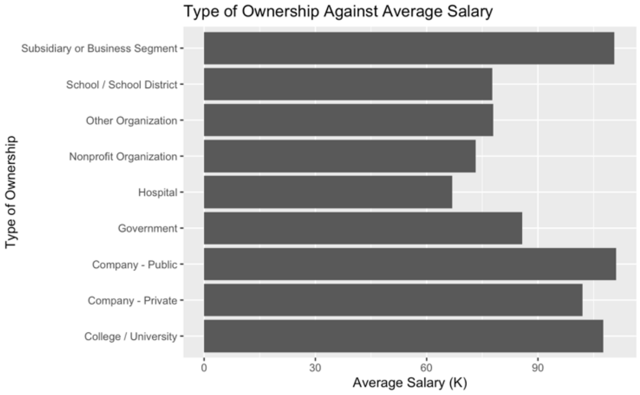
In the graph below, I have used the geom_boxplot. It shows the average, lower quartile, and upper quartile of the salary. In general, the average salary is pretty much the same. It also does not show much difference in the upper quartile and lower quartile salaries. Besides, it does not show a clear trend. However, the difference in the range of the salary is quite obvious. For a big company, the difference between the highest and the lowest salary is above 200k. The difference is slowly shrinking while the company size becomes smaller. Therefore, we find out the range of data scientist salary is wider in a big company. If one aims to get the highest salary way above the industry average, one will have a higher chance to achieve it if one works in a big company.



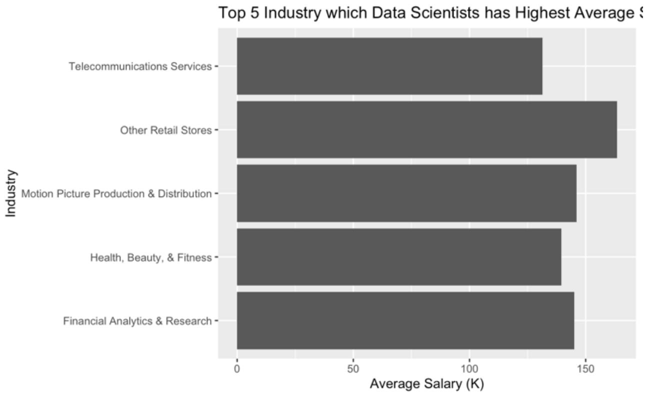
In the following graph, I aim to investigate how company history affects data scientists' salaries. Most of the companies are founded after the 1950s. Also, the points are more clustered between 1980 to 2000s. This only shows that there are more companies that started their business during this period. The salary is average all over the period and a little bit higher than average for those companies that started after 2000 since more points cluster around 100-150K.



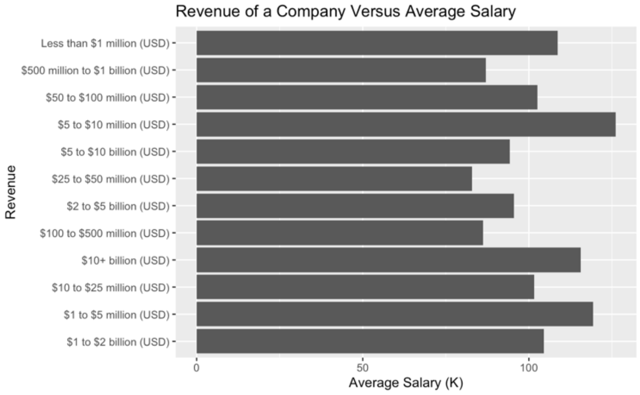
For the factor "type of ownership", working in a business segment and public companies got more pay in the big picture. We could interpret that data science is helping some companies make more money. It is fair enough to get higher pay if the data scientists solve a business problem. In another aspect, working in college/university still get a lot of money because data scientist is contributing their value to research.



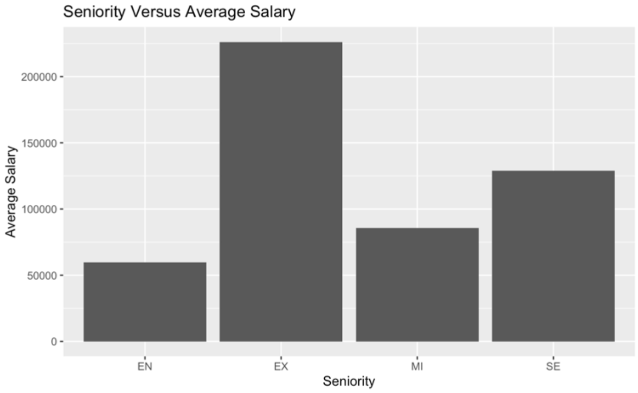
Then, we are trying to look into the industry of the company. Since there are too many categories for the industry, I have filtered out the top 5 industries in which data scientists earn the most. It is no surprise that the highest average salary belongs to retail and Finance which always emphasize their sales.



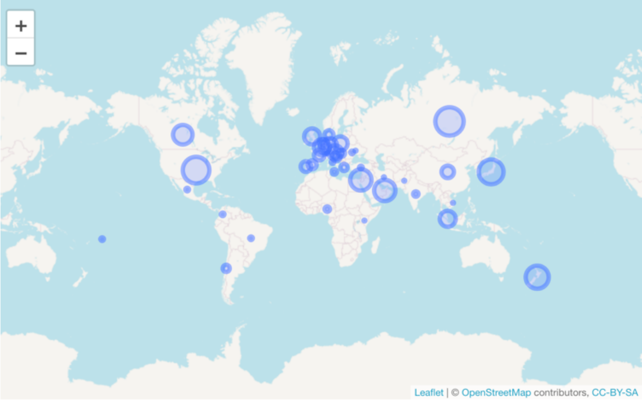
After we know companies that focus on sales and income can be able to pay more for data scientists' salaries, we can try to find the proof from the graph below. Surprisingly, the companies that earn the most are not those companies that pay the most for data scientists. That is interesting that companies earning $5M to $10M pay the most for the data scientists. We may interpret that some of the data scientists work in start-up companies that are growing. Therefore, they contribute more and get more shares from that. Another thing maybe companies earning the most is world-known companies which they have set a standard price for hiring each employee. That may be the reason why companies earning the most may not always pay data scientists for more.



Furthermore, we can start to investigate on a personal level. We have discussed the skillsets in the previous sections. And now we can slightly talk about how seniority affects the salary. The graph below represents our common knowledge which is higher seniority gets paid more. This is just a reference for us as a student. We could see our future through this graph. For a fresh graduate, we generally can get around $60000 dollars for a year. If we want to earn more until more than $200k per year. We have to aim for being an executive.



At last, I have plotted the average salary of each country on the map by using leaflet. The radius of the circle is proportional to the average salary of the data scientists. The bigger the circle, the higher salary of the data scientists. From the map, we can spot a few larger circle which is the USA, Russia, Japan, and Canada. Therefore, students may have a reference to this data and choose where they could move for a better-paying job.



So far, we have investigated how different factors affect data scientists' salaries. In general, I will say if a data scientist wants to get a higher salary. The first thing they have to consider is to improve their skills. Since we have learned the basic programming skills from our degree. However, a data scientist can explore more different tools such as Keras and PyTorch for advanced machine learning. These tools also enhance data scientists' knowledge of Artificial intelligence. These make the data scientist more competitive. Another aspect that we have investigated is the company. However, the factors do not affect much the salary. This point of having this information is to give students a reference for which company might have a better work environment. Besides, they could choose to get into some business-related industry for getting a better salary. Still, students have to think about their future goals if the industry matches their expectations. Besides, the data shows that some countries, such as Russia, the USA, and Canada pay more for data scientists. This also gives students a reference for whether they want to stay in Australia after graduation.

Conclusion

From the data exploration, I have got some views on both company and personal aspects. From the company perspective, I have investigated how rating, size, type of ownership, industry, and revenue affects data scientists' salaries. Companies which have 4 marks are paying higher for the salary. Data scientists can get more pay and a higher potential to get a high-level position while they are working in the large companies working in the business or financial industry. In terms of companies' revenue, companies earning $5M to $10M pay the most for the data scientists which is a surprising result. It means companies earning more are not those who pay the most for their employees.

On a personal level, data science students can keep working on improving their skills by learning advanced machine learning tools. This can boost their salary. Another thing is the seniority, we have got a reference for students about how much they can get at a particular level. Their salary could be 4 fold more than entry-level if they can become the executive of the company.

Besides, in a global view, we could also compare the data scientists' average salaries in different countries. Those companies in developed countries such as the USA, Canada, and Russia could pay more for data scientists. If the student is ambitious to get a higher-level job, the success rate in these countries is higher.

Reflection

Throughout the whole project, I have tried to answer most of the questions. From a company aspect, I could summarize and plot the data for a few factors, such as company size, history, and revenue. The only factor that I have not investigated is the projects that the companies are doing since I realized that I do not have any data related to that. However, this is still an important factor that students have to investigate when they are finding their job.

From a personal aspect, I have investigated how skillsets and experience affect salary. The only thing that I did not investigate is education. It is because the education data is mostly NA and the existing data is all postgraduate level. There is no point to analyse the data so I did not investigate it.

In a global view, we could compare the salary between countries and get the expected result. The map that I finally generated also match my expectation. One thing is I did not calculate the salary after being compensated for living. In dataset 2, it has converted all the salaries into USD after considering the conversion rate. Comparing salaries of different countries in the same currency will be easier for us to visualize the data. Besides, I do not have enough data for calculating the cost of living for each country. Therefore, I do not answer this part of the question.

Bibliography

Nikhil,B. (2021) *Data scientist salary*. Kaggle. https://www.kaggle.com/nikhilbhathi/data-scientist-salary-us-glassdoor?select=data_cleaned_2021.csv

Saurabh, S. (2021) *Data Science Jobs Salaries Dataset*. Kaggle. https://www.kaggle.com/datasets/saurabhshahane/data-science-jobs-salaries/discussion

Situ, W., Zheng, X.Y., Daneshmand, M. (2017) Predicting the Probability and Salary to Get Data Science Job in Top Companies*. 2017 Industrial and Systems Engineering Research Conference.* https://www.proquest.com/openview/dca60d29b2cc1067fe5dd5c6cf8ed168/1?pq-origsite=gscholar&cbl=51908

RDocumentation. Ggplot2. https://www.rdocumentation.org/packages/ggplot2/versions/3.3.5

Countries.csv. https://developers.google.com/public-data/docs/canonical/countries_csv