CSCE 5320 Project

Visualizing Current Data Science Salary Trends: Key Insights for Job Seekers

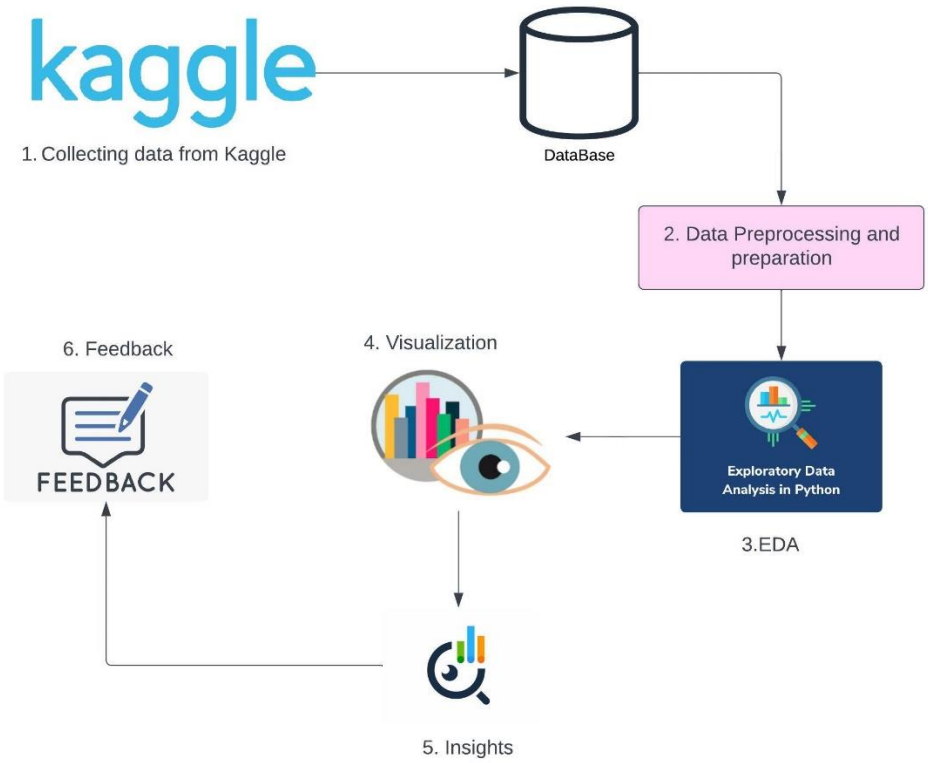
Team Members

Sai Krishna Meduri – 11664578

Sreecharan Vanam– 11544661

Contents

- Workflow Diagram



The above diagram illustrates how acquisition of the data has been carried out in our project which took place on Kaggle. Kaggle is one of the most prominent hub for dataset and the first right stop of our operation, in which we obtain the dataset consists of many job listing from a various data science professionals categorised into different roles from data cleaning analyst to machine learning engineer. The dataset provides a comprehensive set of information about the job such as job titles, salary, years of experience and also geographical data. The process of extracting these data is systematic and detailed, so data science projects can learn from this moves and try it to figure out the trends and patterns behind the salary in data science.

Then, once the dataset is harvested, we move to the data preprocessing and preparation stage, where data transforms from a raw state into a cleaned state, ready to be explored in detail. This preprocessing includes tasks such as data cleaning (to ensure the removal of anomalies, such as missing values, duplicates or irrelevant data that skew our results) and salary harmonisation (standardising values recorded in different global currencies to a single currency [USD] for greater fair, and consistent, comparison). Significant work also goes into formatting and structure, which includes categorisation of job titles and the preparatory work that ensures consistency within the organisation of experience levels to make the data as clear as possible. This level of detail is required at this stage because it paves the way for us to perform an insightful and meaningful EDA later. If the EDA is to be truly insightful, it's

critical to have a dataset that's granular, complete and faithful to the truth. This preprocessing and preparation allows us to meet our objective of both finding and sharing the truth about pay and trends.
Overall, these steps form the foundation of the project pipeline, before the final stages of analysis and visualisation. They are essential to our overarching goal of presenting a coherent, fact-based and informative description of data science salary patterns that can help job-seekers make career decisions..

Related Work

Exploring data science salary trends is a major and needed research, data science salary analysis is a synthesis of industry analysis and academics. It's most notable among hundreds of writings published every day. [1] Data Science 2023 Review: Trends and Salary Expectations contains a very accurate description of the market in terms of its components, drivers and focuses on changes in relations, expectations and roles of data scientists today and in the near future. It's the first of its kind not only with numbers but with its content which narrates the history of the role, duties and expectations of data scientists today in comparison with those a few years ago. This content was written both for those who are already practising data science and for beginners and graduates who will soon join the industry.

This exploration of [2] 'Data Science AI Professionals Salary and Hiring Trends' encodes the mounting significance of Artificial Intelligence (AI) to the job market in analytical ways: the author tracks the dynamic interaction between technological development in AI and the balance of supply and demand of talent in the field of data science as those factors transform into a 'salary map' whose current version emphasises the premium on sharp skills and the bad cop in recruiting these workers: job seekers must heed.

In 'Salary Trends: What to Expect in Data Analytics Roles in 2024', [3] the authors offer a kind of navigational aid for the uncertain seas of the future data science job market. Their forecasts about what the data science industry might look like in the future arm professionals with expectations about how their roles might progress. With this information, they might be able to assess whether their skill set changes are aligned with what the industry expects. By taking aim at future events, the 2024 study extends the visibility of the planning horizon for professionals and organisations.

At even a deeper level, [4] 'Trends in Data Science Salaries: An Exploratory Data Analysis Journey' and [5] 'EDA and Visualizations on Data Science Salaries using Python' show love for the sector's analytics core by putting EDA to work unpacking the salaries dataset. These pieces affirm the power (and necessity) of visualisation for revealing underlying patterns and making findings more widely accessible; they make the case for the fact that data visualisations are as important as the data itself for understanding trends in pay.

In the studies conducted [6] and [7] provide two further sophisticated predictive approaches and a meta-analysis respectively. [6] goes as far as developing prediction models, that is, models that go further than recording what is observed and predict salaries instead, and [7] provides a literature review that links scientific-technical skills to salaries. Between them, they provide the

academic rationale of the field: the empirical data analyses as well as the theoretical models are necessary to get a full understanding of the salaries landscape in data science.

▪ Data Abstraction

The data set that we will be using for this project is provided online. It entails the salary of data scientists. This dataset can be downloaded from Kaggle. The dataset is in a format called CSV, which is Comma-Separated Values. This kind of data set can be read and processed using an array of data processing software.

Type and Attributes:

The dataset's structure is tabular, where each row represents a unique entry of a data science job and the other attributes described below:

work_year: The year of the job data entry.

job_title: The title of the job position.

job_category: The category to which the job belongs, such as Data Engineering, Data Analysis, or Machine Learning.

salary_currency: The currency in which the salary is provided.

salary: The salary amount in the currency specified.

salary_in_usd: The salary amount converted to United States Dollars (USD) for standardization.

employee_residence: The country of residence of the employee.

experience_level: The level of experience required for the job, categorized into levels such as Entry-level, Mid-level, Senior, and Executive.

employment_type: The nature of employment, which could be Full-time, Part-time, Contract, or Freelance.

work_setting: The setting of the work environment, whether Remote, In-person, or Hybrid.

company_location: The location of the company offering the job.

company_size: The size of the company, often categorized into small, medium, or large scale.

```
RangeIndex: 9355 entries, 0 to 9354
Data columns (total 12 columns):
 #   Column              Non-Null Count  Dtype 
---  ------              --------------  ----- 
 0   work_year           9355 non-null   int64 
 1   job_title           9355 non-null   object
 2   job_category        9355 non-null   object
 3   salary_currency     9355 non-null   object
 4   salary              9355 non-null   int64 
 5   salary_in_usd       9355 non-null   int64 
 6   employee_residence  9355 non-null   object
 7   experience_level    9355 non-null   object
 8   employment_type     9355 non-null   object
 9   work_setting        9355 non-null   object
 10  company_location    9355 non-null   object
 11  company_size        9355 non-null   object
dtypes: int64(3), object(9)
memory usage: 877.2+ KB
```

All this above data attributes forms the basis of our analysis, and each type of attribute in it provides a different perspective from which we can grasp the salary trends in data science. It contains both categorical and numerical data types, which can be investigated from all possible levels, from simple univariate statistics to complex multivariate visualisation.

Detailed Description:

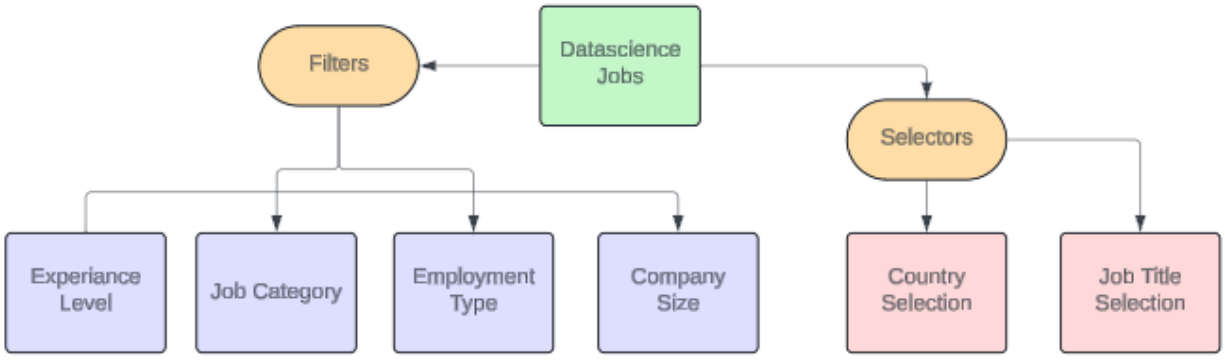
The dataset was collected from job postings and a survey for the data science industry from Kaggle. It consists of many roles from different countries and it includes compensation packages as well as description of the job. The dataset has records from 2023, from multiple job titles and different categories in different countries. There are 9355 entries..

The data are obtained from job postings on general job boards and surveys from the data science community on Kaggle.

Volume: The dataset contains 9,355 rows, each corresponding to an individual job record.

Variety: It contains fields such as Job title, Job category, Salary in local currency and USD, Employee residence, Experience level, Employment type, Work setting., Company location, and Company size.

Feature Design Diagram:



This is a Feature Design Diagram featuring an overall dashboard-oriented flow of user exploration of data science job market.

The "Datascience Jobs" dataset is featured at the center of everything, presenting this dataset as the core of the dashboard from which the user can then refine the view via different filters and selectors. From the "Filter" node, branches several attributes which are "Experience Level", "Job Category", "Employment Type" and "Company Size" which in turn aid the user in drilling down the dataset based on these attributes. Other branches originate from the "Selectors" node, consisting of two main drop-down menus: "Country Selection" And "Job Title Selection" which allow the user to narrow the dataset on these granular level and locations or jobs. This User-Flow facilitates an immersive and interactive user experience, allowing a tailored interactivities of the user to present the data in accordance with their specific analysis requirements while at the same time crowdsourcing information and job preferences to allow more people-centric outcomes in the data-based analysis and inferences.

Data Transformation:

The dataset was cleaned and prepared to move into an analysis phase. Numerous models of training data didn't converge. Mainly, we were missing meaningful information that a monkey could have generated. Initially, we found missing or partially-complete data, which we checked and filtered out of the dataset, going so far as to impute certain values when appropriate. Since, the data is from all over the world, we needed to normalise salary figures across different geographies and local currencies to USD, using exchange rates from year 2023. This was a critical step since it gave a common baseline to compare salaries - comparable to other data fields as it was now not so vast. Also, we aggregated many small fields into bigger sums like mean salary for certain job titles or in each country to provide generalised data for analysis and reduction. For example, personnel and company types, number of years of experience, and company sizes were converted into some common values using percentage and grouping data from various salaries. Finally, we converted categorical variables like job titles, skills, and experience years into an encodable format, similar to numeric values, so that it was useful in terms of estimating statistical models and visualisation. The final dataset has sufficient

information from various training data sources, or salaries, allowing the model to inspect it in significant detail.

▪ Task Description

Task: Understanding and Analyzing Data Science Salary Trends

Target:

We define our primary audience as students, emerging professionals into data science, and career switchers. Also, the findings can give an edge to recruiters and HR people to develop competitive salary packages.

Actions:

The main actions to be performed in this project involve the following:

Data Collection: Acquiring a rich dataset from Kaggle about attributes of data science job postings.

Data Processing: Cleaning and preprocessing the data to ensure consistency and accuracy, thus making it suitable for analysis.

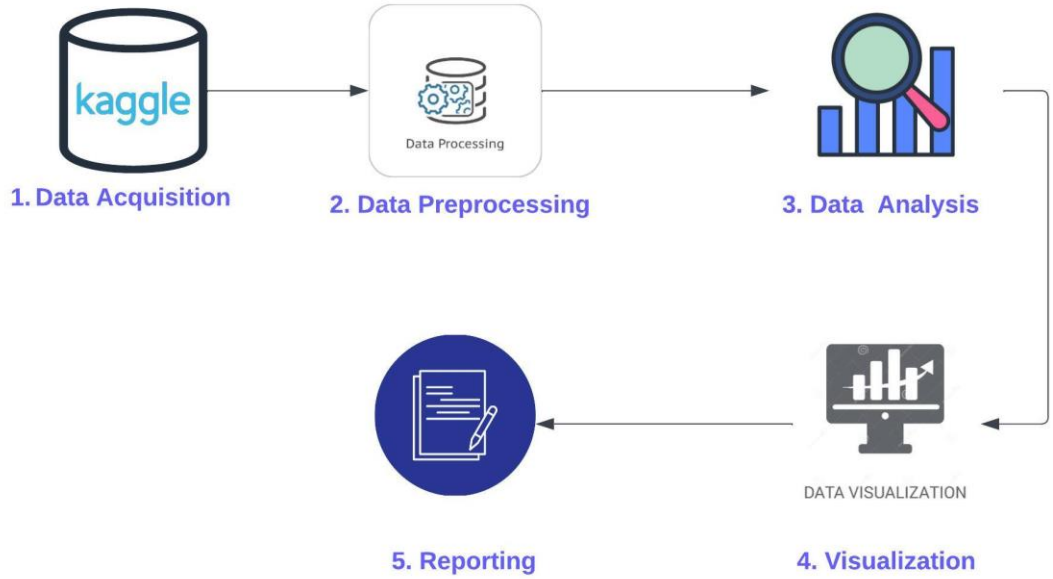
Data Analysis: Performing exploratory data analysis (EDA) to identify trends, patterns, and outliers within the dataset.

Visualization: Producing clear visualisations that enable you to see immediately how salaries vary by job title, years of experience and city or state.

Interpretation and Reporting: Building an understanding about the insights the visualisations offer with respect to an actionable query, and creating a report based on those insights for the target audience.

These actions in our project attempt to transform the raw complexity of the data into graphic narratives that expose the trends in data science salaries.

Workflow Diagram:



The above Task workflow shows the essential steps:

1. Data Acquisition
2. Data preprocessing
3. Data analysis
4. Visualization
5. Reporting

The Task Workflow Diagram depicts a standardised procedure which involves processing raw data into usable information by passing through several steps of analysis. It begins with Data Acquisition which extract data from particular source. In this situation, data gathered from Kaggle - a website which provide huge datasets for public. This is crucial step because analysis will be associated with the area of data acquisition. Subsequently comes Data Preprocessing which is the step of cleaning, filtering and structuring of raw data before that actual analysis takes place. Outdated, duplicated or irrelevant data should be removed to keep the dataset free from any contamination.

After this, the second stage, we start Data Analysis, where we employ statistical tricks and algorithms to discover trends, anomalies and relationships in the pre-processed data. This stage is important to winnow down mountains of data into meaningful distillates. Next, we start Visualisation to convert the findings of Data Analysis into performance charts, tables and other graphical forms to allow stakeholders at all levels to understand and process the data. The

importance of effective visualisation lies in the fact that it is an essential way of bringing out the salient features of the data.

The final step is Reporting, in which the visualisations and insights are formalised into a report or presentation to communicate what was found and reached, and to aid decision making and strategic road-mapping. It's an entire workflow that must be repeated until it delivers meaningful conclusions. Data science is an generally an iterative process.

▪ Tools Description

Data Processing & Visualization Tools:

This project implementation needed plenty of specialized tools to process the dataset, generate visuals, and evaluate data science salary trends.

Python: Python was used primarily to analyse data; it has particularly strong libraries for manipulating data and is very common in the data science domain. The syntax is quite student-friendly (which is important since we're using it with students at all levels).

Pandas: The Pandas library was used to accomplish these tasks through data reading and handling (from the CSV file), cleaning up, restructuring and reshaping the data into a format ready for analysis.

Matplotlib and Seaborn: These libraries were used in order to visualise the data. Matplotlib was a starting point, providing a framework for building graphs, but Seaborn, which is built upon Matplotlib, proved to be able to produce more pleasing and complex visualisations than Matplotlib, using less code.

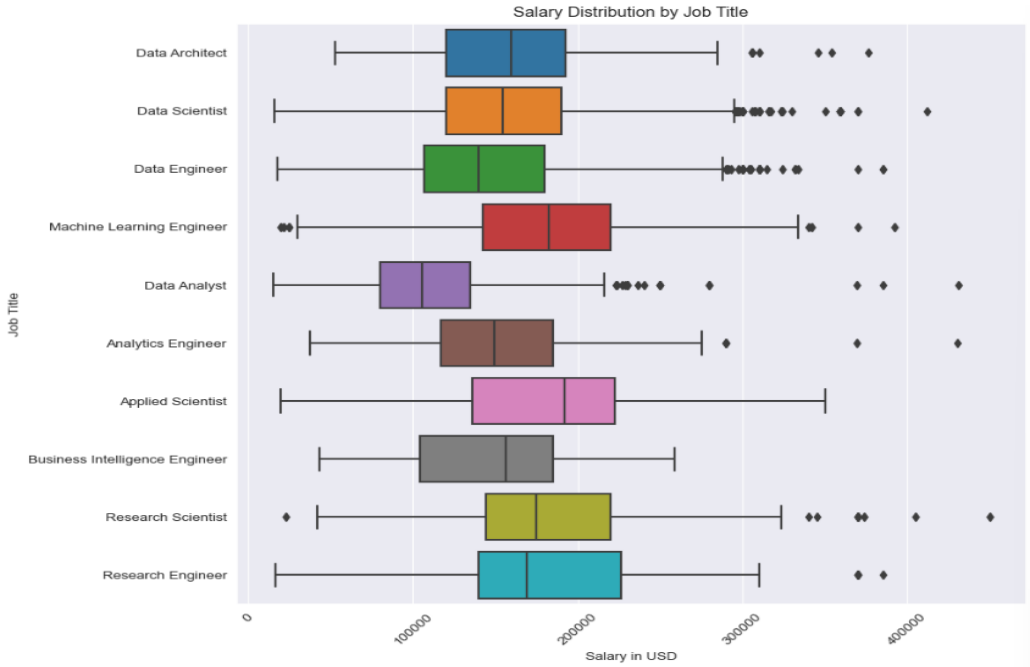
Jupyter Notebook: We've used Jupyter Notebooks to make the coding interactive and enable immediate feedback and visualisation throughout.

Kaggle: Kaggle has one of the largest open-source databases for data science and machine learning, and offers small cash prizes to developers worldwide as they work to compete in data-driven solutions for popular data problems.

All the above tools play a vital role in managing different aspects and stages of the project and are essential.

Visualization Graphs

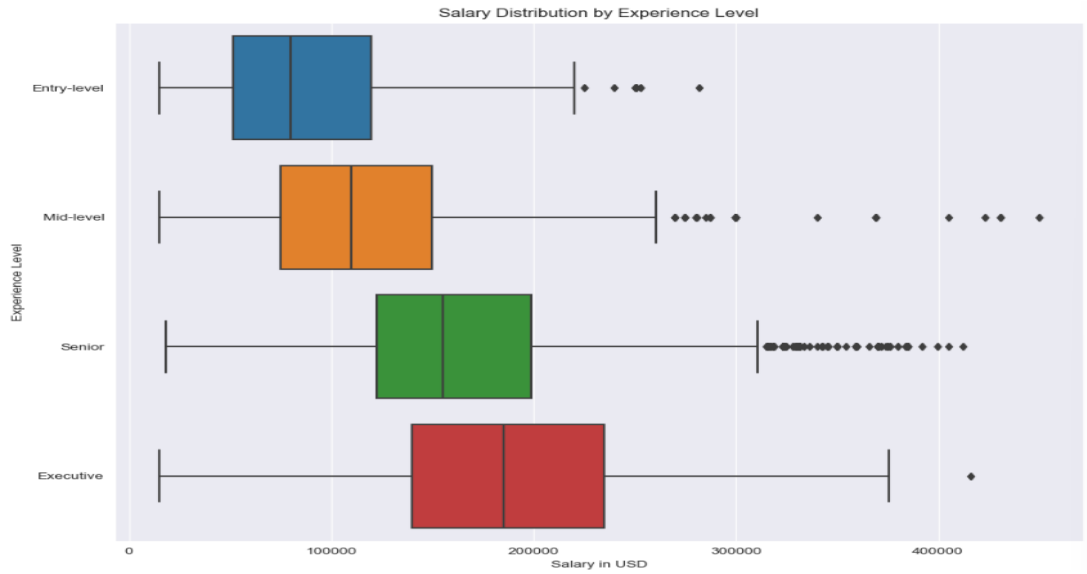
1. Salary Distribution by Job Title:



Style: Box plot with different colors representing distinct job roles within the data field and titles on y-axis whereas x-axis represents Salary in USD.

Description & insights: Some jobs also have a higher salary dispersion that suggests a strong variance based on company size, location, or specific required skills. In some job roles, the salary levels show some outliers, obviously in the very highest pay levels, likely reflecting a very high demand skills or seniority or working in a high paying region.

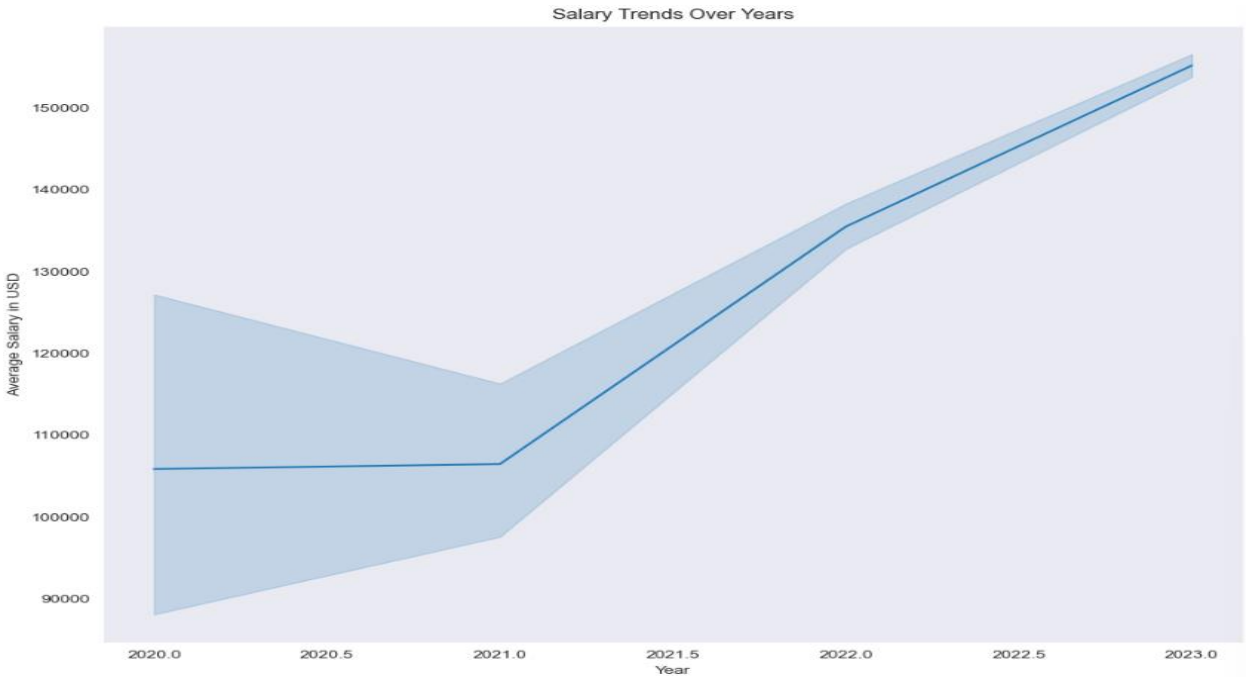
2. Salary Distribution by Experience Level:



Style: Box plot with four different colors representing distinct experience levels within the data field and experience levels on y-axis whereas x-axis represents Salary in USD.

Description & insights: There is a clear trend that salaries are higher with growing experience levels. The lowest median salary is for entry-level jobs, with progressive increasing for mid-level, senior and executive roles showing also the relevance of experience in data science jobs compared to other job profiles, with a large difference in the salaries between senior and executives roles.

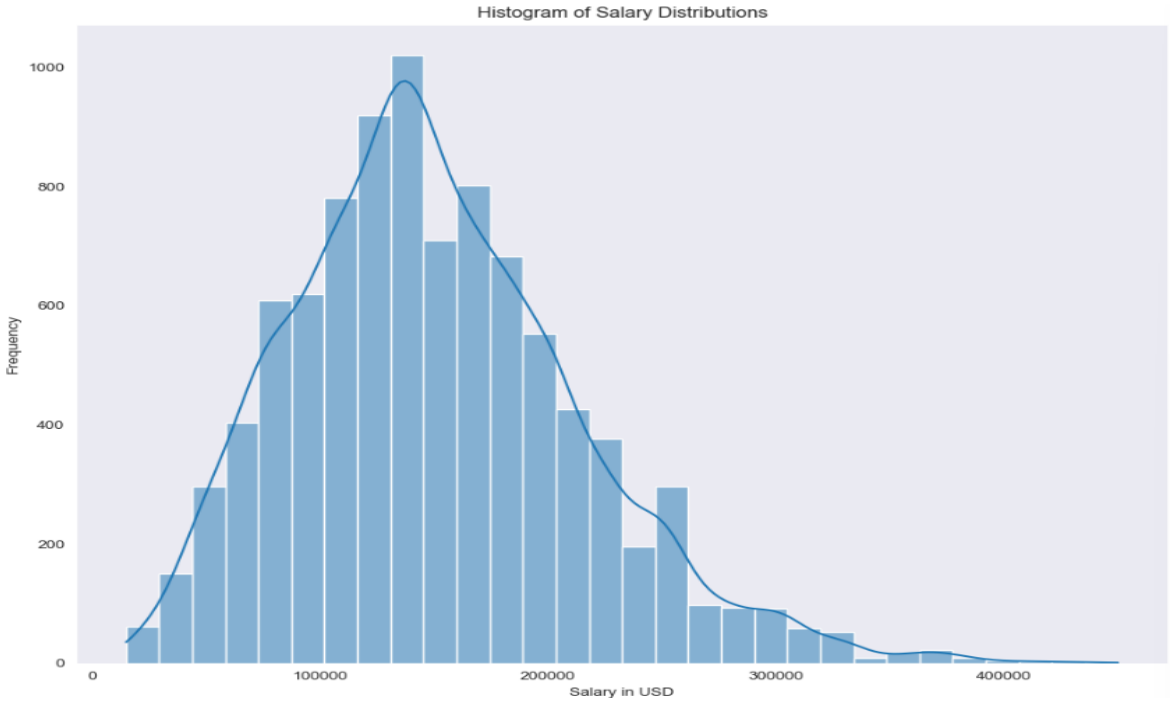
3. Salary Trends Over Years:



Style: Line plot with bounding higher and lower averages, the line represents estimated mean. We use simple blue line and light blue for the bounds for cleaner representation.

Description & insights: This line plot demonstrates that there was an upward trend in the average salary in data science. This is partly due to the fact that data science roles are becoming more important in the past years and there is also an increasing demand for professionals who possess this set of skills, which led to an increase in the economic value of these skills.

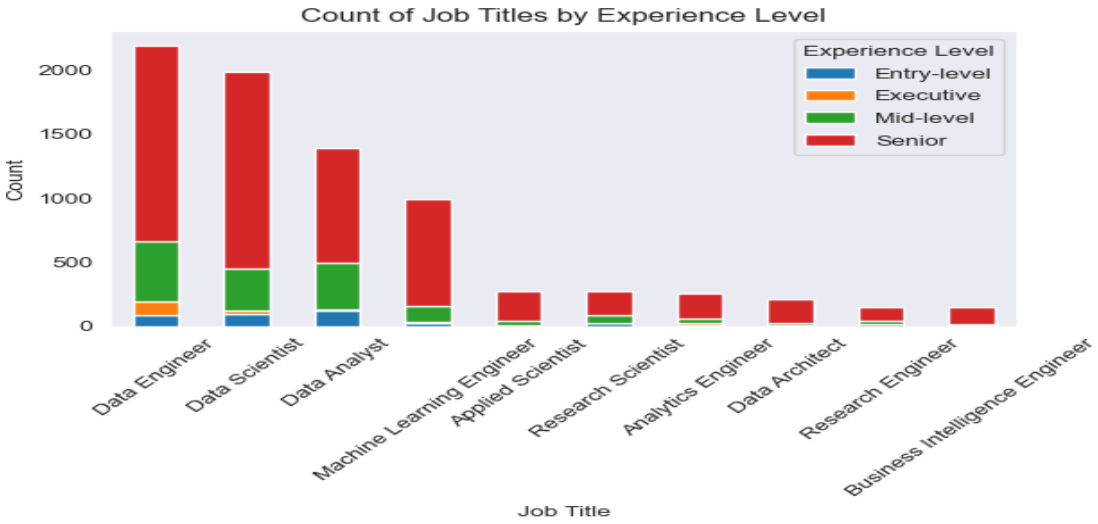
4. Histogram of Salary Distributions:



Style: Histogram with overlaying line that represents the trends as the salary increases towards the x-axis, simple blue color has been selected to represent the bars.

Description & insights: The histogram displays the pay structure for various roles in data science. It is evident that the concentration of salaries in a particular range. Also, there is an evident long tail to the right, indicating that there is high salary in this field, although the majority falls around a central high range, but there are exceptional cases.

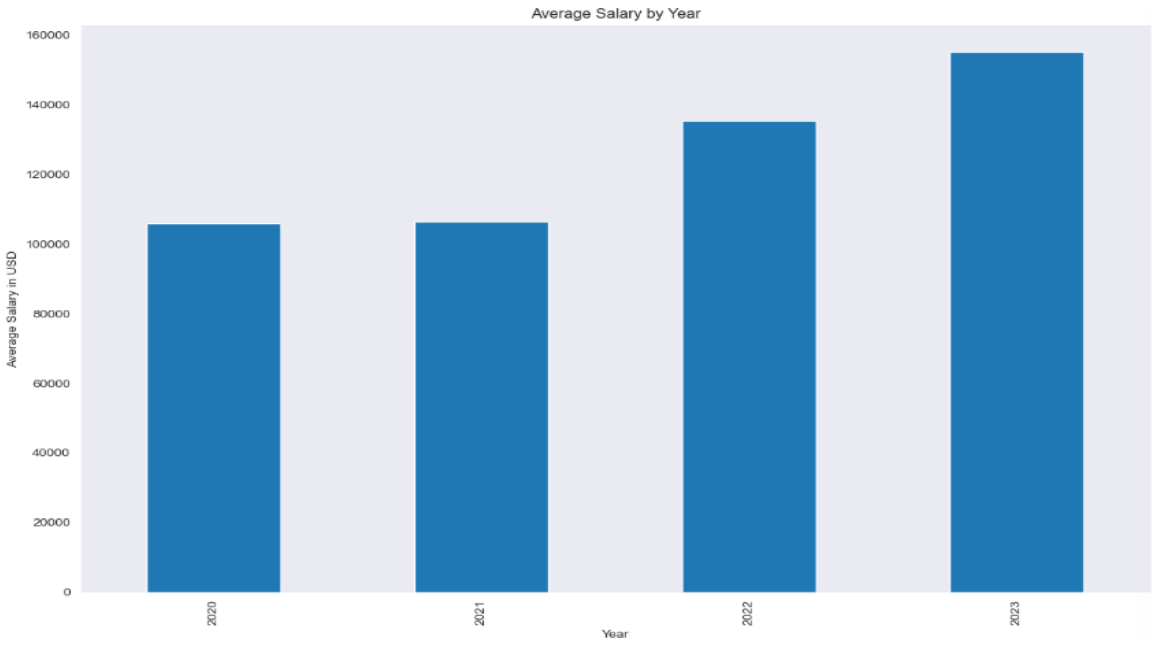
5. Count of Job Titles by Experience Level:



Style: Stacked bar graph representing all the experience levels with distinct colors across the popular roles and the count of jobs available, the selection of stacked bar graph is very appropriate because it conveys more information in a single graph.

Description & insights: This bar graph shows the distriubition of the jobs, between different experience levels of the same jobs, in percentage. Job level is on the Y axis, whereas experience is shown in categories on the X axis. Based on the end result of this graph it can be said that in which level wich job is more common. It will give an idea about the career advancement in datascience, and the typical path of the paths it has , and the level where it is common in some jobs.

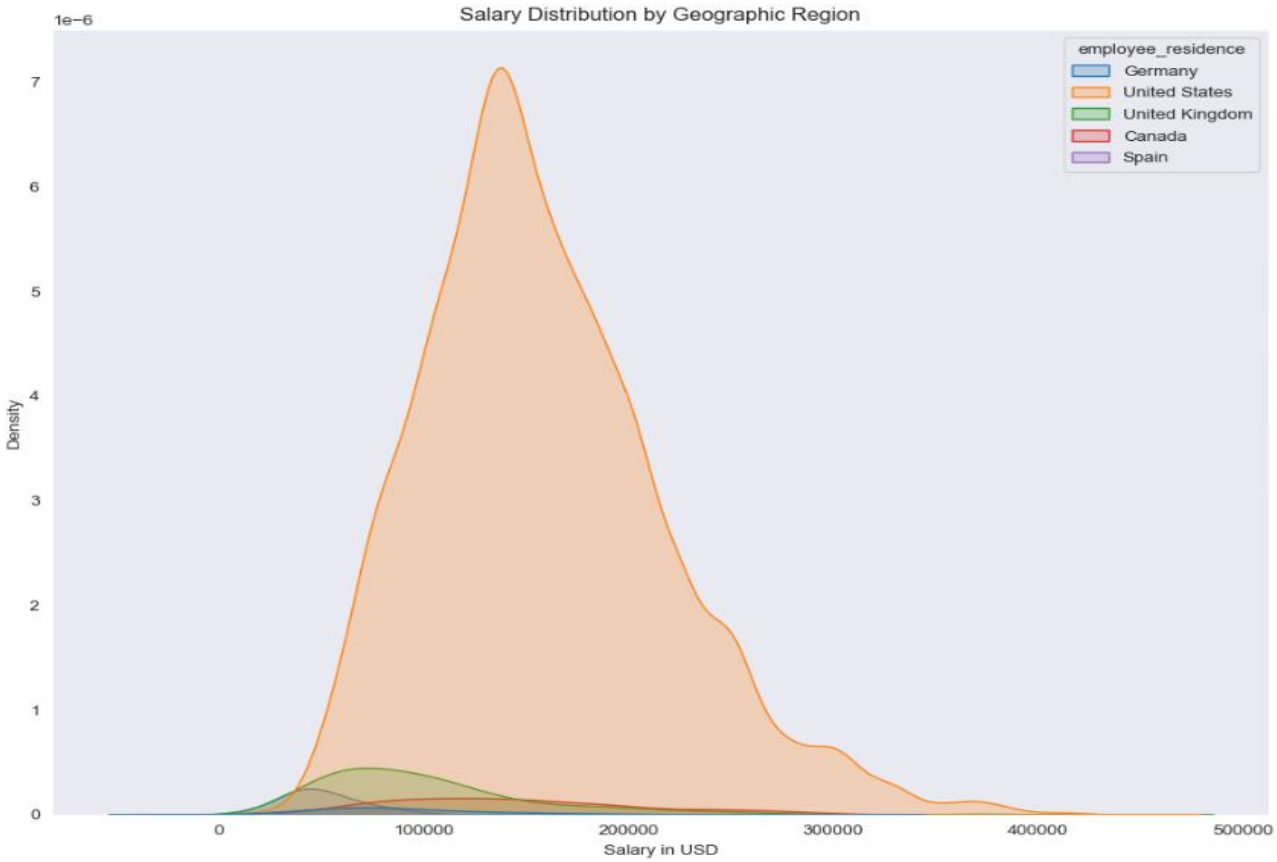
6. Average Salary by Year:



Style: This bar graph shows the average salary trends by year as we can see the salary increases towards the x-axis, simple blue color has been selected to represent the bars.

Description & insights: The bar graph illustrates how average salaries changed through the years listed in this data and shows the steady increase through the years. This supports the statement that Data science professionals' value are going up for the reason that data driven decisions gaining more importance for different industries.

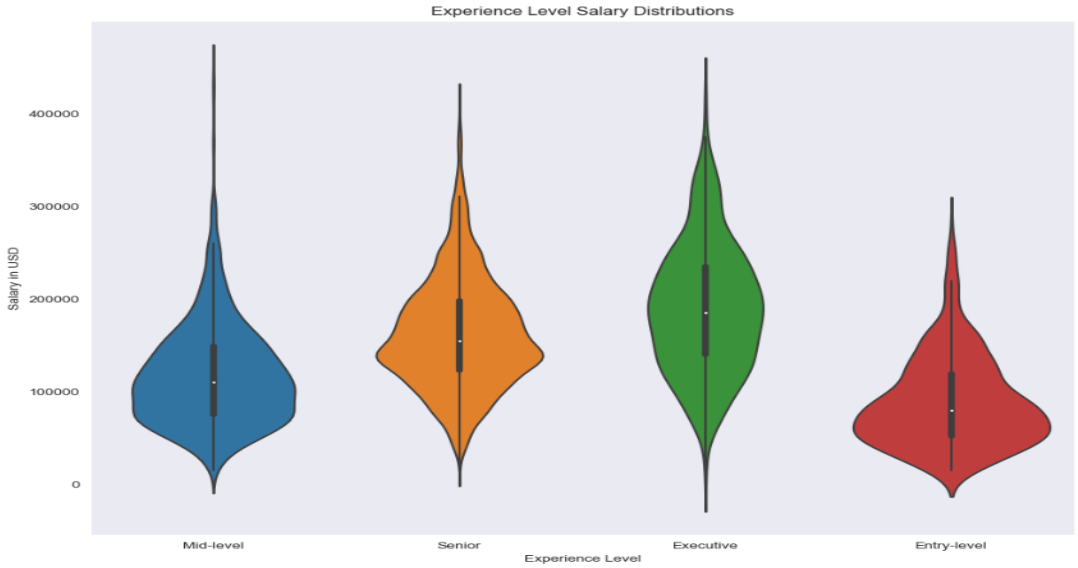
7. Salary Distribution by Geographic Region:



Style: Smoothed representation of salary distributions across different regions, offering a clear view of how salary densities vary geographically.

Description & insights: The above figure illustrates the kernal density distribution comparing the salary of data scientists in five geografical area, Germany, the US, the UK, Canada and Spain. It is evident from the given data, the salary disribution in the US is wider as well as higher compared to other countries. This might indicate that the demand for the skills of data science in the US is higher, This might also be induced by a number of workers being within the US. This might also reflect a larger economy of the US.

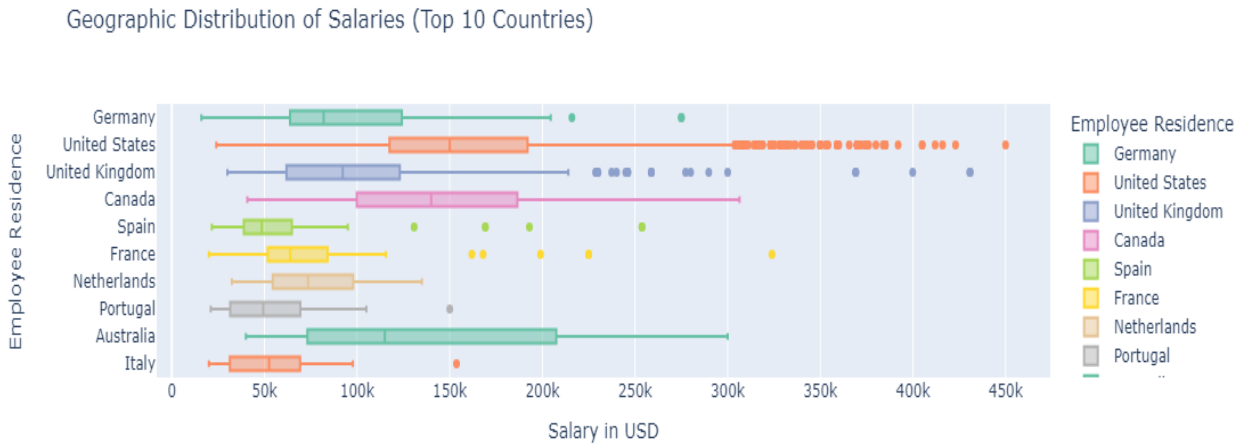
8. Experience Level Salary Distributions:



Style: Violin plot with distinct colors representing different experience levels along the salary distribution.

Description & insights: Violin plots present the salary distributions by experience in more detail compared to box plots, by combining kernel density estimates of the distributions with box plots structure and whiskers. They reveal the spread of the salaries within each experience level, as well as the skewness of the distribution due to the presence of high-paying senior roles at the right tail of each distribution.

9. Interactive Geographic distribution of salaries:

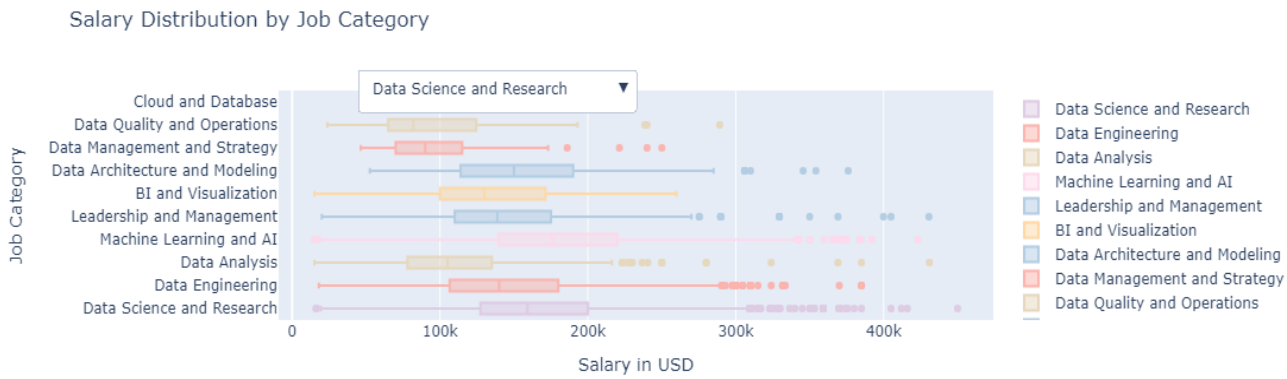


Style: The above box plot displays the distribution of salaries across top 10 countries. Each country has been assigned to different color helping in distinguishing them, plot shows median, quartiles and outliers aswell.

Description & insights:

Salaries are spread across a horizontal axis providing a clear view of each countries distribution. Data points that are far from the distribution are outliers which are dots. The plot shows that united states have wide range of salaries with highest figures showing the competitive market, other countries when compared only Australia shows a good range of salaries apart from other countries.

10. Interactive Salary Distribution by Job Category:



Style: The above box plot shows the salary distribution across different job categories in our dataset, this has an interactive drop down to inspect each of the category.

Description & insights:

Each category has been assigned a distinct color to differentiate from each other, We can see that Data Science and Research offers the high paid salaries over all the categories with data management and quality operations with less, data engineering shows less variation indicating more stable market with consistent salary expectations.

11. Interactive Average Salary by Job title

Style: The bar chart represents average salary in the US based on data science job categories, where each color represents a different range in salaries.

Description & insights: The bar plot shows a color legend to easily translate the salary range, the applied scientist job seems to have higher demand and data science, machine learning engineer follow closely which aligns with high value skillsets, data analyst being the less averagely paid job in the market due to most people are getting into these roles making it common.

▪ Story Telling

Chapter 1: Life

We delve into the narrative of data scientists and analysts who are navigating the complexities of the job market. The protagonists of our story are aspiring and seasoned professionals in the data science field, facing the challenge of understanding their worth and finding suitable employment.

Who: Data scientists and analysts, from fresh graduates to seasoned experts, form the core community in need.

What: They are confronted with the task of navigating a volatile job market with varying compensation and opportunities.

When: The timeline is set in the current year, amid economic fluctuations and rapid technological advancements.

Where: 1) The community thrives in a digitally connected world, often in urban settings where tech industries flourish. 2) The problem unfolds in the global job market, particularly within the tech and data-centric sectors.

Why: The issue arises from a lack of transparency in salary norms and expectations, compounded by the diverse range of roles and responsibilities in the data science domain.

How: The challenge manifests through the complex dynamics of job hunting and career progression, where data and insights are key to making informed decisions.

Chapter 2: Data

Here, we connect the extensive datasets to the real-world challenges identified previously, engaging in data management and processing to reveal underlying patterns and insights.

Who: The dataset encompasses a wide spectrum of data science professionals, with a particular focus on those actively seeking employment or career advancement.

What: It records job titles, salaries, locations, and employment types, capturing the economic landscape of the data science profession.

When: The data encapsulates a snapshot upto year 2023, providing a cross-sectional glimpse into the industry.

Where: Data is sourced globally, with a focus on regions with significant tech industries, potentially underrepresenting areas with less tech presence.

Why: The data was collected to empower professionals with knowledge about current salary trends and market demands.

How: Through surveys and aggregations from job postings, the data set was compiled to offer transparency and facilitate better career decisions.

Chapter 3: Users

This Chapter brings to life the interaction between the user and the application or visualization tool designed to empower them with actionable insights.

Who: The users are data science professionals, from job seekers to HR managers, looking to understand the market better.

What: The application serves as an analytical tool, presenting visualizations of salary trends, job distribution, and market demands.

When: It can be utilized anytime, with real-time updates offering the most current market snapshots.

Where: The tool is accessible through various platforms, such as web-based dashboards or mobile applications, catering to the user's preference.

Why: Visualization aids users in making informed decisions about career moves and hiring practices.

How: Users engage with the application to filter and select data relevant to their needs, utilizing insights to guide negotiations, identify skills gaps, or shape educational pursuits.

This structured narrative framework effectively communicates the journey from the problem faced by the data science community to the data-driven solution provided by the application, highlighting the real-world impact of data analysis and visualization.

- Project Management

✓ Work completed.

Task	Responsibility	Contribution
Dataset Acquisition and Cleaning	First, we downloaded an enormous dataset from Kaggle about jobs, their titles, salaries, seniority, and geographic location. Next, we performed the initial preprocessing such as missing values handling and variables.	Sai Krishna Meduri: 40% Sreecharan Vanam: 60%
Data Analysis and Visualization	Performed the exploratory analysis of the data to identify the patterns and trends. Created several charts and graphs to depict salaries by each title, tenure and location.	Sai Krishna Meduri: 50% Sreecharan Vanam: 50%
Preparing the Final report	Writing the final report with all the sections and summarizing the interpretations	Sai Krishna Meduri: 60% Sreecharan Vanam: 40%

References/Bibliography

[1] Data Science 2023 Review: Trends and Salary Expectations

[2] Data Science & AI Professionals Salary and Hiring Trends

[3] Salary Trends: What to Expect in Data & Analytics Roles in 2024

[4] Trends in Data Science Salaries: An Exploratory Data Analysis Journey

[5] EDA and Visualizations on Data Science Salaries using Python

[6] A. Kaur, D. Verma and N. Kaur, "Utilizing Quantitative Data Science Salary Analysis to Predict Job Salaries," 2022 2nd International Conference on Innovative Sustainable Computational Technologies (CISCT), Dehradun, India, 2022, pp. 1-4

[7] Tee, Zhen & Raheem, Mafas. (2022). Salary Prediction in Data Science Field Using Specialized Skills and Job Benefits -A Literature Review. 70-74.