Final Project Report

SDSC5002 Exploratory Data Anlys & Visua

1. Basic information of project

1.1 Project name

Data Analyst Jobs Visualization

1.2 Team name

Edelweiss

1.3 Team member

Yuhong Fang, Yifan Wang, Ran Wang, Shichen Wang, Tingxuan Chu, Zhe Wang

2. Project background

During the pneumonia epidemic in COVID-19, many people lost their jobs, and employment has always been a topic of concern to many people all over the world.

Data analyst is a hot employment direction in recent years. With the advent of the era of big data and artificial intelligence, companies pay more and more attention to the mining and analysis of data resources. By analyzing the data of market environment, competitors and their own operating conditions, companies can adjust their business direction in time, thus being invincible in the market competition.

Therefore, data analysis has become the direction that many people hope to change careers. We hope that this project will help more people find jobs with higher salaries and better company ratings according to their skills, or help those who want to change careers as data analysts to know what skills need to be improved to find a job as a data analyst.

3. Data cleaning and exploration

3.1 Data acquisition

The dataset was retrieved from Kaggle and and contained 15 attributes related to Data Analyst Jobs and 2257 sample data.

3.2 Parameters

Attribute	Data type	Description
Easy Apply	bool	whether the job is easy to apply
Competitors	string	the competitors of the company
Founded	string	the establish time of the company
sector	string	area the company belongs to
Rating	float	the rating of the company
Headquarters	string	location of the headquarters of the company
Size	string	The quantity of the employee
Type of ownership	string	type of the company ownership
Revenue	string	revenue of the company
Salary Estimate	float	estimated salary of this job
Company name	string	name of the company
Location_city	string	located city of the company
Location_state	string	located state of the company

3.3 Data filtering and preprocessing

	Missing Values	% of Total Values
Easy Apply	2173	96.4
Competitors	1732	76.9
Founded	660	29.3
Industry	353	15.7
Sector	353	15.7
Rating	272	12.1
Headquarters	172	7.6
Size	163	7.2
Type of ownership	163	7.2
Revenue	163	7.2
Salary Estimate	1	0.0
Company Name	1	0.0

The "Salary Estimate" is an interval value, so we take the average value as the new value of salary according to '"lower bound'" and "upper bound".

We split "Location" into "location_city" and "location_state", which is convenient for drawing a map on tableau.

"Easy Apply" and "Competitors" have too many missing value, and the amount of data is too small to give credible analysis. Therefore, we do not analyze these two variables in this project.

There are too many classifications of "Industry" and "city", which are too detailed, so we choose "sector" and "state" to study the influence of industry and geographical location.

"Company Name" is not representative and there are too many companies, so we don't use it as a predictive variable to study the relationship between "company name" and "salary".

Because the data of "job title" has already been processed, it is a filtered job about data analysts, so the number of data analysts exceeds 70%. There are too few Data scientist, data engineer and business analyst to support data analysis.

Because all the information except "job title", "job description", "salary" is about the company, and the same company will publish multiple jobs, which may include both high-paid and low-paid jobs, which will bring a huge wage gap, but it cannot be reflected in other variable. Therefore, we can't know what kind of company can provide higher salary through a regression model. Therefore, we studied the relationship between each variable and salary.

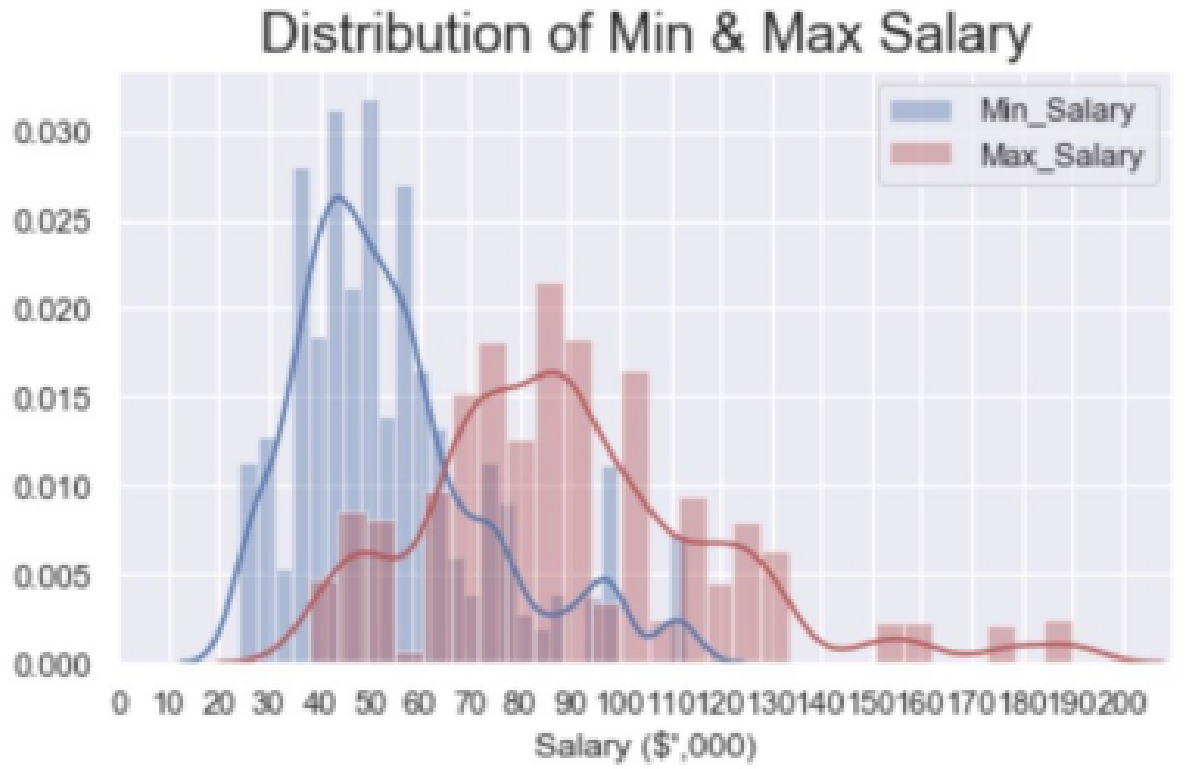
4. Explore Analysis

4.1 Problem modeling

Through the analysis of these 15 variables in the data set, the most important factor for most people when looking for a job should be income level, so we take income as a response variable to find the relationship between other variables and income level, so as to give relevant suggestions on how to find a high-paying data analyst job.

4.2 explore analysis for parameters

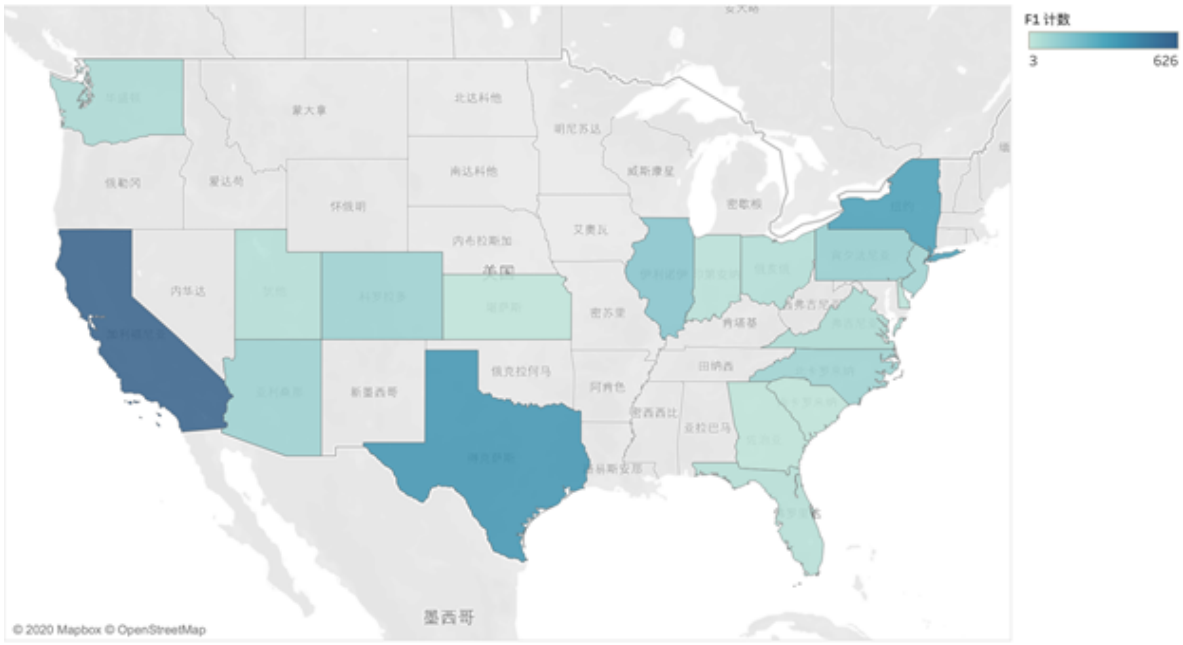
Salary:



This plot represents the distributions of lower bound and upper bound of salaries. The median of min_salary is about 45, and the median of max_salary is about 85. The min_salary is more concentrated and centered in a small range. The max_salary is more spread and flat. There also have some extremely high max_salary that is way larger than others.

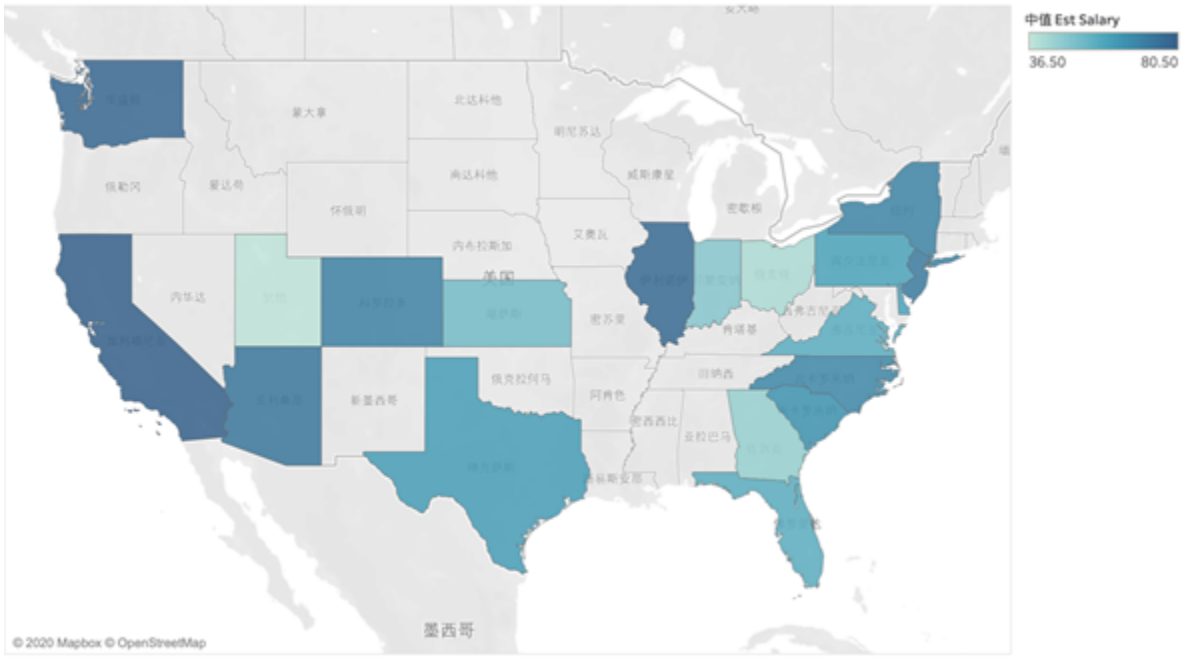
State:

<State vs. Count of Jobs>



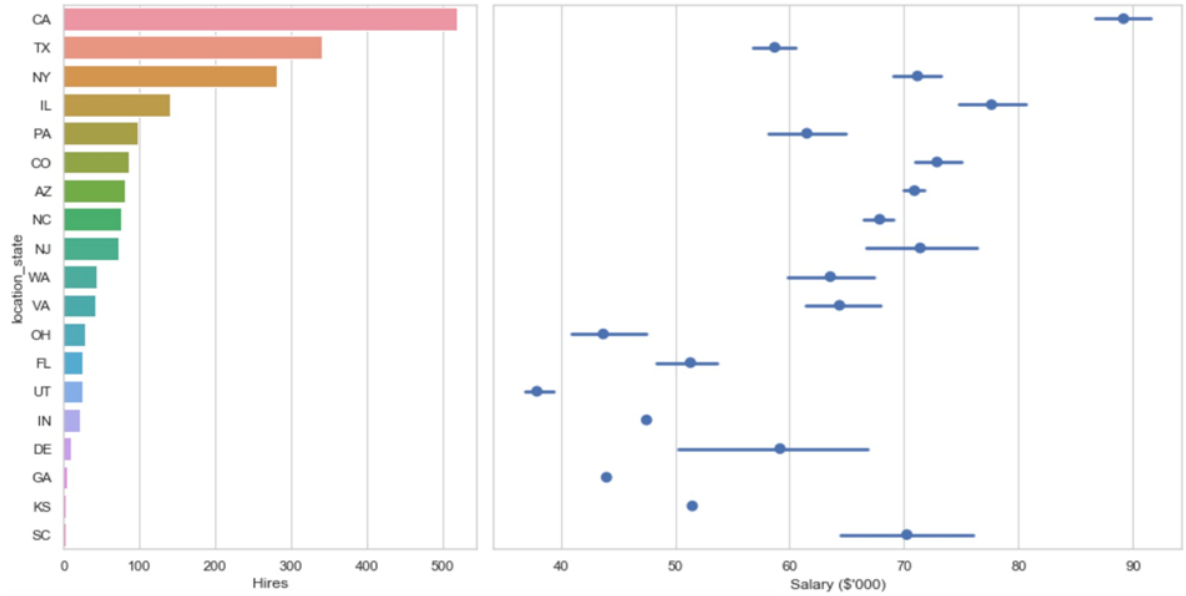
基于 经度(自动生成) 和 纬度(自动生成) 的地图。 颜色显示 F1 计数。 为 Location State 显示了详细信息。

<State vs. Median of Estimated Salary>



基于 经度(自动生成) 和 纬度(自动生成) 的地图。 颜色显示 Est Salary 中位数。 为 Location State 显示了详细信息。

From this map, we can know that the region provides the most jobs is California. The areas with higher salary levels are also concentrated in the western and eastern US regions. In Seattle, California, Chicago, New York have a good foundation for the IT financial industry. So they are gathered here.



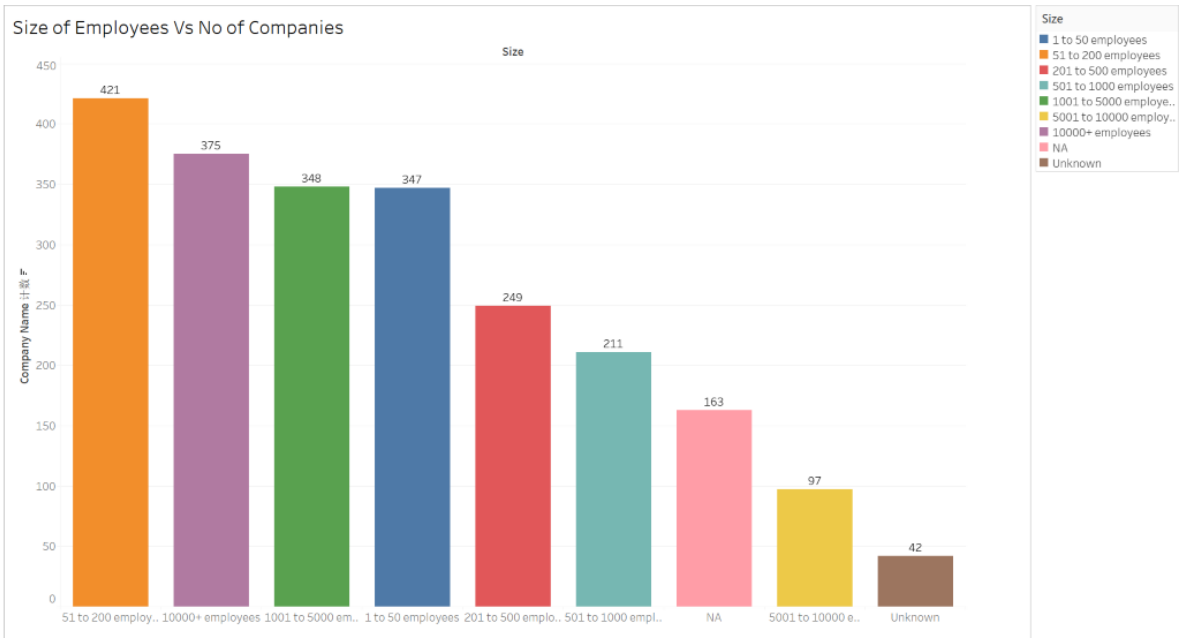
From this chart, California is the largest Silicon Valley in the United States. So it can provide more IT and data analysis positions.

Austin has a well-known Texas Science and Industry Park, and all major US IC manufacturers have established chip design centers in the city. So this area needs to hire more data analysts. But we can see that the salary level in the chip design industry is not high.

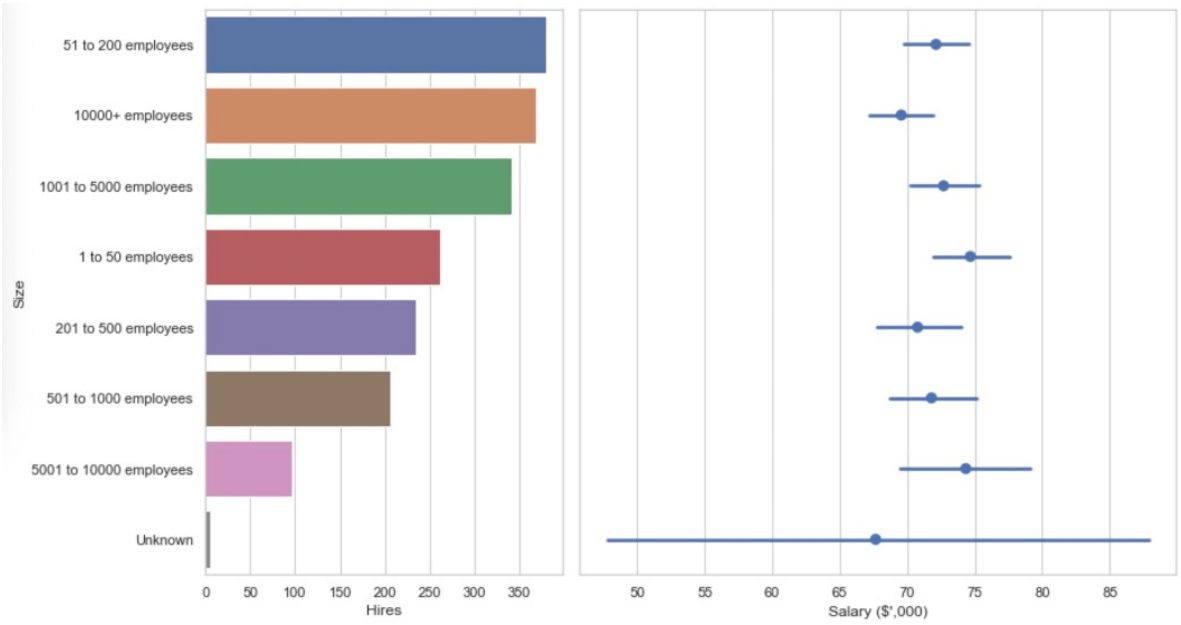
We can also find that Delaware is a very interesting region. There are relatively few people employed in this state. The reason is that the state implements preferential tax policies, so most registered companies only have offices in Delaware symbolically, thus avoiding higher taxes.

California has many jobs and the highest salary. A large number of high-quality jobs are concentrated around San Francisco and Los Angeles.

Size:

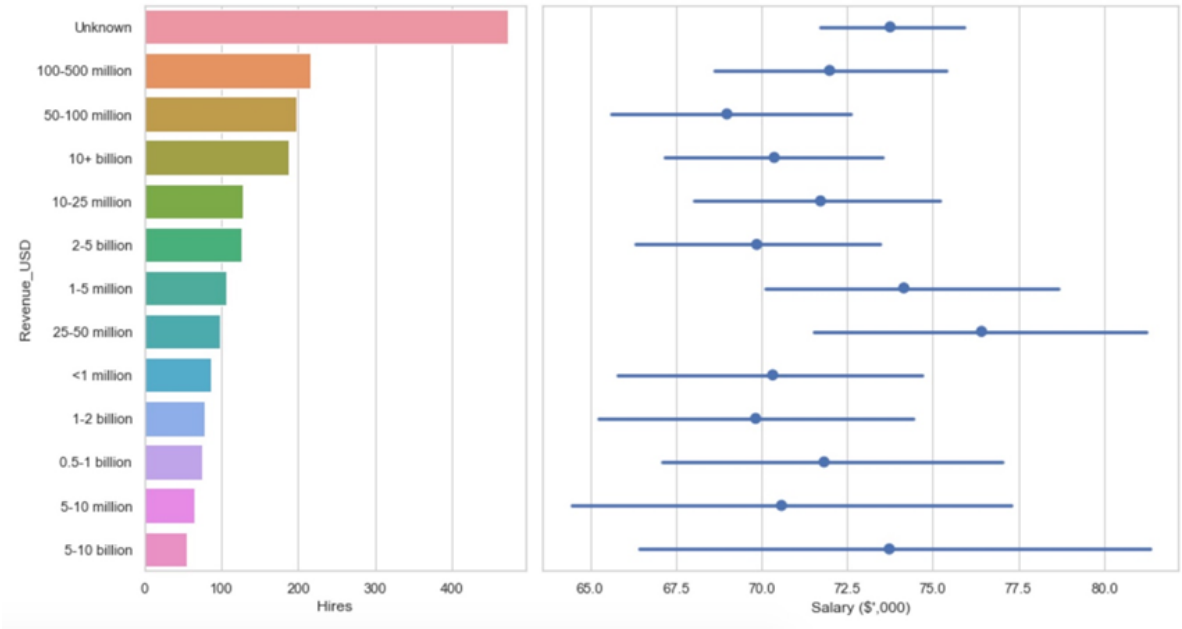


There are 421 companies with the 51 to 200 employees are the most common companies that are hiring data analyst. Companies with small size is hiring data analyst and not just big companies do. Some people may think that company with small size may do not need a data analyst because they do not have enough data to analize, but the data tells us that they do need. By comparing the ratio of large companies and small companies in the real life, the number of small companies are much more than large companies, so actually large companies have a higher percentage to hire data analysts than small companies.



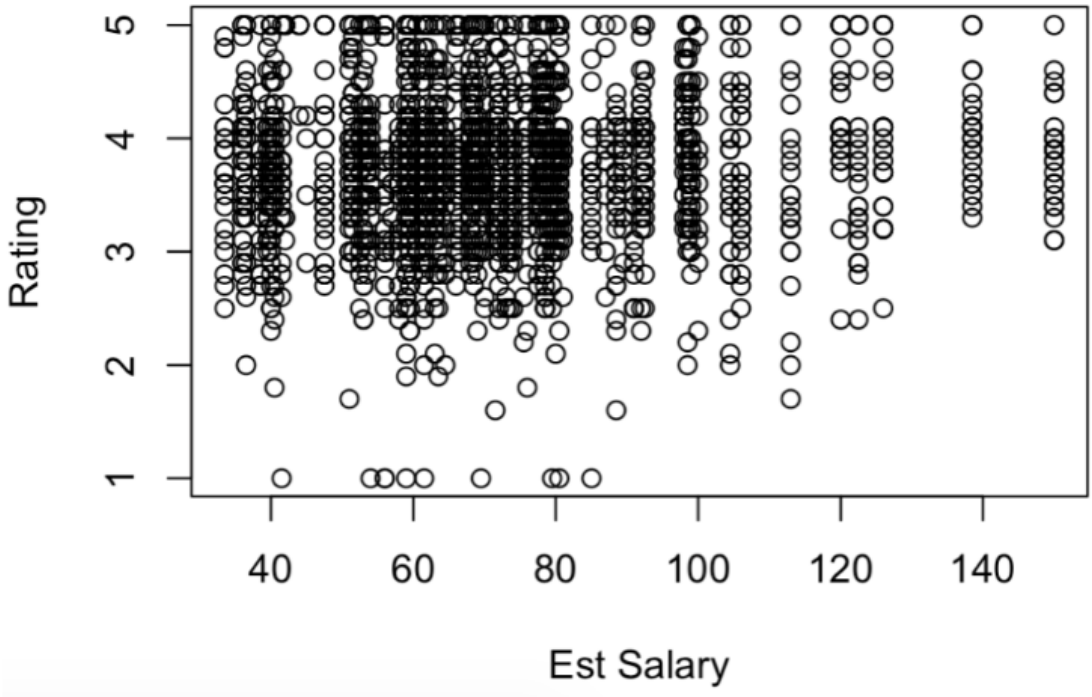
By our common sense, big companies usually provide more higher salary because they have larger scale and also have stronger ability to earn profits than others. So we keep this opinion until we observe the plot that shows the relationship between salary and size and revenue. All sizes companies provide similar salary, which means salary has no relationship with the size of a company, they are independent to each other.

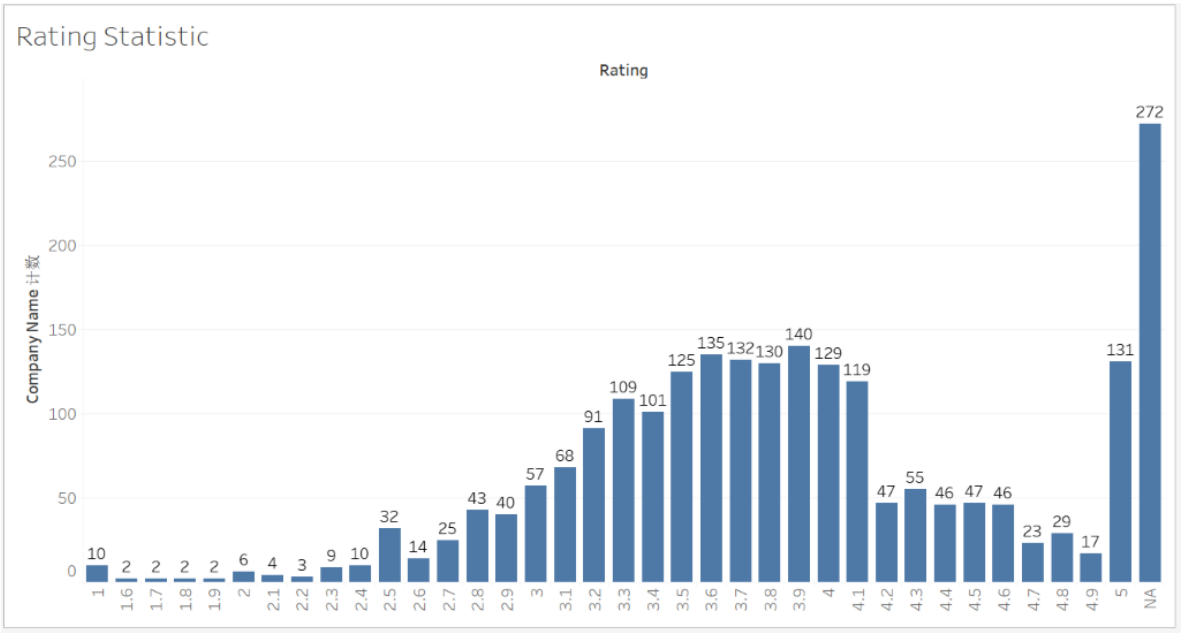
Revenue:



Now we talking about the revenue. We expect to see the companies with high revenue will provide a data analyst job with a higher salary because the company earn more profit because of the employees, so they should pay more to their emplyees. But actually, by observing the plot, we can see that there is no direct positive correlation between revenue and salary, that is to say, companies with high revenue will not provide higher salaries to data analysts. The company with the revenue between 20 and 50 million dollars provide the highest average salary job.

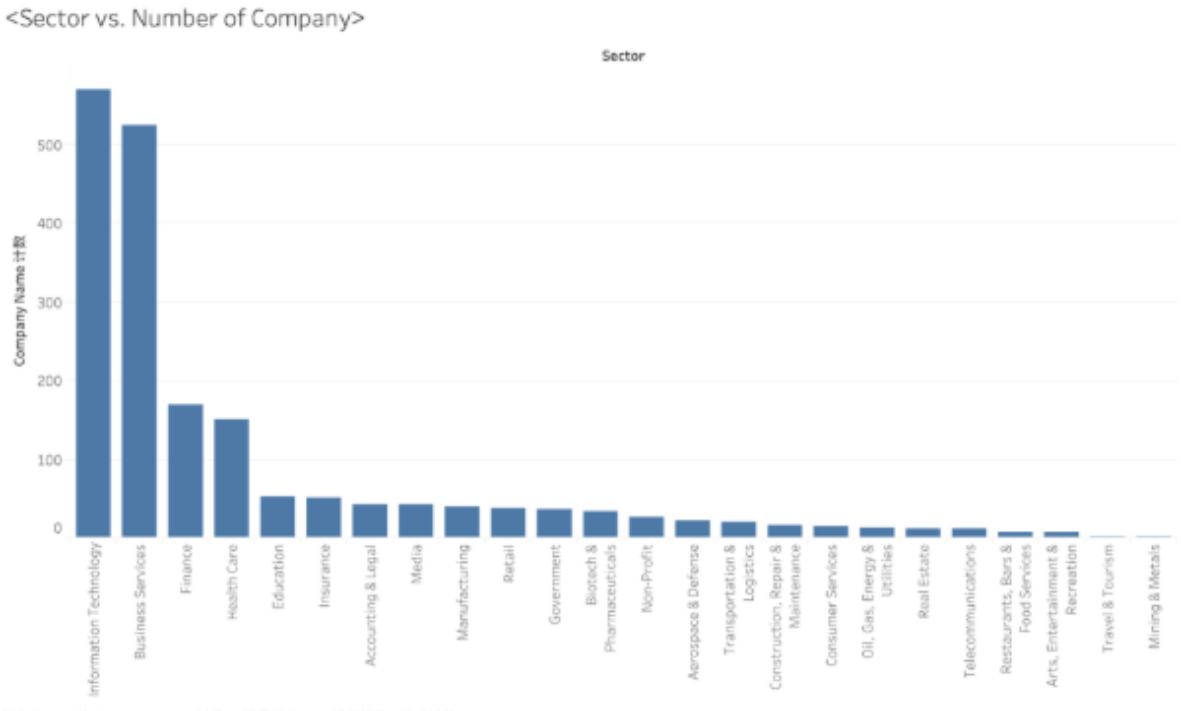
Rating:



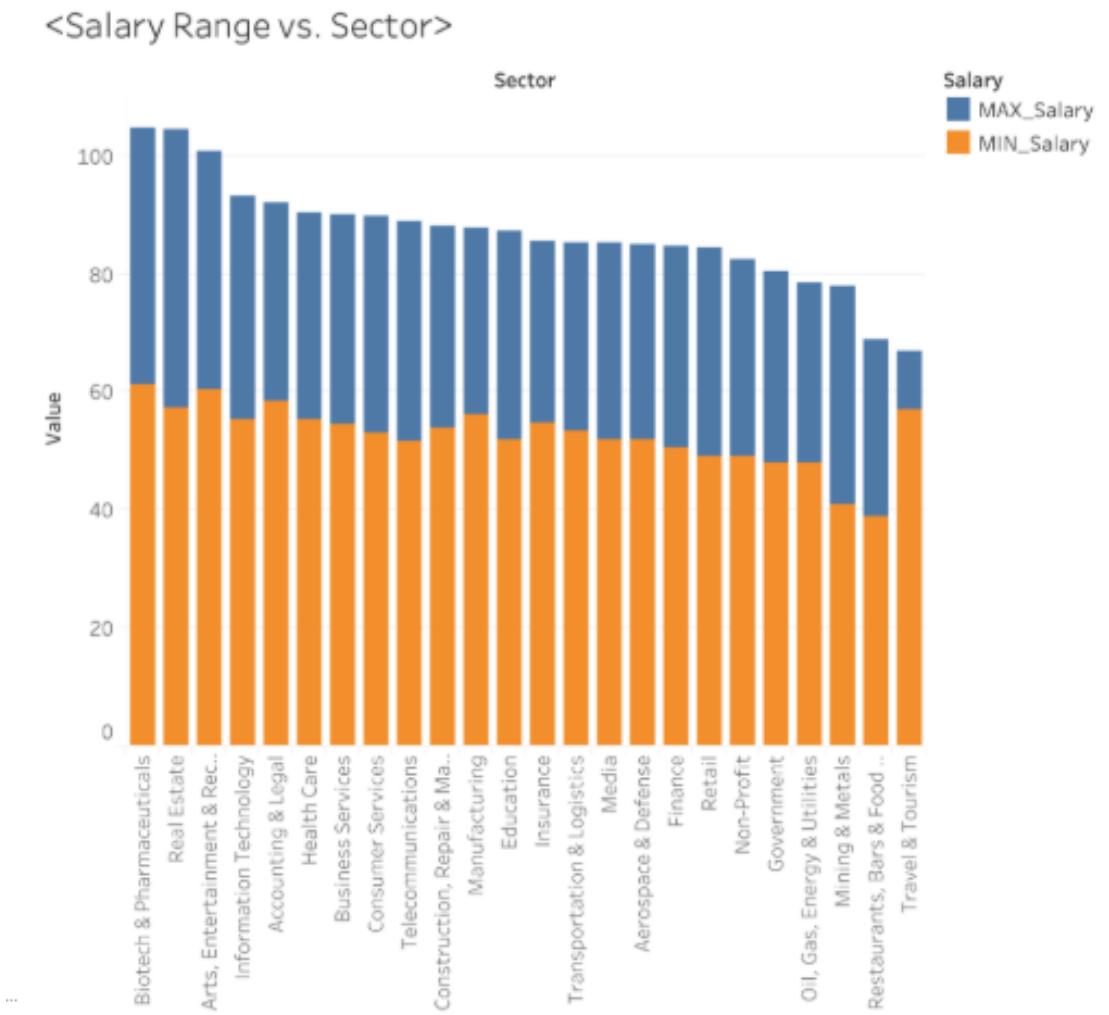


The variable "rating" itself conforms to the law of normal distribution. We analyzed the correlation between rating and salary, and found that the correlation was only 0.0425, which showed that there was no obvious linear relationship between them.

Industry & Sector:



IT and Business Service provide the most jobs for data analysts. Explain the importance of data to these two industries and their dependence on data analysts. The main reason is that these two industries have more competition and broader market, so they pay more attention to data resources.

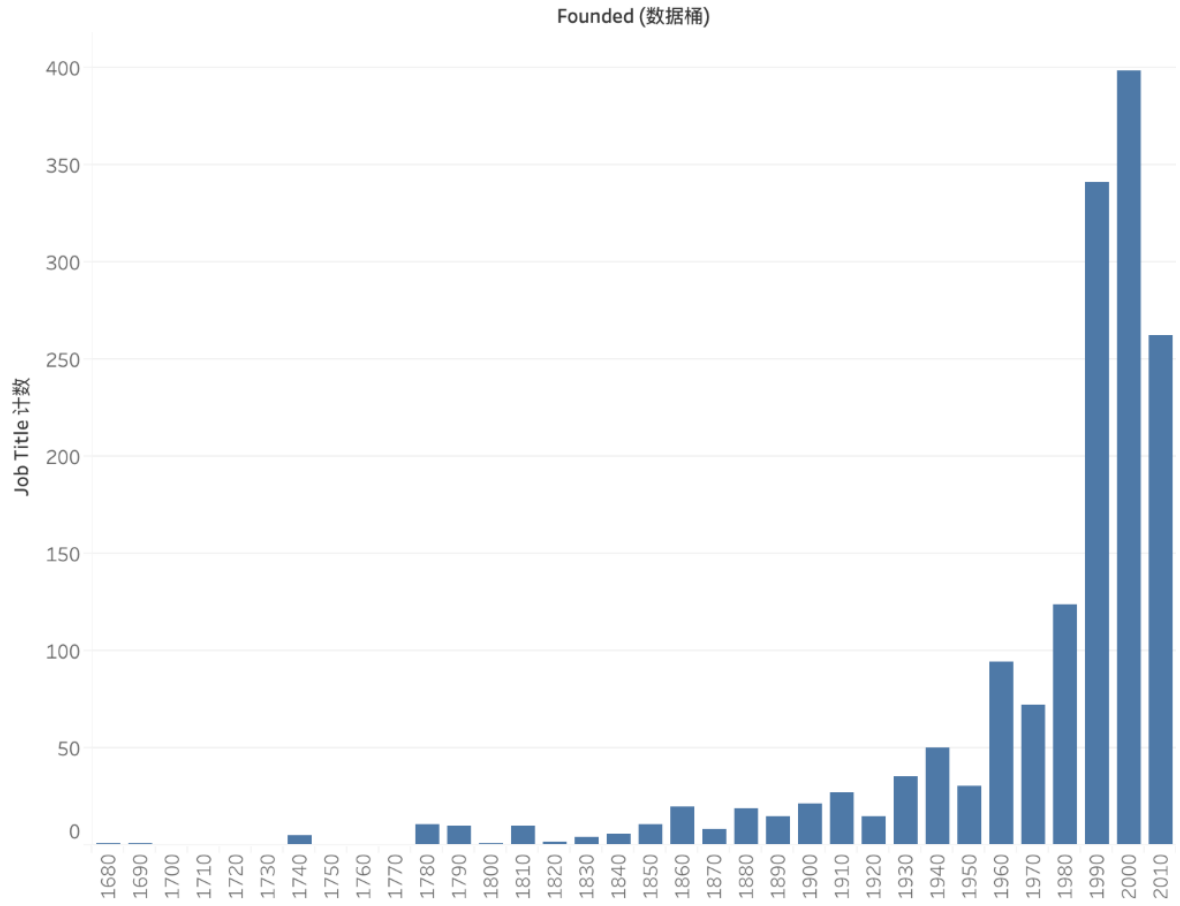


Travel&Tourism has a very small salary range. There is not much difference in salary range in other industries, but there is not a small difference in overall salary level.

Founded:

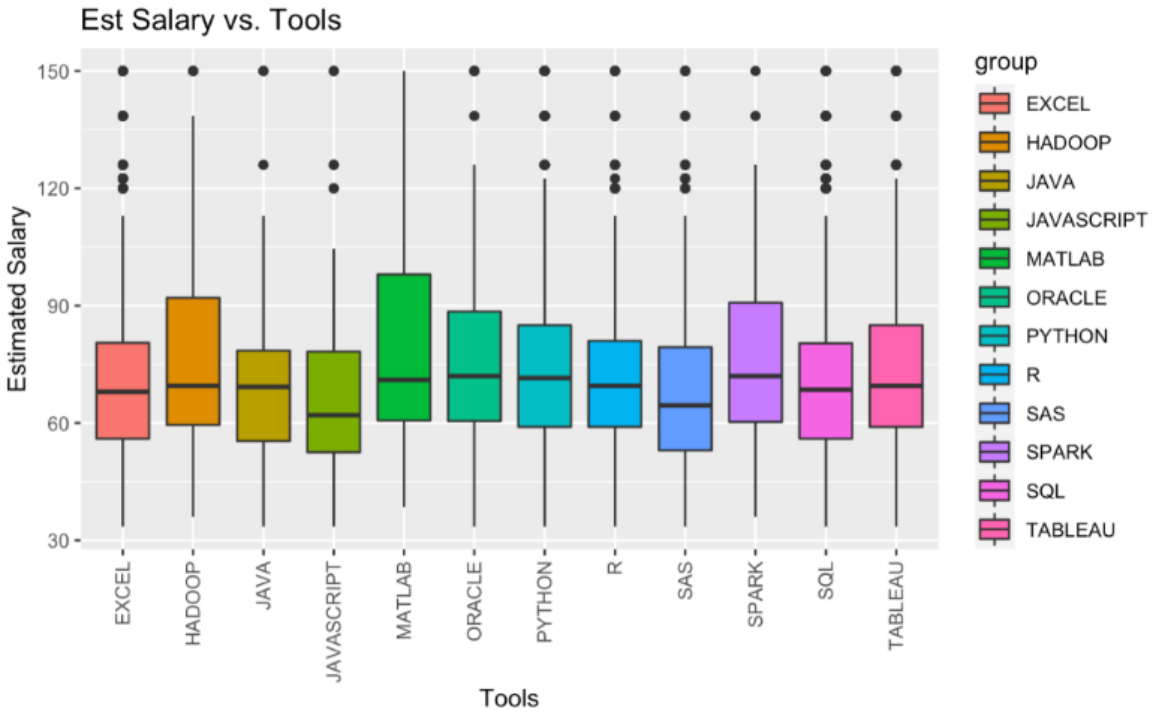
The figure below shows the number of jobs provided by companies established every 10 years. Companies established in 1990s and 2000s provided a large number of job opportunities. The company established in 2010s also provided a lot of jobs, but it may provide fewer jobs than the first two decades because of its short development time.

<Founded vs. Number of Jobs>



每个 Founded (数据桶) 的 Job Title 计数。数据按 Founded 进行筛选，这会排除 -1。

Job desription:



The above figures are statistics of the number of times job skills are mentioned in the salary description, and we show them as a word cloud. Among them, the top three skills mentioned the most are R language, sql and Excel. In other words, if you want to change careers as a data analyst, these three skills are basic essential skills.

The tools used have little impact on wages. Each industry has certain requirements for the tools to be used. There is no better tool in the absolute sense. Because the tool is suitable for different work scenarios and job types, there is no difference in salary.

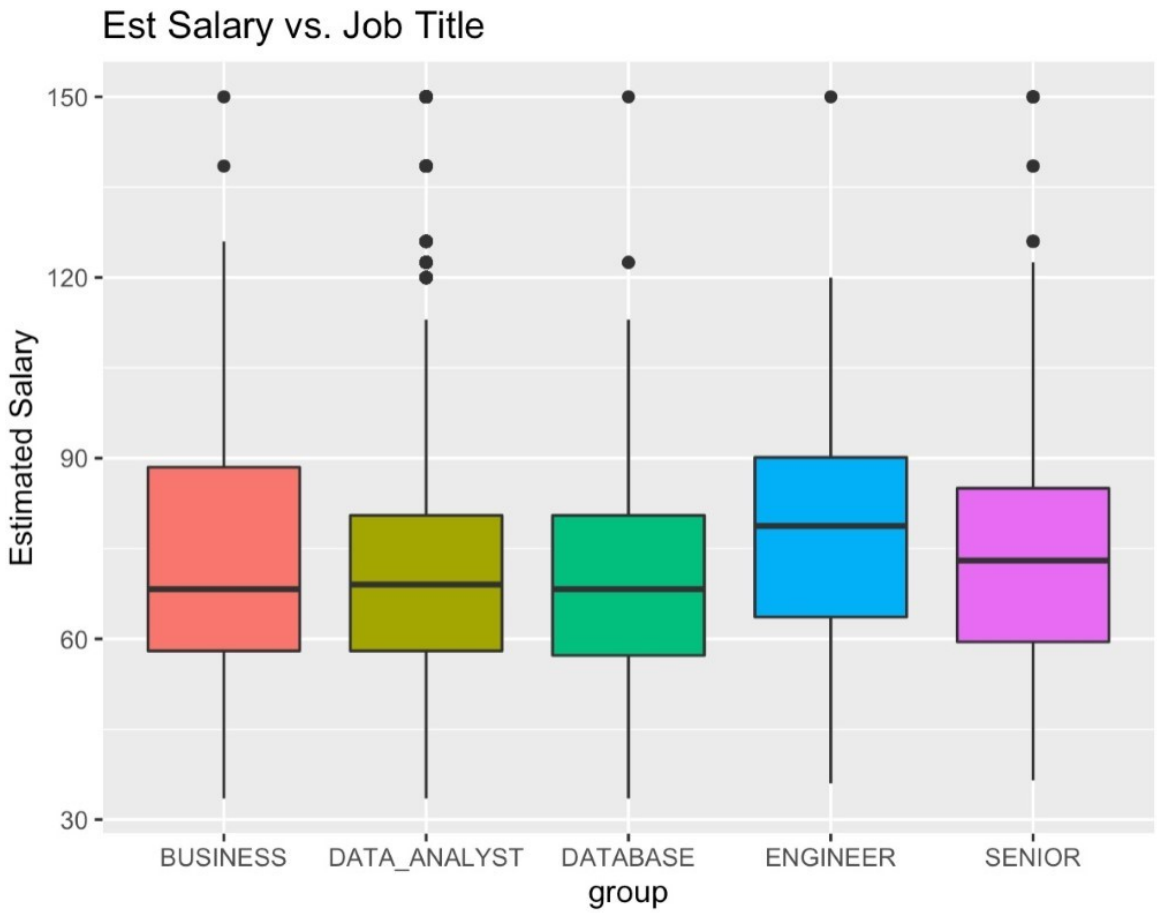
But as you can see from the figure, the data analysis industry offers a lower salary for mastering JavaScript skills, probably because other new tools can replace it.

If the candidate has more skills, such as programming and processing SQL, and analyzing data visually. This is a bonus for job search. Every industry needs a certain number of compound talents.

If an employee can have the skills of two employees, then the enterprise will provide a higher salary level.

Job title:

Because the data is pre-processed, the jobs about data analyst are screened out, so more than 70% of the jobs are data analyst. There are not many data scientists, data engineers and business analysts. The most popular keywords include senior, junior, healthcare, data management, marketing, financial, product, security and so on. By understanding and distinguishing the work of data analysts, the keywords are integrated and five representative keywords are selected again to observe whether they are directly related to salary.



We frequently extracted a few words from job titles, trying to find the relationship between job title and salary. The salary of an engineer is obviously higher than other positions. The wage gap between other positions is not obvious. But this is also the basis for reference by job seekers.

5. Interpretation & conclusion

Through the analysis of 9 variables, we draw the following important conclusions:

1) At present, the industries wITh the most jobs and the highest income level of data analysts are it and financial services.

2) At present, the areas with the most jobs and the highest income level of data analysts are near California, Los Angeles and San Francisco.

3) It is not easy to estimate the income level of this position only from the company size, rating, income and establishment time. Because there is no obvious linear relationship between them.

4) For the position of data analyst, the necessary skills are R language, SQL, Excel, python and tableau, which are also important skills and requirements.

Therefore, Job seekers need to improve their communication skills and project management level. A good data analyst needs to communicate with other departments of the enterprise, and summarize thefeedback results, so as to help the analysis to be related to the feedback results, so as to put forward better suggestions, improve solutions, and make the enterprise have higher benefits. It is particularly important to have basic knowledge of statistical mathematics, which needs to be used frequently in the analysis process, so as to quickly analyze problems and draw conclusions. The basic programming level and the use of software are indispensable, which will improve the efficiency of analysis and be very helpful for making obvious and exquisite reports.