

Data Management and Visualization

Assignment

Names:	Kashmira Chawan & Deval Oza
Student Numbers:	21261109 & 21260176
Programme:	MCM
Module Code:	CA682
Assignment Title:	Data Visualization
Submission Date:	25/11/2021
Module Coordinator:	Dr Suzanne Little

Declaration on Plagiarism

We declare that this material, which I/we now submit for assessment, is entirely my own work and has not been taken from the work of others, save and to the extent that such work has been cited and acknowledged within the text of my work. I/We understand that plagiarism, collusion, and copying are grave and serious offences in the university and accept the penalties that would be imposed should I engage in plagiarism, collusion or copying. I/We have read and understood the Assignment Regulations. I/We have identified and included the source of all facts, ideas, opinions, and viewpoints of others in the assignment references. Direct quotations from books, journal articles, internet sources, module text, or any other source whatsoever are acknowledged, and the sources cited are identified in the assignment references. This assignment, or any part of it, has not been previously submitted by me/us or any other person for assessment on this or any other course of study.

I/We have read and understood the referencing guidelines found at <http://www.dcu.ie/info/regulations/plagiarism.shtml>, <https://www4.dcu.ie/students/az/plagiarism> and/or recommended in the assignment guidelines

Name: Kashmira Chawan

Date: 25/11/2021

Name: Deval Oza

Date: 25/11/2021

Topic: Analysis of Data Science Jobs across the USA

Abstract

According to New York Times, Data Science is the "Hot new field that promises to revolutionize industries from business to government, health care to academia." There are a variety of jobs/roles under the data science umbrella. In our project, we have visualized top data science roles one can choose. A Dataset which we used has the jobs in the USA. This visualization will help anyone who is looking to get a job in the data science field. People from various and different background wants to get into data science. Although jobs in the data science field are present in all sectors, we aimed to know which sector has maximum roles. According to our visualization with no surprise, it was the Information Technology with most roles. Anyone who is a working professional or just graduated or a student who is planning to make a career in data science, is concerned about which job title they should apply for? What will be the salary for that role? For the same, we tried to find a solution by plotting the graph for minimum and maximum salary ranges for 5 job titles. Through it, we came to know about 3 top roles with maximum salary were Data Engineer, Data Scientist, and Data Analyst. Sometimes getting into a top company is difficult. Therefore, we decided to know the average salary for the top 3 roles in the newly founded company. The USA being a large country, there might be the question of which location will be the best to apply for these roles. So Next up, we decided to get more insight into these 3 top roles. Based on positive ratings we understood that Chicago is popular for Data Analyst, Texas for Data Engineer, and New York for Data Scientist.

Data Description

We collected three different .csv datasets from Kaggle. Links for the datasets are given below:

All these files have common column names with 15 columns each and different numbers of rows.

[Data Analyst Jobs | Kaggle](#) - 8 MB in size with 2253 rows.

[Data Science Position Analysis | Kaggle](#) - 14.96 MB in size with 3909 rows.

[Data Engineer Jobs | Kaggle](#) - 9 MB in size with 2528 rows.

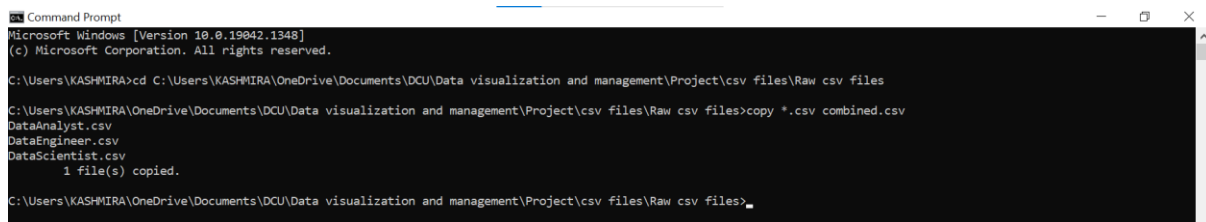
Considering the rows mentioned above we had a total 8690 of rows. Data Types present were string, float, and integers. Datasets had the following columns/attributes:

Job Title, Salary Estimate, Job Description, Ratings, Company Name, Location, Headquarters, Size, Founded, Type of Ownership, Industry, Sector, Revenue, Competitors, and Easy Apply.

In our opinion, Datasets have volume as a characteristic of Big Data. We integrated all three files before processing and cleaning process making it 31.96 MB of a dataset. Also, it had a variety of data such as Salary in \$, negative values, and special characters which we were eliminated in later stages.

Data Exploration, Processing, Cleaning, and/or Integration

The dataset collected is three different .csv files. All files were having the same column name. Integration of files was done locally in Command Prompt.



```
Microsoft Windows [Version 10.0.19042.1348]
(c) Microsoft Corporation. All rights reserved.

C:\Users\KASHMIRA>cd C:\Users\KASHMIRA\OneDrive\Documents\DCU\Data visualization and management\Project\csv files\Raw csv files
C:\Users\KASHMIRA\OneDrive\Documents\DCU\Data visualization and management\Project\csv files\Raw csv files>copy *.csv combined.csv
DataAnalyst.csv
DataEngineer.csv
DataScientist.csv
        1 file(s) copied.

C:\Users\KASHMIRA\OneDrive\Documents\DCU\Data visualization and management\Project\csv files\Raw csv files>
```

All the further steps are performed on Google Colab using Python.

The Combined CSV file was loaded in google drive and pandas and numpy libraries were imported.

Cleaning:

- Data had some -1 value, which means it had no entry for that specific job title. It was replaced by NaN.
- Attribute 'Easy apply' had True and Nan. Nan place was filled by 'False' to understand the data properly.
- Primary quality checks for null values were done using *isnull()* function.
- 'Company Name' had unnecessary '\n' was present. It was removed using *str.split()* function.

Processing:

- From 'Size' column we just required minimum size for that we removed data using *mydata['min_size'] = mydata['Size'].str.split(' ').str[0]*.
- Column location had State and city names included. We separate that column into two new columns named State and City.

```
[ ] mydata["State"] = mydata["Location"].agg(lambda x: x[-2:])
    mydata["City"] = mydata["Location"].agg(lambda x: x[:-4])
```

- The Founded column had a float value, which was not our requirement, so it was converted to integer datatype.
- Next, The salary column had special characters such as \$ and unwanted string. We aimed to know the minimum and maximum salary only, we

required separate values for min and max salary. Python code was run for the same as below:

```
[ ] df_salary = mydata['Salary'].str.split("-", expand=True,)

minimum_salary = df_salary[0]
minimum_salary = minimum_salary.str.replace('K', ' ')

maximum_salary = df_salary[1].str.replace('(Glassdoor est.)', ' ')

maximum_salary = maximum_salary.str.replace('(', ' ')
maximum_salary = maximum_salary.str.replace(')', ' ')
maximum_salary = maximum_salary.str.replace('K', ' ')
maximum_salary = maximum_salary.str.replace('Employer est.', ' ')
maximum_salary = maximum_salary.str.replace('Per Hour', ' ')

maximum_salary = maximum_salary.str.replace('$', ' ').fillna(0).astype(int)
minimum_salary = minimum_salary.str.replace('$', ' ').fillna(0).astype(int)

mydata['Salary_minimum'] = minimum_salary
mydata['Salary_maximum'] = maximum_salary

mydata
```

- A new column named 'salary average' was created by performing a calculation on the minimum and maximum columns.
- Unwanted columns were dropped from a dataframe.

Exploration:

Attributes with which we worked are Job Title, Salary Minimum, Salary Maximum, Salary Average, City, State, Ratings, Sector, Founded, Min Size.

As it was Data Science jobs analysis, knowing which role has more count was important. From it, we came to know Data Engineer, Data Analyst and Data Scientist had more count.

```
[ ] a = mydata["Job Title"].value_counts()[:10]
a
```

Data Engineer	729
Data Analyst	656
Data Scientist	282
Senior Data Engineer	140
Senior Data Analyst	137
Big Data Engineer	114
Software Engineer	93
Senior Data Scientist	91
Machine Learning Engineer	64
Sr. Data Engineer	48

Name: Job Title, dtype: int64

From the above result, we decided to have insight for the top 3-5 roles only through visualization.

Visualization

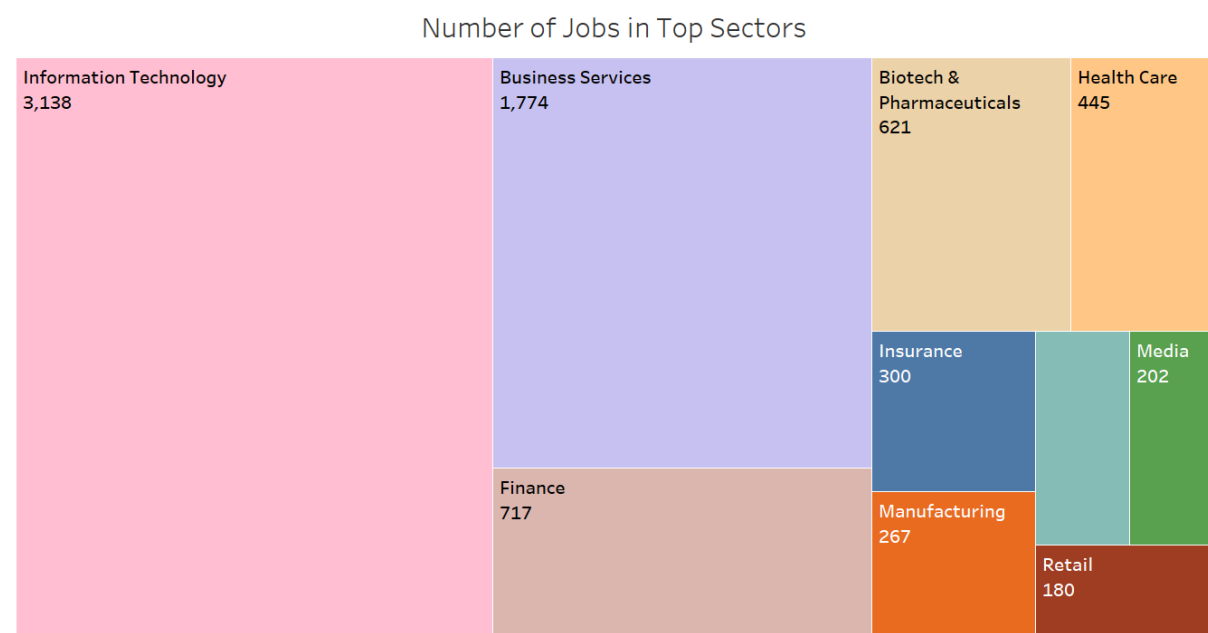
Visualization 1- Number of Jobs in Top Sectors

Choice of Chart

For our first visualization, we used a Treemap chart. The Treemap shown below illustrates the count of jobs in top sectors. We filtered the sector by getting the top 10 sectors by the minimum size of any company. The Size of the rectangle explains that there are a greater number of jobs present in the Information Technology sector and very few in the Retail sector.

Design Choice

We used light colours for the sectors with greater size and dark colours for the smaller sector. Dark colour would highlight the sectors and show their existence in the chart. First, we thought to give shades of one colour (e.g., Blue) but it will not make every sector stand out so decided to give each rectangle different colour.



We started our visualization by considering which sector has the maximum job title present thus even if anyone is not interested in the Information Technology sector, they can look for some other sector to apply.

Visualization 2- Salary range for top 5 roles

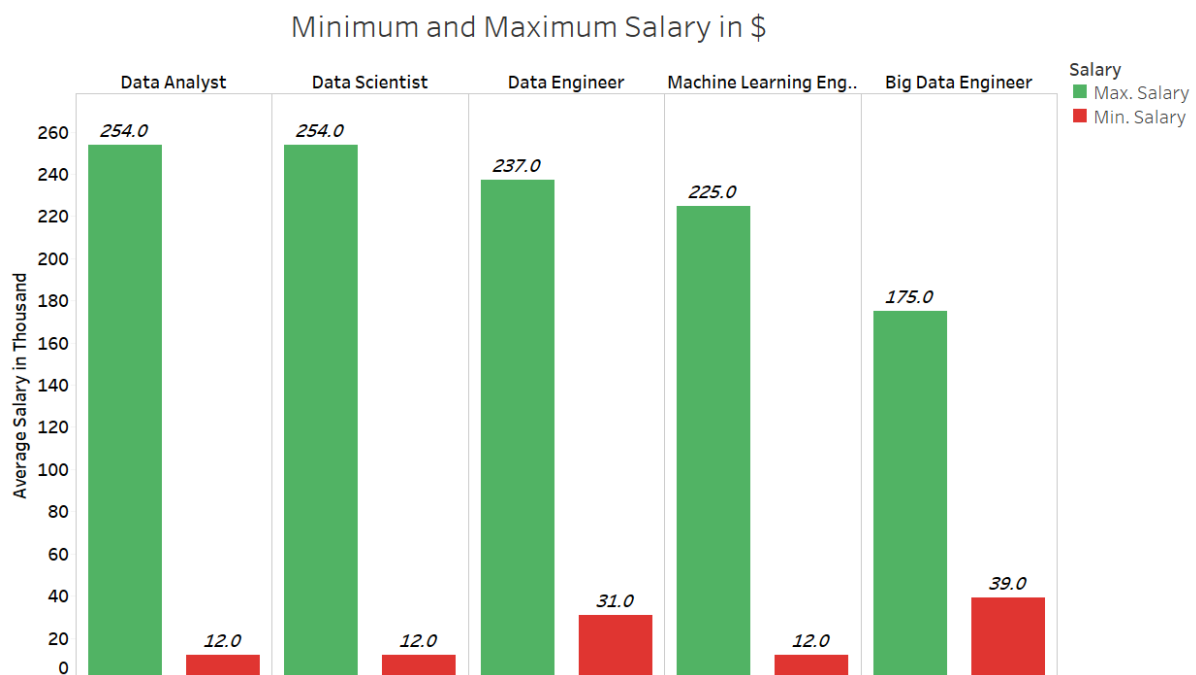
Choice of Chart

For the second visualization, we thought to present data in bar graph. Bar works effectively when working with numeric data and categorical data. We filtered the job titles by top roles concerning rating count. The Vertical bar graph is easier to understand the low and high range so instead of using a horizontal bar graph we guessed for a vertical graph.

Design Choice

We used bright colours to make it attractive and easy to understand the range. For minimum salary, the red colour will highlight it more than compared to green and as people are more interested in lower salary value it will draw their attention over it at very first. Green itself is a very pleasant and natural colour, hence thought of using it for maximum salary.

Job titles are displayed on the top y-axis so that right after looking at the salary range user can look up to job title and figure it out. Salary is written on top of the bar to make it simpler to get the exact salary value.



This graph will help people know the salary range for the top 5 roles and make the decision or look up to their dream role.

Visualization 3- Maximum Average Salary of Top 3 Roles in Companies founded between 2010-2019

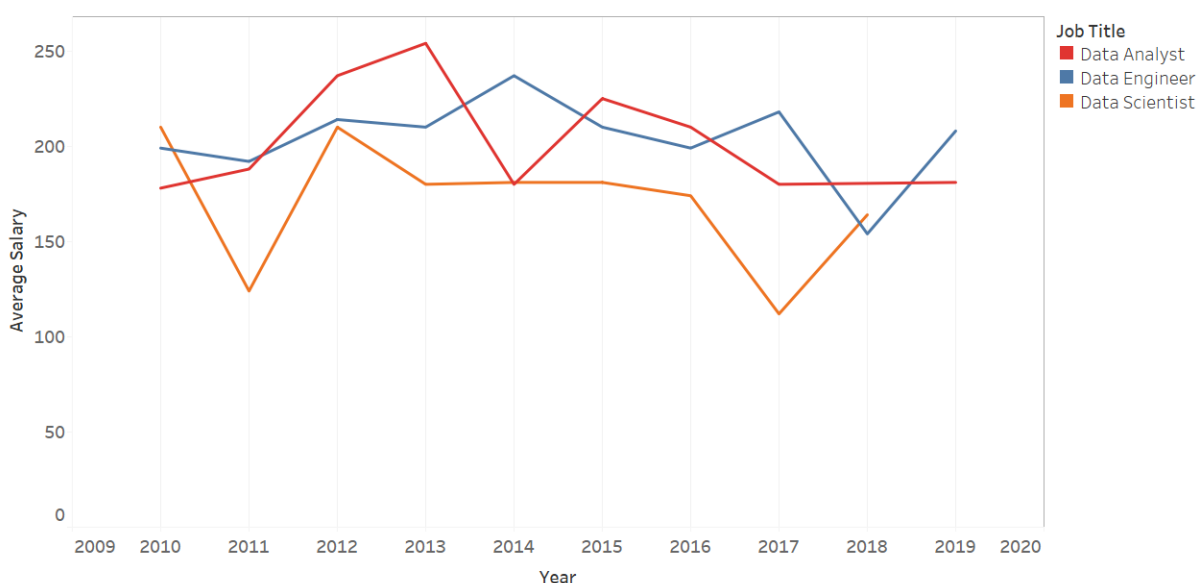
Choice of Chart

In this graph, years were involved. However, using a line graph was a convenient way to visualize data over the years. Dataset included year on which the company was founded. If anyone is struggling to get into old, big, and popular companies, they can see the average salary for a recently established company through this graph. Therefore, a filter was applied on the founded measure and the 2010 to 2019 range was selected. In this visualization, categorical values were supposed to be shown. For that purpose, we created a parameter consisting of Data Analyst, Data Engineer, and Data Scientist, and the same parameter was shown for the user to select their choice of the role which they would like to know about.

Design Choice

Red, Orange, and Blue are bright, which makes them more appealing. We used red for Data Analyst because it is very popular and most talked about among youngsters. However, using the rich, wild, and bright red was matching that role. Data Engineer is the coolest job in the data science branch. Therefore, Blue was the choice for it. We feel that it is tough to get the Data Scientist role and orange is not so common colour we talk about. Thus, used to display Data Scientist's average salary trend. Using the parameter drop-down list will display the average salary for every year with salary labelled and selecting an item from colour legend will highlight the specific job title only.

Maximum Average Salary of Top 3 Roles in Companies founded between 2010-2019



From graph there is one conclusion that data scientist role was not present in any company for the year 2019. The line graph will help job searchers to know the trend of average salary over the years and predict future trends as well. It will help to understand the scope of specific roles in the future.

Visualization 4- Popular Job Title in Cities as per Positive Ratings

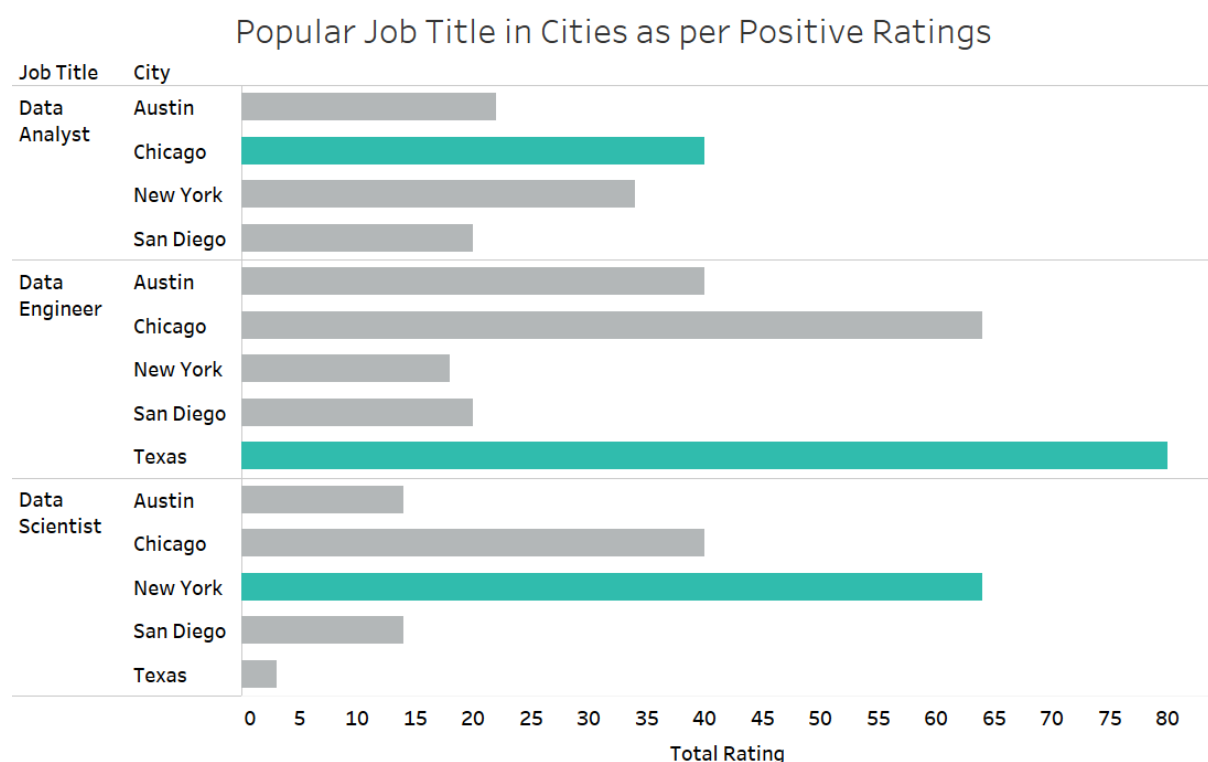
Choice of Chart

For the next visualization categorical values from two different columns were included, and one numerical value so the best option was Bar Graph. Here we represented a horizontal bar graph because just one categorical value is to be determined. We needed the top 3 roles popular in cities, so we filtered cities with a count of job titles. the rating was filtered from 3-5 range which is positive ratings according to us.

Design Choice

On Y-axis we took job title and city (location). The reason we did not present it on X-axis as it was easier to read labels horizontally, that is from Y-axis. On X-axis, we plotted ratings count.

In this visualization, we wanted to show only one city popular for the top 3 job titles. City and job title were grouped so that cities with a maximum rating can only be coloured differently. For that, we used light teal for all top cities to make it visible at one glance and others were given light Gray to make contrast with teal.



Looking at the graph anyone could understand which is the best city for specific role.

Tools and Libraries used

Loading, Cleaning, Processing, and Exploration were done in Google Colab. The Programming language that we used was Python. Libraries used were *pandas* and *numpy*.

Visualization was completely done on the tableau.

ScreenCast- https://drive.google.com/file/d/1wR-AN1HBvMjWr_6ttoYNR6Q15lvoKzSK/view?usp=sharing

Visualization Dashboard- [Profile - kashmira.chawan | Tableau Public](#)

Source Code- [DVM Project Source - Google Drive](#)

Conclusion

After carefully visualizing the data, we can say that there is wide scope for the top 3 data science roles in the USA concerning salary and location that we choose. While dealing with the dataset we wanted to combine files using python but were unable to do it, instead of that we did it locally. We think doing it in python would have made our files integrate properly with no redundancies. There were many NaN values in the dataset and removing that was leading to fewer rows. There would have been some other methods to it which we were incapable to achieve.

We both stay on-campus, in the same Postgrad apartment blocks so we decided to do it together as it would be better to go late nights at each other's place to do an assignment. Regarding the project, the coding part was done by both of us to improve code and dataset. It was our first-time using tableau so we both were keen to learn it. We decided to take up 2 graphs each and that is how we completed our project.

References

Dataset links:

[Data Analyst Jobs | Kaggle](#)

[Data Science Position Analysis | Kaggle](#)

[Data Engineer Jobs | Kaggle](#)

An article we came across before beginning our assignment work:

[Top 9 Job Roles in the World of Data Science for 2021 \(mygreatlearning.com\)](#)

For Bar Graph:

<https://youtu.be/z5CcOoTYDxM>