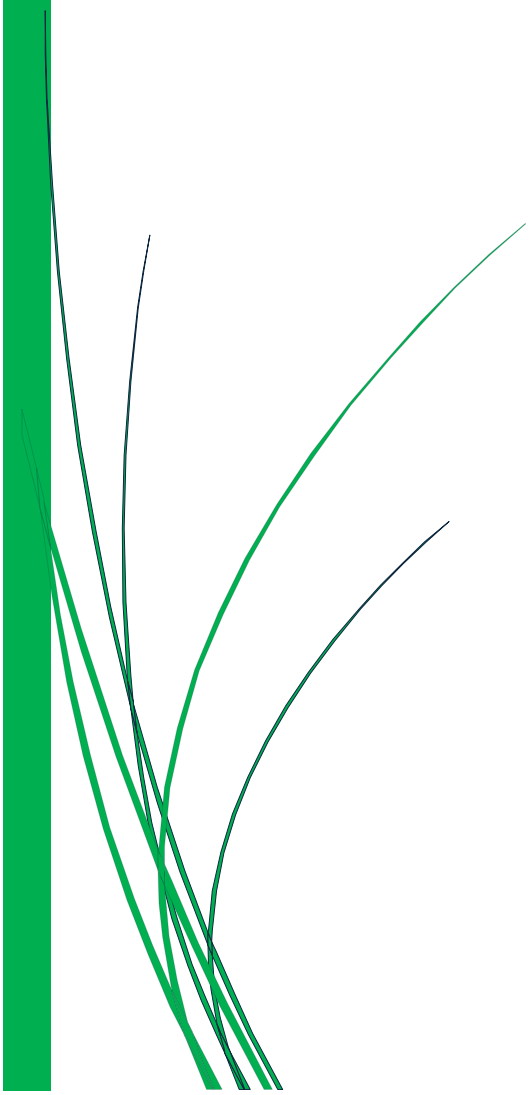
5/1/2024

FINAL REPORT

DV304.3



Group – E
PGJ Lakshani. (25180)
HHN Peiris. (25011)

Abstract

In the dynamic field of data science, understanding job market dynamics is pivotal. This project delves into Glassdoor's dataset via Kaggle, offering comprehensive insights into the multifaceted landscape of data science-related jobs. With a focus on job titles, salary distributions, and top hiring companies, the project aims to illuminate the diverse roles within the sector. It addresses challenges like varying job descriptions, emerging trends, and evolving skill demands through meticulous analysis. The dataset, comprising 15 essential variables, provides a detailed snapshot of the industry's nuances, including company details, revenue, and competitor information.

The project unfolds its narrative through a storytelling data visualization dashboard, employing tools like Power BI or Tableau. By integrating theoretical insights, and raw data transformation using data blending, cleaning, and advanced visualization techniques, the project navigates the complex realm of data science job opportunities. It acknowledges potential challenges, such as the dataset's focus on the United States, ensuring transparent reporting and reliable outcomes despite limitations.

The primary goals encompass understanding the data science job market dynamics, identifying salary trends, and offering valuable insights to both job seekers and companies. The project's systematic timeline adheres to set milestones, ensuring a comprehensive approach from initiation to the completion of the proposal, data analysis, and the preparation of an effective data visualization dashboard. By aligning with guidelines and leveraging this dataset, the project intends to contribute nuanced insights, simplifying the data science job landscape for stakeholders in the US market and crafting a compelling narrative through an interactive and informative dashboard.

Table of Contents

1. Introduction

1.1 Background study of the selected sector & reason for selecting the proposed sector.

The data science sector is witnessing unprecedented growth, with diverse job opportunities emerging across industries. As data science undergraduates, exploring this sector offers a firsthand understanding of the evolving demands and trends within the field. The choice of this sector aligns with the increasing importance of data-driven decision-making and the significance of data science professionals in today's job market.

This dataset provides a unique opportunity to observe the variety of data science-related jobs, identify the top companies actively hiring for these roles, and gain insights into salary distributions in the US job market. Exploring these dimensions will contribute to a holistic understanding of the data science job landscape.

In the era of data-driven decision-making, the field of data science plays a pivotal role in shaping the landscape of various industries. The vast amount of job-related information available on platforms like Glassdoor provides a unique opportunity to delve into the trends and dynamics of the data science job market. The selected sector, focused on data science job postings, not only aligns with the growing demand for skilled professionals but also allows for an in-depth exploration of factors influencing job opportunities, salary ranges, and industry preferences. By harnessing the power of data visualization, this project aims to not only uncover the variety of data science-related roles but also to identify the top industries actively seeking professionals in this domain and their salary ranges in the US.

1.2 Problems related to the selected area.

The data science sector, while flourishing, presents intricate challenges for both employers and prospective professionals. The abundance of diverse job titles, constantly shifting salary ranges, and the dynamic nature of skill requirements contribute to the complexity of this domain. This project acknowledges these challenges and aims to offer a solution through a thorough analysis. By examining the details of the data science job market, including the vast array of job titles, identifying key hiring companies, and understanding salary distributions in US, the project strives to provide valuable insights that address the multifaceted challenges associated with navigating this dynamic landscape.

1.3 Objectives of the Project

The primary objectives are,

• Conducting an in-depth analysis of data science job postings.

• Uncovering trends in job titles, salary distributions, and company attributes.

- o Investigate the salary estimation variable to gain insights into the distribution of salaries for different data science job roles in the US.
- o Identify and showcase the top companies in the US actively hiring for data science roles, considering factors such as company size, industry, and location.
- o Analyze the dataset to understand the geographical distribution of data science jobs, emphasizing the states and cities of the US with high job demand.
- o Explore the dataset to identify profiles of companies based on variables such as ratings, headquarters location, and company age.

• Developing an interactive data visualization dashboard to facilitate exploration and understanding.

1.4 Expected Limitations on the visualization process pertaining to the selected area.

While the dataset presents a valuable resource, limitations may arise from variations in job descriptions and potential biases in user-generated content on Glassdoor.

- • The dataset focuses on U.S. data science job postings. Findings may not be universally applicable to global job markets.
- • Visualization quality depends on accurate and complete data. Possible gaps or inaccuracies in Glassdoor data may affect the precision.
- • Data captures a specific timeframe, potentially missing real-time changes. Insights may not fully reflect the latest developments in the dynamic field.
- • Subjectivity or variations in categorizations (job titles, industries) may cause confusion. Careful consideration is needed for potential issues in categorization.

To mitigate these challenges, the project will employ strict data-cleaning practices and transparently report any constraints encountered during the visualization process.

1.5 Proposed Work Schedule

Task	Start Date	End Date
1. Initiating the project	1/30/2024	5/1/2024
2. Data collection	1/30/2024	2/6/2024
3. Completion of proposal	2/6/2024	2/11/2024
4. Analysis	2/13/2024	3/13/2024
5. Preparation of the visualization dashboard	3/13/2024	4/25/2024
6. Final report completion	4/25/2024	5/1/2024
7. Presentation	5/1/2024	5/2/2024

2. Literature Review

2.1 Introduction to the dataset selected.

The dataset under consideration, obtained from Kaggle that consists of the data of web-scraped job posts from Glassdoor, contains a rich compilation of information on data science-related job postings.

Link: https://www.kaggle.com/datasets/rashikrahmanpritom/data-science-job-posting-on-glassdoor/data?select=Uncleaned_DS_jobs.csv

- Overview of Dataset Structure

With 15 columns and 672 records, the dataset is structured to capture essential details about each job posting. From job titles and salary estimations to company information and industry specifics, the dataset's comprehensive structure positions it as a valuable resource for in-depth analysis.

- Diversity of Data Science Roles

An initial exploration of the dataset underscores its value in showcasing the variety of data science-related positions available. Roles such as Data Scientist, Data Engineer, and Senior Data Scientist feature prominently, indicating the multifaceted nature of opportunities within the field. This diversity forms a critical aspect of the dataset's contribution to understanding the intricacies of data science job postings.

- Geographic Distribution

The dataset sheds light on the geographic distribution of data science jobs, emphasizing key locations such as San Francisco, CA, New York, NY, Washington, DC, and Boston, MA. This regional breakdown underscores the widespread demand for data science professionals across major cities and states in the United States. Such insights prove valuable for both job seekers and companies seeking to establish or expand their presence in the field.

2.2 Table for Variables, their definitions and sources

Uncleaned dataset Variable.

Variable	Definition	Source
Index	A series of labels that identify each row	Kaggle
Job Title	Title of the job posting	Kaggle/Glassdoor
Salary Estimation	Salary range for the job	Kaggle/Glassdoor
Job Description	Detailed description of the job	Kaggle/Glassdoor
Rating	Rating of the job posting	Kaggle/Glassdoor
Company Name	Name of the hiring company	Kaggle/Glassdoor
Location	Location of the company	Kaggle/Glassdoor
Headquarters	Location of the company's headquarters	Kaggle/Glassdoor
Size	Total number of employees in the company	Kaggle/Glassdoor
Founded	Founded Year of the Company	Kaggle/Glassdoor
Type of Ownership	Company ownership type (e.g., public, private)	Kaggle/Glassdoor
Industry	Industry in which the company operates	Kaggle/Glassdoor
Sector	Sector of the company (e.g., Technology, Healthcare)	Kaggle/Glassdoor
Revenue	Total revenue of the company	Kaggle/Glassdoor
Competitors	Competitors of the company	Kaggle/Glassdoor

The cleaned dataset now contains essential variables like 'Job Title,' 'Job Description,' 'Rating,' 'Company Name,' 'Location,' 'Headquarters,' 'Size,' 'Type of Ownership,' 'Industry,' 'Sector,' 'Revenue,' 'Min Salary K$,' 'Max Salary K$,' 'Avg Salary Estimate,' 'Years in Market,' 'Location State,' 'Same State,' 'Job Role,' 'Job Seniority,' and skill-based categorizations.

Cleaned dataset Variable.

Variable	Definition	Source
Job Title	Title of the job posting	Kaggle/Glassdoor
Job Description	Detailed description of the job	Kaggle/Glassdoor
Rating	Rating of the job posting	Kaggle/Glassdoor
Company Name	Name of the hiring company	Kaggle/Glassdoor
Location	Location of the company	Kaggle/Glassdoor
Headquarters	Location of the company's headquarters	Kaggle/Glassdoor
Size	Total number of employees in the company	Kaggle/Glassdoor
Type of Ownership	Company ownership type (e.g., public, private)	Kaggle/Glassdoor
Industry	Industry in which the company operates	Google/Glassdoor
Sector	Sector of the company (e.g., Technology, Healthcare)	Google /Glassdoor
Revenue	Total revenue of the company	Kaggle/Glassdoor
Min Salary K$	Minimum salary in thousands (K$)	Created (Python)
Max Salary K$	Maximum salary in thousands (K$)	Created (Python)
Avg Salary Estimate	Average salary estimates	Created (Python)
Years in Market	Number of years the company has been in the market	Created (Python)
Location State	State where the company is located	Created (Python)
Same State	Indicates if the location and headquarters are in the same state	Created (Python)
Job Role	Role categorization based on job title	Created (Python)
Job Seniority	Seniority level of the job	Created (Python)
Excel	Presence of Excel skills in the job description	Created (Python)
Sql	Presence of SQL skills in the job description	Created (Python)
Python	Presence of Python skills in the job description	Created (Python)
power_bi	Presence of Power BI skills in the job description	Created (Python)
Tableau	Presence of Tableau skills in the job description	Created (Python)
Scikit	Presence of Scikit skills in the job description	Created (Python)
Spark	Presence of Spark skills in the job description	Created (Python)

Moving forward, the cleaned and transformed dataset will serve as the foundation for our subsequent analyses, including exploratory data analysis (EDA) and the development of an interactive data visualization dashboard. This progress emphasizes our commitment to delivering insightful and meaningful findings regarding the data science job landscape.

3. Data Preparation Process

3.1. Data Blending & Integration

Data blending is the process of combining and integrating multiple datasets to create a more comprehensive and enriched dataset. It involves merging data from different sources or datasets, aligning them based on common attributes, and creating a unified dataset that captures a more holistic view of the information. This technique is particularly useful when dealing with diverse datasets that individually contain valuable insights.

In the case of our selected dataset, data blending wasn't explicitly required. The dataset "Data Science Job Posting on Glassdoor" obtained from Kaggle already provided a consolidated view of job-related information, including job titles, company details, and salaries.

Our focus in this project was primarily on data cleaning, transformation, and reduction to ensure the dataset's accuracy, consistency, and relevance for subsequent analysis. While data blending might be crucial in scenarios involving disparate datasets, the completeness of our chosen dataset made it unnecessary for this specific project.

3.2. Data Cleaning

Data cleaning, also known as data cleansing or data scrubbing, is a crucial step in the data preparation process. Its primary goal is to ensure that datasets are accurate, consistent, and free of errors or inconsistencies.

Uncleaned dataset: https://1drv.ms/u/s!AkI7kbf9N08hgo1gW-ekLz6WTz4FCA?e=hcgPMC

Python Notebook (Including the data preprocessing): https://1drv.ms/u/s!AkI7kbf9N08hgo1fHhLgBLWlKJtOoQ?e=df6Qwa

Cleaned dataset: https://1drv.ms/u/s!AkI7kbf9N08hgo1l5tcOD56ul5g2xg?e=6k9007

3.2.1. Handling Data Type.

As part of our rigorous data preparation process, we have meticulously verified and confirmed that each variable in our selected dataset is assigned the correct data type. This crucial step in data quality assurance ensures that our dataset is accurate, reliable, and ready for meaningful analysis.

3.2.2. Data Formatting.

In the data formatting phase, our attention is directed toward refining the 'Job Title' column for improved clarity and consistency. Key steps include identifying and standardizing seniority indicators (e.g., replacing 'Sr.' with 'sr.'), removing extraneous information enclosed in parentheses, and carefully addressing special characters. By ensuring uniformity and eliminating unnecessary elements, this meticulous formatting process enhances the interpretability and structure of the 'Job Title' column, paving the way for more focused and meaningful analyses.

```
In [6]: # Make the replacement
        job_data["Job Title"]=job_data.loc[:,"Job Title"].str.replace("(Sr.)","sr.")

In [7]: # Delete the rest of the instances using a regex
        job_data["Job Title"]=job_data.loc[:,"Job Title"].str.extract('([^()]+)')

In [8]: #Replace the special characters with an empty value by defining a regex pattern
        job_data['Job Title'] = job_data['Job Title'].str.replace(r'[^a-zA-Z0-9-,/\s]','',regex=True)

In [9]: # Check if there are missing values in the "Job Title" column
        job_data.loc[job_data.loc[:,"Job Title"]=="-1"]
```

Using regex, we refine the 'Company Name' column by removing delimiters and subsequent numerical values, focusing on extracting only the company names. This targeted approach enhances the precision and clarity of company names within the dataset, ensuring a streamlined and interpretable representation of roles.

```
In [58]: job_data['Company Name'] = job_data.loc[:,'Company Name'].str.replace(r"\n\d+(\.\d+)?",'',regex=True)
         job_data.loc[:,"Company Name"][:10]

Out[58]: 0           Healthfirst
         1               ManTech
         2       Analysis Group
         3              INFICON
         4   Affinity Solutions
         5          HG Insights
         6              Novartis
         7                iRobot
         8        Intuit - Data
         9   XSELL Technologies
         Name: Company Name, dtype: object
```

3.2.3. Handling Missing Data.

In our dataset, missing values are currently represented by the value -1 in various columns, including the 'rating' column. The 'rating' column also contains legitimate values of -1, making a direct replacement with a different non-numeric representation challenging. Changing the representation to something like 'N/A' could potentially alter the column's data type, posing challenges for its future usability.

```
In [11]: # Check if there are missing values in the "Job Description" column
         job_data.loc[job_data.loc[:,"Job Description"]=="-1"]
```

Out[11]:

index	Job Title	Salary Estimate	Job Description	Rating	Company Name	Location	Headquarters	Size	Founded	Type of ownership	Industry	Sector	Revenue	Competitors
351	351	Scientist	(Glassdoor est.)		mission is to buil...	-1.0	Ventures		CA	-1	-1	-1	-1	
357	357	Data Scientist	122K−146K (Glassdoor est.)		Job Overview: The Data Scientist is a key memb...	-1.0	Hatch Data Inc		San Francisco, CA	-1	-1	-1	-1	
358	358	Data Scientist	122K−146K (Glassdoor est.)		Job Overview: The Data Scientist is a key memb...	-1.0	Hatch Data Inc		San Francisco, CA	-1	-1	-1	-1	
359	359	Data Scientist	122K−146K (Glassdoor est.)		Job Overview: The Data Scientist is a key memb...	-1.0	Hatch Data Inc		San Francisco, CA	-1	-1	-1	-1	
360	360	Data Scientist	122K−146K (Glassdoor est.)		Job Overview: The Data Scientist is a key memb...	-1.0	Hatch Data Inc		San Francisco, CA	-1	-1	-1	-1	
361	361	Data Scientist	122K−146K (Glassdoor est.)		Job Overview: The Data Scientist is a key memb...	-1.0	Hatch Data Inc		San Francisco, CA	-1	-1	-1	-1	
362	362	Data	122K−146K (Glassdoor		Job Overview: The Data	-1.0	Hatch Data		San Francisco,	-1	-1	-1	-1	

```
In [14]: # Check if there are missing values in the "Rating" column
         job_data.loc[job_data.loc[:,"Rating"]==-1.]
```

```
In [ ]:
```

```
In [121]: # Replace missing values in the 'Rating' column with the mean of that column
          mean_rating = round(job_data['Rating'].mean(), 1)
          job_data.loc[job_data["Rating"] == -1, "Rating"] = mean_rating
```

```
In [122]: # Display the unique values in the "Rating" column after replacement
          job_data['Rating'].unique()
```

```
Out[122]: array([3.1, 4.2, 3.8, 3.5, 2.9, 3.9, 4.4, 3.6, 4.5, 4.7, 3.7, 3.4, 4.1,
                 3.2, 4.3, 2.8, 5. , 4.8, 3.3, 2.7, 2.2, 2.6, 4. , 2.5, 4.9, 2.4,
                 2.3, 4.6, 3. , 2.1, 2. ])
```

In the dataset, records with a value of "-1" in the "Headquarters" column are indicative of missing information. Subsequently, these records also lack values in other critical columns such as "Size," "Founded," "Type of Ownership," "Industry," "Sector," "Revenue," and "Competitors." To ensure the integrity and reliability of our analysis, we have opted to remove these records from the dataset.

index	Job Title	Salary Estimate	Job Description	Rating	Company Name	Location	Headquarters	Size	Founded	Type of ownership	Industry	Sector	Revenue	Competitors

```
In [43]: # Check if there are missing values in the "Headquarters" column
job_data.loc[job_data.loc[:,"Headquarters"]=="-1"]
```

Job Title	Salary Estimate	Job Description	Rating	Company Name	Location	Headquarters	Size	Founded	Type of ownership	Industry	Sector	Revenue	Competitors
ELISA RESEARCH SCIENTIST (CV-15)	$90K–109K (Glassdoor est.)	Covaxx, a subsidiary of the UBI Group, has joi...	-1.0	Covid-19 Search Partners	Hauppauge, NY	-1	-1	-1	-1	-1	-1	-1	-1
Machine Learning Engineer	$101K–165K (Glassdoor est.)	Overview\n\nRadical Convergence is a fast-pace...	-1.0	Radical Convergence	Reston, VA	-1	-1	-1	-1	-1	-1	-1	-1
Data Scientist	$90K–124K (Glassdoor est.)	Full Time Data Scientist\n\nThe Drive Media is...	5.0	The Drive Media, Inc.\n5.0	New York, NY	-1	Unknown	-1	Unknown	-1	-1	Unknown / Non-Applicable	-1
Data Scientist	$122K–146K (Glassdoor est.)	About Our AI/ML Team\n\nOur mission is to buil...	-1.0	Point72 Ventures	Palo Alto, CA	-1	-1	-1	-1	-1	-1	-1	-1

```
In [48]: # Remove records with "-1" in the "Headquarters" column
new_job_data = job_data[job_data["Headquarters"] != "-1"]
```

```
In [50]: # Check if there are any null value in "Headquarter"column
new_job_data.loc[job_data.loc[:,"Headquarters"]=="-1"]
```

Out[50]:

index	Job Title	Salary Estimate	Job Description	Rating	Company Name	Location	Headquarters	Size	Founded	Type of ownership	Industry	Sector	Revenue	Competitors

In the dataset, instances where the 'Founded' column holds a value of '-1' are indicative of missing information regarding the establishment year of certain companies. To address this, we conducted additional research using online sources, notably Google, to obtain accurate founding year information for a subset of more than 60 records. This manual verification process allowed us to replace the placeholder '-1' values in the 'Founded' column with the actual founding years, enhancing the precision and completeness of our dataset.

```
In [113]: # Check if there are any null value in "Founded"column
new_job_data.loc[new_job_data.loc[:,"Founded"]==-1]
```

Out[113]:

	index	Job Title	Salary Estimate	Job Description	Rating	Company Name	Location	Headquarters	Size	Founded	Type of ownership	Industry
69	69	Data Scientist - Machine Learning	$79K–131K (Glassdoor est.)	Headquartered in Brisbane, Calif., CareDx, Inc...	2.5	CareDx\n2.5	Brisbane, CA	Brisbane, CA	1 to 50 employees	-1	Company - Private	Biotech & Pharmaceuticals
112	112	Data Scientist (TS/SCI)	$99K–132K (Glassdoor est.)	Are you an intelligence analyst looking to fur...	3.5	Maxar Technologies\n3.5	Herndon, VA	Westminster, CO	5001 to 10000 employees	-1	Company - Public	Aerospace & Defense
162	162	Geospatial Data Scientist	$101K–165K (Glassdoor est.)	Maxar is seeking a seasoned Geospatial Data Sp...	3.5	Maxar Technologies\n3.5	Chantilly, VA	Westminster, CO	5001 to 10000 employees	-1	Company - Public	Aerospace & Defense
193	193	Data Scientist	$56K–97K (Glassdoor est.)	Job Description\nClient JD below:\n\nWe need a...	5.0	SkillSoniq\n5.0	San Francisco, CA	Jersey City, NJ	Unknown	-1	Company - Public	-1
195	195	Data Scientist	$56K–97K (Glassdoor est.)	About Joby\nLocated in Northern California, th...	4.3	Joby Aviation\n4.3	San Carlos, CA	Santa Cruz, CA	51 to 200 employees	-1	Company - Private	-1
...	...	...	...	...	...	...	...	...	...	...	...	...

```
# Replace the "Founded" yr column
new_job_data.loc[new_job_data["Company Name"] == "CareDx", "Founded"] = 1998
new_job_data.loc[new_job_data["Company Name"] == "Maxar Technologies", "Founded"] = 2017
new_job_data.loc[new_job_data["Company Name"] == "SkillSoniq", "Founded"] = 2016
new_job_data.loc[new_job_data["Company Name"] == "Joby Aviation", "Founded"] = 2009
new_job_data.loc[new_job_data["Company Name"] == "Comtech Global Inc", "Founded"] = 2006
new_job_data.loc[new_job_data["Company Name"] == "Qurate Retail Group", "Founded"] = 1998
new_job_data.loc[new_job_data["Company Name"] == "SolutionIT, Inc", "Founded"] = 1989
new_job_data.loc[new_job_data["Company Name"] == "Clear Ridge Defense", "Founded"] = 2015
new_job_data.loc[new_job_data["Company Name"] == "ChaTeck Incorporated", "Founded"] = 2000
new_job_data.loc[new_job_data["Company Name"] == "Encode, Inc", "Founded"] = 2000
new_job_data.loc[new_job_data["Company Name"] == "Surya Systems", "Founded"] = 1997
new_job_data.loc[new_job_data["Company Name"] == "Predictive Research Inc", "Founded"] = 1999
new_job_data.loc[new_job_data["Company Name"] == "Sprezzatura Management Consulting", "Founded"] = 2011
new_job_data.loc[new_job_data["Company Name"] == "Descript", "Founded"] = 2017
new_job_data.loc[new_job_data["Company Name"] == "Better Hire", "Founded"] = 2016
new_job_data.loc[new_job_data["Company Name"] == "Tygart Technology, Inc", "Founded"] = 1992
new_job_data.loc[new_job_data["Company Name"] == "Advanced Bio-Logic Solutions Corp", "Founded"] = 2002
new_job_data.loc[new_job_data["Company Name"] == "Central Business Solutions, Inc", "Founded"] = 2000
new_job_data.loc[new_job_data["Company Name"] == "Unicom Technologies INC", "Founded"] = 2005
new_job_data.loc[new_job_data["Company Name"] == "Trovetechs Inc", "Founded"] = 2005
new_job_data.loc[new_job_data["Company Name"] == "PETADATA","Founded"] = 2016
new_job_data.loc[new_job_data["Company Name"] == "Capio Group", "Founded"] = 2014
new_job_data.loc[new_job_data["Company Name"] == "Colony Brands", "Founded"] = 1982
new_job_data.loc[new_job_data["Company Name"] == "Kollasoft Inc", "Founded"] = 2005
new_job_data.loc[new_job_data["Company Name"] == "Capio Group", "Founded"] = 2007
new_job_data.loc[new_job_data["Company Name"] == "Advance Sourcing Concepts", "Founded"] = 1964
new_job_data.loc[new_job_data["Company Name"] == "Microagility", "Founded"] = 2003
new_job_data.loc[new_job_data["Company Name"] == "Conch Technologies, Inc", "Founded"] = 2004
new_job_data.loc[new_job_data["Company Name"] == "Rainmaker Resources, LLC", "Founded"] = 2011
new_job_data.loc[new_job_data["Company Name"] == "B4Corp", "Founded"] = 2010
new_job_data.loc[new_job_data["Company Name"] == "WCG (WIRB-Copernicus Group)", "Founded"] = 2012
new_job_data.loc[new_job_data["Company Name"] == "PROPRIUS", "Founded"] = 1993
new_job_data.loc[new_job_data["Company Name"] == "Latitude, Inc", "Founded"] = 2019
new_job_data.loc[new_job_data["Company Name"] == "TECHNOCRAFT Solutions", "Founded"] = 2013
```

However, it is noteworthy that some records still lack information in critical columns such as 'Type of Ownership,' 'Industry,' 'Sector,' 'Revenue,' and 'Competitors.' Recognizing the importance of comprehensive data for robust analysis, we have decided to remove these records from the dataset. By doing so, we aim to ensure the integrity and reliability of our analyses by working with a completer and more accurate subset of the data, where essential attributes are present for a meaningful evaluation of the companies under consideration.

In [134]:
```
# Remove records where the "Founded" column has a value of -1
new_job_data = new_job_data[new_job_data["Founded"] != -1]
```

In a parallel effort to enhance the quality and completeness of our dataset, a similar cleaning process has been applied to the 'Industry' and 'Sector' columns. Similar to the 'Founded' column, instances where these columns contained missing, or placeholder values have been addressed. For this purpose, we conducted comprehensive online research to identify and replace missing or inaccurate information regarding the industry and sector affiliations of the respective companies.

In [138]:
```
# Check if there are any null value in "Industry"column
new_job_data.loc[new_job_data.loc[:,"Industry"]=="-1"]
```

Out[138]:

Salary timate	Job Description	Rating	Company Name	Location	Headquarters	Size	Founded	Type of ownership	Industry	Sector	Revenue	Competitors
–131K ssdoor est.)	Who is Cenlar?\n\nYou are.\n\nEmployee-owners ...	2.6	Cenlar\n2.6	Ewing, NJ	Ewing, NJ	1001 to 5000 employees	1958	Company - Private	-1	-1	100to500 million (USD)	-1
–165K ssdoor est.)	Job Number: 10202\nGroup: Cosma International\...	3.5	Magna International Inc.\n3.5	Birmingham, AL	Aurora, Canada	10000+ employees	1957	Company - Public	-1	-1	$10+ billion (USD)	Bosch, Lear Corporation, Faurecia
–97K ssdoor est.)	Job Description\nClient JD below:\n\nWe need a...	5.0	SkillSoniq\n5.0	San Francisco, CA	Jersey City, NJ	Unknown	2016	Company - Public	-1	-1	Unknown / Non-Applicable	-1
–97K ssdoor est.)	About Joby\nLocated in Northern California, th...	4.3	Joby Aviation\n4.3	San Carlos, CA	Santa Cruz, CA	51 to 200 employees	2009	Company - Private	-1	-1	Unknown / Non-Applicable	-1

[31]:
```
new_job_data.loc[new_job_data["Company Name"] == "Cenlar", "Sector"] = "Finance"
new_job_data.loc[new_job_data["Company Name"] == "Magna International Inc", "Sector"] = "Telecommunications"
new_job_data.loc[new_job_data["Company Name"] == "Descript", "Sector"] = "Information Technology"
new_job_data.loc[new_job_data["Company Name"] == "Comcast", "Sector"] = "Telecommunications"
new_job_data.loc[new_job_data["Company Name"] == "Advance Sourcing Concepts", "Sector"] = "Information Technology"
new_job_data.loc[new_job_data["Company Name"] == "Tygart Technology, Inc", "Sector"] = "Information Technology"
new_job_data.loc[new_job_data["Company Name"] == "Advanced Bio-Logic Solutions Corp", "Sector"] = "Biotech & Pharmaceuticals"
```

In the 'Revenue' column of our dataset, certain entries are represented by placeholder values such as '-1' or 'Unknown / Non-Applicable,' indicating missing or undisclosed revenue information. As part of our data cleaning process, we have implemented a systematic approach to enhance the quality and coherence of the dataset.

```
In [150]: # Inspect the revenue column
          job_data['Revenue'].value_counts()

Out[150]: Unknown / Non-Applicable            213
          $100 to $500 million (USD)           94
          $10+ billion (USD)                   63
          $2 to $5 billion (USD)               45
          $10 to $25 million (USD)             41
          $1 to $2 billion (USD)               36
          $25 to $50 million (USD)             36
          $50 to $100 million (USD)            31
          $1 to $5 million (USD)               31
          -1                                   27
          $500 million to $1 billion (USD)     19
          $5 to $10 million (USD)              14
          Less than $1 million (USD)           14
          $5 to $10 billion (USD)               8
          Name: Revenue, dtype: int64
```

To address this, we have replaced these placeholder values with a standardized 'n/a' (not applicable) designation. This strategic substitution not only ensures uniformity in the representation of missing revenue data but also facilitates more transparent and consistent analyses.

```
In [158]: def revenue_cleanup(revenue):
              return revenue.replace("-1", "n/a").replace("Unknown / Non-Applicable", "n/a")

          new_job_data['Revenue'] = revenue_cleanup(new_job_data['Revenue'])
```

```
In [159]: new_job_data['Revenue']=new_job_data.loc[:,'Revenue'].str.extract('([^()]+)')
          new_job_data['Revenue'].value_counts()

Out[159]: n/a                           183
          $100 to $500 million           94
          $10+ billion                   63
          $2 to $5 billion               45
          $10 to $25 million             40
          $1 to $2 billion               36
          $25 to $50 million             36
          $50 to $100 million            31
          $1 to $5 million               29
          $500 million to $1 billion     19
          $5 to $10 million              14
          Less than $1 million           11
          $5 to $10 billion               8
          Name: Revenue, dtype: int64
```

3.2.4. Handling Unwanted Data

As a crucial step in data preprocessing, we carefully address unwanted or noise data that may impede the clarity and efficiency of our analyses. In our selected dataset, we identified two columns, namely 'Index' and 'Competitors,' which are deemed unnecessary for our analytical goals.

```
In [206]: # Drop the index column
          new_job_data.drop('index',axis=1,inplace=True)
```

```
In [201]: # Inspect the competitors column
          new_job_data['Competitors'].value_counts()

Out[201]: -1                                                              438
          Roche, GlaxoSmithKline, Novartis                                 10
          Los Alamos National Laboratory, Battelle, SRI International       6
          Leidos, CACI International, Booz Allen Hamilton                   6
          MIT Lincoln Laboratory, Lockheed Martin, Northrop Grumman         3
                                                                          ...
          DHL Supply Chain, UPS, FedEx                                     1
          Pfizer, GlaxoSmithKline                                          1
          Square, Amazon, Apple                                            1
          Lumentum Operations, Keysight Technologies, O-Net Technologies   1
          Genomic Health, Myriad Genetics, The Broad Institute             1
          Name: Competitors, Length: 109, dtype: int64

In [202]: np.round((438/610)*100,decimals=2)

Out[202]: 71.8

In [203]: #The percentage is too high so we drop this column
          new_job_data.drop('Competitors',axis=1,inplace=True)
```

3.2.5. Feature Engineering

As part of the feature engineering phase, we recognize the potential insights that can be derived from the 'Salary Estimate' column. To unlock more granular information, we have transformed this single column into three distinct features, namely 'Min Salary K$,' 'Max Salary K$' and 'Average Salary Estimate'.

```
In [212]: #Inspect the salary estimate column
          new_job_data['Salary Estimate'][100:200]

Out[212]: 100    $99K-$132K (Glassdoor est.)
          101    $99K-$132K (Glassdoor est.)
          102    $99K-$132K (Glassdoor est.)
          103    $99K-$132K (Glassdoor est.)
          104    $99K-$132K (Glassdoor est.)
                           ...
          201    $79K-$106K (Glassdoor est.)
          202    $79K-$106K (Glassdoor est.)
          203    $79K-$106K (Glassdoor est.)
          204    $79K-$106K (Glassdoor est.)
          205    $79K-$106K (Glassdoor est.)
          Name: Salary Estimate, Length: 100, dtype: object
```

```
In [220]: #Obtain the minimum salary in the range
          new_job_data['Min Salary K$'], new_job_data['Salary Estimate'] = zip(*new_job_data['Salary Estimate'].apply(extract_values))

In [221]: #Obtain the maximum salary in the range
          new_job_data['Max Salary K$'], new_job_data['Salary Estimate'] = zip(*new_job_data['Salary Estimate'].apply(extract_values))

In [222]: #Drop the old salary range column
          new_job_data.drop('Salary Estimate', axis=1, inplace=True)

In [226]: # Obtain the average salary estimate, handling non-finite values
          new_job_data['Avg Salary Estimate'] = np.round((new_job_data['Min Salary K$'] + new_job_data['Max Salary K$']) / 2, decimals=0)
          new_job_data['Avg Salary Estimate'] = new_job_data['Avg Salary Estimate'].astype('Int64')
          new_job_data.head()
```

As a forward-looking step in data preprocessing, we have introduced a novel feature to our dataset – the 'Years in Market' column. This newly created column encapsulates the number of years a company has been actively present in the job market. Calculated from the 'Founded' column, this information serves as a valuable metric for assessing the longevity and experience of companies offering job opportunities.

```
In [342]: # obtain the years the company has been in the market and drop the founed column
          current_year = datetime.now().year
          new_job_data['Years in Market'] = current_year - new_job_data['Founded']
          new_job_data.drop('Founded',axis=1,inplace=True)
          new_job_data.head()
```

3.2.6. Splitting Columns

As part of the data cleaning process, we have undertaken a strategic transformation of the 'Location' column in our selected dataset. This involved splitting the 'Location' column into two distinct features: 'Location State' and 'Same Location'.

The 'Location State' column now represents the state abbreviation corresponding to each job posting, providing a more granular geographical insight. Simultaneously, we introduced a new binary column, 'Same Location' which indicates whether the job posting's location matches the headquarters' location.

In [335]:
```
# Inspect the location column
new_job_data['Location'][:10]
```

In [336]:
```
# Check the counts of each unique location state
new_job_data.loc[:,"Location"].apply(lambda x: x.split(",")[-1]).value_counts()
```

In [337]:
```
# Create the state column
new_job_data['Location State'] = new_job_data['Location'].apply(lambda x: x.split(',')[-1].strip())
```

In [338]:
```
# Replace the inconsistencies with their correct state abbreviation
def clean_location(location):
    state_mapping = {
        "California": "CA",
        "Texas": "TX",
        "Utah": "UT",
        "New Jersey": "NJ",
        "Remote": "n/a",
        "United States": "n/a"
    }

    return state_mapping.get(location, location)

new_job_data['Location State'] = new_job_data['Location State'].map(clean_location)
```

In [339]:
```
new_job_data["Location State"].value_counts()
```

In [340]:
```
# Check if the location of the job and the HQ are in the place
new_job_data['Same State'] = (new_job_data['Location'] == new_job_data['Headquarters']).astype(int)
```

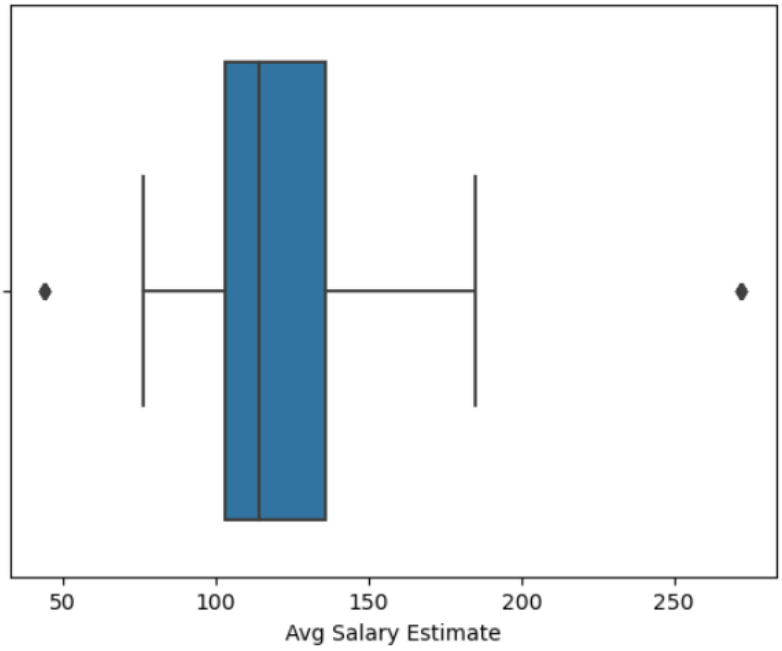
3.2.7. Outliers

In the final stage of our data cleaning process, we conduct a comprehensive assessment of potential outliers within the dataset. To visually inspect the distribution of the 'Avg Salary Estimate' column, we employ a box plot, a graphical representation that allows us to identify potential outliers based on their position outside the interquartile range (IQR).

Having visualized the data, we apply a principled approach to outlier removal by leveraging the IQR method. This method assists us in distinguishing values that significantly deviate from the overall trend, allowing us to pinpoint potential anomalies. In this particular instance, a careful examination reveals that there are a modest 36 outliers within the 'Avg Salary Estimate' column.

In [343]:
```
import seaborn as sns
import matplotlib.pyplot as plt

# Example: Box plot for 'Avg Salary Estimate'
sns.boxplot(x=new_job_data['Avg Salary Estimate'])
plt.show()
```



In [351]:
```
Q1 = new_job_data['Avg Salary Estimate'].quantile(0.25)
Q3 = new_job_data['Avg Salary Estimate'].quantile(0.75)
IQR = Q3 - Q1

# Identify outliers
outliers = ((new_job_data['Avg Salary Estimate'] < Q1 - 1.5 * IQR) | (new_job_data['Avg Salary Estimate'] > Q3 + 1.5 * IQR))

# Count the number of outliers.
outliers.sum()

# Remove outliers and create a new DataFrame
filtered_job_data = new_job_data[~outliers]

# Display the resulting DataFrame without outliers
filtered_job_data
```

3.3. Data Transformation

In the data transformation phase, our objective is to refine and enhance the dataset to facilitate meaningful analysis. We employ a two-fold approach to role and seniority definition, and skill□based categorization. Leveraging custom functions, we classify job roles and seniority levels based on 'Job Title' and 'Job Description.' Simultaneously, we create dummy columns for key skills, streamlining the dataset for subsequent exploratory data analysis (EDA). These transformations lay the foundation for a more granular and insightful analysis of the job market landscape, empowering data-driven decision-making.

3.4. Data Reduction

In simplifying our dataset, we aimed to make it more manageable and relevant. We removed unnecessary columns like 'Index' and 'Competitors,' focusing on the key information. This not only makes our dataset more efficient but also ensures we're working with the most important data for our analysis. Our approach aligns with best practices, creating a streamlined dataset for a clearer understanding of job market trend.

4. Methodology

4.1. Introduction

In our methodology, we employed Python for data cleaning, ensuring the dataset's accuracy and reliability. The cleaned dataset will serve as the foundation for our analysis. To visualize insights effectively, we plan to utilize Tableau, a powerful tool for creating interactive dashboards.

4.2. Type of Data to be collected and data sources

We collected job-related data from Glassdoor using the Kaggle dataset 'Data Science Job Posting on Glassdoor.' This dataset includes essential information like job titles, company details, salaries, and more, offering a comprehensive view of the job market.

The cleaned dataset consists of 565 records with columns such as Job Title, Job Description, Rating, Company Name, etc. We collected this data from reputable sources related to job postings, ensuring reliability.

4.3. Data collection tools and plan

Python played a pivotal role in our data-cleaning process. Leveraging libraries such as Pandas, BeautifulSoup, and requests, we systematically cleaned, transformed, and organized the dataset. Our plan prioritized accuracy and efficiency to create a reliable foundation for analysis.

4.4. Methods, techniques, and tools used to visualize the data

To analyze the dataset, we'll apply statistical methods and techniques, including descriptive statistics and exploratory data analysis (EDA). Tableau will be our primary tool for visualizing data trends and creating interactive dashboards that offer a user-friendly and insightful experience.

5. Data Analysis

The dataset contains information about various job roles, salary estimates, company details, and required skills. The goal of the analysis is to gain insights into the dataset and understand the relationships between different variables.

Jupyter Notebook Link: https://1drv.ms/u/s!AkI7kbf9N08hgo1fHhLgBLWlKJtOoQ?e=2qHOuN

5.1. Descriptive Statistics

```
In [71]: # Describe the dataset
         filtered_job_data.describe()
```

Out[71]:

	Rating	Min Salary K$	Max Salary K$	Avg Salary Estimate	Years in Market	Same State	excel	sql	python	power_bi	tableau	scikit	
count	567.000000	567.000000	567.000000	567.000000	567.000000	567.000000	567.000000	567.000000	567.000000	567.000000	567.000000	567.000000	567.00
mean	3.865961	97.010582	144.107584	120.525573	38.165785	0.444444	0.437390	0.541446	0.735450	0.045855	0.186949	0.105820	0.27
std	0.601255	24.315721	32.315156	26.934595	39.917238	0.497343	0.496502	0.498719	0.441483	0.209356	0.390215	0.307879	0.44
min	2.000000	56.000000	97.000000	76.000000	5.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00
25%	3.500000	79.000000	121.000000	103.000000	13.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00
50%	3.800000	91.000000	132.000000	114.000000	24.000000	0.000000	0.000000	1.000000	1.000000	0.000000	0.000000	0.000000	0.00
75%	4.300000	112.000000	163.000000	136.000000	45.500000	1.000000	1.000000	1.000000	1.000000	0.000000	0.000000	0.000000	1.00
max	5.000000	145.000000	225.000000	185.000000	243.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.00

1. **Rating**

Central Tendency

The mean rating of job postings of different data science job roles offered by companies is approximately 3.87. It indicates a moderate to good rating on average.
The median rating (3.80) suggests that ratings are evenly distributed around this central value.

Measures of Variability

The standard deviation (0.60) indicates that the ratings vary moderately around the mean. This suggests that while some job postings have high ratings, others have lower ratings.

Measures of Position

Minimum: The minimum rating among the job postings about data science-related job roles at different companies is 2.0.
25th Percentile (Q1)- 25% of the job postings have a rating of 3.50 or lower.
Median (50th Percentile or Q2)-50% of the job postings have a rating of 3.80 or lower.
75th Percentile (Q3) -75% of the companies have a rating of 4.30 or lower.
Maximum - The maximum rating among the job postings offering data science job roles is 5.0.

2. **Min Salary K$ and Max Salary K$**

Central Tendency

The average minimum salary offered is around $97k, while the average maximum salary is about $144k.
The mean and median salaries provide insight into the typical salary range for data science positions.

Measures of Variability

The standard deviation for the minimum salary is approximately $24.32K. So, on average, individual minimum salaries deviate from the mean by around $24.32K.
The standard deviation for the maximum salary is approximately $32.32K. On average, individual maximum salaries deviate from the mean by around $32.32K.

Measures of Position of Min Salary

25th Percentile (Q1) - 25% of the job roles offer a minimum salary of $79,000 or lower.
Median (50th Percentile or Q2) - 50% of the job roles offer a minimum salary of $91,000 or lower.
75th Percentile (Q3) - 75% of the job roles offer a minimum salary of $112,000 or lower.

Measures of Position of Max Salary

25th Percentile (Q1)- 25% of the job roles offer a maximum salary of $121,000 or lower.
Median (50th Percentile or Q2)- 50% of the job roles offer a maximum salary of $132,000 or lower.
75th Percentile (Q3)- 75% of the job roles offer a maximum salary of $163,000 or lower.

3. **Avg Salary Estimate**

Central Tendency

The average salary estimate for data science positions is approximately $120.53K. This provides a general idea of the expected salary range for these roles.

Measures of Position

25th Percentile (Q1)-25% of the job roles have an average salary estimate of $103,000 or lower.
Median (50th Percentile or Q2)- 50% of the job roles have an average salary estimate of $114,000 or lower.
75th Percentile (Q3)- 75% of the job roles have an average salary estimate of $136,000 or lower.

Measures of Variability

The standard deviation (approximately $26.93K) suggests the extent of variability in salary estimates, indicating that there may be significant differences in compensation estimates across different job postings.

4. **Years in the Market**

Central Tendency

The mean and median years in the market indicate the average and middle value for how long companies offering data science-related jobs have been operating.
The mean years of the companies in the market is approximately 38.17 years.

Measures of Variability

The standard deviation (approximately 39.92 years) suggests a high degree of variability in the number of years companies have been in the market. This indicates that some companies are relatively new while others have been established for a considerable period.

Measures of Position

Minimum - The minimum number of years a company has been in the market offering data science job roles is 5 years.
25th Percentile (Q1) -25% of the companies offering job roles have been in the market for 13 years or fewer.
Median (50th Percentile or Q2) -50% of the companies offering job roles have been in the market for 24 years or fewer.
75th Percentile (Q3) -75% of the companies offering job roles have been in the market for 45.5 years or fewer.
Maximum - The maximum number of years a company has been in the market offering data science job roles is 243 years.

5. **Binary Variables (Same State, Excel, SQL, Python, Power BI, Tableau, Scikit, Spark)**

These variables represent the presence or absence of specific skills or attributes in job postings.

5.2. Relationship Analysis

- **Pair plot**

Drawing a pair plot using the variables in the dataset allows us to visualize the relationships between different pairs of variables simultaneously. Each scatterplot in the pair plot matrix represents the relationship between two variables, while the diagonal displays the distribution of each variable.

- Main Diagonal shows the distribution of each variable along the diagonal. It gives us insights into the distribution of variables like rating, minimum salary, maximum salary, average salary estimates, years in the market, and binary variables(skills) like Excel, SQL, Python, Power BI, Tableau, Scikit, and Spark.

- The scatterplots outside the diagonal show the relationship between two variables.

  If there's a positive linear relationship between two variables, points on the scatterplot will tend to follow an upward trend.

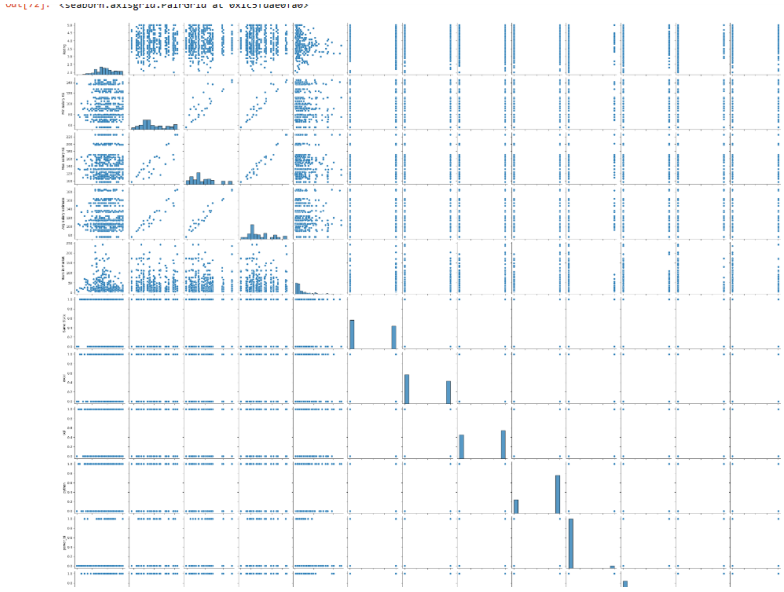
  E.g: - Max Salary K$ and Avg Salary Estimate
  Min Salary K$ and Avg Salary Estimate
  Min Salary K$ and Max Salary K$

  If there's a negative linear relationship, points will tend to follow a downward trend.
  If there's no clear pattern, points will be scattered randomly.

- The density and clustering of points in scatterplots can give us an idea about the strength and direction of the correlation between variables.

  For example: -

  A tight cluster of points suggests a strong correlation.
  A spread-out distribution of points suggests a weak correlation.
  If points form a diagonal line, it indicates a perfect correlation.



- **Heatmap**

In a heatmap depicting correlation, the Pearson correlation coefficient is used to quantify the strength and direction of the linear relationship between two numerical variables. Based on the correlation values in this heat map,

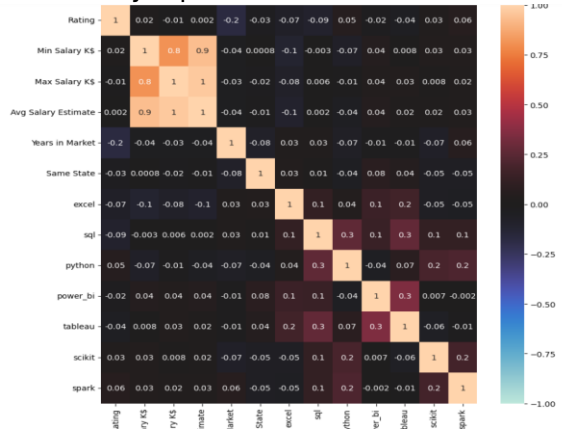
- The correlation between Min Salary K$ and Max Salary K$ is 0.8.
- The correlation between Min Salary K$ and Avg Salary Estimate is 0.9.
- The correlation between Max Salary K$ and Avg Salary Estimate is 1.
- Other correlation values between variables are close to zero.

Min Salary K$ and Max Salary K$ - The correlation coefficient of 0.8 indicates a strong positive linear relationship between the minimum and maximum salary offered for data science roles. This means that as the minimum salary increases, there's a strong tendency for the maximum salary also to increase, and vice versa. It suggests that job roles in the postings with higher minimum salaries tend to have higher maximum salaries as well, and vice versa, indicating consistency or proportionality in salary offerings.

Min Salary K$ and Avg Salary Estimate - The correlation coefficient of 0.9 suggests a very strong positive linear relationship between the minimum salary offered and the average salary estimate for the data science roles. This implies that there's a high consistency between the minimum salary stated in job role postings and the average salary estimates. When the minimum salary is high, the average salary estimate tends to be high as well, and vice versa.

Max Salary K$ and Avg Salary Estimate - The correlation coefficient of 1 indicates a perfect positive linear relationship between the maximum salary offered and the average salary estimate for the data science roles. This means that the maximum salary and average salary estimate move in perfect synchronization. It suggests that when the maximum salary offered is higher, the average salary estimate tends to be higher as well, and vice versa.

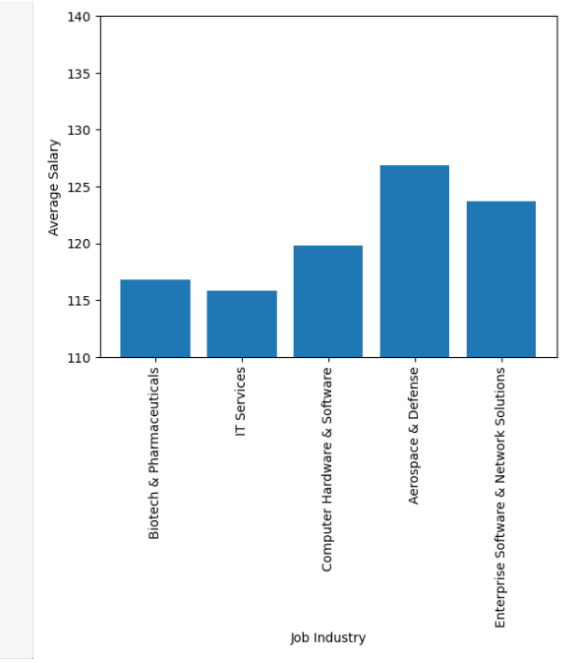
Correlation values close to zero imply that there is little to no linear relationship between the corresponding variables. For example, a correlation coefficient between the rating of the job posting and the maximum salary offered is 0.002. It suggests a very weak positive linear relationship, nearly close to zero. This implies that there is almost no association between the rating of the job posting and the maximum salary offered for data science roles. In practical terms, it means that the rating of the job posting, as provided on platforms like Glassdoor, does not significantly influence the maximum salary offered for those positions. Therefore, variations in job ratings are unlikely to impact the salary expectations for data science roles.



5.3. Industry Analysis

- **Bar plot**

A bar plot was used to visualize the average salary estimates across different industries. Biotech & Pharmaceuticals and IT Services emerged as the top-paying industries.



5.4. Salary Analysis

- **Boxplot**

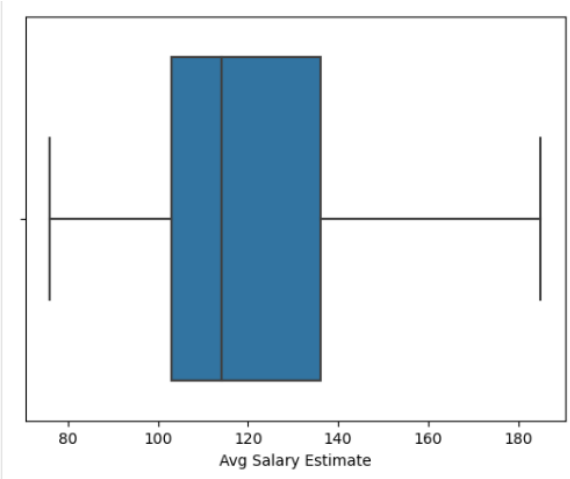
Minimum Value-The lowest value observed in the 'Avg Salary Estimate' column is $76K. It indicates the lowest reported Avg Salary Estimate for data science job positions in the dataset.

First Quartile (Q1) -25% of the data points fall below this value. In this dataset, Q1 of Avg Salary Estimate column is $103K.

Median (Q2) - The median salary estimate of $114.0K represents the midpoint of the dataset. This means that half of the data points have salary estimates below $114.0K, and the other half have estimates above it. It provides a measure of the central tendency of the salary distribution.

Third Quartile (Q3) -In this dataset, Q3 of Avg Salary Estimate is $136K. The third quartile value of $136K indicates that 75% of the data points have salary estimates below this value.

Maximum Value -The highest value observed in the 'Avg Salary Estimate' column is $185K. The maximum Avg Salary Estimate of $185K represents the highest calculated Avg Salary Estimate salary for data science positions in the dataset. It indicates the upper limit of Avg Salary Estimate and highlights the potential earning ceiling for data science professionals.



- **Simple Linear Regression**

Simple linear regression analysis was conducted to analyze the relationship between the number of years a company has been in the market and the average salary estimate. The coefficient of determination (R^2) was found to be approximately -0.038, indicating a weak negative correlation. The root mean squared error (RMSE) was approximately 24.70.

5.5. Skill Analysis

- **Pivot Table**

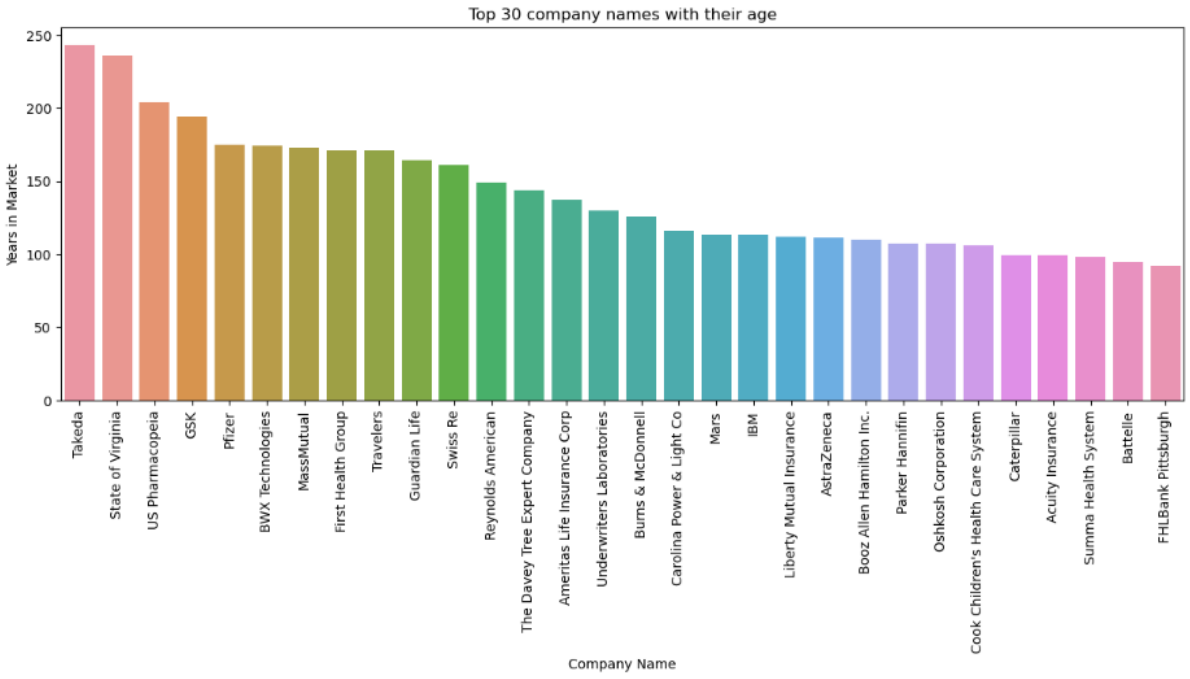
A pivot table was created to analyze the percentage of job descriptions mentioning specific skills (e.g., excel, python, SQL). It provides insights into the demand for various skills across different job roles.

Job Role	excel	power_bi	python	scikit	spark	sql	tableau
Data Analyst	55.77	9.62	51.92	0.00	9.62	73.08	51.92
Data Engineer	47.73	11.36	79.55	0.00	40.91	84.09	13.64
Data Scientist	43.04	4.20	81.10	13.39	31.23	53.02	17.85
Machine Learning Engineer	18.18	0.00	72.73	27.27	15.15	36.36	0.00
Other	49.12	0.00	38.60	0.00	14.04	31.58	8.77

5.6. Company Analysis

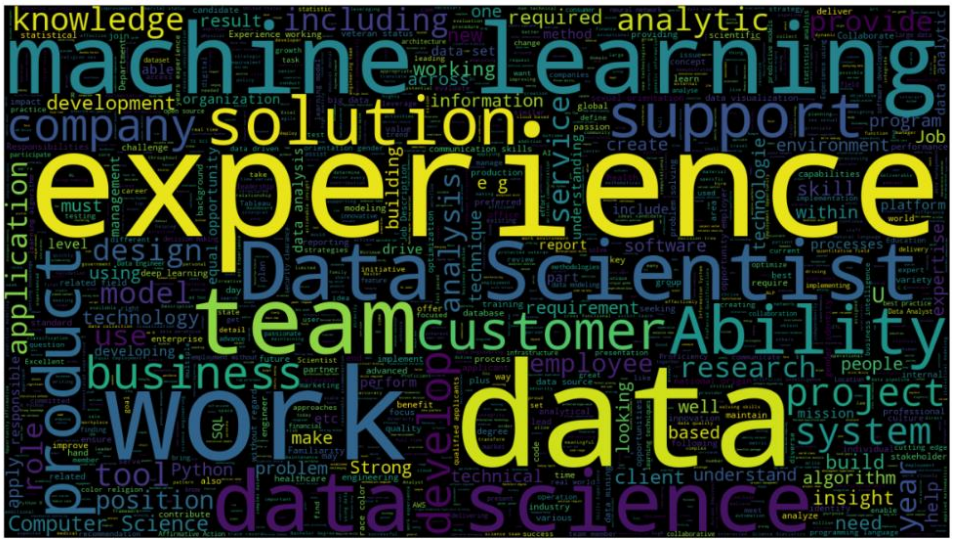
- **Bar plot**.

A bar plot was used to visualize the top 30 company names with their respective ages in the market. It helps identify the oldest companies in the dataset.



5.7. Job Description Analysis

- **Word Cloud.**



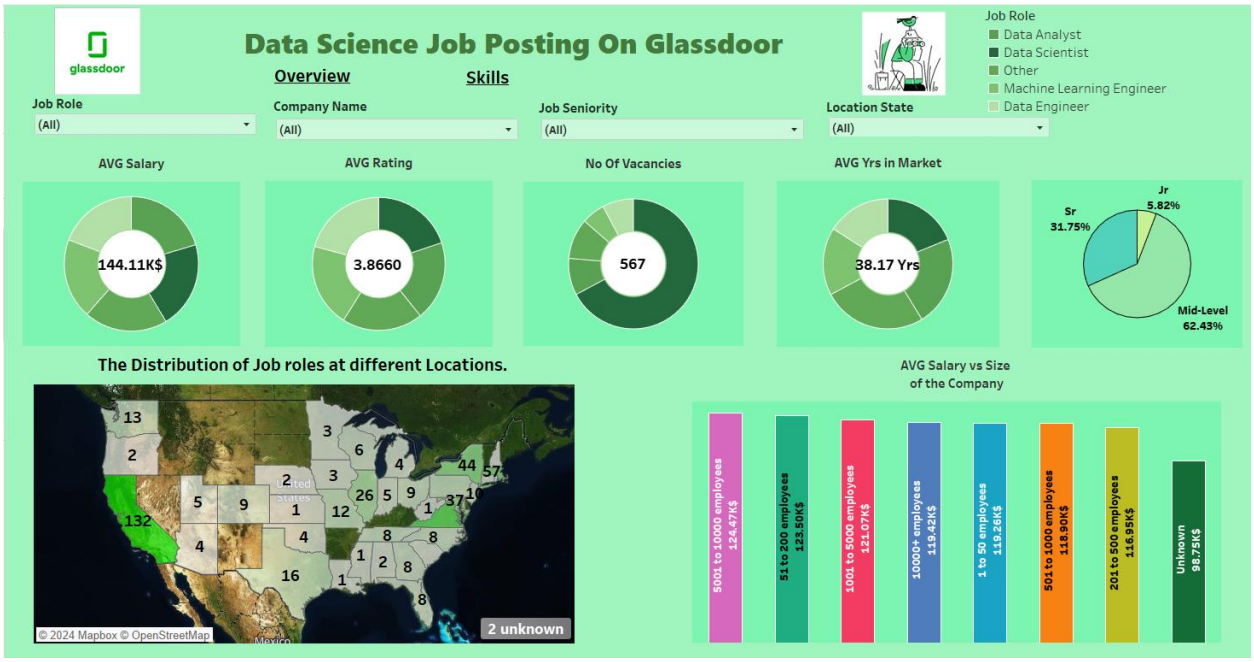
Larger words tend to represent terms with greater prominence in the "Job Description" text you analyzed. Words like "experience", "Machine Learning", "data", "Scientist", "Python", and "problem" are all relatively large, suggesting they appear frequently in the job descriptions.

Smaller words correspond to less frequent terms.

6. Dashboard Analysis

6.1. Overview of the Dashboard Layout

Tableau workbook link:
https://public.tableau.com/views/Data_Science_Job_Posting_On_Glassdoor_Group_E/Overview?:language=en-US&publish=yes&:sid=&:display_count=n&:origin=viz_share_link

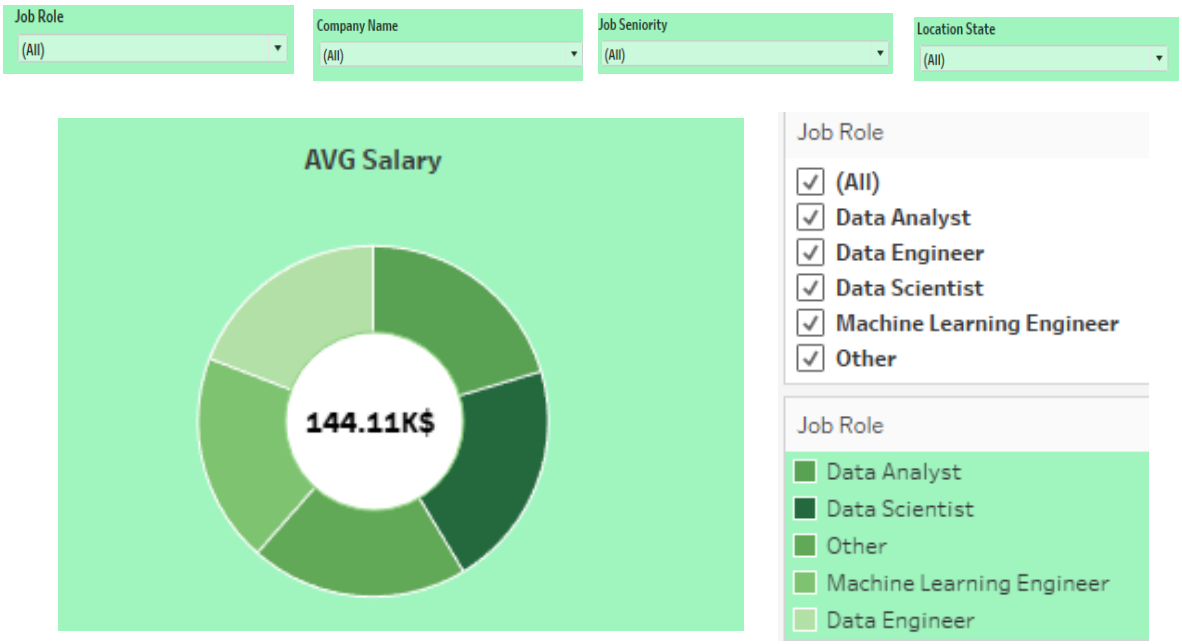


6.1.1. Dashboard 1 – Overview

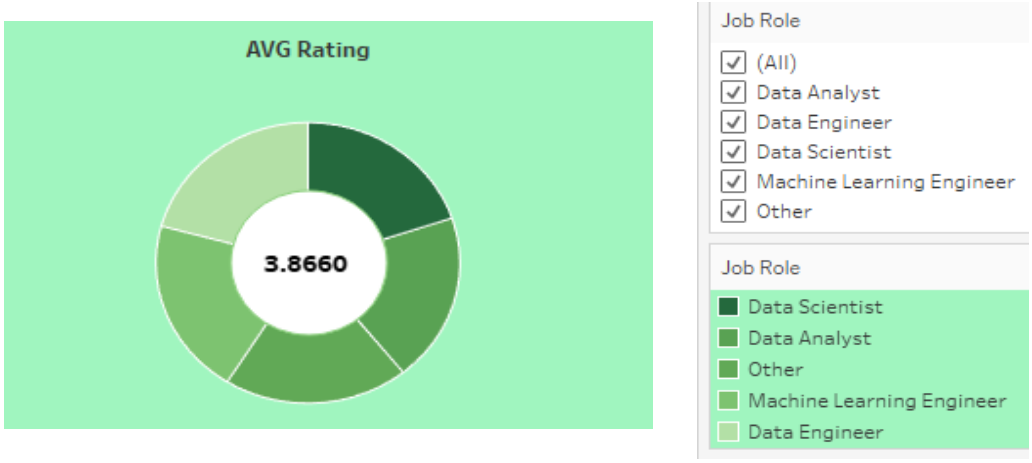
6.1.1.1.1. Data Visualization Using Different Graphics

Doughnut Charts

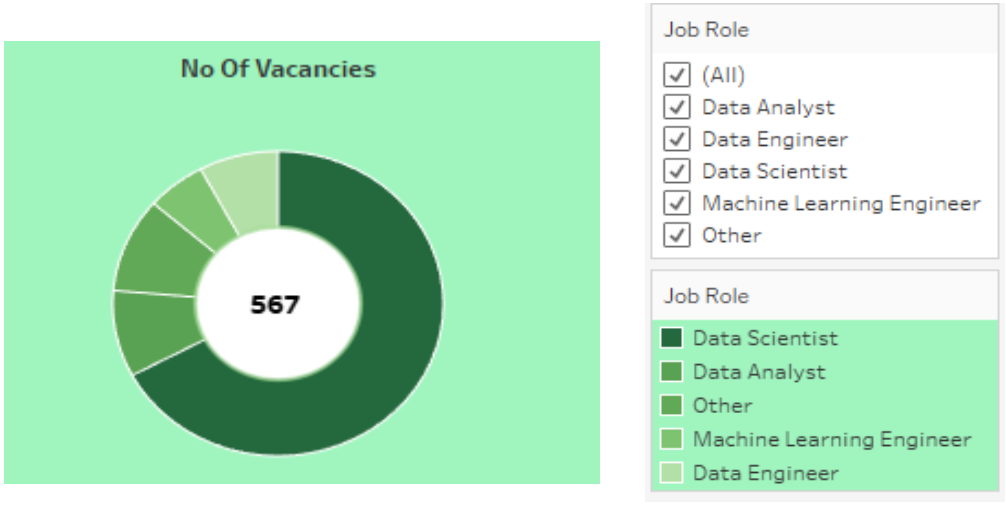
1. **Average Salary**
    - This doughnut chart provides a visual representation of the average salary across different job roles.
    - When users select specific filters such as Job Role, Company Name, Job Seniority, or Location State, the average salary doughnut chart updates to reflect the average max salary within the filtered subset of data. This allows users to see how the average salary varies based on their chosen criteria.



2. **Average Rating**
    - This doughnut chart displays the average rating of job roles based on Glassdoor ratings. It helps in understanding the quality or reputation of job roles in the dataset.
    - Like the average salary chart, the average rating doughnut chart dynamically adjusts to display the average rating of job roles within the filtered dataset. Users can assess the reputation or satisfaction level of job roles and companies based on Glassdoor ratings that match their selected criteria.

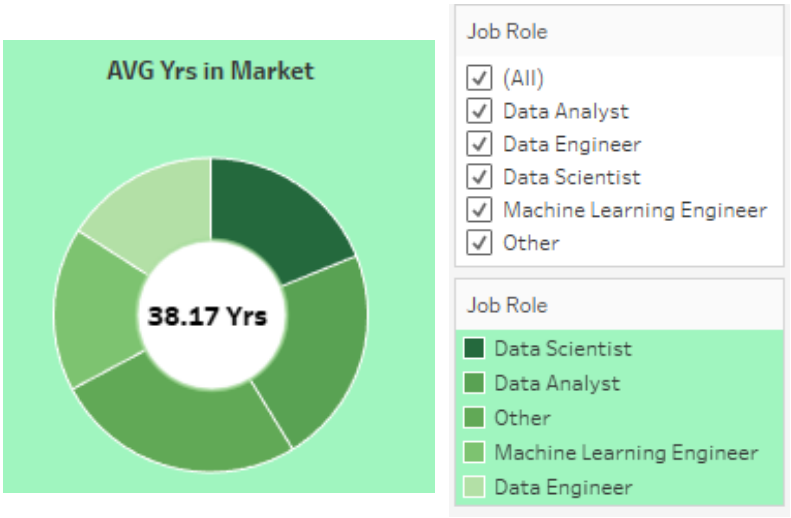


3. **Number of Vacancies**
    - This doughnut chart illustrates the distribution of job vacancies across 5 categories of main job roles. It gives an overview of the demand for different roles in the market.
    - This chart shows the distribution of job vacancies within the filtered subset of data. Users can observe the availability of job opportunities based on their chosen filters (job roles, companies, seniority levels, and location states) helping them understand the demand for specific roles or within specific companies.



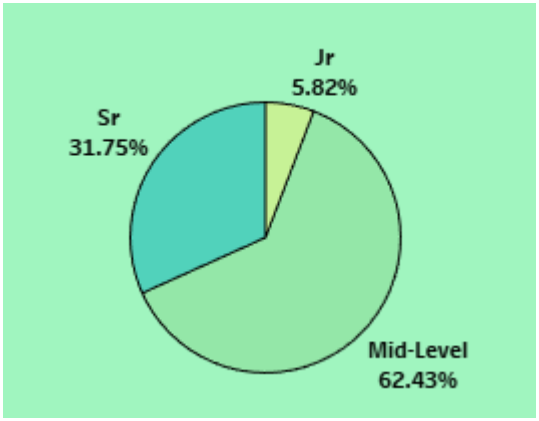
4. **Average Years in Market**

- This doughnut chart shows the average number of years companies related to 5 main job roles have been in the market. It provides insights into the longevity or stability of companies related to the data science job market in the dataset. This also dynamically updates with the filter.



Pie Chart

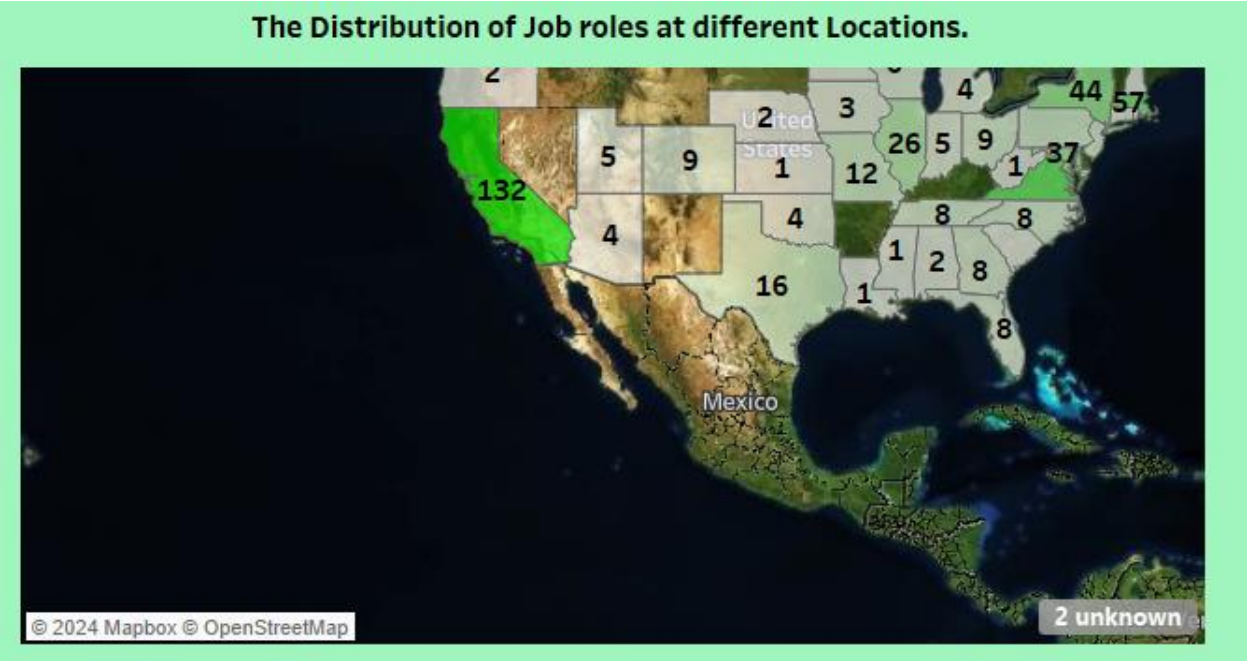
5. **Percentage of Seniority**
   This pie chart represents the distribution of job seniority levels in the dataset. It helps in understanding the proportion of different seniority levels of job roles in the job postings.



Satellite Map

6. Distribution of Job Roles by Location

This satellite map displays the geographical distribution of different job roles across various locations in the US. Also, each location marker on the map includes the count of job roles available in that area. Users can visually explore where specific job roles are concentrated and assess the job market density in different regions.



Bar Chart

7. Average Salary vs Size of Company

In this bar chart, each bar represents a range of company sizes (e.g., 5001-10000 employees) along the x-axis. The height of each bar corresponds to the average salary offered by companies within that size range. Users can easily compare the average salary levels across different company sizes. The average salary value is explicitly written on each bar to provide clear and precise information for easy interpretation.

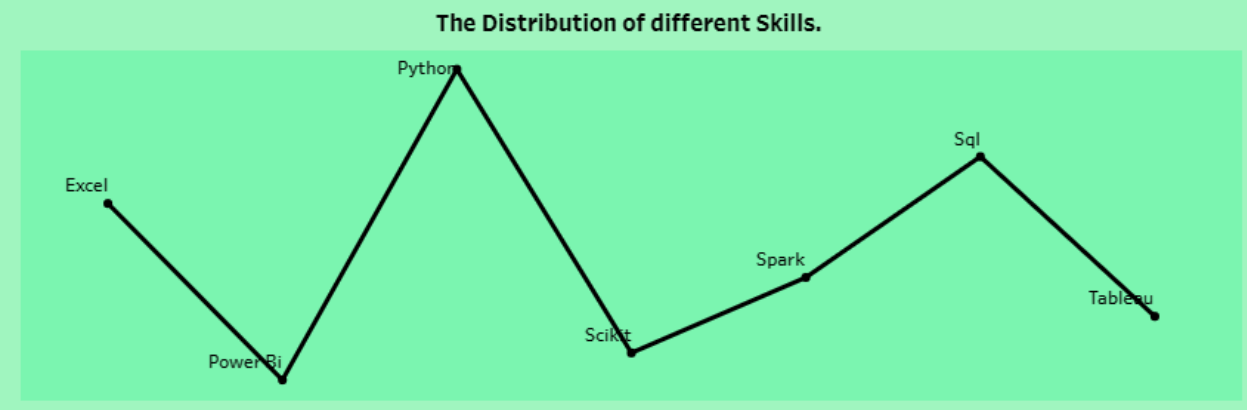
6.1.2. Dashboard 2 – Skills

Similar to the overview dashboard, the skills dashboard also includes 4 filters for Job Role, Company Name, Job Seniority, and Location State. These filters enable users to refine their analysis based on specific criteria of interest.

Line Chart

8. **Distribution of Different Skills**

This line chart visualizes the distribution of different skills (e.g., Excel, Power BI, SQL, Python, Scikit, Spark, Tableau) across job postings. Users can observe how the prevalence of each skill varies within the job roles with filters and identify which skills are most commonly required for data science roles.

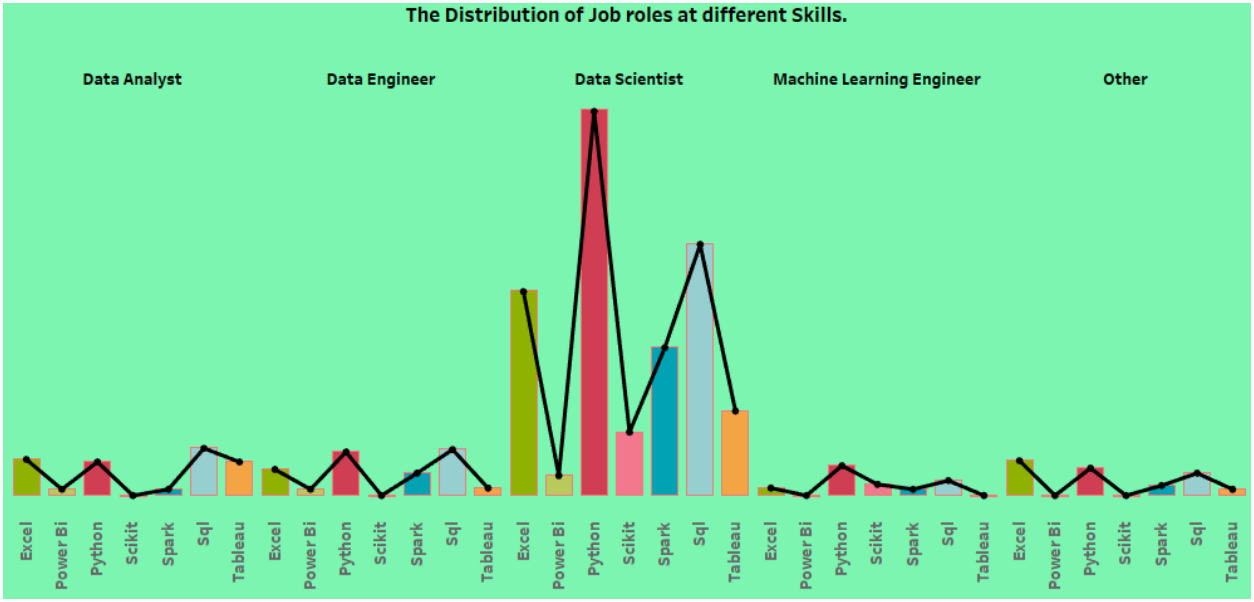


Combination Chart (Bar Chart with the Line Chart)

9. Distribution of Job Roles at Different Skills

This combination chart offers a comprehensive view of how job roles correspond to different skills mentioned in job postings. Each bar in the bar chart represents a specific skill, and the height of the bar indicates the count of times that skill is mentioned in job postings according to the job roles.

The line chart overlays the bars connecting the middle point of the top of each bar. It functions as a trend line, illustrating the change in the distribution of skills across different job roles. Users can visually identify the patterns between specific skills and job roles by observing how the trend line intersects or trends alongside the bars. This visualization provides valuable insights into the relationship between skills and job roles & helps users understand which skills are most in demand for various data science positions and guides them in their career pursuits.

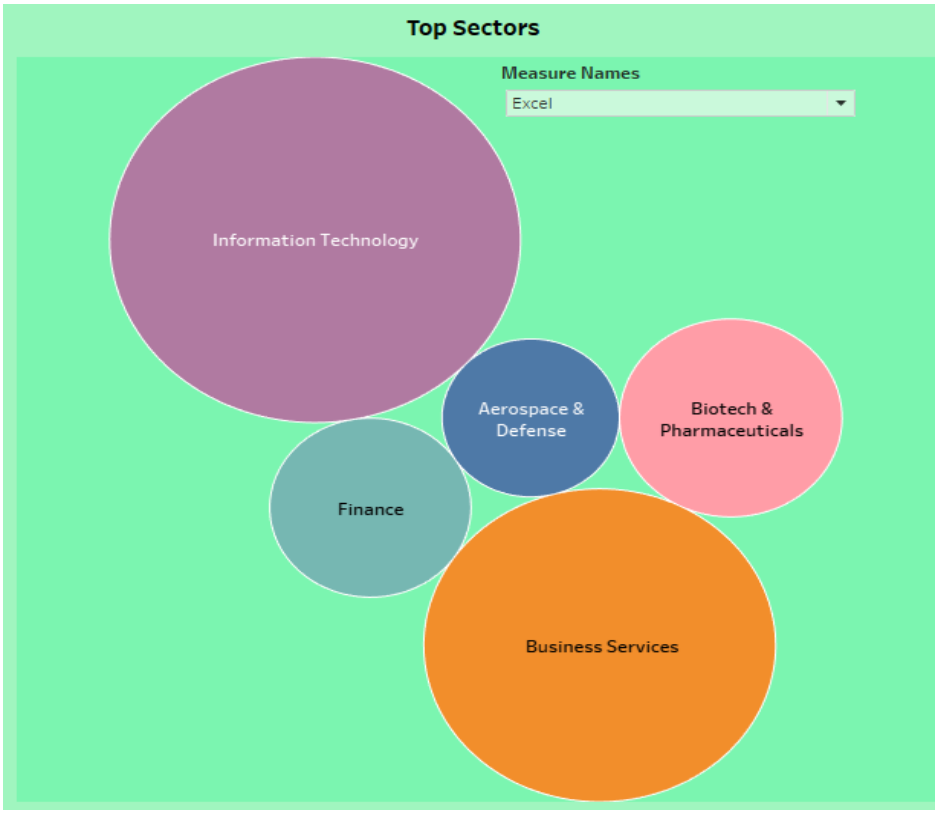


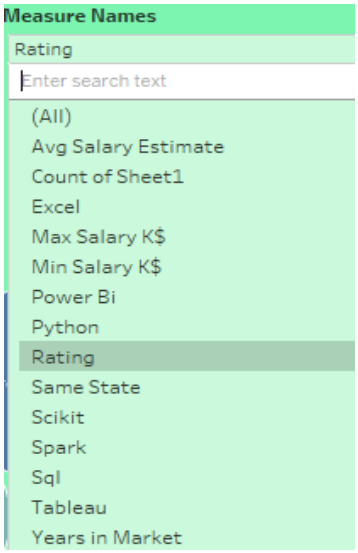
Bubble Chart

10. Top Sectors

This bubble chart displays the top sectors in which data science job roles are frequent. Each bubble represents a sector, with the size of the bubble indicating the relative frequency of job roles within that sector.

Users can quickly identify which sectors offer the most opportunities for data science professionals to guide their career decisions or job search strategies. Also, users can interact with the chart by applying the filter to select specific skills or salary ranges (minimum, maximum, or average), location states, or years in the market. These filters allow users to customize their analysis and focus on sectors that align with their interests or career objectives. It gives the chance to do their analysis based on skill requirements, salary ranges, geographical locations and the number of years a company has been in the market.





6.2. Interactive Elements and Features

- Filtering Options - Interactive filters for selecting job roles, company names, job seniority, and location state.
- Dynamic Charts - Charts that update based on user selections in filters.
- Tooltip Information- Hover-over tooltips providing additional details on data points.
- Dashboard Navigation Buttons and Inter Dashboard Navigation- Two buttons allowing users to switch between the Overview Dashboard and the Skills Dashboard and the ability to seamlessly transition from the Overview Dashboard to the Skills Dashboard and vice versa, providing users with access to different sets of insights.

7. Conclusion

The project started with a comprehensive analysis of the dynamic landscape of data science careers within the US job market using Glassdoor's dataset sourced from Kaggle. Our goal was to provide nuanced insights into the diverse roles, salary distributions, and top hiring companies within the data science sector. Through meticulous data preprocessing, the exploratory data analysis (EDA), and statistical analysis using Python, we gained invaluable insights into the dataset, laying a robust foundation for analyses and dashboard development.

Significant progress has been achieved by extensive data cleaning, transformation, and reduction processes using Python. These efforts ensured the accuracy, consistency, and relevance of the dataset, preparing it for in-depth analysis. Various Python libraries were used for data preprocessing, EDA, and statistical analysis. It includes Pandas, NumPy, Matplotlib, Seaborn, and Scikit-learn.

After data preprocessing, the cleaned dataset now includes only essential variables such as job titles, salary estimates, company details, and skill-based categorizations to facilitate a comprehensive understanding of the data science job landscape. Exploratory data analysis (EDA) further explored deeper into the dataset, uncovering the key insights regarding descriptive statistics, relationships between variables, industry trends, salary distributions, and skill requirements. Various visualization techniques such as bar plots, line charts, pie charts, bubble charts, box plots, pair plots, and heatmaps were used to represent the data effectively and get actionable insights.

Following the data preprocessing, EDA, and statistical analysis, the final dashboard was developed using Tableau. The interactive dashboard consists of two main sections, Overview and Skills. It offers users a seamless and intuitive platform to explore and visualize the dataset's insights. The overview dashboard provides a holistic view of job roles, salary distributions, company attributes, and geographical distributions, facilitating informed decision-making for job seekers and companies alike. The skills dashboard offers a focused exploration of the skills required for data science roles, enabling users to identify the relationships between specific skills and job roles and recognize emerging trends within the sector.

In conclusion, this project has sought to contribute valuable insights into the dynamic landscape of data science careers in the US serving as a valuable resource for individuals navigating the job market and organizations with storytelling through a dashboard. We aim to empower stakeholders with actionable insights to drive informed decisions using the power of data visualization.

8. List of References

‘Data Pre-Processing for Machine Learning Models using Python Libraries’ (2020) International Journal of Engineering and Advanced Technology, 9(4), pp. 1995–1999. Available at: https://doi.org/10.35940/ijeat.d9057.049420.

Sojasingarayar, A. (2022) Data Preprocessing in Python, Medium. Available at: https://medium.com/@abonia/data-preprocessing-in-python-1f90d95d44f4.

www.youtube.com. (n.d.). Tableau Desktop Crash Course | Tableau training for beginners. [online] Available at: https://www.youtube.com/watch?v=-Aj8IlC0IEA [Accessed 31 Mar. 2024].

Tableau in Two Minutes - Tableau Basics for Beginners. (2018). YouTube. Available at: https://www.youtube.com/watch?v=jEgVto5QME8.

Jupyter Notebook Link: https://1drv.ms/u/s!AkI7kbf9N08hgo1fHhLgBLWlKJtOoQ?e=2qHOuN

Uncleaned dataset: https://1drv.ms/u/s!AkI7kbf9N08hgo1gW-ekLz6WTz4FCA?e=hcgPMC

Python Notebook (Including the data preprocessing): https://1drv.ms/u/s!AkI7kbf9N08hgo1fHhLgBLWlKJtOoQ?e=df6Qwa

Cleaned dataset: https://1drv.ms/u/s!AkI7kbf9N08hgo1l5tcOD56ul5g2xg?e=6k9007

Tableau workbook link: https://public.tableau.com/views/Data_Science_Job_Posting_On_Glassdoor_Group_E/Overview?:language=en-US&publish=yes&:sid=&:display_count=n&:origin=viz_share_link