MIS 6380- Data Visualization

Data Science job positions in the US

Group 7

Yadullah Haider

Sharvari Vikas Jadhav

Nishant Rana

Palaniappan Sathappan

Galuge Hansani Wijewardana

**Objective**

Demand for data science skills is growing exponentially and according to the job sites it is a great time to be a data scientist entering the job market. LinkedIn Workforce Report August 2018 states that, there is a shortage of approximately 151,717 people with data science skills particularly in cities such as New York city, San Francisco and Los Angeles.

The objective of the project is to provide better understanding of the data science job market in the United States, so that those who are looking for data science jobs can have an idea about the data science positions, required skills and the cities which has shortages for those positions.

When applying for data science job it is advantageous to know where to apply, the positions that are available, and the tools, skills needed for those positions. By using the information, we can narrow down the industries and the companies that have data scientist jobs and the states the has the most job opportunities. Also, we can answer the questions like who gets hired, what kind of talent do employers want when they are hiring a data scientist. The job seekers can target those companies and states or cities which has more job opportunities and even they can find the most suitable position according to their skills.

**Dataset**

The dataset on data science job positions in the United States was downloaded from www.kaggle.com. which is based on the job postings in Indeed on August 2018.

The data consist of position, company, description, reviews and the location. There were 7000 records in the dataset before the cleaning. The data science job positions include data scientist, data analyst, data engineer, research analyst, director, vice president, machine learning engineer and so on. The dataset consists with the well know companies such as Amazon, KPMG, Facebook, Walmart, Bank of America, and some unpopular companies like SunTrust, Lab126 and Honeywell as well.

The description field gives the details of the job descriptions. This field contains the required skills and tools that the recruiters are looking for a particular position. Moreover, the dataset shows the number of reviews for each job position. The location consists of the city, state and the postal code where the companies are located at.

**Methods Used**

We have initially used two datasets namely “US Population Data” and “Data Science Dataset” from Kaggle. Going ahead, we first cleaned the data using Excel. Cleaning of the data required removing missing values and redundant information. Using Python notebook, we split the columns of Country, Postal Code and State accordingly. Also aggregated the various positions into 20 major unique positions. Data Imputation was done for “Reviews” column of which some fields were initially blank.

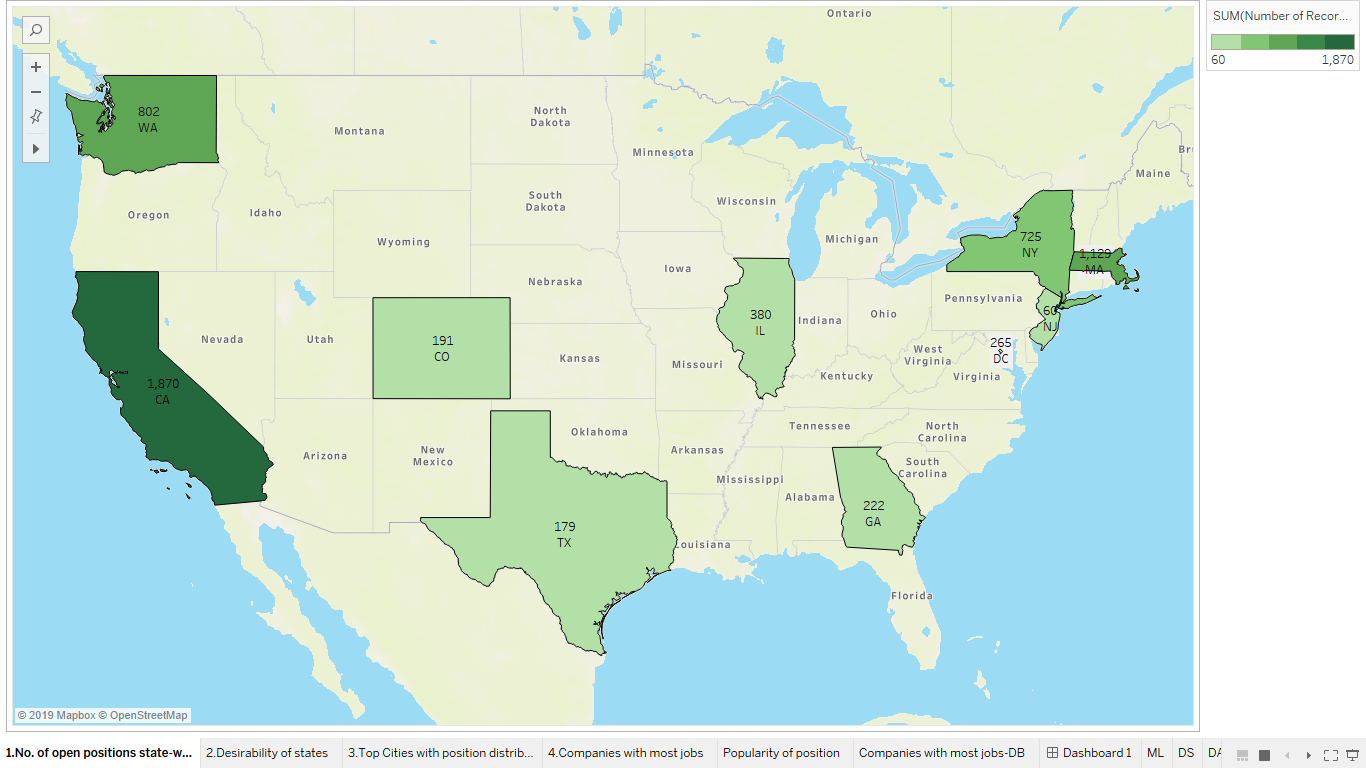
Next, we merged the data using Tableau on StateCode field. In tableau, we created additional calculated fields such as “PopulationRate”, “PositionRate” and “Population>Positions”.

Also, certain custom shapes were added for each Job Title, the method used here was to add different pictures for every Job Title/Position in the shapes folder of the Tableau Repository on the local computer. And from Tableau, it appeared in the “Shape” palette.

**Insights**

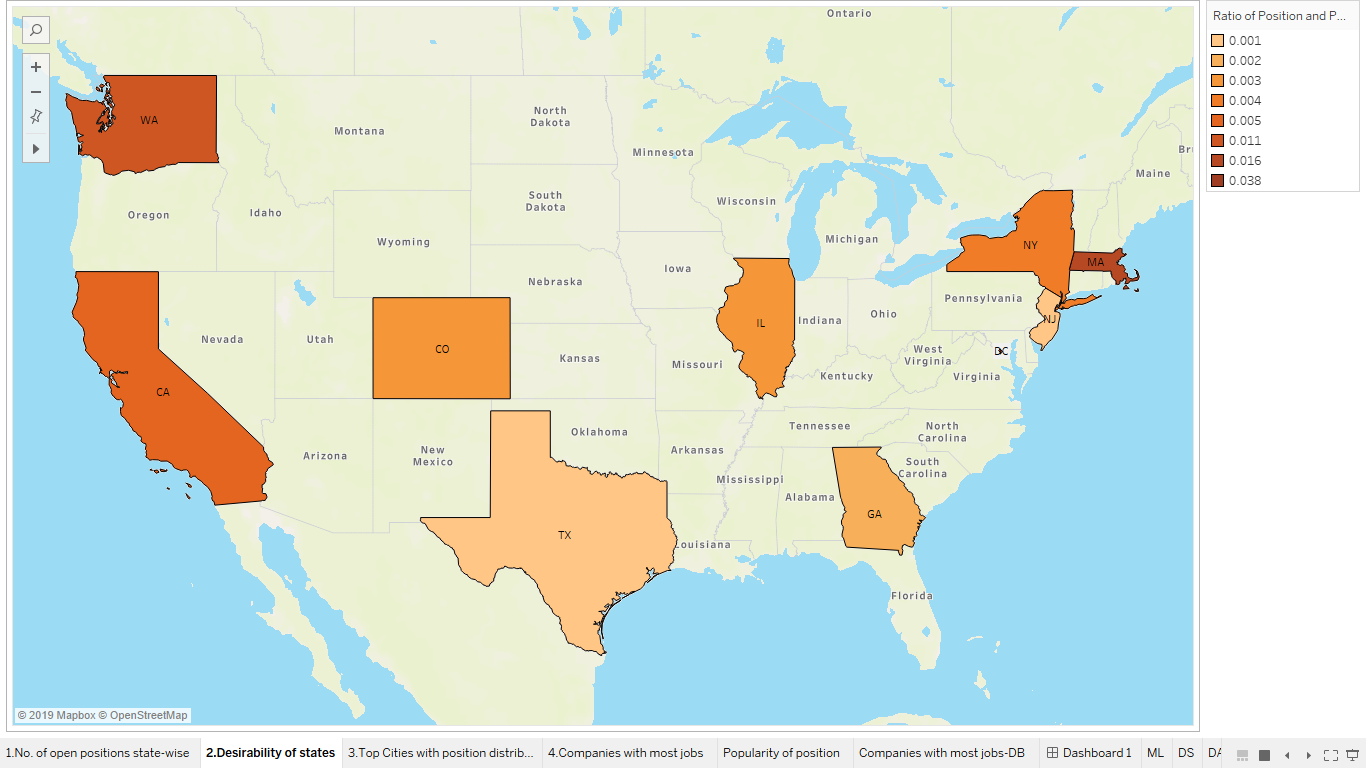
The first visualization shows the states that has shortages in the data science job positions. We can clearly see that California has the most job opportunities followed by Massachusetts, Washington and New York. States such as Colorado, Texas and New Jersey have few open positions in data science jobs. The following table shows the states with the number of open positions in descending order.

|  |  |
| --- | --- |
| **State** | **Number of Open positions** |
| CA | 1,870 |
| MA | 1,129 |
| WA | 802 |
| NY | 725 |
| IL | 380 |
| DC | 265 |
| GA | 222 |
| TX | 179 |
| NJ | 60 |



The job seekers can focus on the states like California, Massachusetts and Washington as there are more data science job opportunities.

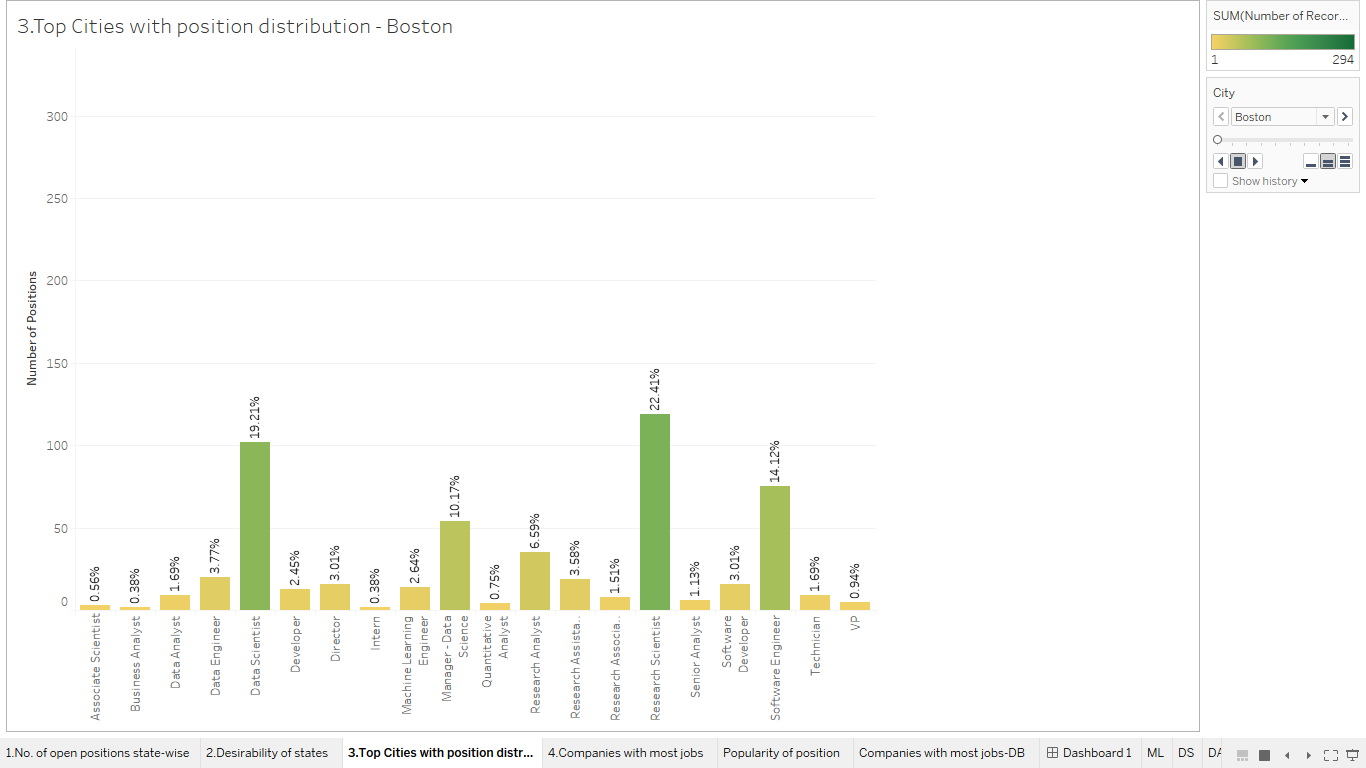
However, the population of the states are different, and it is better to consider the number of job positions accordingly. This helps job seekers to know their chance to get selected for a position. The second map is created according to the number of positions/ population ratio.

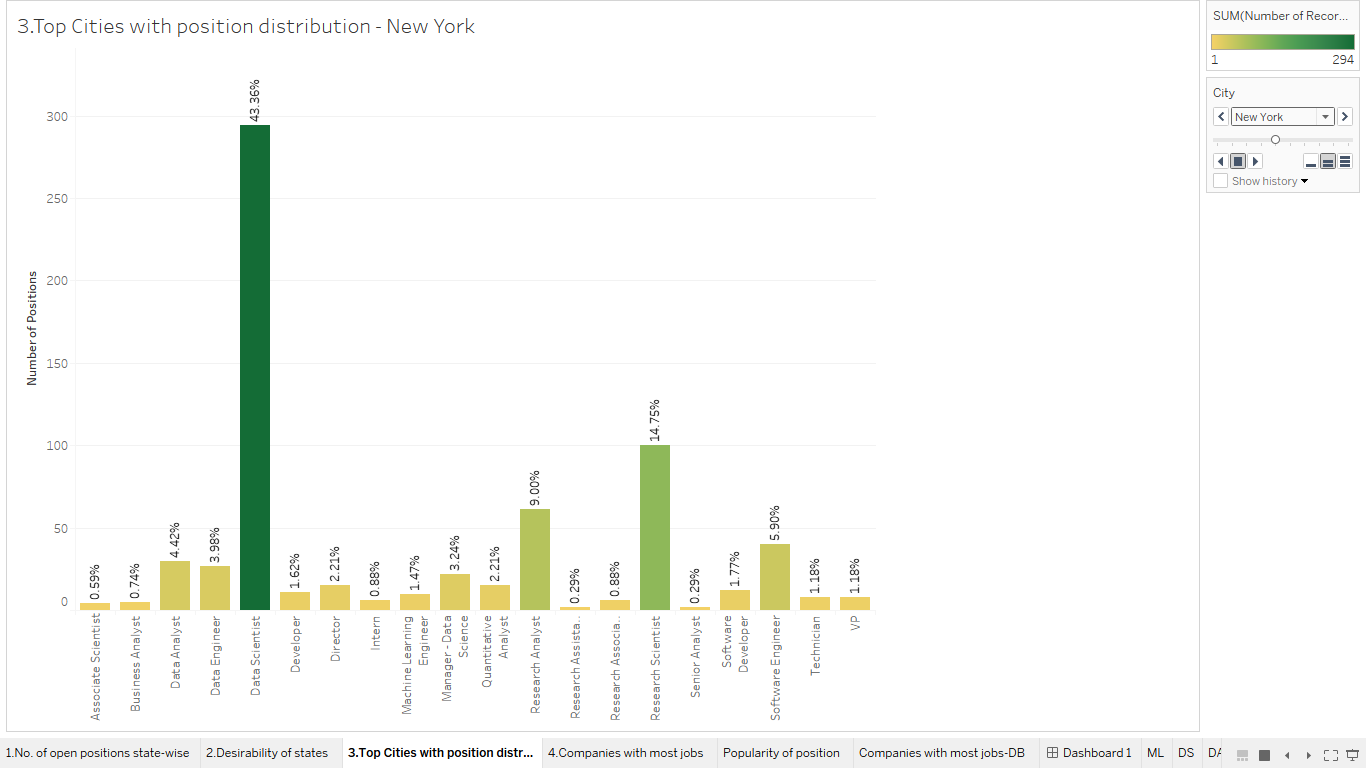


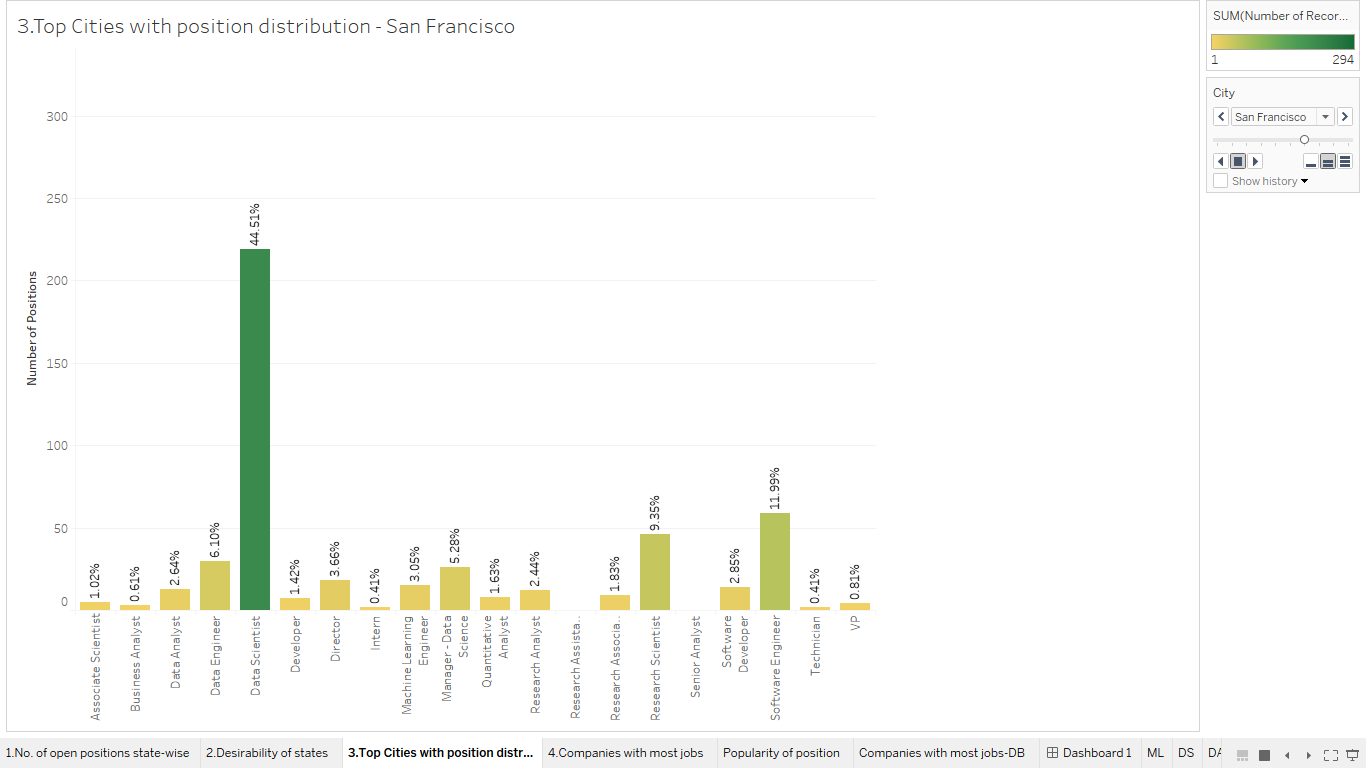
Since, District of Columbia has low population, it has the highest ratio of number of positions to population, whereas the states with largest population such as Texas has lower ratio. The summary of the number of positions/ population ratio according to the state are as follows.

|  |  |
| --- | --- |
| **State** | **Number of Open positions** |
| DC | 0.038 |
| MA | 0.016 |
| WA | 0.011 |
| CA | 0.005 |
| NY | 0.004 |
| CO | 0.003 |
| IL | 0.003 |
| GA | 0.002 |
| NJ | 0.001 |
| TX | 0.001 |

After identified the states with most jobs, we can narrow down the search to the cities. By using a bar chart, we can identify the data science job positions available in the job market. We added another dimension to the chart to see which cities have the shortage of the skilled people. By using this animated visualization, the job seeker can decide what city he or she needs to focus on along with the positions to be applied.





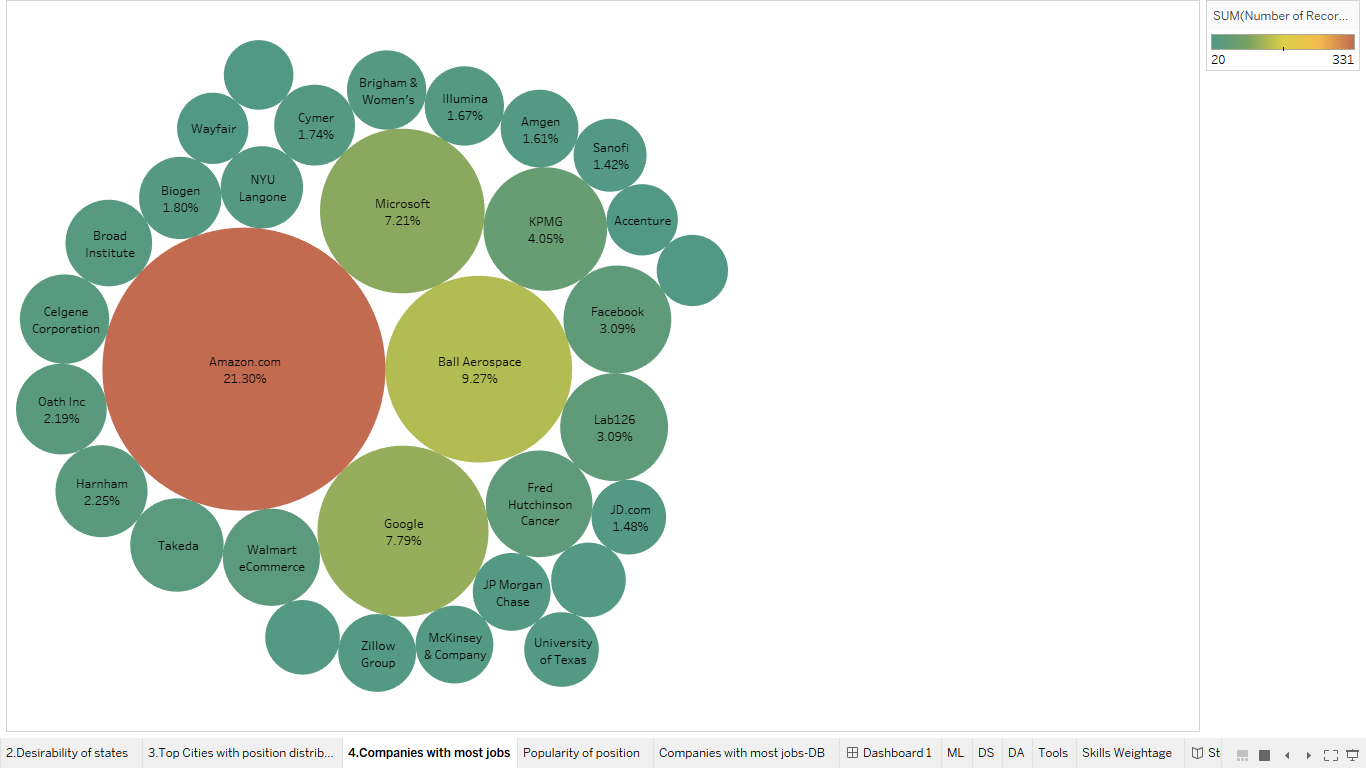


The above visualizations show the distribution of number of positions in three cities- Boston, New York and San Francisco. Likewise, in this animated visualization we have the bar charts for other cities- Cambridge, Chicago, Mountain View, San Diego, Seattle, Sunnyvale and Washington.

San Francisco, Washington, Chicago and New York are the cities with high demand for data scientists while San Diego, Cambridge and Boston have more Research Scientist positions. Moreover, Sunnyvale and Mountain View are the cities which need more software engineers more than other positions.

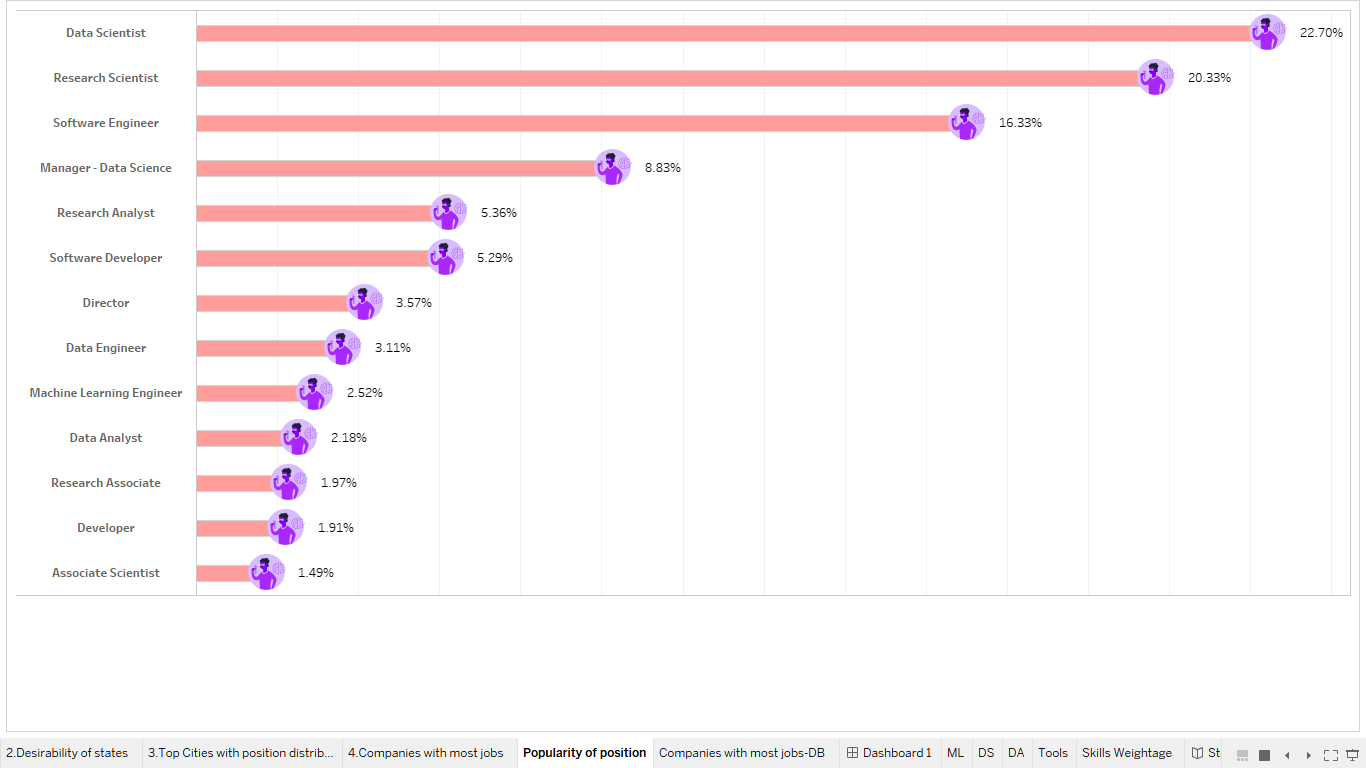
Next, it is important for the job seekers to know that the companies that are hiring for data science jobs. We have created a packed bubble chart to identify the companies that are demanding for data science positions. The bubble size indicates the number of positions available in that company. The companies which are popular for IT job positions such as Amazon, Facebook, Microsoft and Google have the greatest number of open job positions, but unfamiliar companies such as Lab126, Ball Aerospace, Fred Hutchinson Cancer Research Center and Harnham also have considerable number of job opportunities.

In 2018 Amazon was expected to hire 331 (21.3%) people for data science jobs while Ball Aerospace falls behind with 144 jobs followed by Google, Microsoft and KPMG with 121, 112, and 63 positions available respectively.

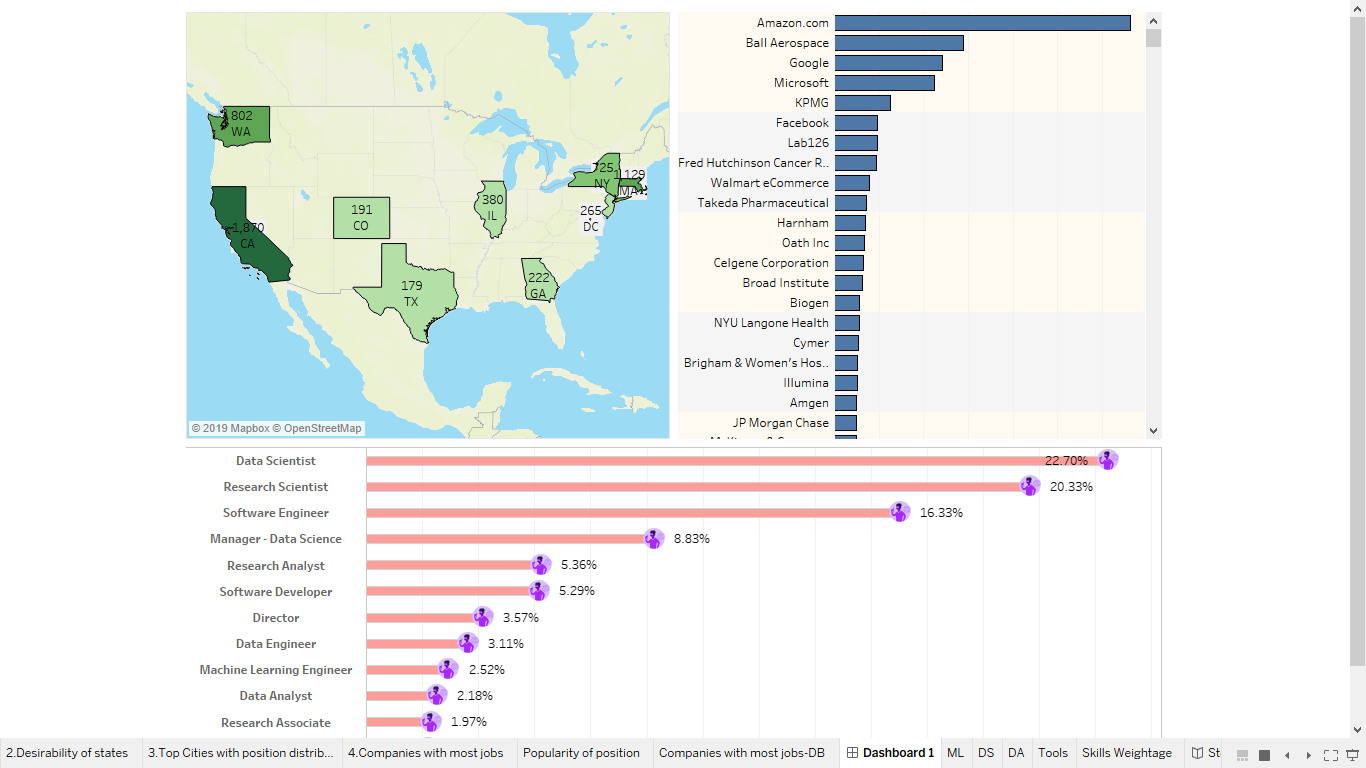


One of the most important insight that the job seekers are looking for is the data science job positions available in the job market. Overall, the most demanding data science job positions are data scientist, research scientist and software engineer. The following table shows the most demanding job positions with the number of opportunities.

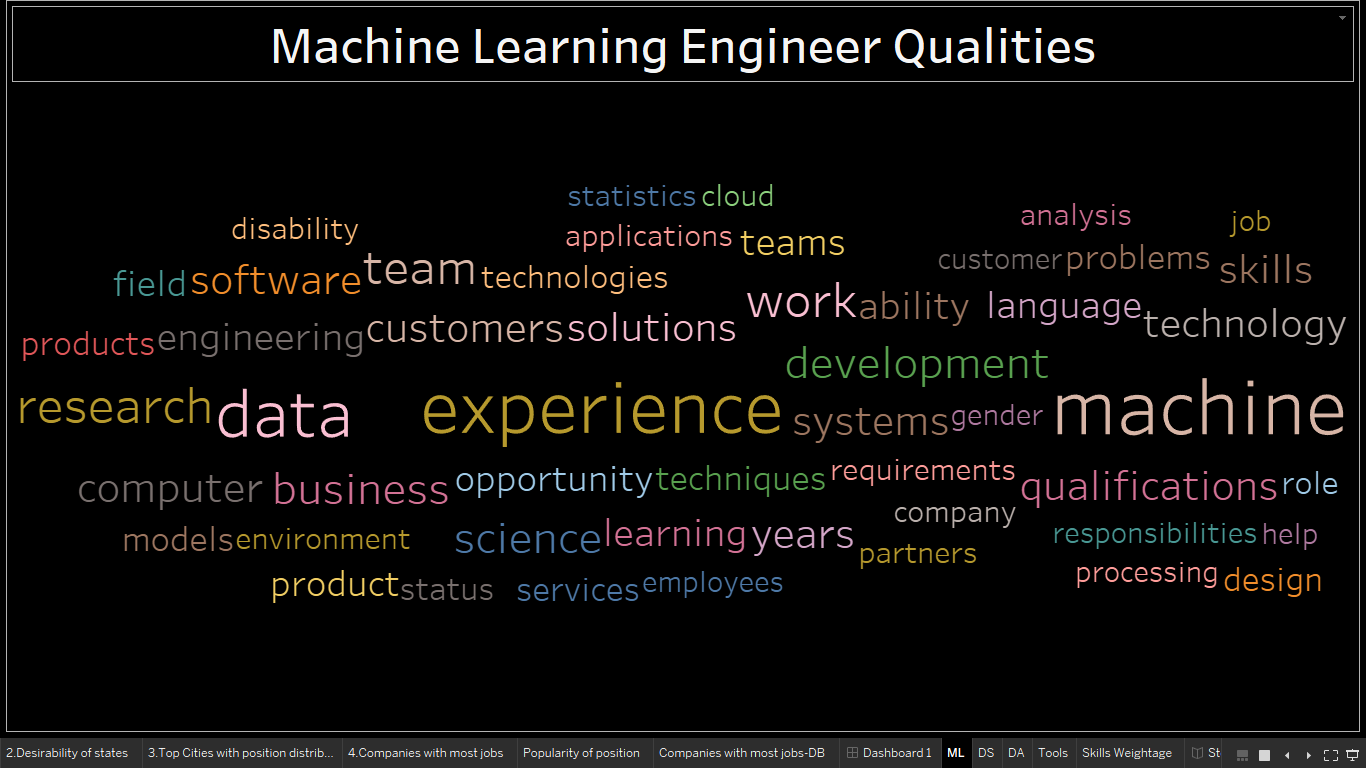
|  |  |
| --- | --- |
| **Position** | **No of Openings** |
| Data Scientist | 1,322 |
| Research Scientist | 1,184 |
| Software Engineer | 951 |
| Manager- Data Science | 514 |
| Research Analyst | 312 |
| Software Developer | 308 |
| Director | 208 |
| Data Engineer | 181 |
| Machine Learning Engineer | 147 |
| Data Analyst | 127 |
| Research Associate | 115 |

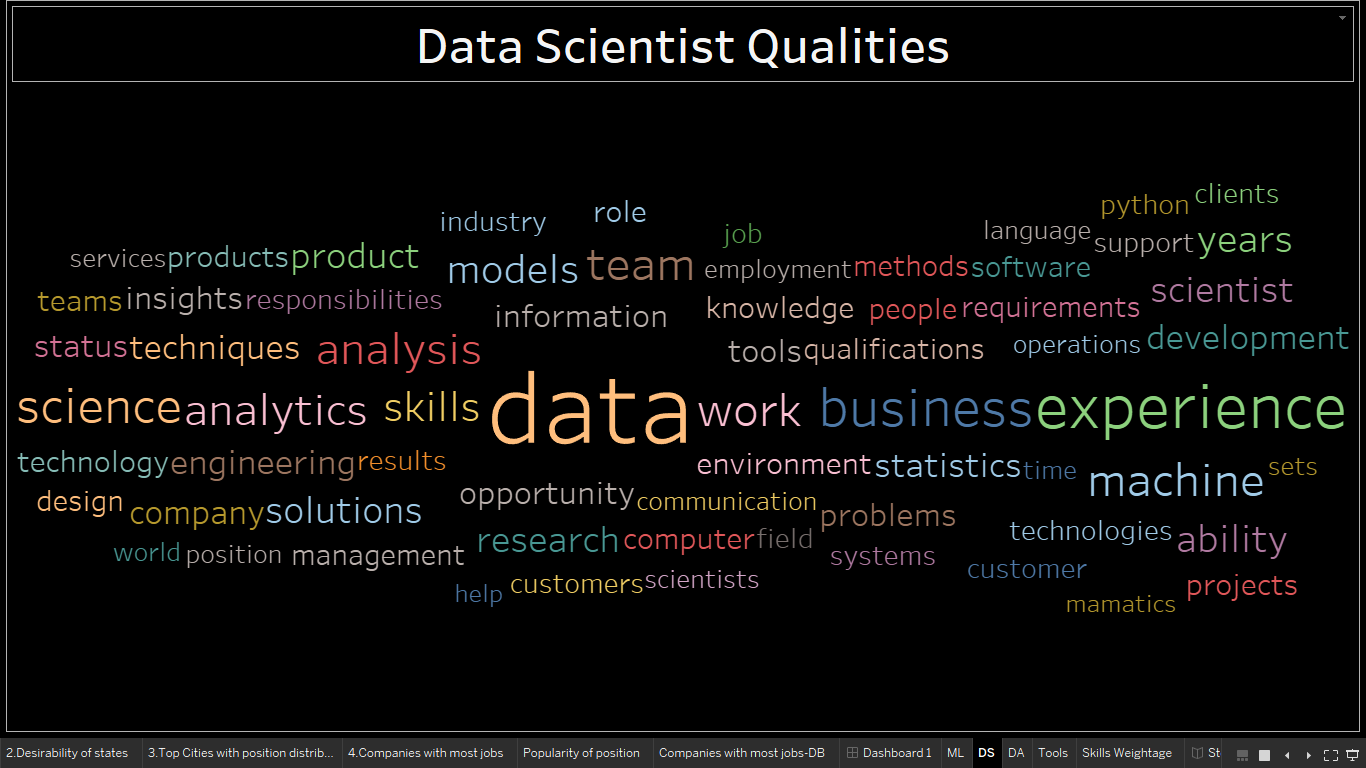


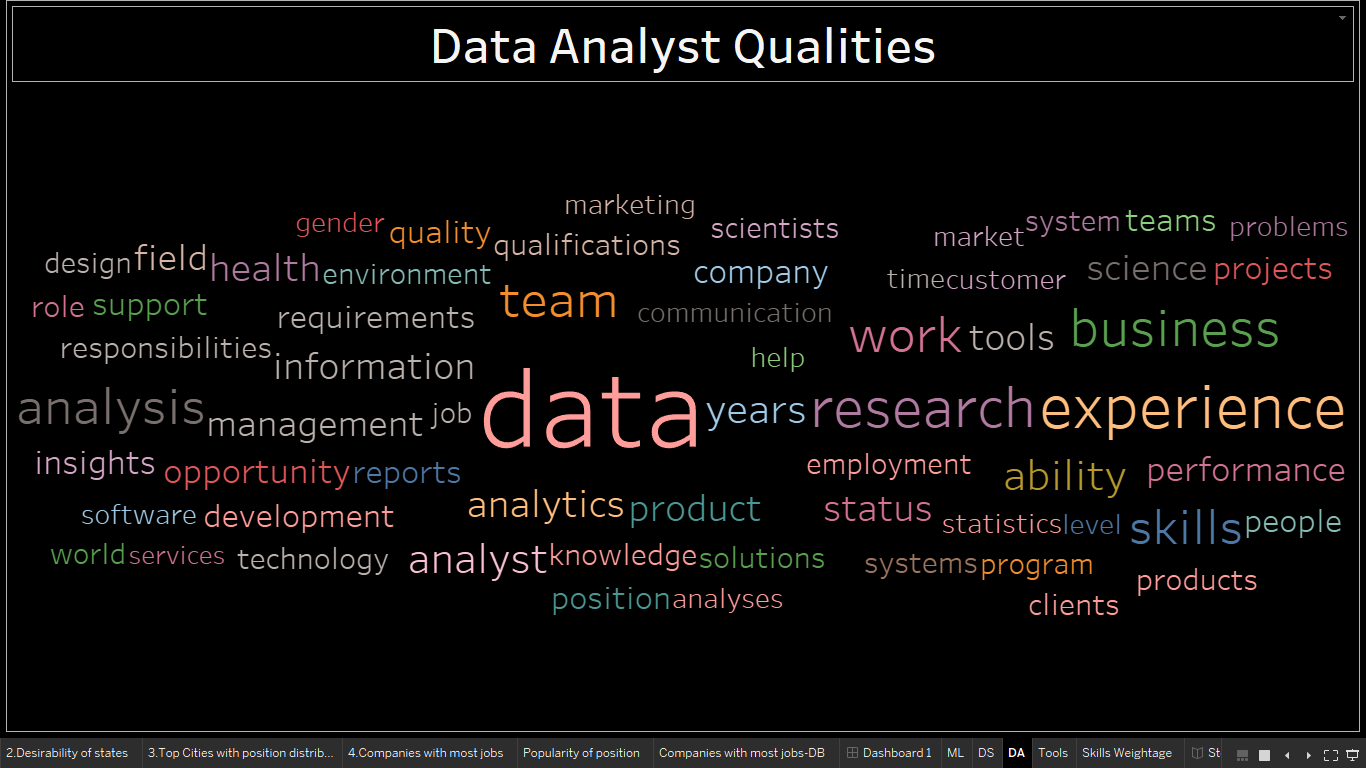
For easy comparison between the company, state and the position we created a dashboard. By selecting a position, we can clearly see who (companies) are hiring people for that position and which state needs people for that position. Likewise, if we select a company we can see where it located at and the job positions that they are hiring for.



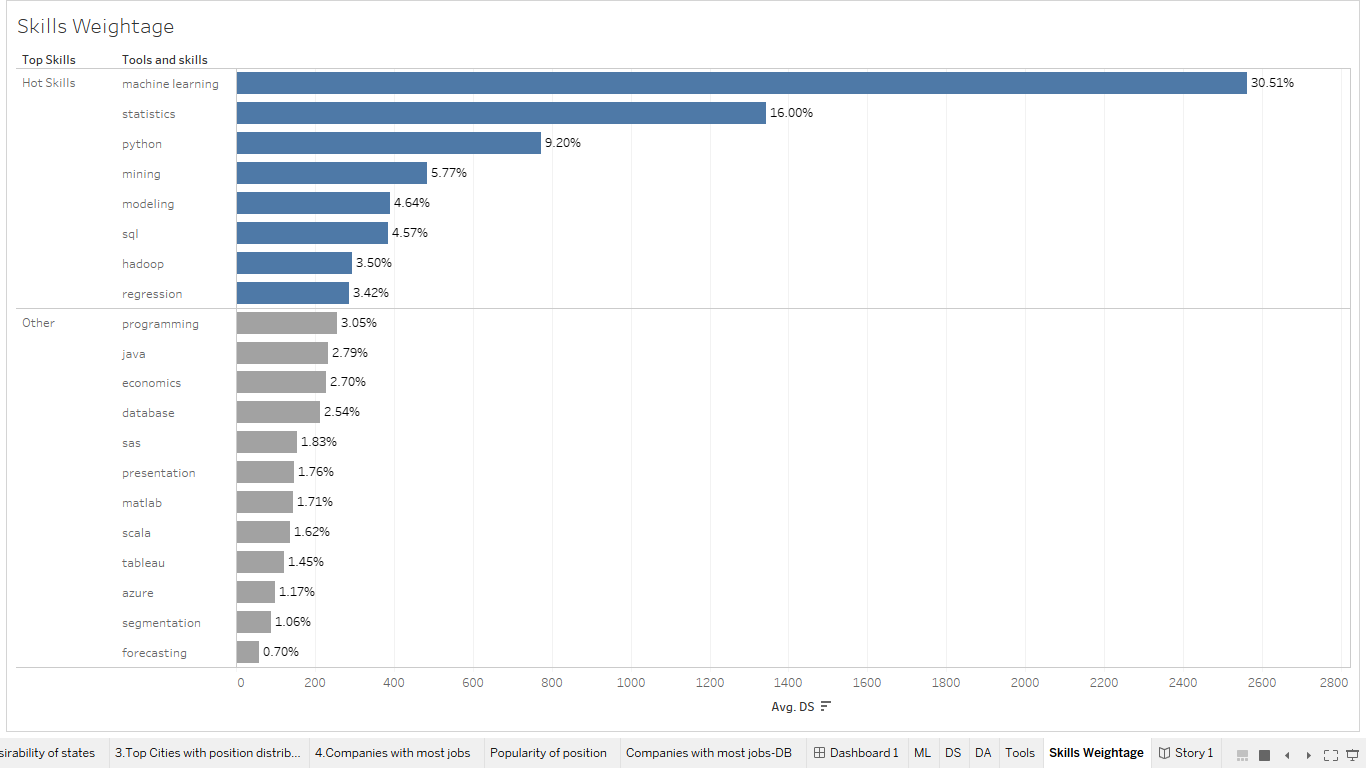
It is essential for a job seeker to know the skills required for each job. We selected three popular jobs to find the skills and tools that the recruiters are looking for. Using the word-clouds that we created in tableau shows the key words used in the job descriptions. For Machine learning engineer, experience is the most important quality that the companies are looking for where for data scientist, and data analyst it is the knowledge on data. As font size of the quality increases, the importance of the quality increase.



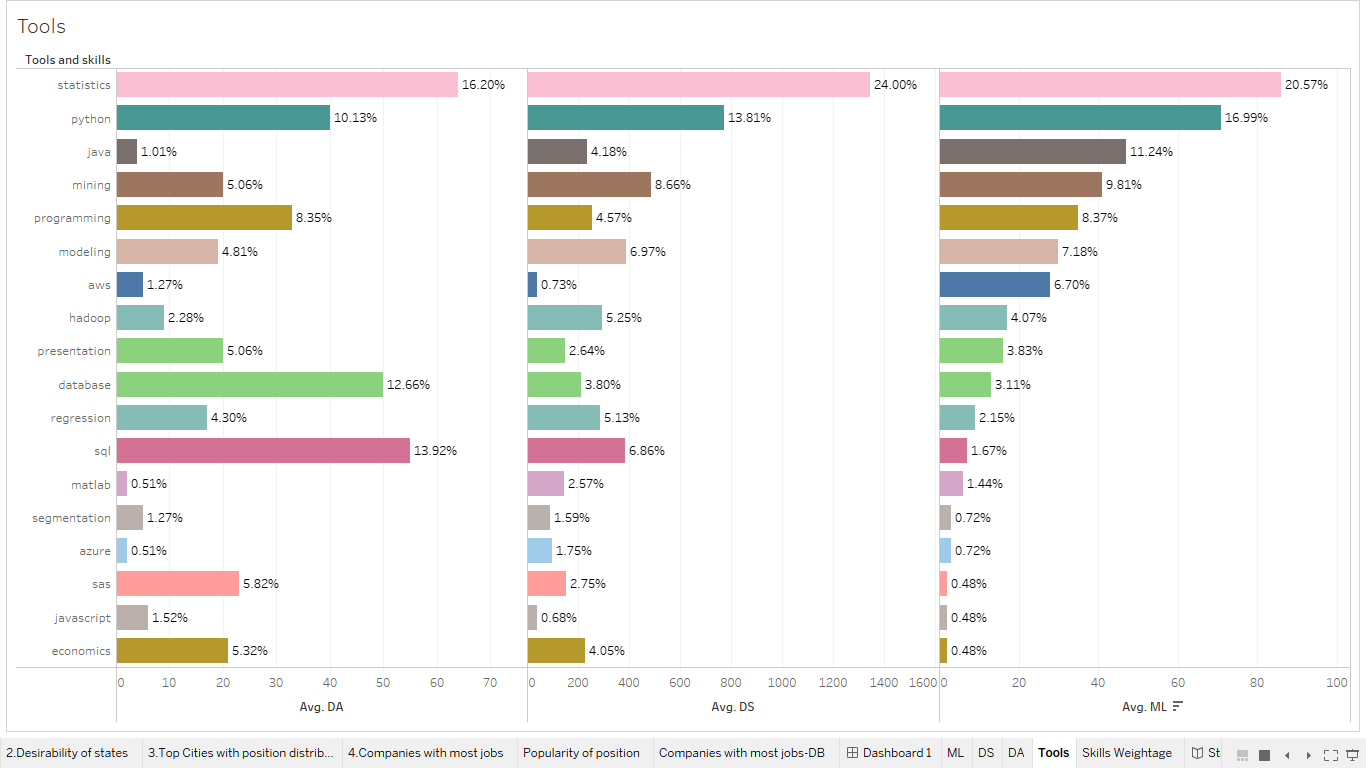




Finally, more precisely we can find the tools and skills that the data science job seekers need to have. Machine learning, statistics, Python, mining modeling, sql and Hadoop are the top skills that are required by the companies. However, programming, presentation and forecasting skills and tools such as Java, SAS and Tableau also in the list.



Apart from that we categorized the skills according to the popular job positions that we considered above and the visualization is as shown below.



All three positions require statistics knowledge at most. Data analysts need sql and Python knowledge next while for Machine learning engineer needs to have more knowledge and experience on Python and Java with programming and modeling skills.

**Conclusion**

In conclusion, the jobs seekers can target to apply for jobs in California, New York and Massachusetts as those states have more job opportunities, particularly in cities like New York city, San Francisco and Boston. Also, consider looking for jobs not only in large companies like Amazon, Google and Microsoft but also consider the unpopular companies as well. As the next step of looking for the data science job, the job seekers can look for the most demanding positions and decide which position he or she is willing to fill. Apart from that referring to the visualization on the required skills, the job seeker can identify which job is fit for his or her skills and the knowledge on the tools. Moreover, if anyone needs to be part of the data science workforce in the US, that person needs to obtain the knowledge on statistics, mining, forecasting and experience with the tools such as Python, SAS, Java and SQL.