MIT-ADT UNIVERSITY
PUNE, INDIA
A leap towards World Class Education

Mini Project

(21MSST321)

on

Data Science Job Salaries Analysis

Submitted
to

School of Engineering & Science
Department of Applied Science and Humanities
as a part of Curriculum

by

Mohan L. Bodake
(MITU21MSDS0006)
M. Sc. Applied Statistics (Data Science)
2022-23

Guided by
Dr. Ashok Kumar
M.Sc. Applied Statistics,
School of Engineering & Science
MIT ADT University, Pune
28th December, 2022

MIT ART, DESIGN AND TECHNOLOGY UNIVERSITY, PUNE (INDIA)

SCHOOL ENGINEERING & SCIENCE

DEPARTMENT OF APPLIED SCIENCE AND HUMANITIES

M.SC.APPLIED STATISTICS (DATA SCIENCE)

CERTIFICATE

This is to certify that the work incorporated in the mini project entitled "DATA SCIENCE JOB SALARIES ANALYSIS" Submitted by Mr. Mohan L. Bodake of Second Year M.Sc. Applied Statistics (Data Science) has satisfactorily completed the mini project for the academic year 2022-23 as per the university rules.

Dr. Ashok Kumar **Supervisor** M.Sc. Applied Statistics	**Dr. Pratibha Jadhav** Program Coordinator M.Sc. Applied Statistics	**Prof. Dr. Haribhau Bhapkar** Head Department of Applied Sciences & Humanities

DECLARATION

I hereby declare that, the project entitled is an outcome of my own efforts under the guidance of Dr. Ashok Kumar The mini project is submitted to the MIT Art, Design and Technology University, Pune (India). For the partial fulfilment of the Master of Applied Statistics (Data Science) examination 2022-2023.

I also declare that this mini project report has not been previously submitted to any other university.

Date: 28th Dec 2022

Place: Loni Kalbhor, Pune Mohan L. Bodake

ACKNOWLEDGEMENT

I would like to thank all teaching and non-teaching staff and all my colleagues. I would like to express my profound gratitude towards Dr. Ashok Kumar for his valuable guidance for completion of this project.

I am also thankful to **Dr. Pratibha Jadhav, Prof. M.S. Kasture , Dr. Harsh Tripathi** for their timely suggestions and encouragements.

I would like to thank **Prof. Dr. Haribhau Bhapkar,** Head of Department Applied Science & Humanities, **MIT Art, Design and Technology University, Pune (India),** for providing me the necessary facilities.

Mohan L. Bodake

INDEX OF CONTENT

1. INTRODUCTION

1.1 Overview of Data Science

Data science is now an amalgamation of different roles. In other words, the demand for data professionals has increased because the need for different kinds of expertise has increased. As more specific data roles turn up regularly, the wider & more lucrative a career in Data Science becomes. Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from noisy, structured and unstructured data, and apply knowledge from data across a broad range of application domains. Data science is related to data mining, machine learning and big data.

Data science is a "concept to unify statistics, data analysis, informatics, and their related methods" in order to "understand and analyse actual phenomena" with data. It uses techniques and theories drawn from many fields within the context of mathematics, statistics, computer science, information science, and domain knowledge. However, data science is different from computer science and information science. Turing Award winner Jim Gray imagined data science as a "fourth paradigm" of science (empirical, theoretical, computational, and now data-driven) and asserted that "everything about science is changing because of the impact of information technology" and the data deluge. Data Science is kind of blended with various tools, algorithms, and machine learning principles. Most simply, it involves obtaining meaningful information or insights from structured or unstructured data through a process of analyzing, programming and business skills. It is a field containing many elements like mathematics, statistics, computer science, etc. Those who are good at these respective fields with enough knowledge of the domain in which you are willing to work can call themselves as Data Scientist. It’s not an easy thing to do but not impossible too.

You need to start from data, it's visualization, programming, formulation, development, and deployment of your model. In the future, there will be great hype for data scientist jobs. Taking in that mind, be ready to prepare yourself to fit in this world.

1.2 Overview of Exploratory Data Analysis (EDA)

Exploratory data analysis (EDA) is used by data scientists to analyze and investigate data sets and summarize their main characteristics, often employing data visualization methods. It helps determine how best to manipulate data sources to get the answers you need, making it easier for data scientists to discover patterns, spot anomalies, test a hypothesis, or check assumptions. EDA is primarily used to see what data can reveal beyond the formal modelling or hypothesis testing task and provides a provides a better understanding of data set variables and the relationships between them. It can also help determine if the statistical techniques you are considering for data analysis are appropriate. Originally developed by American mathematician John Tukey in the 1970s, EDA techniques continue to be a widely used method in the data discovery process today. The main purpose of EDA is to help look at data before making any assumptions. It can help identify obvious errors, as well as better understand patterns within the data, detect outliers or anomalous events, find interesting relations among the variables.

Data scientists can use exploratory analysis to ensure the results they produce are valid and applicable to any desired business outcomes and goals. EDA also helps stakeholders by confirming they are asking the right questions. EDA can help answer questions about standard deviations, categorical variables, and confidence intervals. Once EDA is complete and insights are drawn, its features can then be used for more sophisticated data analysis or modelling, including machine learning.

There are four primary types of EDA:

- **Univariate non-graphical.** This is simplest form of data analysis, where the data being analyzed consists of just one variable. Since it's a single variable, it doesn't deal with causes or relationships. The main purpose of univariate analysis is to describe the data and find patterns that exist within it.
- **Univariate graphical.** Non-graphical methods don't provide a full picture of the data. Graphical methods are therefore required. Common types of univariate graphics include:
    - Stem-and-leaf plots, which show all data values and the shape of the distribution.
    - Histograms, a bar plot in which each bar represents the frequency (count) or proportion (count/total count) of cases for a range of values.
    - Box plots, which graphically depict the five-number summary of minimum, first quartile, median, third quartile, and maximum.
- **Multivariate non-graphical:** Multivariate data arises from more than one variable. Multivariate non-graphical EDA techniques generally show the relationship between two or more variables of the data through cross-tabulation or statistics.
- **Multivariate graphical:** Multivariate data uses graphics to display relationships between two or more sets of data. The most used graphic is a grouped bar plot or bar chart with each group representing one level of one of the variables and each bar within a group representing the levels of the other variable.

Other common types of multivariate graphics include:

- Scatter plot, which is used to plot data points on a horizontal and a vertical axis to show how much one variable is affected by another.
- Multivariate chart, which is a graphical representation of the relationships between factors and a response.
- Run chart, which is a line graph of data plotted over time.
- Bubble chart, which is a data visualization that displays multiple circles (bubbles) in a two-dimensional plot.
- Heat map, which is a graphical representation of data where values are depicted by colour.

The most common data science tools used to create an EDA is:

- **Python:** An interpreted, object-oriented programming language with dynamic semantics. Its high-level, built-in data structures, combined with dynamic typing and dynamic binding, make it very attractive for rapid application development, as well as for use as a scripting or glue language to connect existing components together. Python and EDA can be used together to identify missing values in a data set, which is important so you can decide how to handle missing values for machine learning.

1.3 Aim & objectives of the study

The purpose of this project is to analyze the salaries of data science jobs. This analysis will focus on the factors that impact salary in data science positions, including geographic location, experience level, and job title. We will use a combination of publicly available salary data and survey data to build a comprehensive picture of salary trends in the data science field. By studying the data, we aim to gain insight into how salaries are determined and how they vary across different regions of the country. Additionally, we hope to identify any opportunities for data scientists to maximize their earning potential. The following aim has discussed in the project study.

- To find out the most job titles in data science domain.
- To study the average salaries by each job title & country.
- To find out how many companies providing the data science job & their locations.
- To find out no. of peoples with all experience level in each job.
- To study the average salaries with all experience level in each job.

2. METHODOLOGY

Collect Historical Data: The first step is to collect historical data on Data science job salaries. This data collected from kaggle.com.

Data Pre-processing: After collecting the historical data, it is important to clean the data. This includes removing any outliers, missing values, and any other data that might be irrelevant.

Data Visualization: After cleaning the data, it is important to visualize the data. This step can be done using various Python libraries such as Matplotlib, Seaborn, Plotly, and others libraries.

Conclusion: After analyzing the data, it is important to draw conclusions on the basis of visualization.

2.1Data Structure

For this project, I used the dataset of Data Science Job Salaries Dataset. There are total 11 features that are used in the project so the overall dataset contains 607 rows and 11 columns. In this project we have considered data from year 2019-2021.

2.2 Data Description

1. Work year: The year the salary was paid.
2. Experience level: The experience level in the job during the year.
3. Employment type: The type of employment for the role.
4. Job title: The role worked in during the year.
5. salary: The total gross salary amount paid.

6. Salary currency: The currency of the salary paid as an ISO 4217 currency code.

7. Salary in usd: The salary in USD.

8. Employee residence: Employee's primary country of residence in during the work year as an ISO 3166 country code.

9. Remote ratio: The overall amount of work done remotely.

10. Company location: The country of the employer's main office or contracting branch.

11. Company size: The median number of people that worked for the company during the year.

Summary Statistics:

	work_year	salary	salary_in_usd	remote_ratio
count	607.000000	6.070000e+02	607.000000	607.00000
mean	2021.405272	3.240001e+05	112297.869852	70.92257
std	0.692133	1.544357e+06	70957.259411	40.70913
min	2020.000000	4.000000e+03	2859.000000	0.00000
25%	2021.000000	7.000000e+04	62726.000000	50.00000
50%	2022.000000	1.150000e+05	101570.000000	100.00000
75%	2022.000000	1.650000e+05	150000.000000	100.00000
max	2022.000000	3.040000e+07	600000.000000	100.00000

2.3 Computational Tools

- Python

Python is a multi-paradigm programming language. It supports multiple programming paradigms, including structured, object-oriented, and functional programming. Python is meant to be an easily readable language. Its formatting is visually uncluttered and often uses English keywords whereas other languages use punctuation. Unlike many other languages, it does not use curly brackets to delimit blocks, and semicolons after statements are allowed but rarely used.

- Jupyter Notebook

Jupyter Notebooks are a spin-off project from the IPython project, which used to have an IPython Notebook project itself. Jupyter ships with the IPython kernel, which allows you to write your programs in Python, but there are currently over 100 other kernels that you can also use.

PYTHON LIBRARIES

Pandas :

Pandas is defined as an open-source library that provides high-performance data manipulation in Python. The name Pandas is derived from the word Panel Data, which means econometrics from multidimensional data. It is used for data analysis in Python.

Matplotlib :

Matplotlib is a python library used to create 2D graphs and plots by using python scripts. It has a module named pyplot which makes things easy for plotting by providing features to control line styles, font properties, formatting axes, etc. It supports a very wide variety of graphs and plots namely - histograms, bar charts, power spectra, error charts, etc.

Seaborn:

Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics. one of an amazing library for visualization of the graphical statistical plotting in Python. Seaborn provides many color palettes and defaults beautiful styles to make the creation of many statistical plots in Python more attractive.

Plotly:

Python Plotly Library is an open-source library that can be used for data visualization and understanding data simply and easily. Plotly supports various types of plots like line charts, scatter plots, histograms, cox plots, etc.

3. EXPLORATORY DATA ANALYSIS

3.1 Analysing the distribution of salary in USD

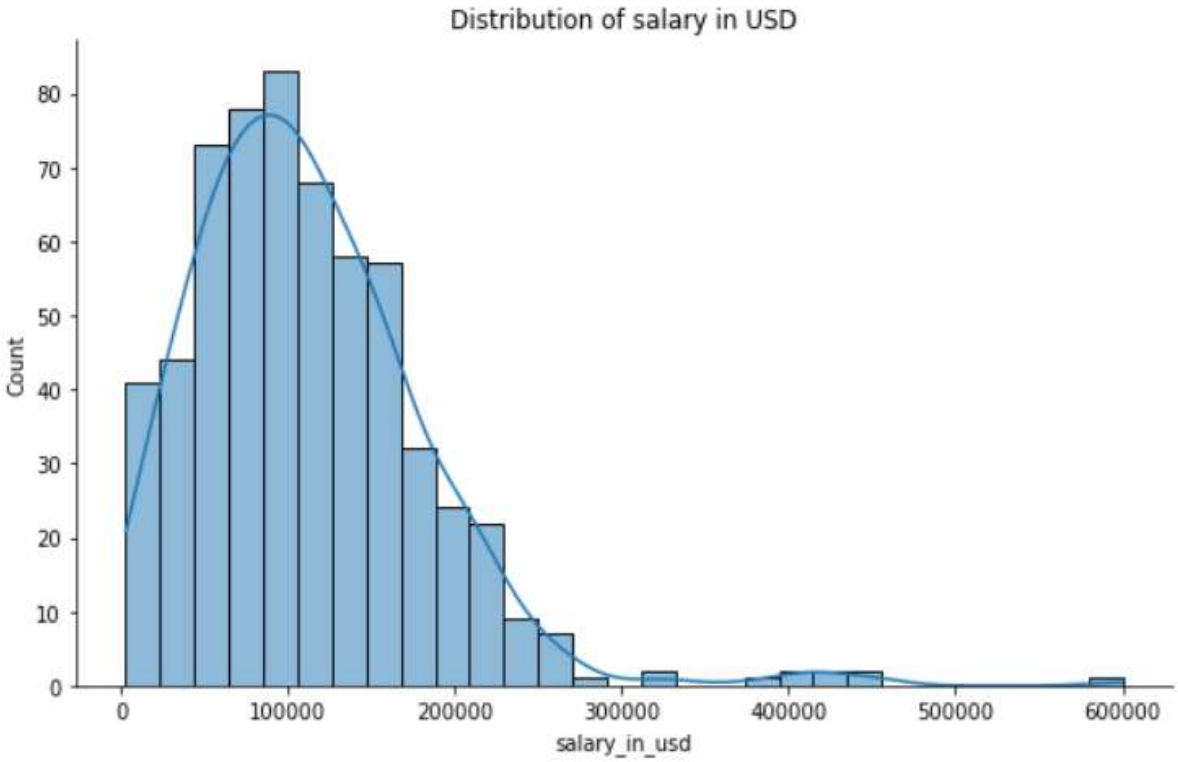

Fig 3.1 Distribution of salary in USD

Its seems like to be the distribution curve of salaries in USD is positively skewed. The positively skewed distribution means most values are clustered around the left tail of the distribution while the right tail of the distribution is longer.

From the density plot of salaries, we conclude that there are more salaries at the lower end of the income spectrum than at the higher end.

3.2 Analysing job titles & top job titles in Data Science

There are total 50 job titles in Data Science. The total job titles & the job counts are describe in the table below;

Job Title	Count	Job Title	Count
Data Scientist	143	Analytics Engineer	4
Data Engineer	132	Machine Learning Developer	3
Data Analyst	97	Machine Learning Infrastructure Engineer	3
Machine Learning Engineer	41	Lead Data Scientist	3
Research Scientist	16	Lead Data Analyst	3
Data Science Manager	12	Data Science Engineer	3
Data Architect	11	Principal Data Engineer	3
Big Data Engineer	8	Computer Vision Software Engineer	3
Machine Learning Scientist	8	Principal Data Analyst	2
Director of Data Science	7	Financial Data Analyst	2
AI Scientist	7	ETL Developer	2
Principal Data Scientist	7	Director of Data Engineering	2
Data Science Consultant	7	Product Data Analyst	2
Data Analytics Manager	7	Cloud Data Engineer	2
Computer Vision Engineer	6	NLP Engineer	1
BI Data Analyst	6	Marketing Data Analyst	1
ML Engineer	6	3D Computer Vision Researcher	1
Lead Data Engineer	6	Machine Learning Manager	1
Data Engineering Manager	5	Lead Machine Learning Engineer	1
Business Data Analyst	5	Head of Machine Learning	1
Applied Data Scientist	5	Finance Data Analyst	1
Head of Data	5	Data Specialist	1
Head of Data Science	4	Data Analytics Lead	1
Data Analytics Engineer	4	Big Data Architect	1
Applied Machine Learning Scientist	4	Staff Data Scientist	1

Table 3.1 Job Titles & Count

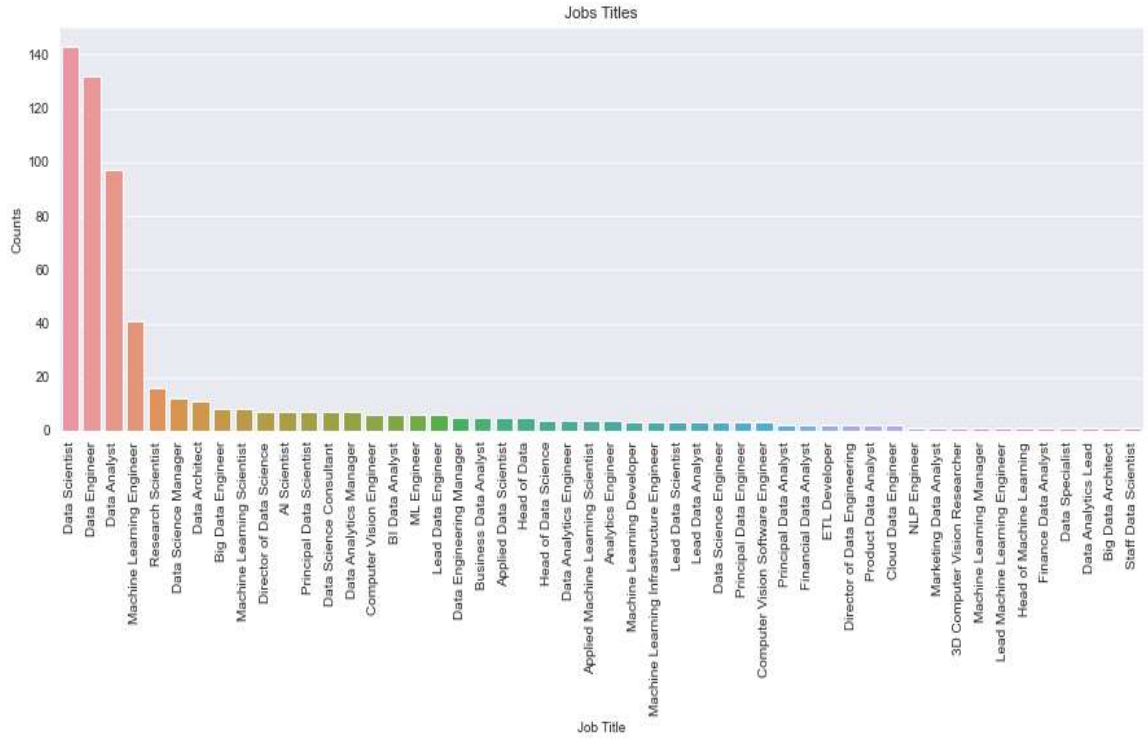

Fig 3.2 Count of Job Titles

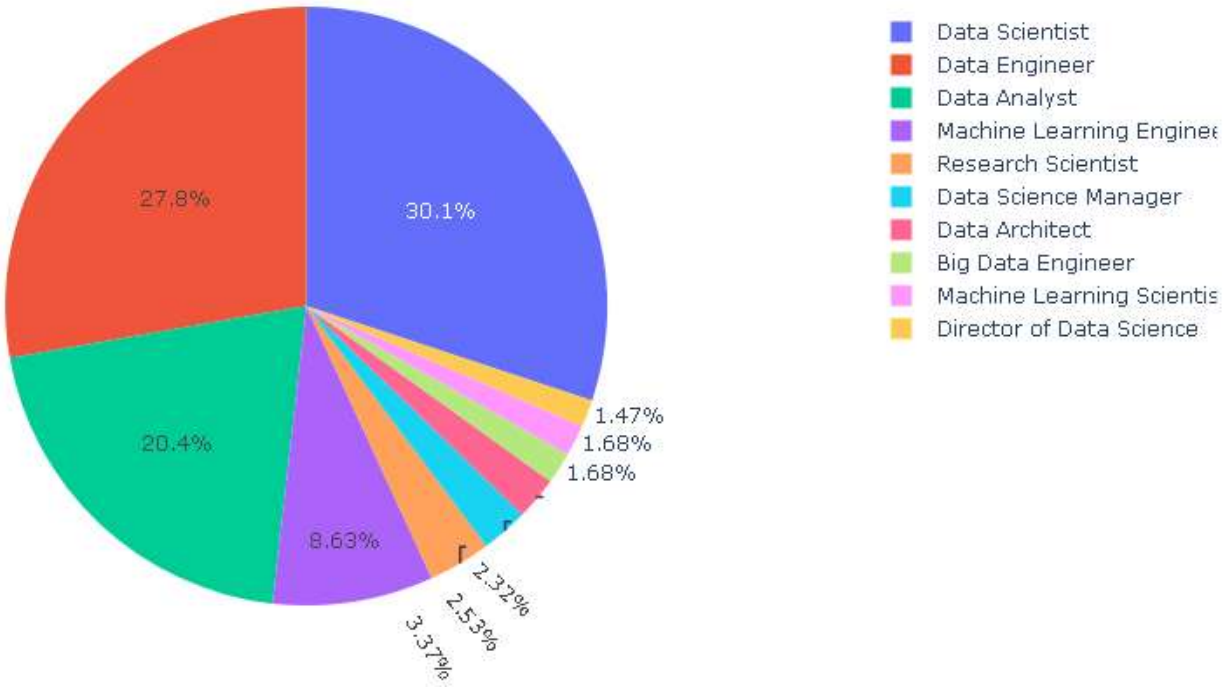

Fig 3.3 Top Job Titles

Most of the job title in this dataset contains ‘Data Scientist’ i.e. 143, followed by Data Engineer & Data Analyst. The least job titles are Staff Data Scientist, Big Data Architect, Data Analytics lead.

3.3 Analysis of Salaries by Job Title

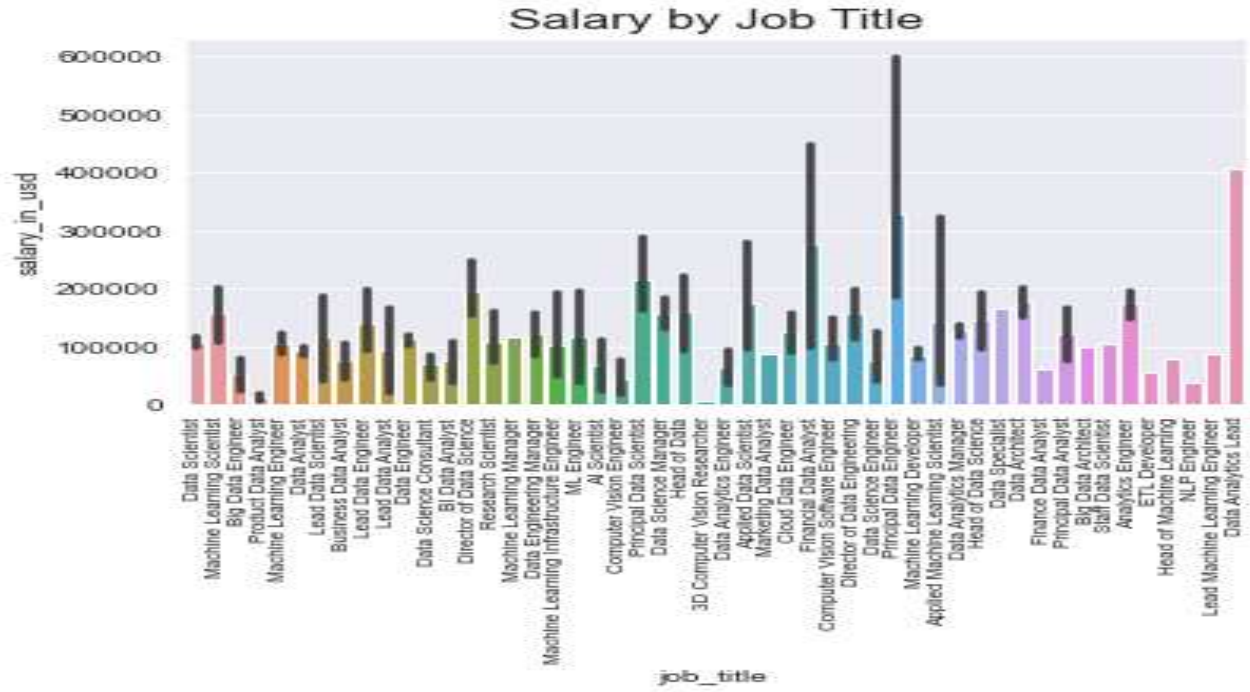

Fig 3.4 Salary by Job Title

The Principal Data Engineer has highest salary i.e. about 6.5L $ per annum & 3D Computer Vision Researcher has lowest salary i.e. 5409$ per annum.

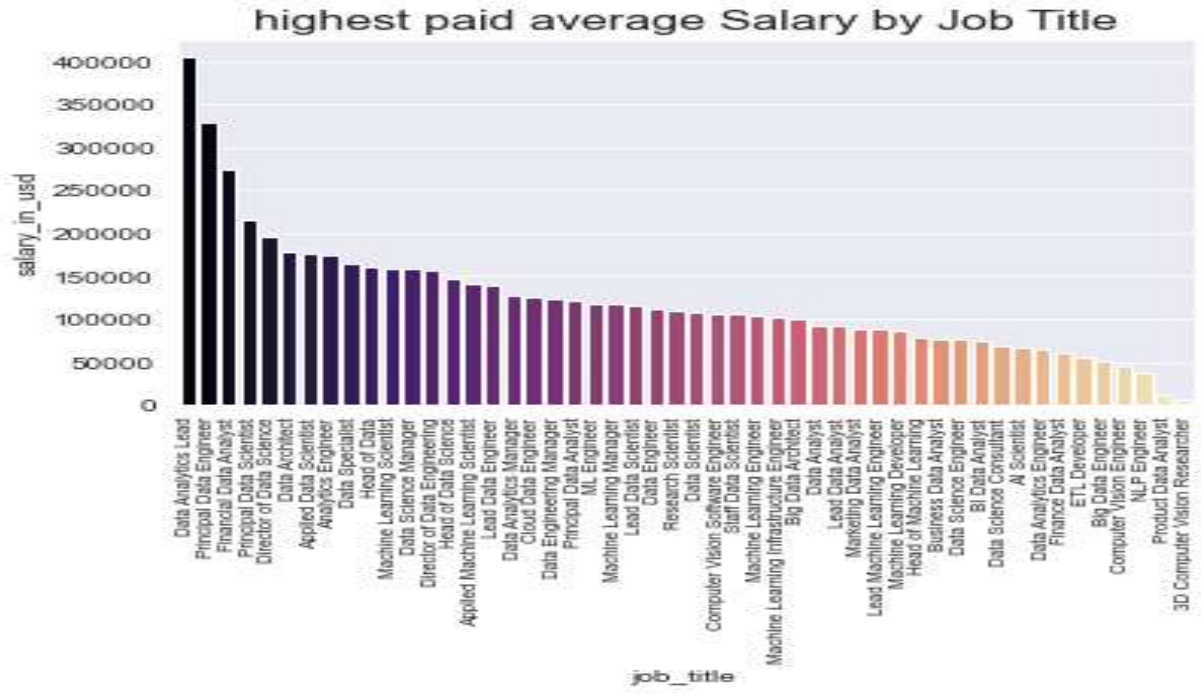

Fig 3.5 Average salary by Job Title

The Data Analyst Lead has highest average salary & 3D Computer Vision Researcher has lowest average salary.

3.4 Analysis of salaries by country

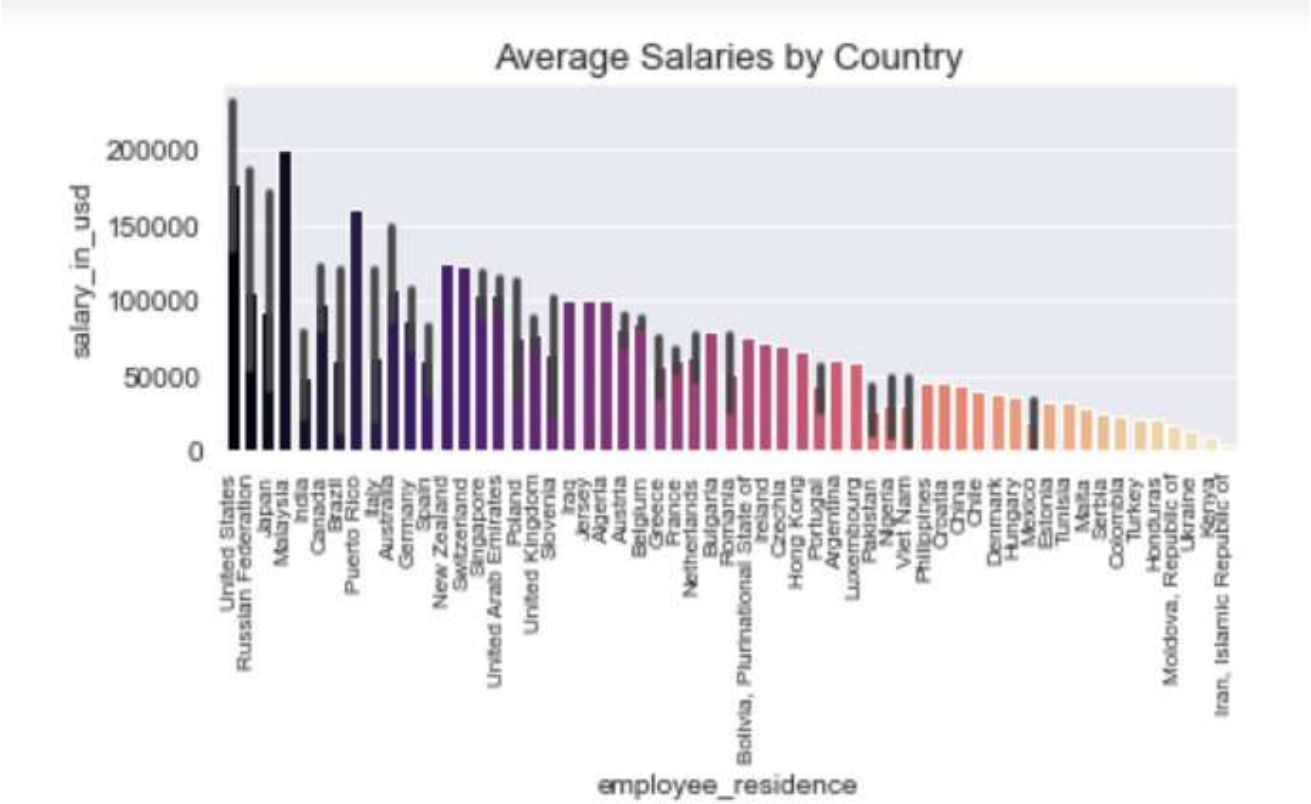

Fig 3.6 Average Salary by Country

United states, Russian federation, Japan, Malaysia these are the countries which are paying high salaries compare to other countries. The countries which are paying low salaries are Iran, Kenya, Ukraine.

3.5 Analysis of experience level of employee

There is four type of experience level viz. senior, mid, entry & executive.

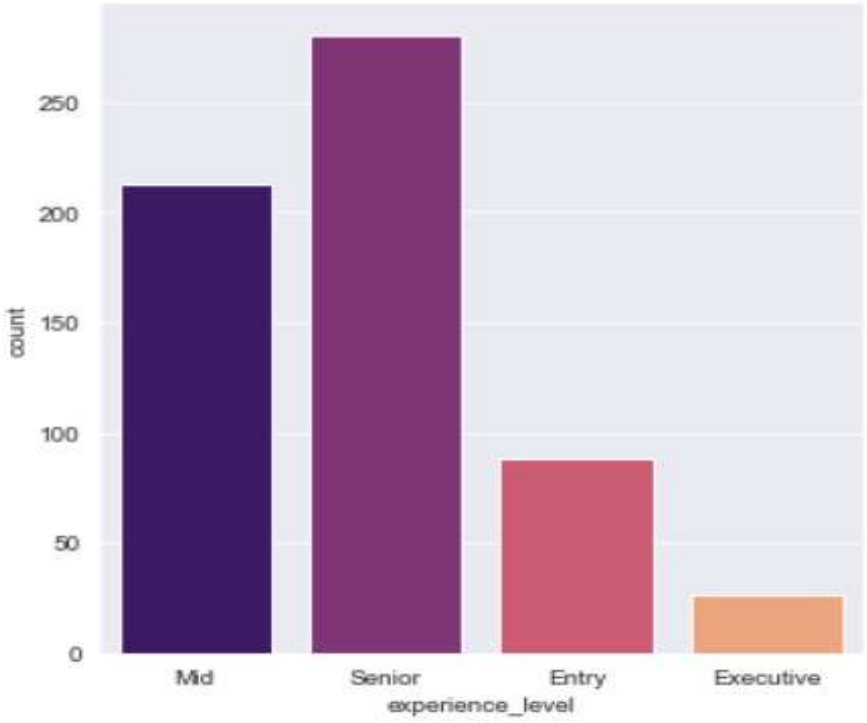

Fig 3.7 Count plot of Experience Level

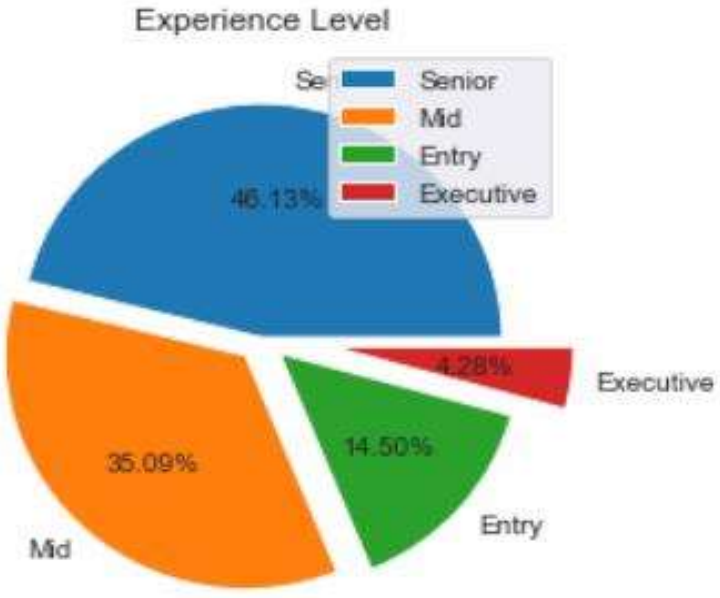

Fig 3.8 Pie Chart of percentage of Experience Level

Most of the employee are Senior level. About 46% employee are senior level & mid-level employee are about 35%

3.6 Analysis of employment type of employee

The employment type means that the employee is doing full time work or having contract based job or any part time job.

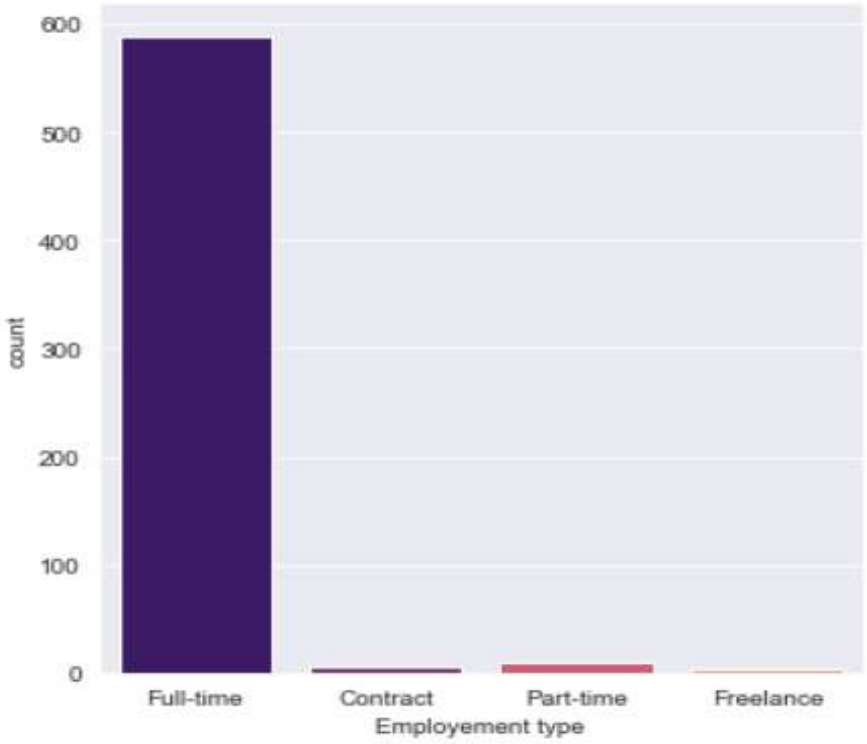

Fig 3.9 Count plot of Employment Type

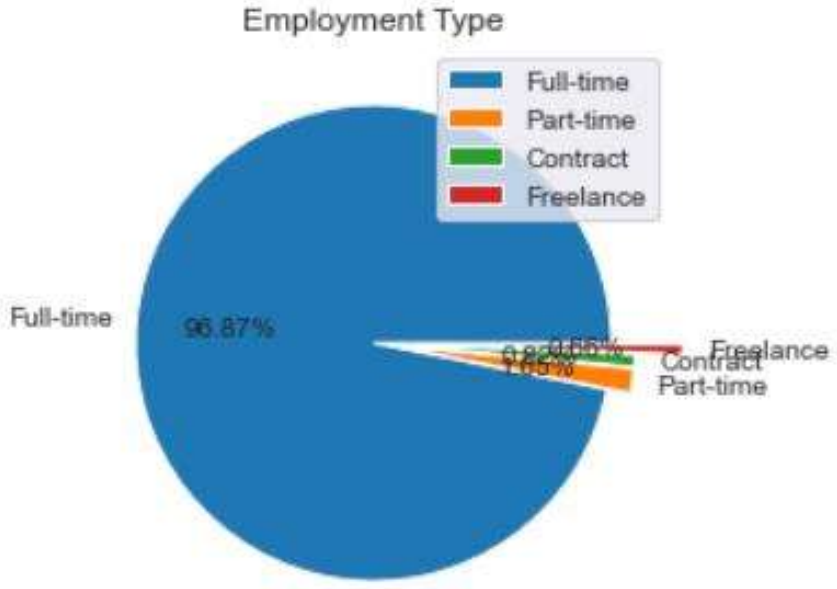

Fig 3.10 Pie chart of Employment Type

Out of 607 employees, 588employee are doing full time work.

3.7 Companies Locations

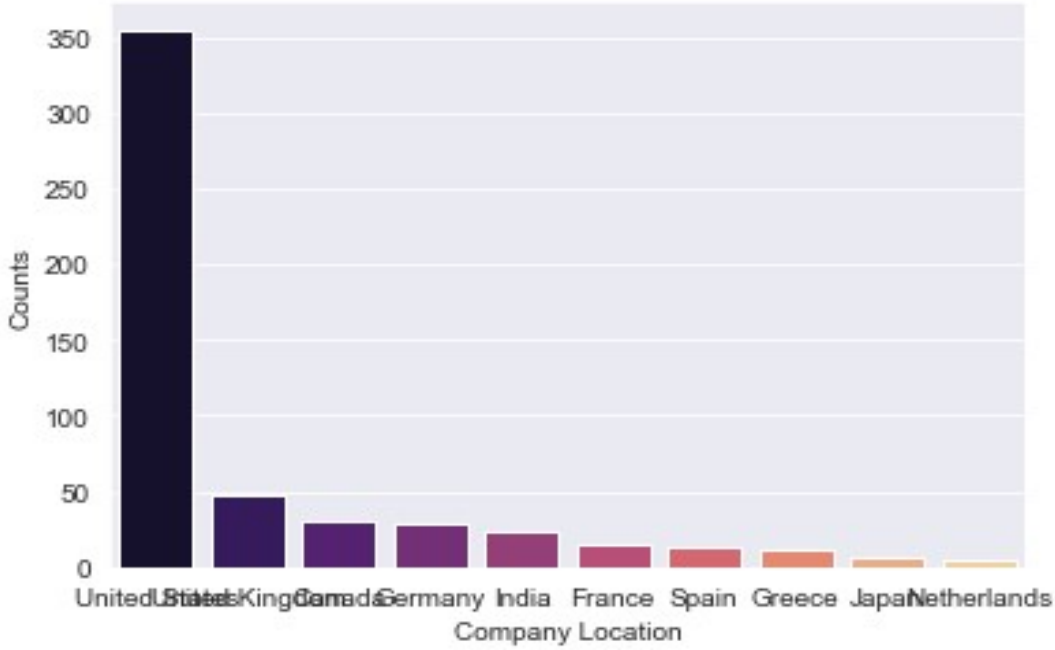

Fig 3.11 Count plot of compny location

Fig 3.12 Tree map of company location

It is observed that most of Data Science Companies (approximately 60%) are located at USA. There are 355 data science companies located in USA.

3.8 Location of companies based on experience level

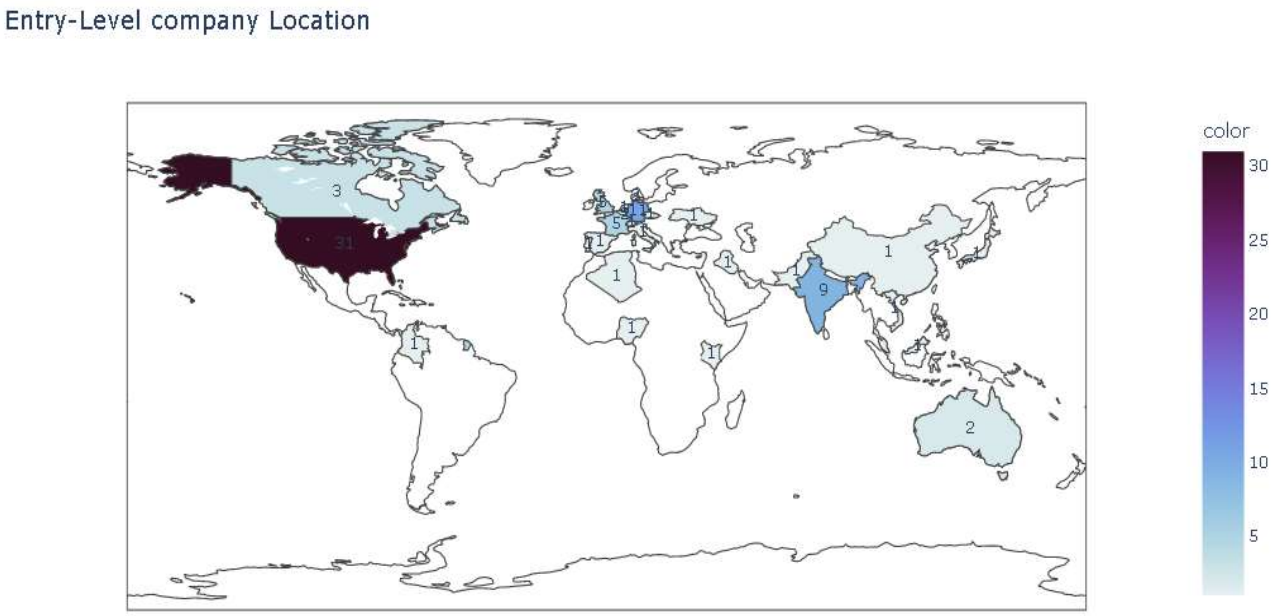

Fig 3.13 Choropleth map of entry-level company

There are 31 Entry-Level companies located in United States of America.

Fig 3.14 Choropleth map of mid-level company

There are 90 Mid-Level companies located in United States of America.

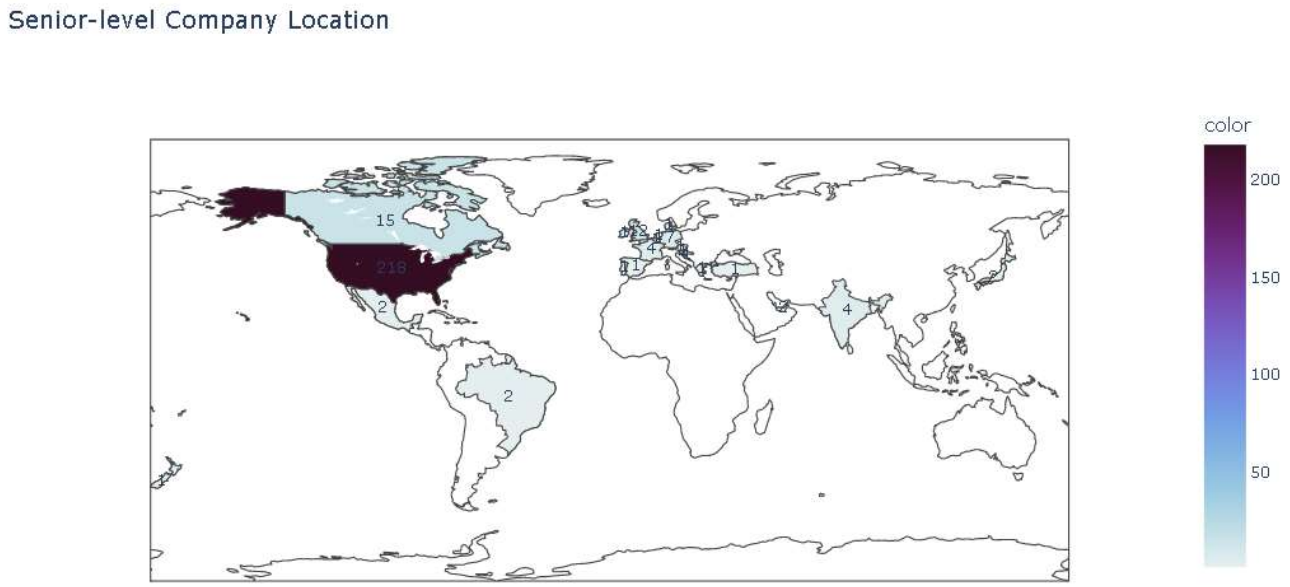

Fig 3.15 Choropleth map of senior-level company

There are most companies i.e. 218 Senior-Level companies located in United States of America.

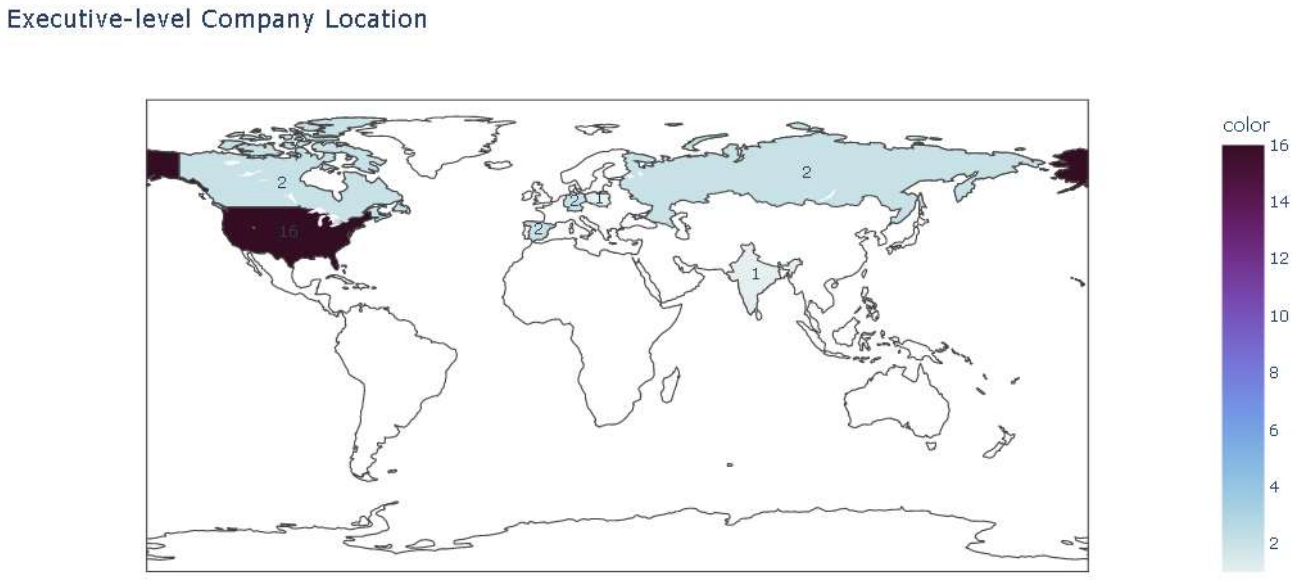

Fig 3.16 Choropleth map of executive-level company

There are 16 Executive-Level companies located in United States of America.

3.9 No. of people with experience in each job

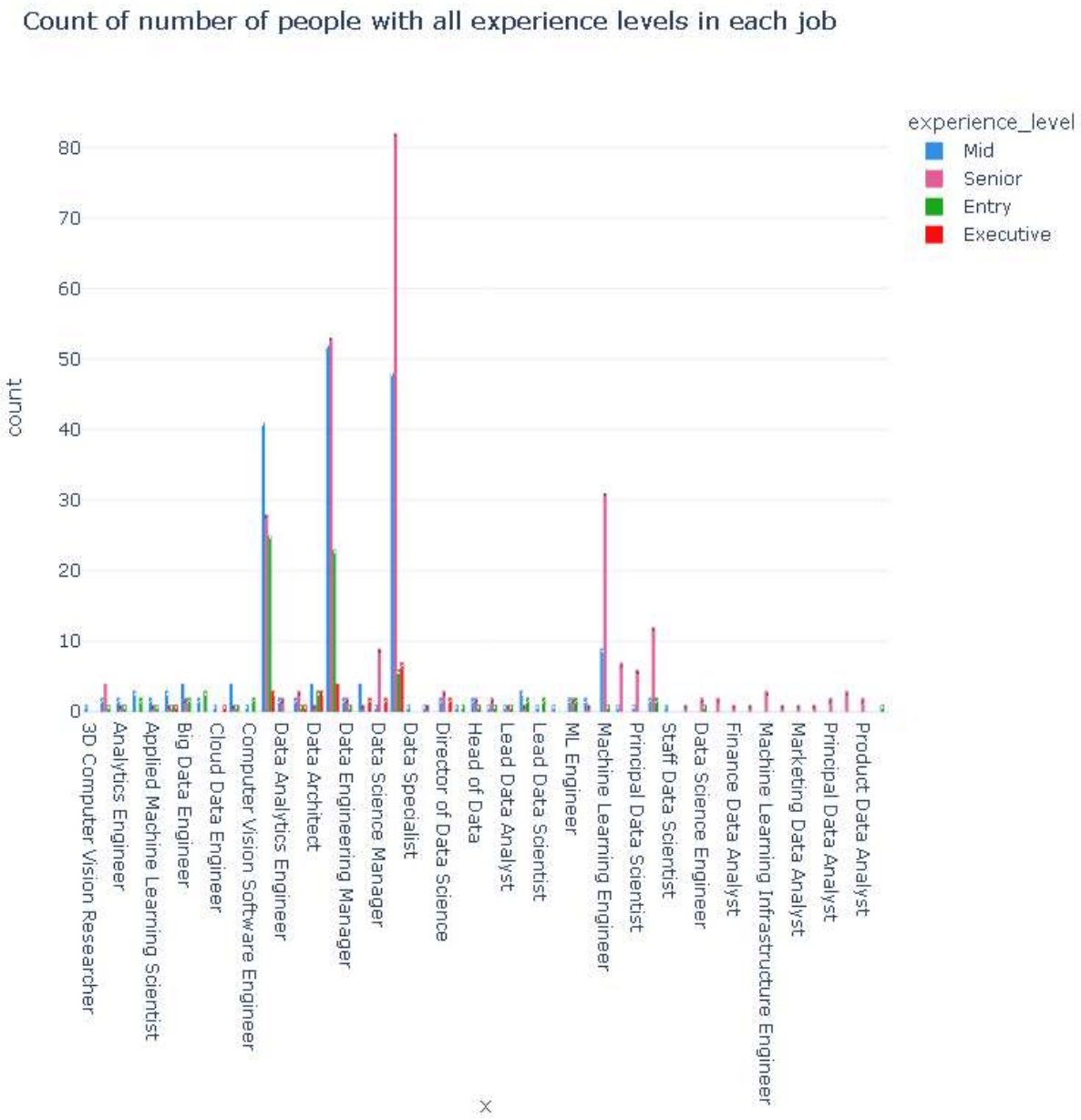

Fig 3.17 Count plot people with experience in each job

Most of the employees are working for the role of Data scientist, Data analyst, data engineer and Machine Learning Engineer.

It is observed that the Data scientist job demands senior level employees as compared to the entry level employees.

For the data analyst position all experience people nearly have same opportunities except for the entry level employees.

3.10 Average salaries with experience level in each job.

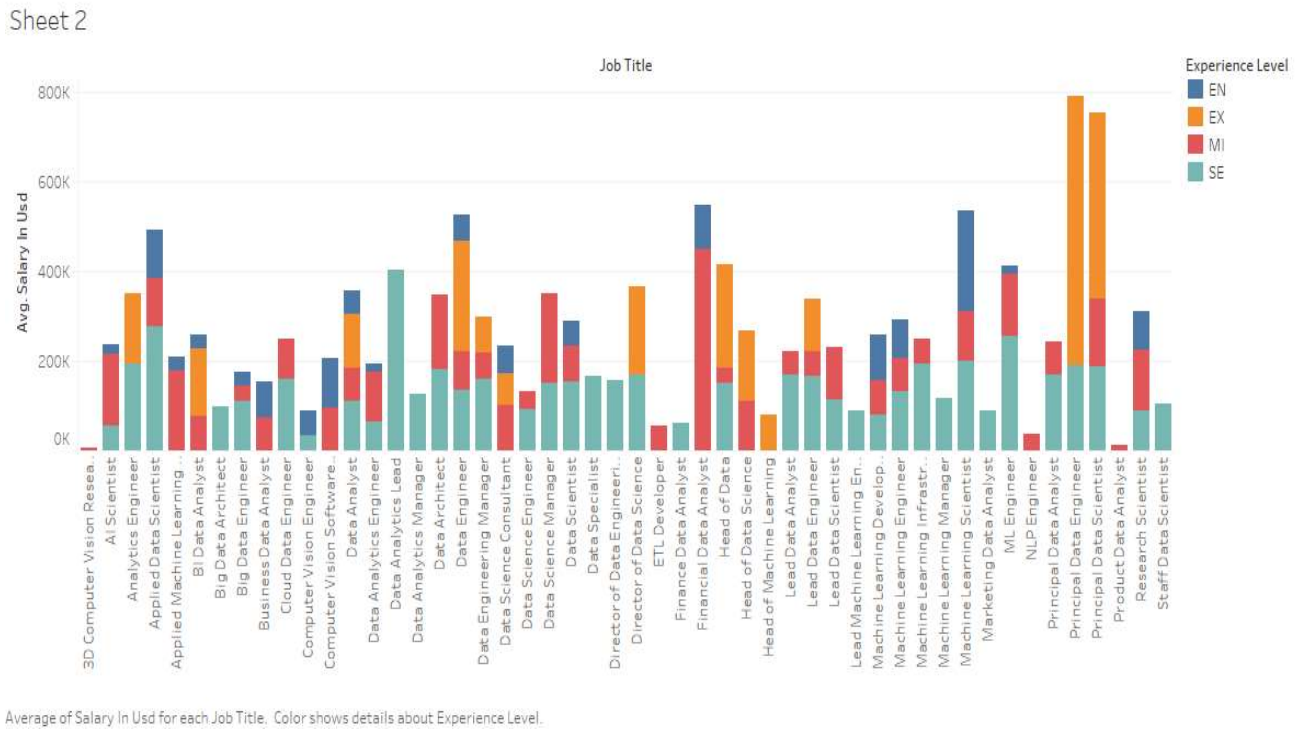

Fig 3.18 Subdivided bar plot of average salary with experience level I each job

For the position of financial data analyst, the difference between the average salary of mid-level employees and entry level employees is big.

The data architect role pays approx. same salary to the mid-level employees and the senior-level employees

The average salary for the data scientist position is good for the senior-level employees but not as good for other experience level employees.

4. Results & Conclusions

- Most of the job title in this dataset contains 'Data Scientist' i.e. 143, followed by Data Engineer & Data Analyst.
- The Principal Data Engineer has highest salary i.e. about 6.5L $ per annum but the Data Analyst Lead has highest average salary.
- United states, Russian federation, Japan, Malaysia these are the countries which are paying high salaries compare to other countries.
- Most of Data Science Companies (approximately 60%) are located at USA. There are 355 data science companies located in USA.

Overall, the analysis of the data science job markets indicates that data science is a highly lucrative field, with salaries ranging from $60,000 to $600,000 per year. The analysis also revealed that the highest paid data science jobs are located USA. Additionally, the analysis found that the most common job titles for data science jobs are Data Scientist, Machine Learning Engineer, and Data Analyst. Therefore, these findings suggest that individuals who are interested in pursuing a career in data science should focus their job search in the areas identified in the analysis.

5. Reference

- https://www.kaggle.com
- Python Programming: 3 books in 1 - Ultimate Beginner's, Intermediate & Advanced Guide to Learn Python Step by Step by Ryan Turner | 9 February 2020
- Data Analytics Using Python by Bharti Motwani, Wiley| **Author: Bharti Motwani**
- https://www.linkedin.com/posts/harsha-chavan_assignment-datascienceinternship-machinelearning-ugcPost-7012479809888882688-6TN7?utm_source=share&utm_medium=member_android