# Data Science Job Trend Analysis

## INFO-I 590 Data Visualization

Dr Aha

Luddy School of Informatics, Computing and Engineering
Indiana University

RAHUL GATTU
NEERAJA KATHA
DHANUSH CHANDRA RAJU

## Table of Contents

# Abstract

With the increase in technology, data has become so powerful and valuable. Many Organizations and governments have been collecting data, maintaining separate servers for them and in the cloud. This data is gathered from across the globe. Data Helps us correct problems in real time. Gives an opportunity to the companies to plan their future and what their focus must be on. With the increase in demand on data, Data Science has become much popular, and jobs related to it are much in demand. Employment of data scientists is projected to grow 36 percent from 2021 to 2031, much faster than the average for all occupations.

About 13,500 openings for data scientists are projected each year, on average, over the decade. Many of those openings are expected to result from the need to replace workers who transfer to different occupations or exit the labor force, such as to retire. One result of the demand for data science talent is the high wages awarded to those who are hired. While the mean salary for a data scientist in the U.S. is $108,660, according to the BLS. But in the San Jose metro area—which also has the highest concentration of data scientist jobs in the U.S.—the mean salary for data scientists is $157,110.

*Keywords*: Data Science, Employment, Data, Data Scientist, Salary

# Introduction

## Motivation

Without the expertise of professionals who turn cutting-edge technology into actionable insights, Big Data is nothing. Today, more and more organizations are opening up their doors to big data and unlocking its power—increasing the value of a data scientist who knows how to tease actionable insights out of gigabytes of data. It's become a universal truth that modern businesses are awash with data. Last year, McKinsey estimated that big data initiatives in the US healthcare system "could account for $300 billion to $450 billion in reduced healthcare spending or 12 to 17 percent of the $2.6 trillion baselines in US healthcare costs". On the other hand, though, bad data is estimated to be costing the US roughly $3.1 trillion a year.

A major change in data science over the past decade is that the need for an ethical dimension to the field is now widely acknowledged, though the topic was rarely mentioned in 2012. The turning point for data science ethics was probably the 2016 U.S. presidential election, in which data scientists in social media (Cambridge Analytica and Facebook in particular) attempted to influence voters and further polarized electoral politics. Since that time, considerable attention has been devoted to issues of algorithmic bias, transparency, and responsible use of analytics and AI.

Some companies have already established responsible AI groups and processes. A key function of them is to educate data scientists about the issues involved in ethical AI. And there is an increased regulation that is being instituted in response to ethical lapses.

## Background

Ten years ago, we published the article "Data Scientist: Sexiest Job of the 21st Century." Most casual readers probably remember only the "sexiest" modifier — a comment on their demand in the marketplace. The role was relatively new at the time, but as more companies attempted to make sense of big data, they realized they needed people who could combine programming, analytics, and experimentation skills. At the time, that demand was largely restricted to the San Francisco Bay Area and a few other coastal cities. Startups and tech firms in those areas seemed to want all the data scientists they could hire. We felt that the need would expand as mainstream companies embraced both business analytics and new forms and volumes of data.

At the time, we defined the data scientist as "a high-ranking professional with the training and curiosity to make discoveries in the world of big data." Companies were beginning to analyze voluminous and less-structured data like online clickstreams, social media, and images and speech. Because there wasn't yet a well-defined career path for people who could program with and analyze such data, data scientists had diverse educational backgrounds. The most common qualification in our informal survey of 35 data scientists at the time was a PhD in experimental physics, but we also found astronomers, psychologists, and meteorologists. Most had PhDs in some scientific field, were exceptional at math, and knew how to code. Given the absence of tools and processes at the time to perform their roles, they were also good at experimentation and invention. It's not that a science PhD was really required to do the work, but rather that these individuals had the rare ability to unlock the potential of data, wading through complex, messy data sets and building recommendation algorithms.

A decade later, the job is more in demand than ever with employers and recruiters. AI is increasingly popular in business, and companies of all sizes and locations feel they need data scientists to develop AI models. By 2019, postings for data scientists on Indeed had risen by 256%, and the U.S. Bureau of Labor Statistics, predicts data science will see more growth than almost any other field between now and 2029. The sought-after job is generally paid quite well; the median salary for an experienced data scientist in California is approaching $200,000.

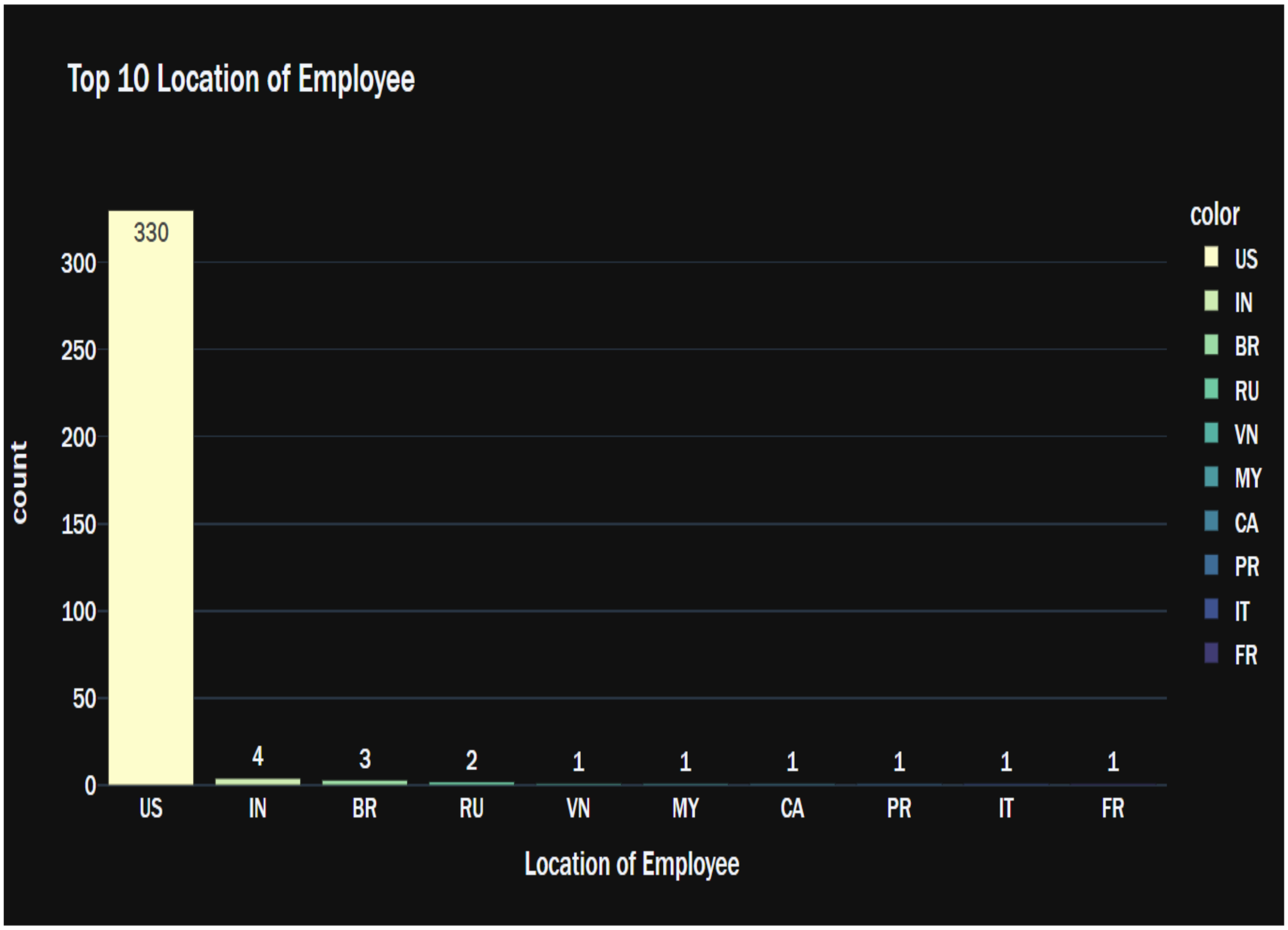
# Objective

With the current market scenario, we have come to know that there are spots filling out for the data scientist roles, when we take COVID-19 into consideration, the number of positions that are being offered had a huge surge since. So, when we take the situation of Graduates and Under Graduates or any job seekers in the field of Data Science. There are many positions in the area of Data Science, only few get popular like Data Scientist, Data Analyst, Machine Learning Engineer, Data Engineer etc. the positions which people know are very much less when compared to what the field is open to. We would like to show the Designations, Median Salary, type of employment, remote working ratio, company size, employment residence, experience.

The outcome we expect of this data set is, showing the job trend in the field of Data Science and the number of positions that are offered for different employment experience. The main objective of this paper is to analyze the data given by ai-jobs, and to visualize the different roles and salaries that are paid to the designations and many other aspects as well.

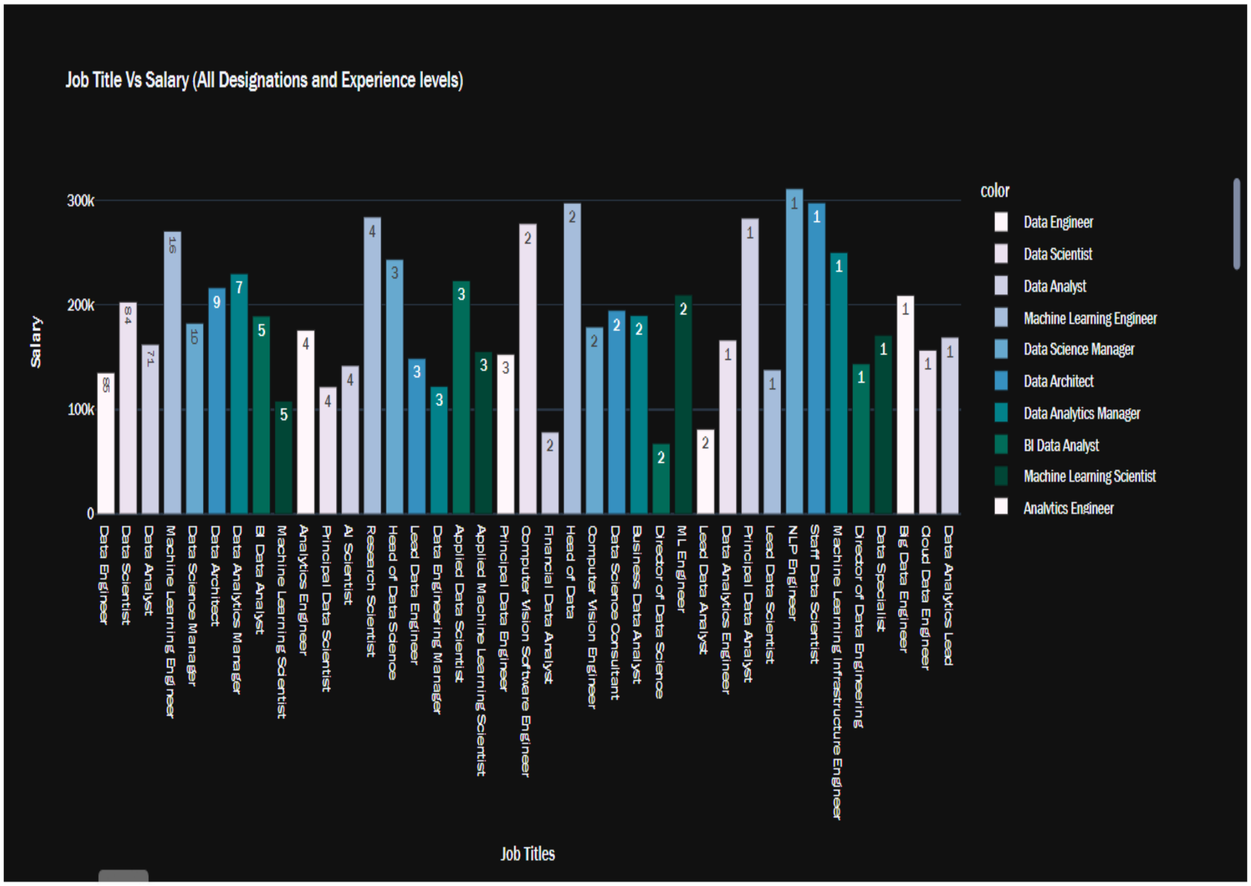
# Existing Visualizations

## Visualizations -1



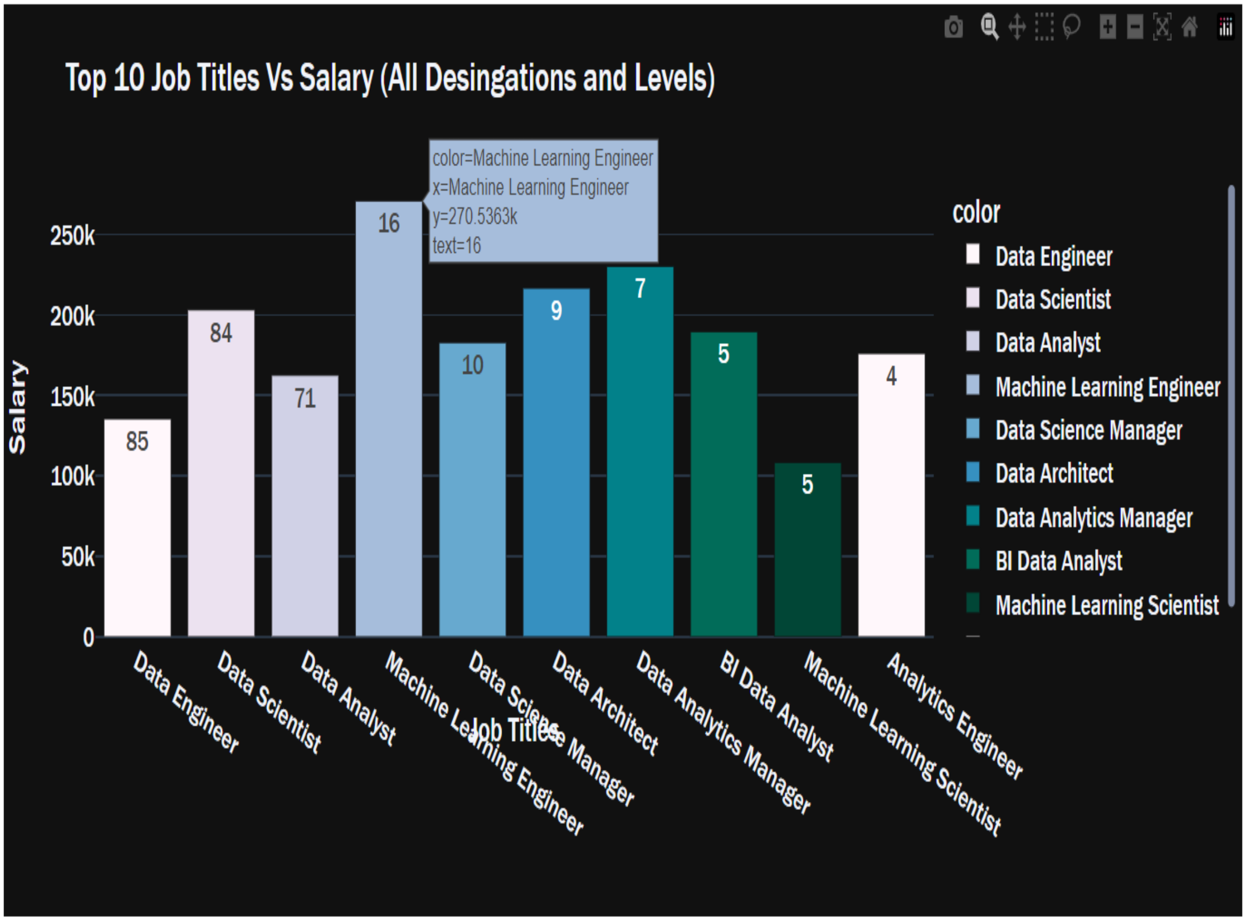
The above visualization states the top 10 locations of the employees across the world. However, a lot of employees reside in United States. This can be because a lot of companies are in the US and from the dataset employees residing in other locations are working remote.

## Visualizations -2



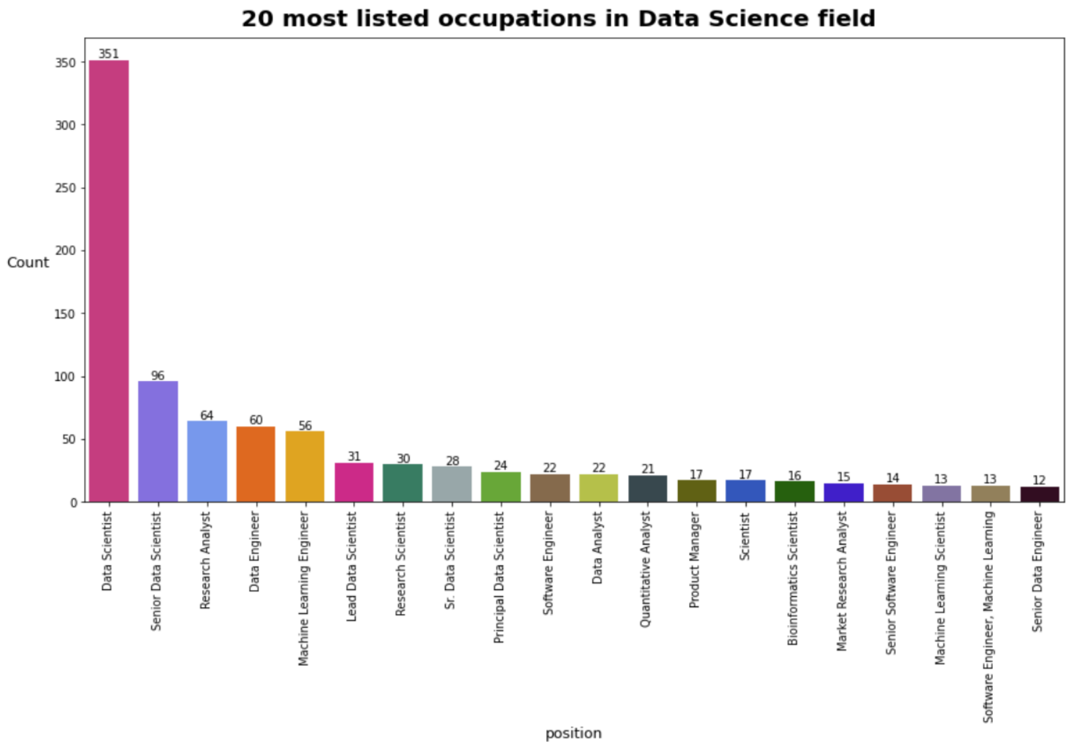
The above visualization shows the Job title vs Salary of all the designations and experience levels. X-axis shows the Job titles and Y-axis shows the mean salary information of the job titles.

## Visualizations -3



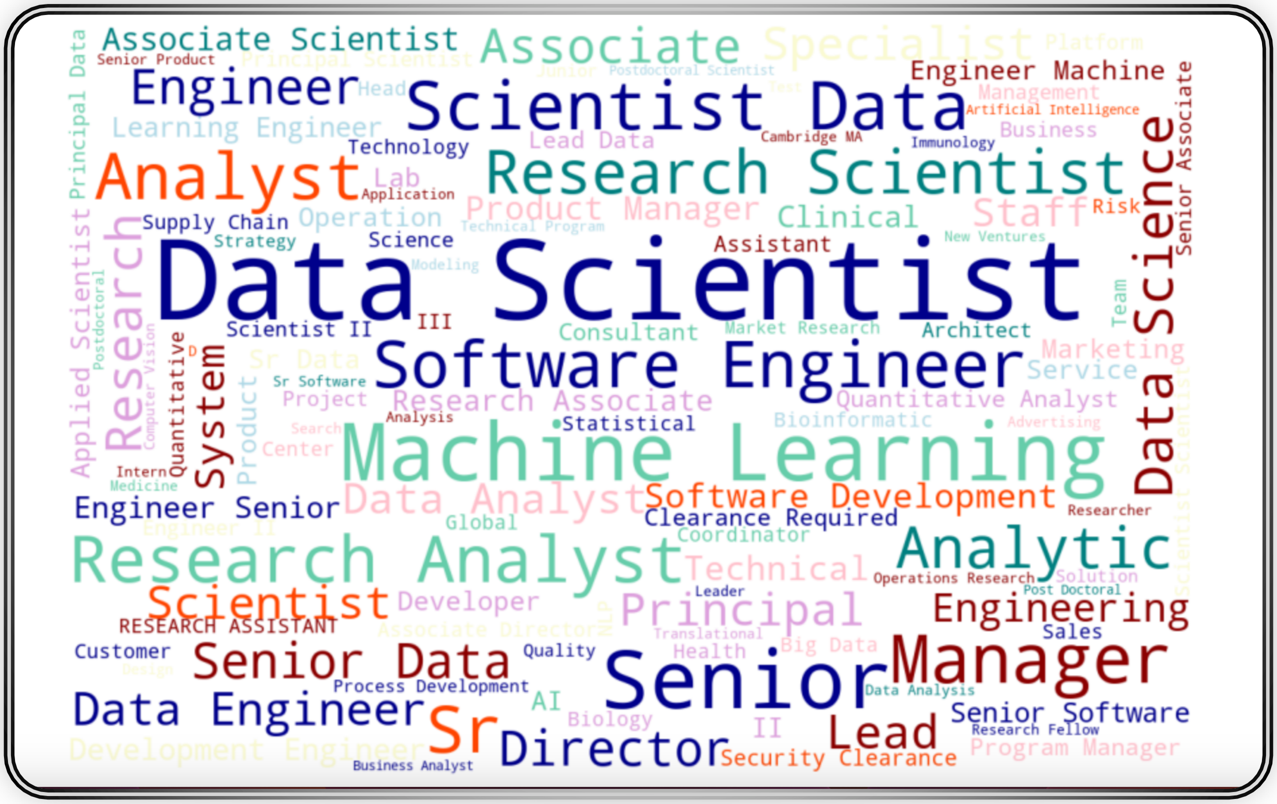
The Visualization shows the Top 10 Job titles that are paying highest mean Salary. On X-axis we see the job titles and on Y-axis we see the mean salary information. This is the interactive visualization created using the plotly library.

## Visualizations -4



The above visualization shows the 20 most occupations in the Data Science field. X-axis shows the positions and count on the Y-axis. Out of top 20 listed occupations Data Scientist role has the highest number of positions and Senior Data Engineer has the lowest number of positions.

Visualizations -5



The word cloud shows all the Job roles in the Data Science field.

- A lot of the existing visualizations are bar graphs and word cloud.

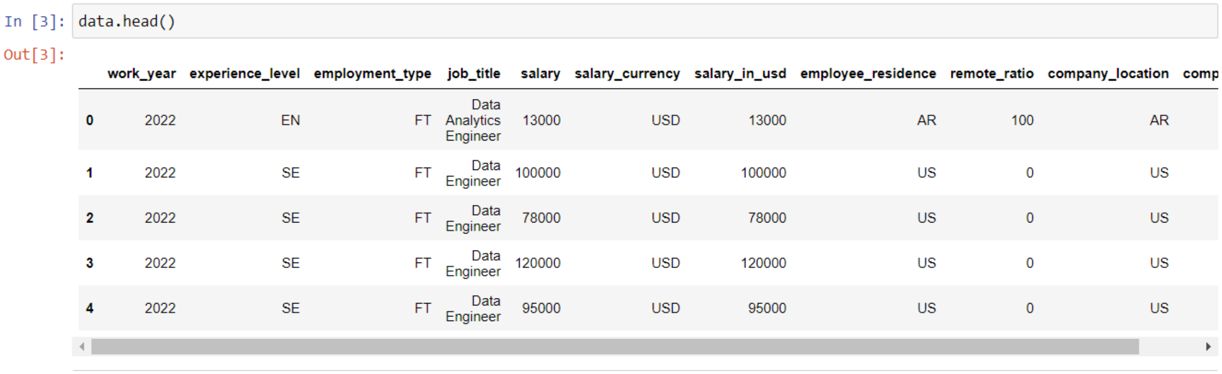
# Data Sources and description

| | | |
|---|---|---|
| **work_year** | The year the salary was paid. | |
| | | |
| **experience_level** | The experience level in the job during the year with the following possible values: | |
| | **EN** | Entry-level / Junior |
| | **MI** | Mid-level / Intermediate |
| | **SE** | Senior-level / Expert |
| | **EX** | Executive-level / Director |
| | | |
| **employment_type** | The type of employement for the role: | |

| | | |
|---|---|---|
| | **PT** | Part-time |
| | **FT** | Full-time |
| | **CT** | Contract |
| | **FL** | Freelance |
| | | |
| **job_title** | The role worked in during the year. | |
| **salary** | The total gross salary amount paid. | |
| **salary_currency** | The currency of the salary paid as an ISO 4217 currency code. | |
| **salary_in_usd** | The salary in USD (FX rate divided by avg. USD rate of respective year via data from BIS). | |
| **employee_residence** | Employee's primary country of residence in during the work year as an ISO 3166 country code. | |
| | | |
| **remote_ratio** | The overall amount of work done remotely, possible values are as follows: | |
| | | |
| | **0** | No remote work (less than 20%) |
| | **50** | Partially remote |
| | **100** | Fully remote (more than 80%) |
| | | |
| **company_location** | The country of the employer's main office or contracting branch as an ISO 3166 country code. | |
| | | |
| **company_size** | The average number of people that worked for the company during the year: | |
| | **S** | less than 50 employees (small) |
| | **M** | 50 to 250 employees (medium) |
| | **L** | more than 250 employees (large) |
| | | |

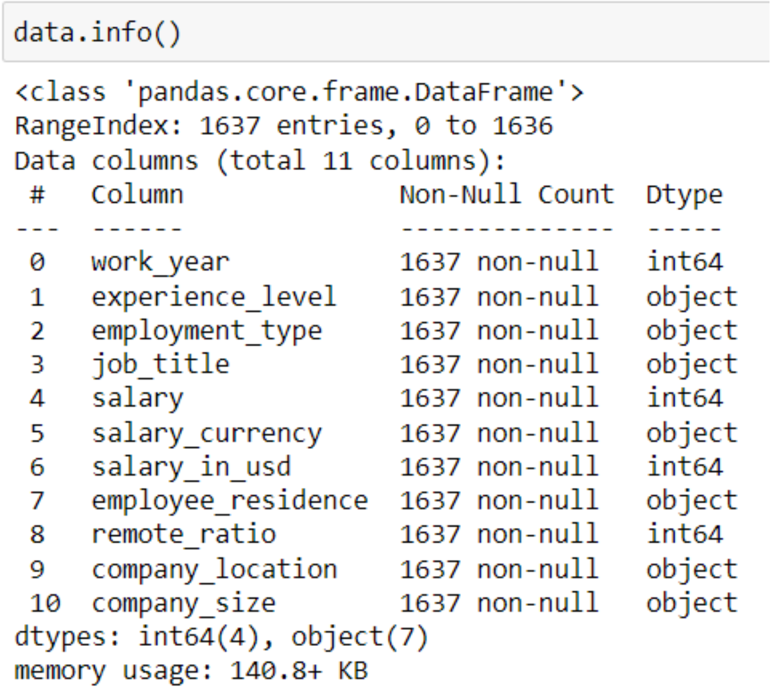
# Results, Insights and methods

## Pre-Processing of the Dataset



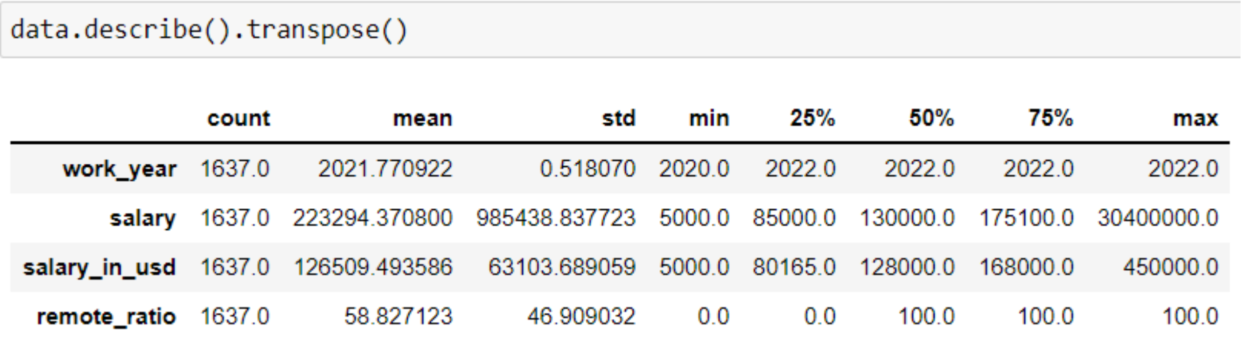
```
In [3]: data.head()
Out[3]:
```

| | work_year | experience_level | employment_type | job_title | salary | salary_currency | salary_in_usd | employee_residence | remote_ratio | company_location | comp |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 2022 | EN | FT | Data Analytics Engineer | 13000 | USD | 13000 | AR | 100 | AR | |
| 1 | 2022 | SE | FT | Data Engineer | 100000 | USD | 100000 | US | 0 | US | |
| 2 | 2022 | SE | FT | Data Engineer | 78000 | USD | 78000 | US | 0 | US | |
| 3 | 2022 | SE | FT | Data Engineer | 120000 | USD | 120000 | US | 0 | US | |
| 4 | 2022 | SE | FT | Data Engineer | 95000 | USD | 95000 | US | 0 | US | |



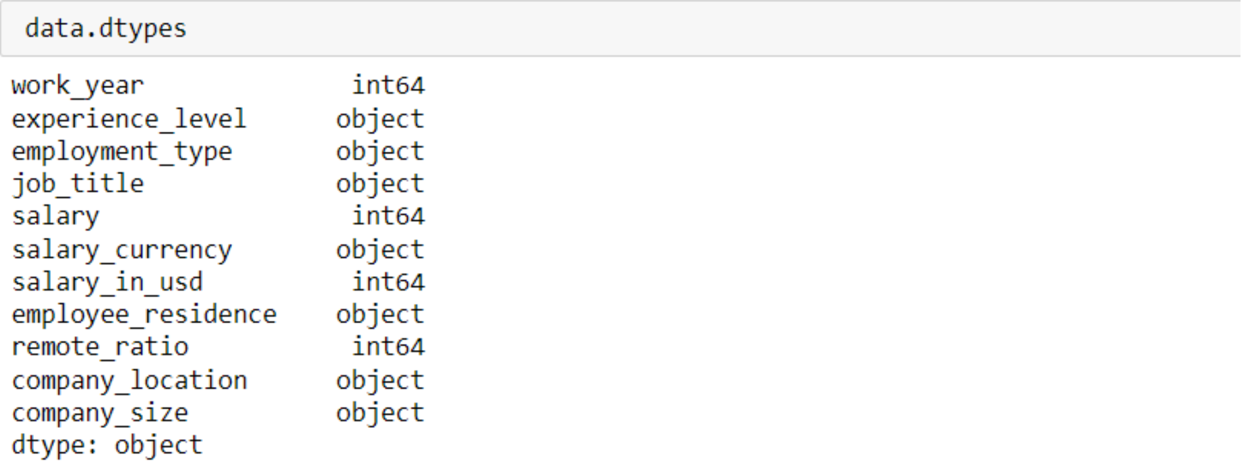
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1637 entries, 0 to 1636
Data columns (total 11 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   work_year           1637 non-null   int64
 1   experience_level    1637 non-null   object
 2   employment_type     1637 non-null   object
 3   job_title           1637 non-null   object
 4   salary              1637 non-null   int64
 5   salary_currency     1637 non-null   object
 6   salary_in_usd       1637 non-null   int64
 7   employee_residence  1637 non-null   object
 8   remote_ratio        1637 non-null   int64
 9   company_location    1637 non-null   object
 10  company_size        1637 non-null   object
dtypes: int64(4), object(7)
memory usage: 140.8+ KB
```



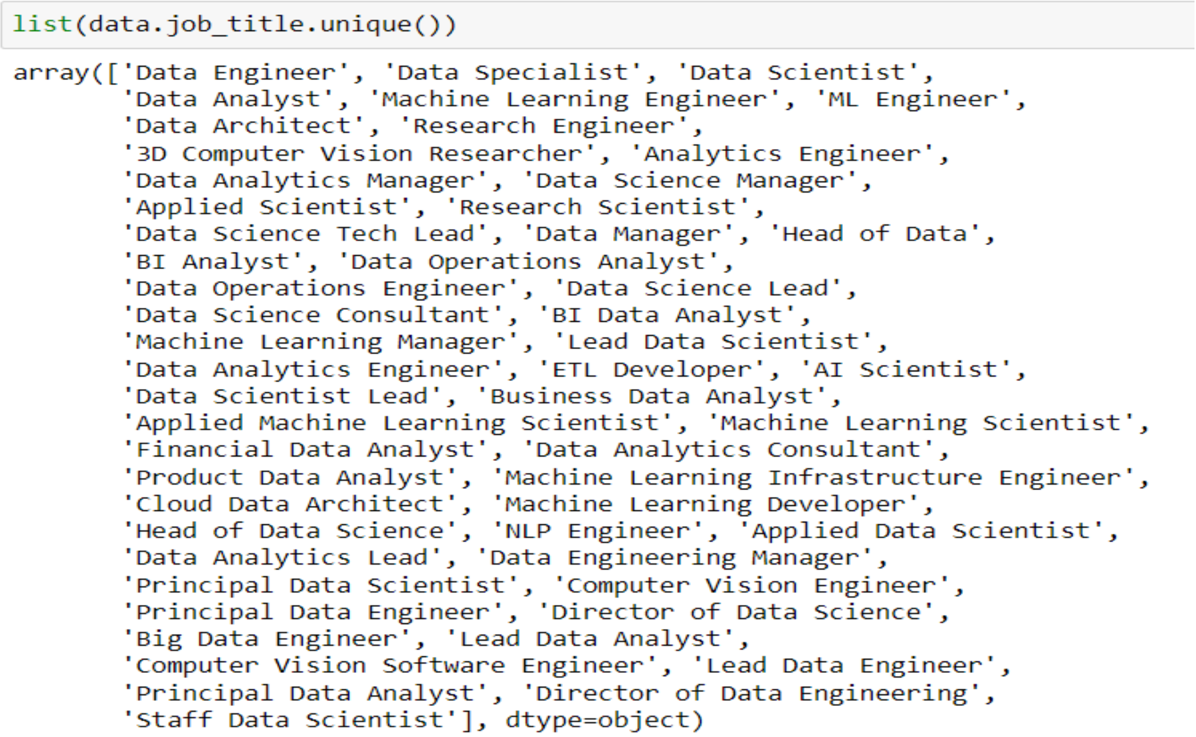
```
data.describe().transpose()
```

|  | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| **work_year** | 1637.0 | 2021.770922 | 0.518070 | 2020.0 | 2022.0 | 2022.0 | 2022.0 | 2022.0 |
| **salary** | 1637.0 | 223294.370800 | 985438.837723 | 5000.0 | 85000.0 | 130000.0 | 175100.0 | 30400000.0 |
| **salary_in_usd** | 1637.0 | 126509.493586 | 63103.689059 | 5000.0 | 80165.0 | 128000.0 | 168000.0 | 450000.0 |
| **remote_ratio** | 1637.0 | 58.827123 | 46.909032 | 0.0 | 0.0 | 100.0 | 100.0 | 100.0 |



```
data.dtypes
```

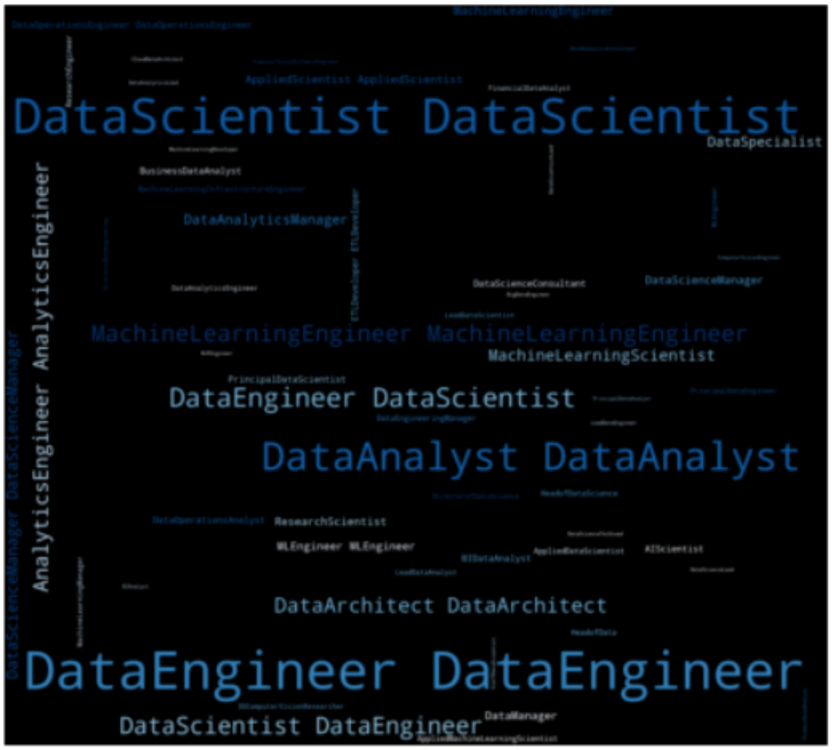
```
work_year              int64
experience_level      object
employment_type       object
job_title             object
salary                 int64
salary_currency       object
salary_in_usd          int64
employee_residence    object
remote_ratio           int64
company_location      object
company_size          object
dtype: object
```



```
list(data.job_title.unique())
```

```
array(['Data Engineer', 'Data Specialist', 'Data Scientist',
       'Data Analyst', 'Machine Learning Engineer', 'ML Engineer',
       'Data Architect', 'Research Engineer',
       '3D Computer Vision Researcher', 'Analytics Engineer',
       'Data Analytics Manager', 'Data Science Manager',
       'Applied Scientist', 'Research Scientist',
       'Data Science Tech Lead', 'Data Manager', 'Head of Data',
       'BI Analyst', 'Data Operations Analyst',
       'Data Operations Engineer', 'Data Science Lead',
       'Data Science Consultant', 'BI Data Analyst',
       'Machine Learning Manager', 'Lead Data Scientist',
       'Data Analytics Engineer', 'ETL Developer', 'AI Scientist',
       'Data Scientist Lead', 'Business Data Analyst',
       'Applied Machine Learning Scientist', 'Machine Learning Scientist',
       'Financial Data Analyst', 'Data Analytics Consultant',
       'Product Data Analyst', 'Machine Learning Infrastructure Engineer',
       'Cloud Data Architect', 'Machine Learning Developer',
       'Head of Data Science', 'NLP Engineer', 'Applied Data Scientist',
       'Data Analytics Lead', 'Data Engineering Manager',
       'Principal Data Scientist', 'Computer Vision Engineer',
       'Principal Data Engineer', 'Director of Data Science',
       'Big Data Engineer', 'Lead Data Analyst',
       'Computer Vision Software Engineer', 'Lead Data Engineer',
       'Principal Data Analyst', 'Director of Data Engineering',
       'Staff Data Scientist'], dtype=object)
```

## Word Cloud



This is the Word Cloud of all the Job titles in the Data Science fields. Job titles with the Number of jobs available in the market with this dataset.

# Insights

## Insight -1

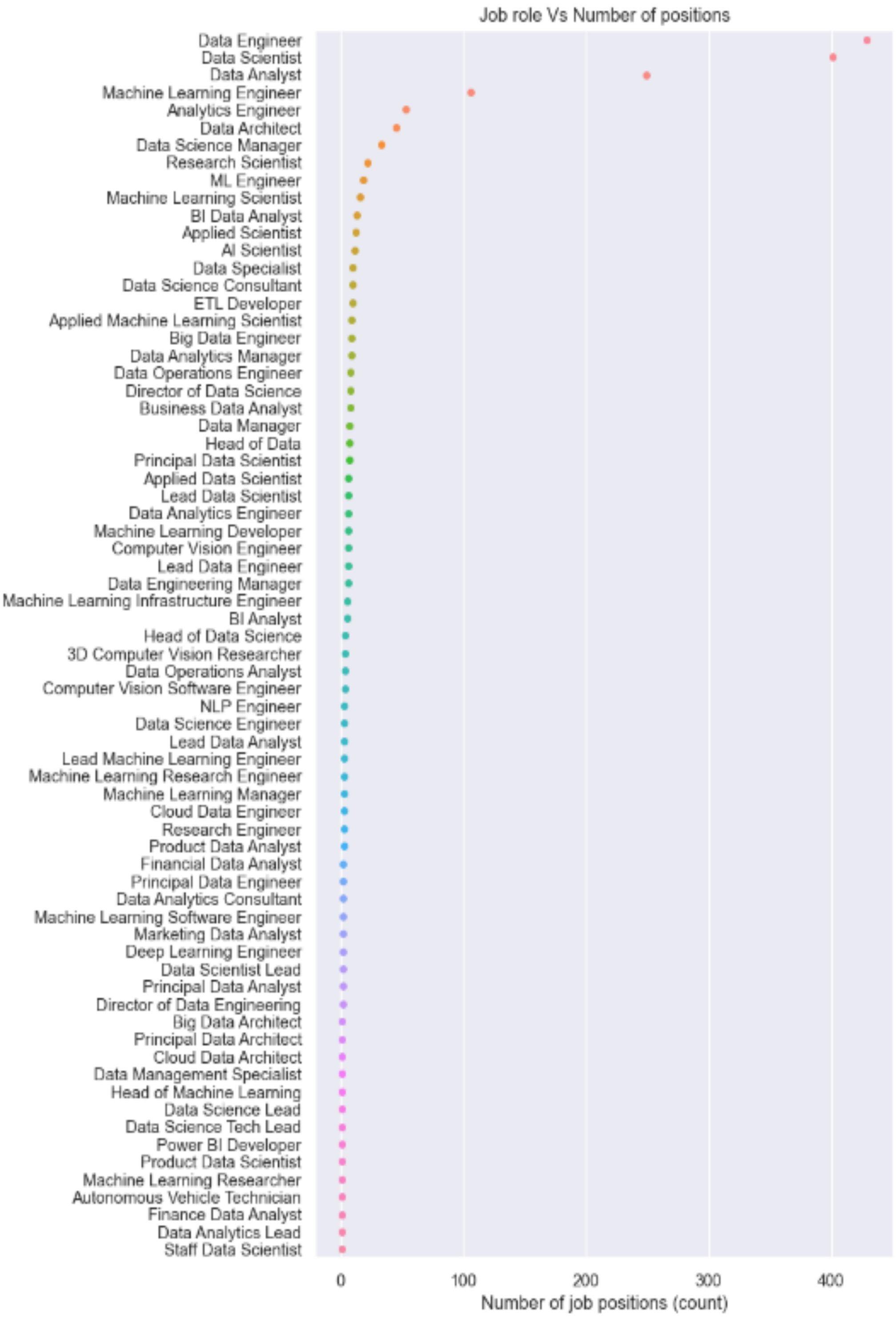
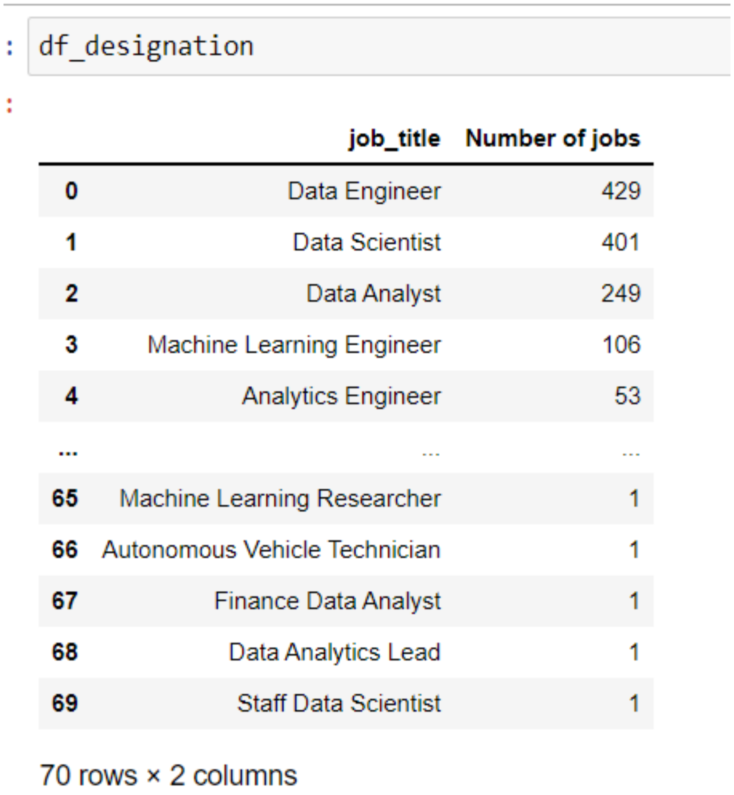


Figure shows the number of Jobs available for the title in the market in all the roles in the United States.

### Insight -2

There are 70 job roles in the Data Science domain including all the experience levels.

From the data set mid-level experience has more positions than senior or junior levels, and executive roles have very few positions and number of jobs in the market.



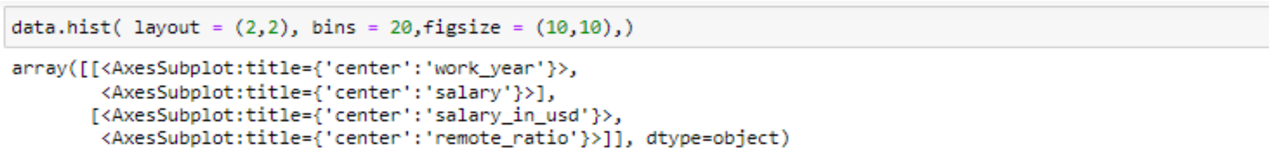
```
: df_designation
```

:

| | job_title | Number of jobs |
|---|---|---|
| **0** | Data Engineer | 429 |
| **1** | Data Scientist | 401 |
| **2** | Data Analyst | 249 |
| **3** | Machine Learning Engineer | 106 |
| **4** | Analytics Engineer | 53 |
| **...** | ... | ... |
| **65** | Machine Learning Researcher | 1 |
| **66** | Autonomous Vehicle Technician | 1 |
| **67** | Finance Data Analyst | 1 |
| **68** | Data Analytics Lead | 1 |
| **69** | Staff Data Scientist | 1 |

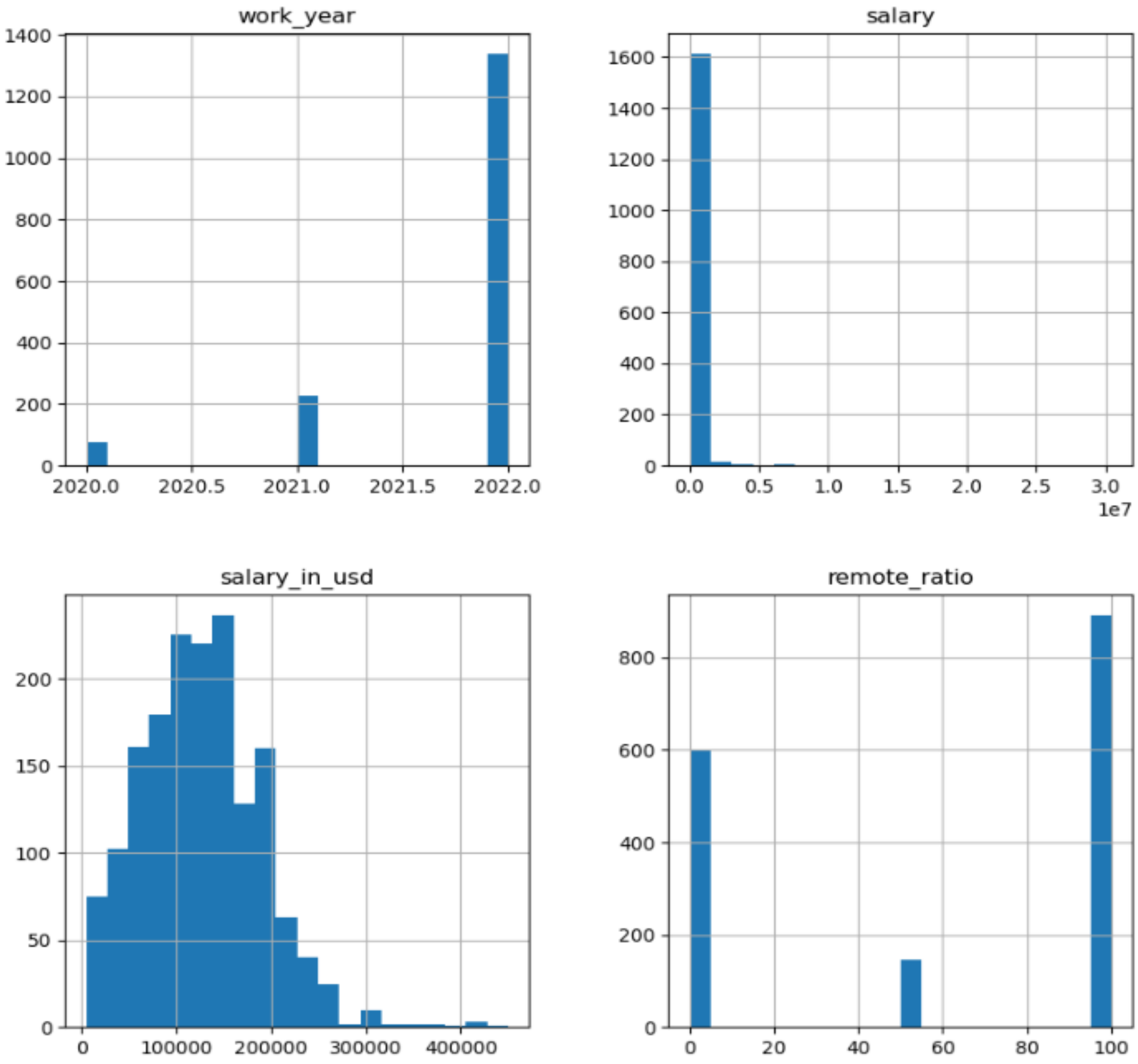
70 rows × 2 columns

## Insight -3



```
data.hist( layout = (2,2), bins = 20,figsize = (10,10),)
```

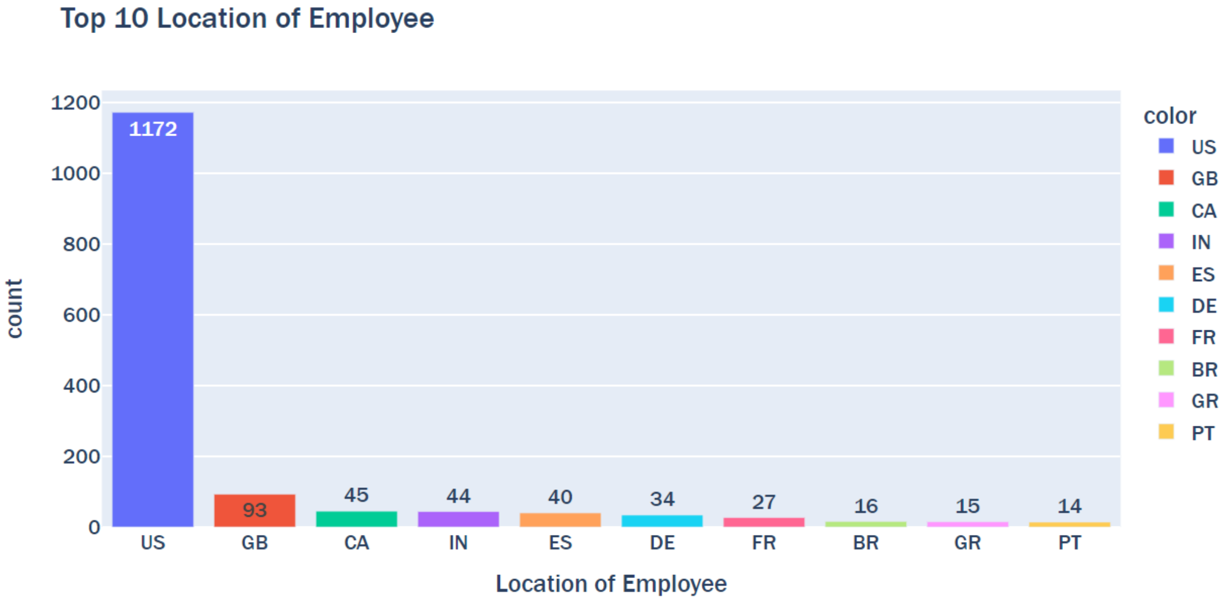
```
array([[<AxesSubplot:title={'center':'work_year'}>,
        <AxesSubplot:title={'center':'salary'}>],
       [<AxesSubplot:title={'center':'salary_in_usd'}>,
        <AxesSubplot:title={'center':'remote_ratio'}>]], dtype=object)
```



The histogram above shows the distribution of the data.

- salary_in_usd is distributed mainly between 100k and 200k.
- The remote_ratio is more at 100 ie., we have more remote jobs than the hybrid or in-office.

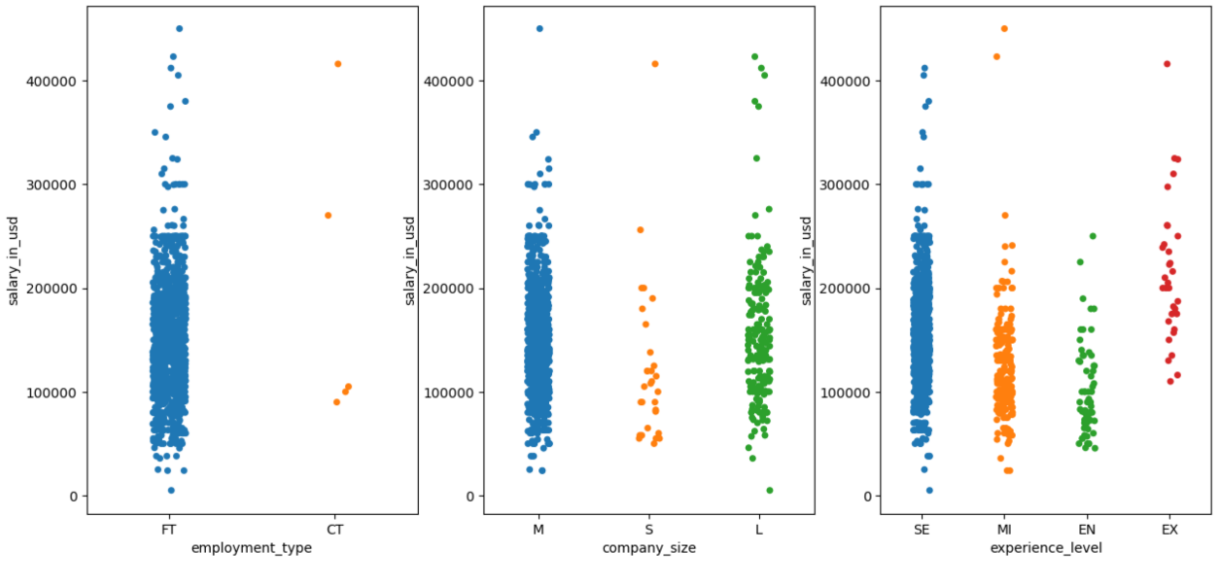
## Insight -4



- The Bar graph showing the top 10 locations of the employees.
- X-Axis showing the location of the employees and
- Y-Axis showing the count (# of jobs in the location)

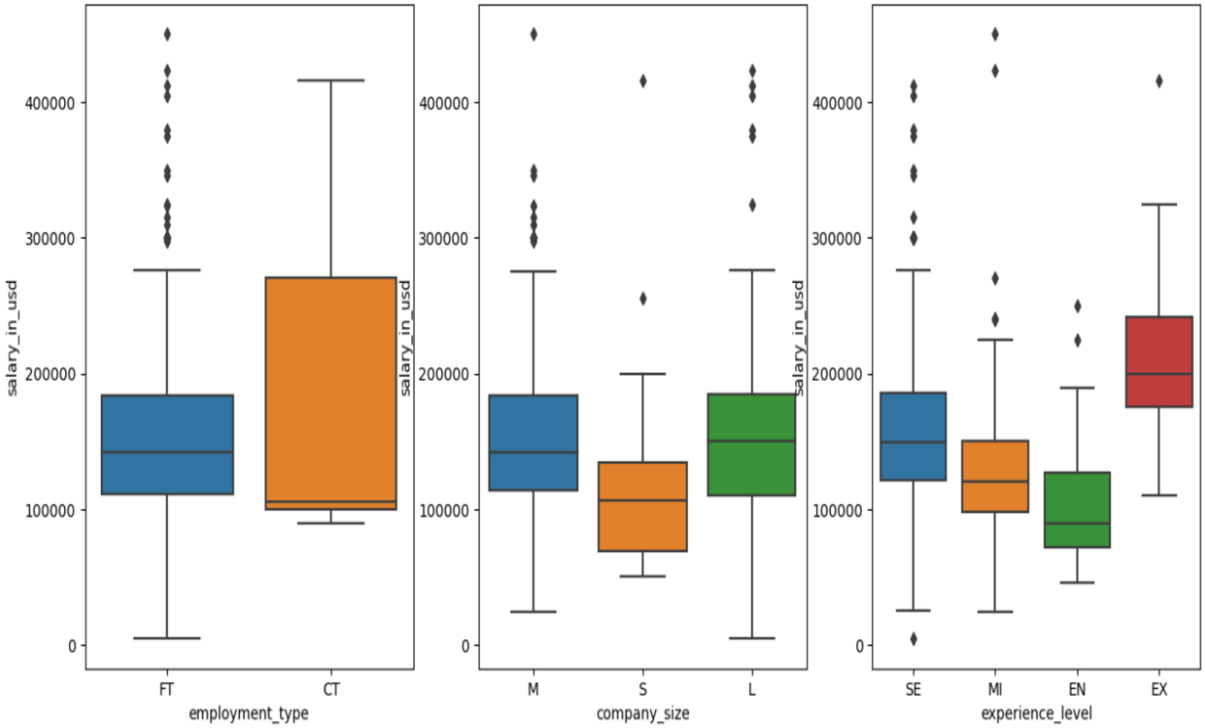
## Insight -5

```
<AxesSubplot:xlabel='experience_level', ylabel='salary_in_usd'>
```

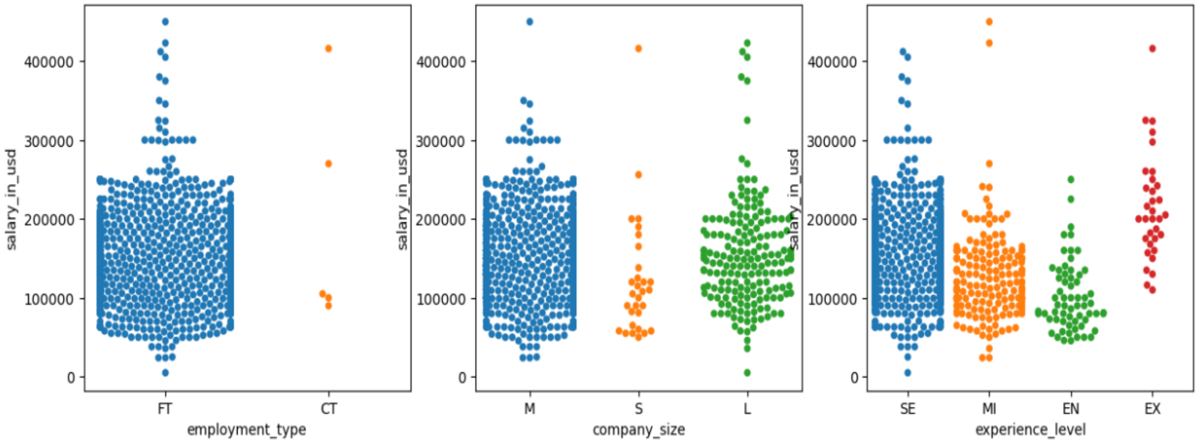


- The Strip chart showing the salary_in_usd in y-axis and
- employee_type, company_size and experience_level on the x-axis.
- In this strip chart we still couldn’t figure out the exact density of the data points.

Therefore, we used box plot to get the five number summary of the data points.



- We observe that in employment_type vs salary_in_usd we see a lot of outliers in FT job and in CT we observe that 25 percentile value is close to the 50-percentile value.
- We see quite a lot of outliers in box plot with company_size on x-axis and experience_level on x-axis.
- We are also visualizing the same data on the bee swarm plot.

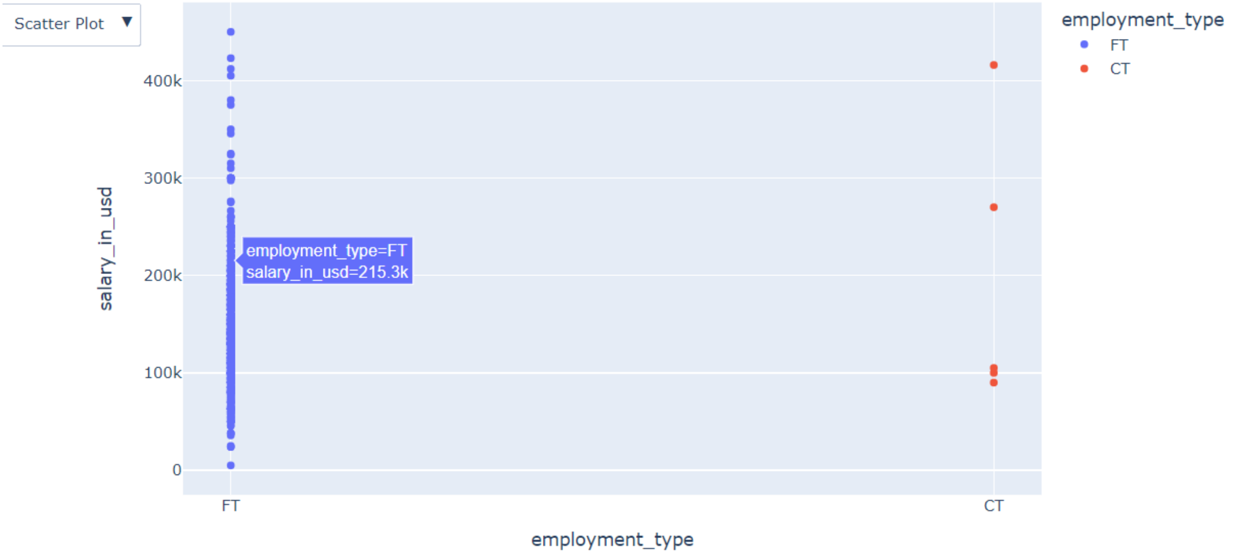


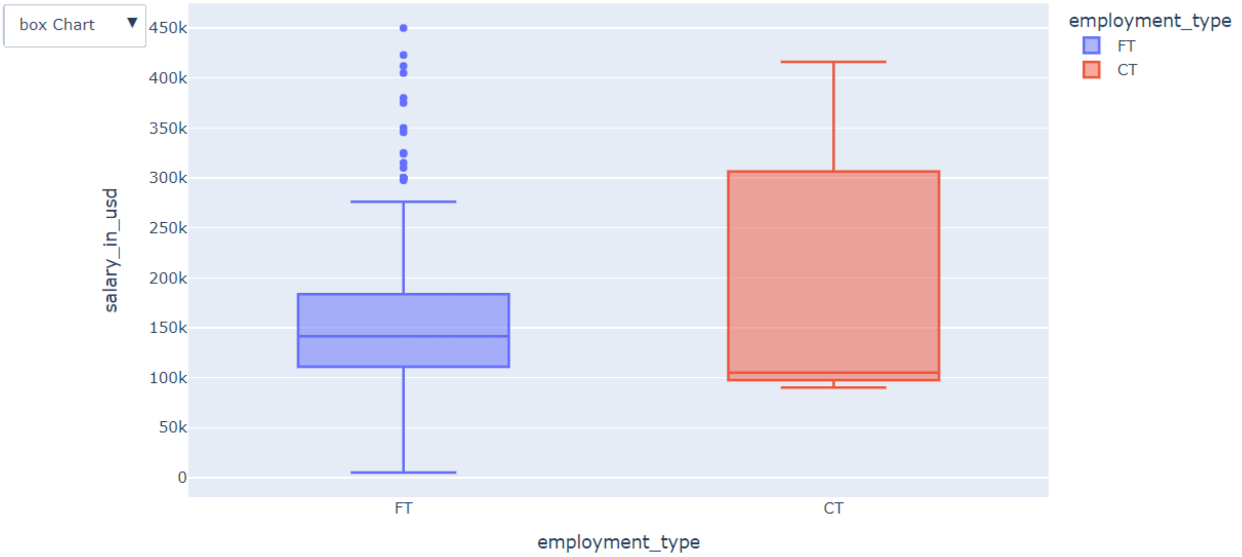
Here, we observe that in above graph. Data do not have the exact band width to fit. After a certain band width, we are not able to visualize the data points. Similarly in the other two plots as well.

## Insight -6

In the following plot we can get both the distribution and statistics of the data points.

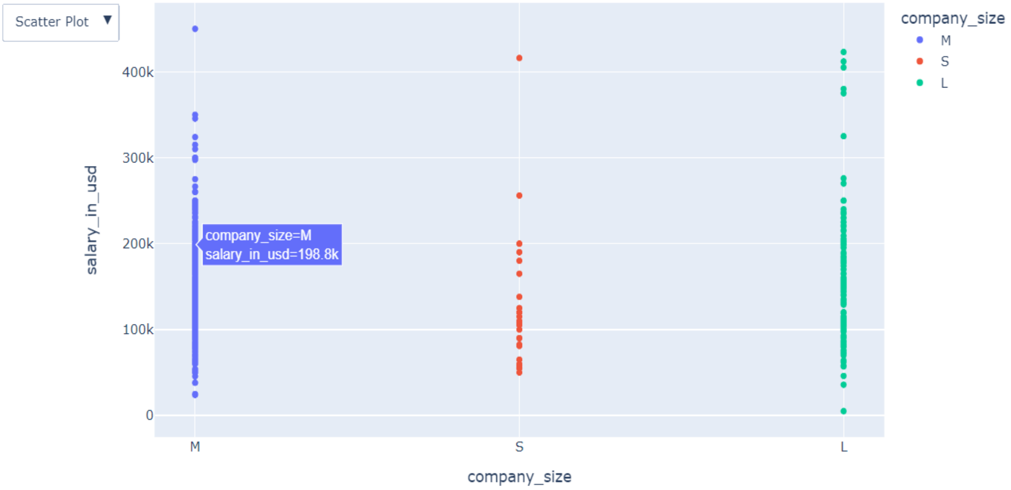
Below are some of the interactive visualizations of same data on different plots.

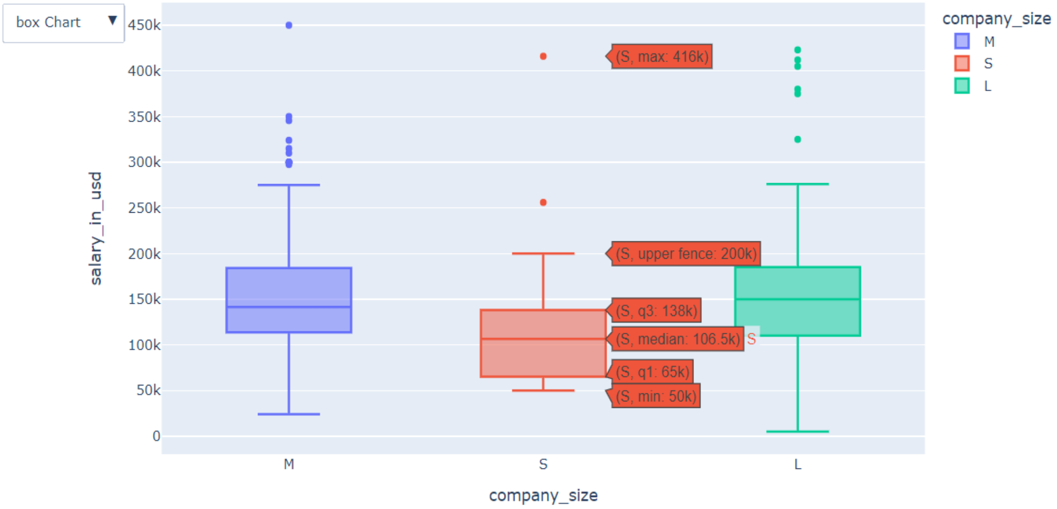




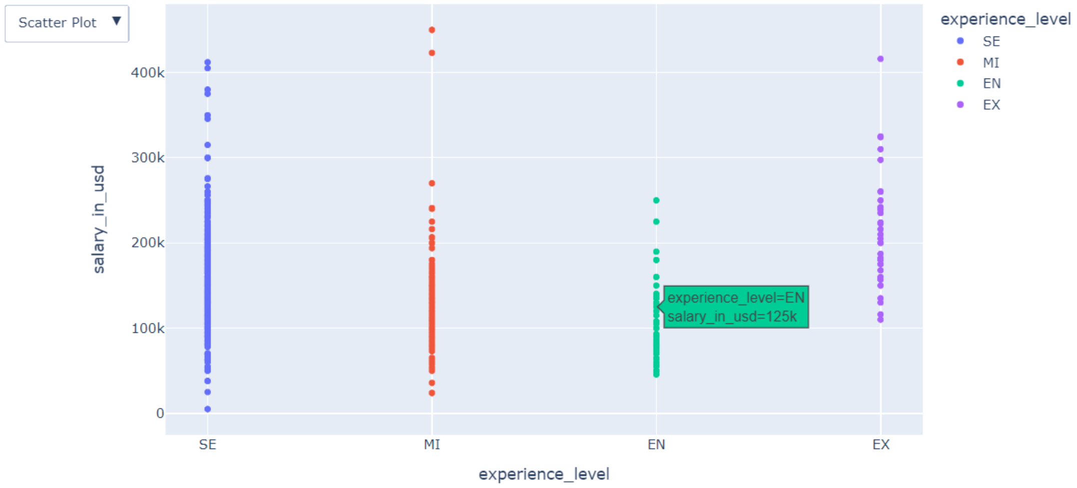
In the above graphs we plotted the same data in three different visualizations. We can see there are a lot of data points between 100k to 200k, but we do not see other statistics.

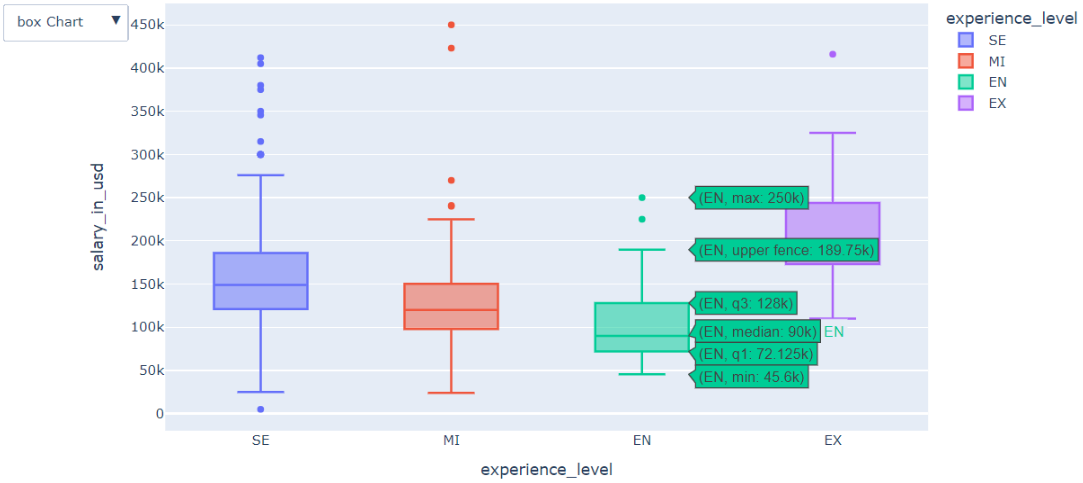
Similarly, we are visualizing company_size on x-axis and salary_in_usd on y-axis.





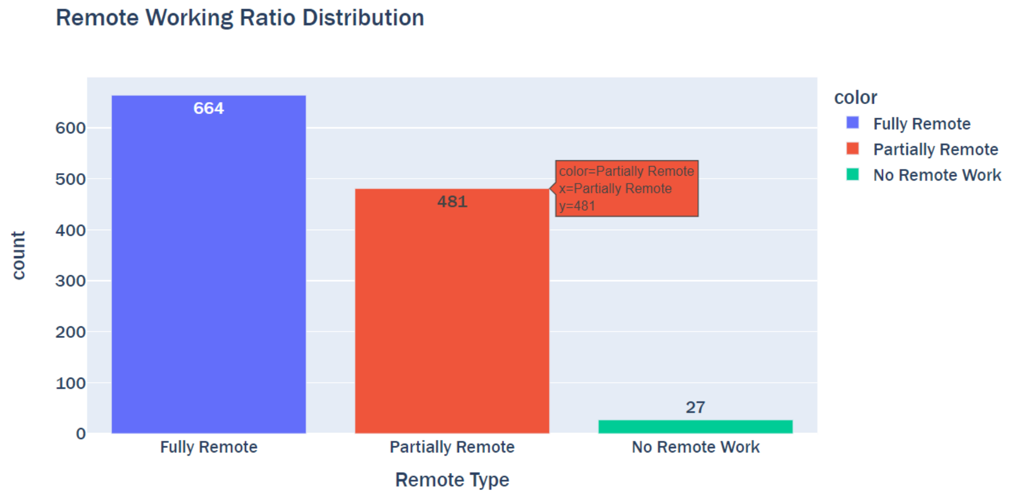
Different visualizations of experience levels on x-axis vs salary on y-axis





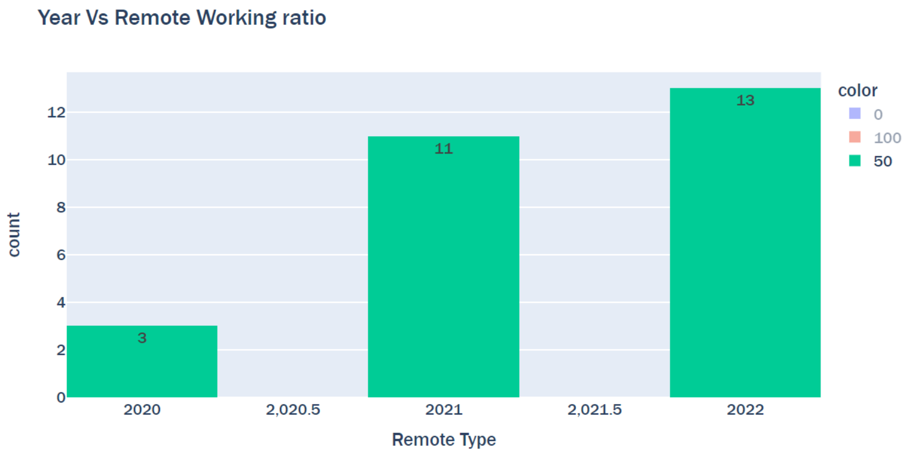
## Insight -7

Visualization showing data distribution of remote type and count (# of available positions)

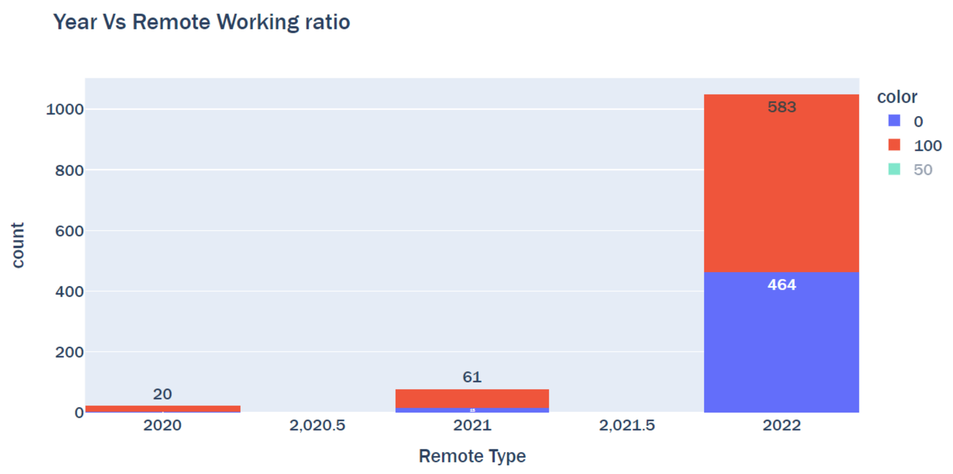


- This is the interactive bar plot between remote type and number of jobs.
- From this we observe that there are more fully remote positions than ‘no remote work’ positions.

The following plot is type of remote positions vs count of it.

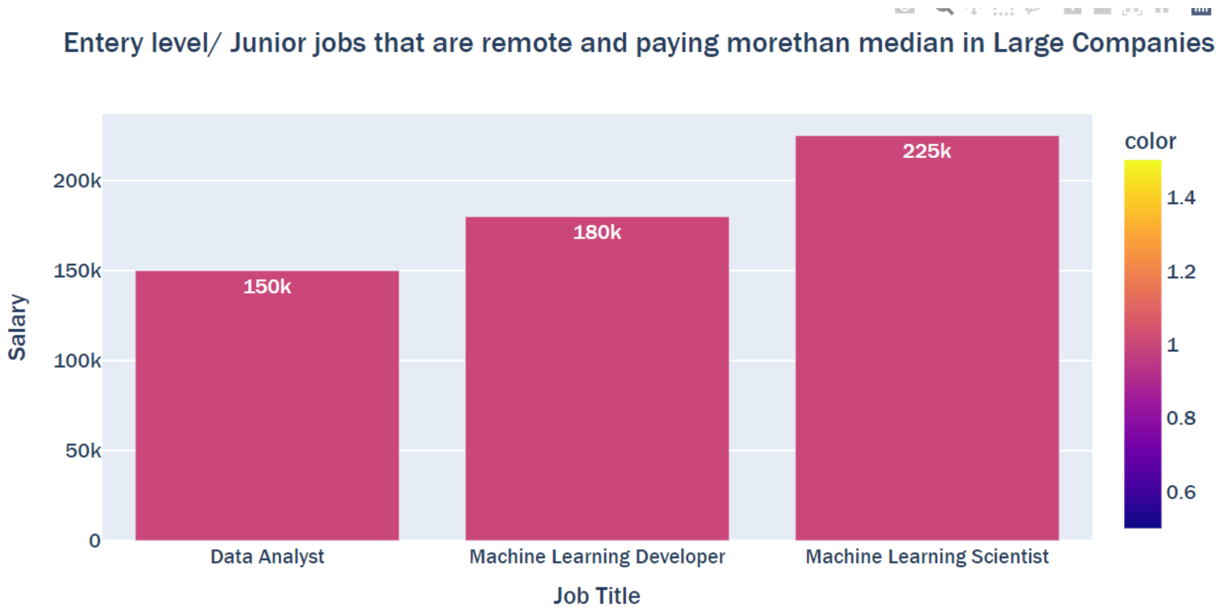


Here we are visualizing fully remote positions and no remote together in number of years and count of it.

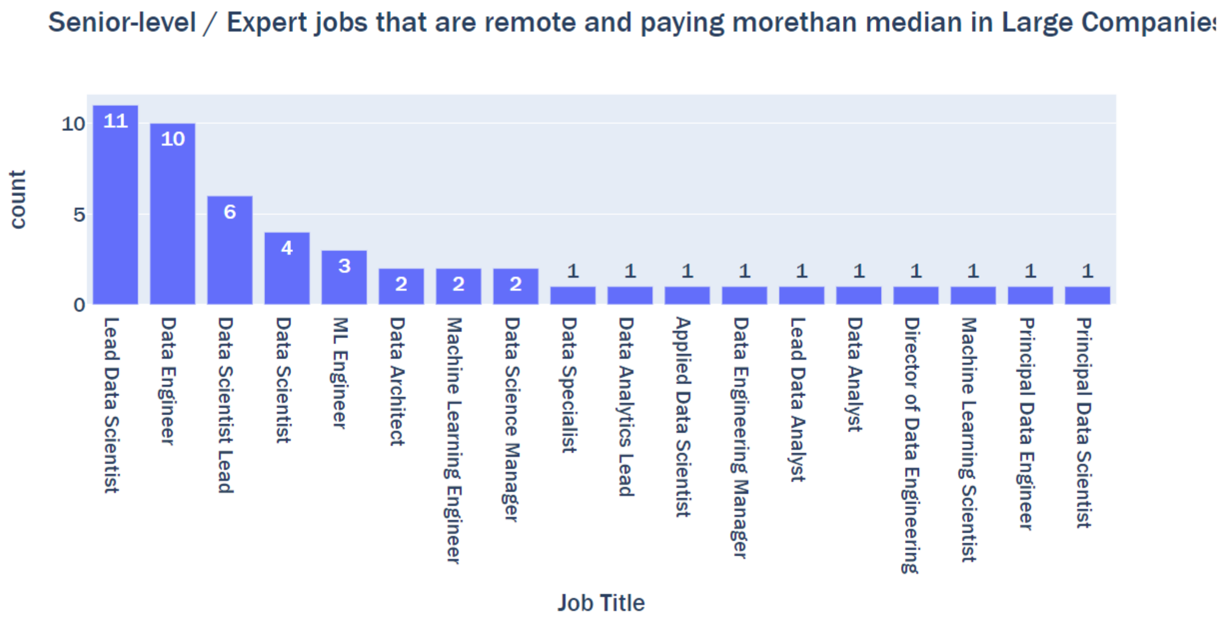


Insight -8

This is the graph for entry level jobs that are remote and paying more than median salary in large companies.



This is the graph for senior level jobs that are remote and paying more than median salary in large companies.

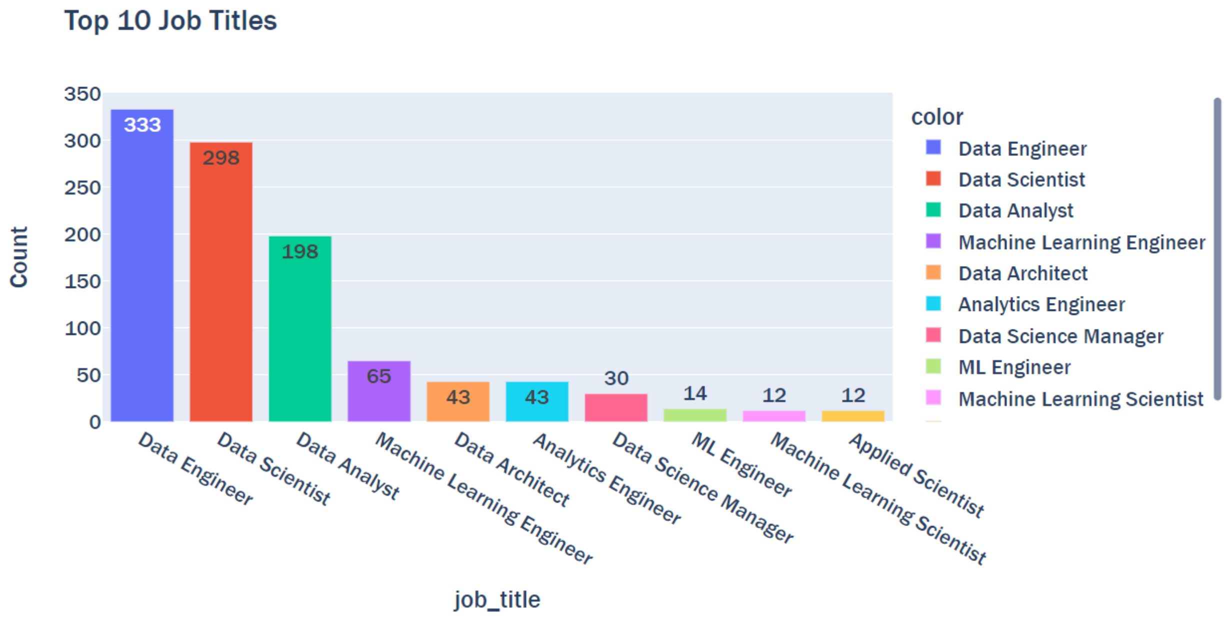


Senior-level / Expert jobs that are remote and paying morethan median in Large Companie:

In the above visualizations top one is the Job title vs median pay for the job role. And the bottom one is the job title on x-axis and count on the y-axis.

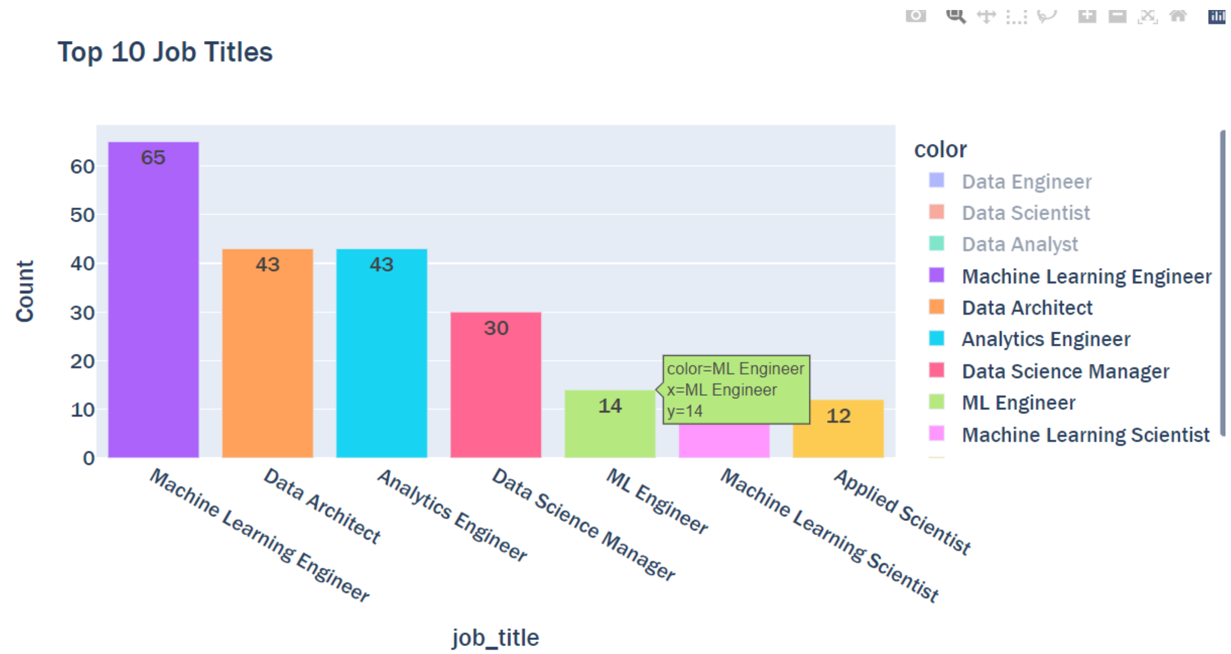
## Insight -9

Top 10 Job titles with respect to availability



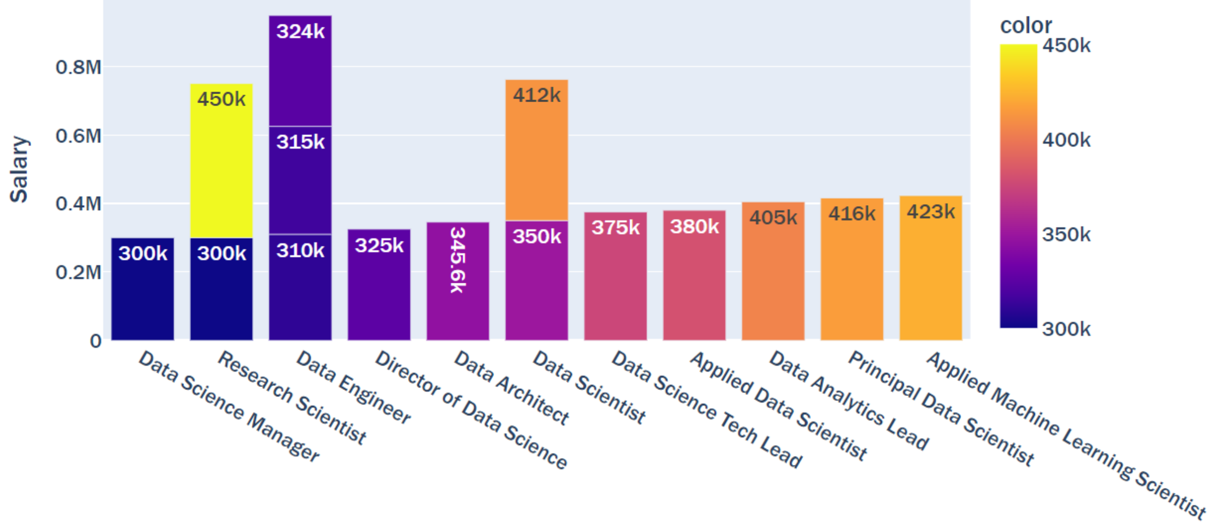
Top 10 Job Titles

This is the visualization with job_title on the x-axis and count on the y-axis. Visualizing top 10 most available jobs in US. We see that the most available job title is Data Engineer, and next is Data Scientist.



Here we are visualizing job_titles with in top 10 job titles excluding top 3 most occurring ones. In this interactive visualization.

In the following visualization we are visualizing some of the high paid job titles in the market.



On x-axis we see job-titles and salary on y-axis. We are distinguishing the titles with different colors using the color. This color changes with respect to the salary distribution.

# Conclusion

- The pay range for most of the jobs is between 100k and 200k.
- Remote positions from 2020 to 2022 have increased drastically with the Covid situation. This has helped the employees to work from anywhere place.
- Full time positions are comparatively more than the other Job Types.
- Data Scientist roles occupies highest job positions with 31 %.
- More Senior lever positions than the Executive level positions.
- Only Full-time positions are available at Executive roles.
- Senior level employees get less salary than mid-level employees.

# Future Work

- Work on large datasets related to Data Science Jobs.
- Hosting this and making our dashboard as open source and work with the live data.
- We also would like to consider the various other data sources if available in the market.
- We will try to sync our dashboard with the glassdoor data and will expand to various other domains and departments.

# References

- https://ai-jobs.net/salaries/download/
- https://www.kaggle.com/code/loka1282/heatmap-data-science-fields-salary-classification
- https://www.kaggle.com/datasets/whenamancodes/data-science-fields-salary-categorization/code?datasetId=2467136
- https://dash.plotly.com/dash-core-components/dropdown
- https://dash.plotly.com/introduction