Job Data Analytics Dashboard
Final project Report

Group F

Data Visualization

DS304.3

HSS Hewage-25073

GLS Chamaka-25772

DMU Nimsara-25024

Abstract

This project focuses on data visualization within the data science domain, specifically analyzing job salaries and other related data. The objective is to provide insights into the dynamic job market landscape, catering to stakeholders such as job seekers, employers, and policymakers. Through the selection of a comprehensive dataset from Kaggle, we create an interactive dashboard using Power BI and Python, highlighting trends in various roles and titles. Challenges such as tool limitations and dataset constraints are acknowledged, with a proposed work schedule outlining the project timeline. Leveraging descriptive analysis, data visualizations, and statistical techniques, the project seeks to empower decision-makers and aid career decisions for aspiring professionals.

Table of Contents

Introduction

Background study and reasons

The goal of this assignment was to select a dataset and analyze and visualize the data for deriving and interpreting insights. For our visualization assignment, we have selected the data science domain and their job salaries.

The underlying reason for selecting this domain was that the job market is dynamic with changes happening in the industry with the evolving market. Job salaries are important for job seekers, employers, and policymakers who want to understand what's happening in the job market.

Problems related to the selected area

The dataset is packed with info, offering a good overview of the hiring scene, including details like company sizes and remote ratio. We want the visualization to be user-friendly, so everyone stakeholders, job seekers, and recruiters can easily grasp what's important. Using charts, graphs, and interactive features, we aim to present a complete view that's not only insightful but also accessible. As undergraduates who are soon to be in the industry, this analysis will help us make our career decisions as well.

Objectives of the project

Our data visualization project is all about crafting a helpful dashboard using Power BI and Python or R, but not building complex models just to get the dataset ready for visualization. We're exploring the above job dataset from Kaggle, aiming to make sense of the job market data and present insights in a visually straightforward manner.

The main goal is to create an interactive dashboard that highlights trends in different roles and titles. We're focusing on things like experience, employee residence, and salary information to provide a clear picture of the job market landscape.

The dataset's depth provides opportunities to empower decision-makers without diving into complex modeling. Visualizing evolving market demands gives valuable insights for strategic decision-making. Job seekers can benefit from visually clear information about industry requirements, helping them make informed career decisions. Recruiters and HR professionals will gain a visually enriched understanding of the synthetic job market, aiding them in making informed choices and staying updated on industry trends.

Expected limitations

We expected certain limitations on part of the dataset and the tools we were using for analysis. Since our dashboarding tool of choice was Power BI, we faced several problems with its workflow and usage. Certain plot types are unavailable for the free tier version that we were using. We were not able to derive any map plots due to this reason. All 3 map plots available in Power BI required a premium license.

We faced the same issue with added plots from AppSource. Certain lot types were not to be used with our subscription. Other limitations include the number of numerical and categorical columns available in the dataset. If we had more data columns, we could have been able to do a much-detailed analysis of our domain.

Proposed Work Schedule

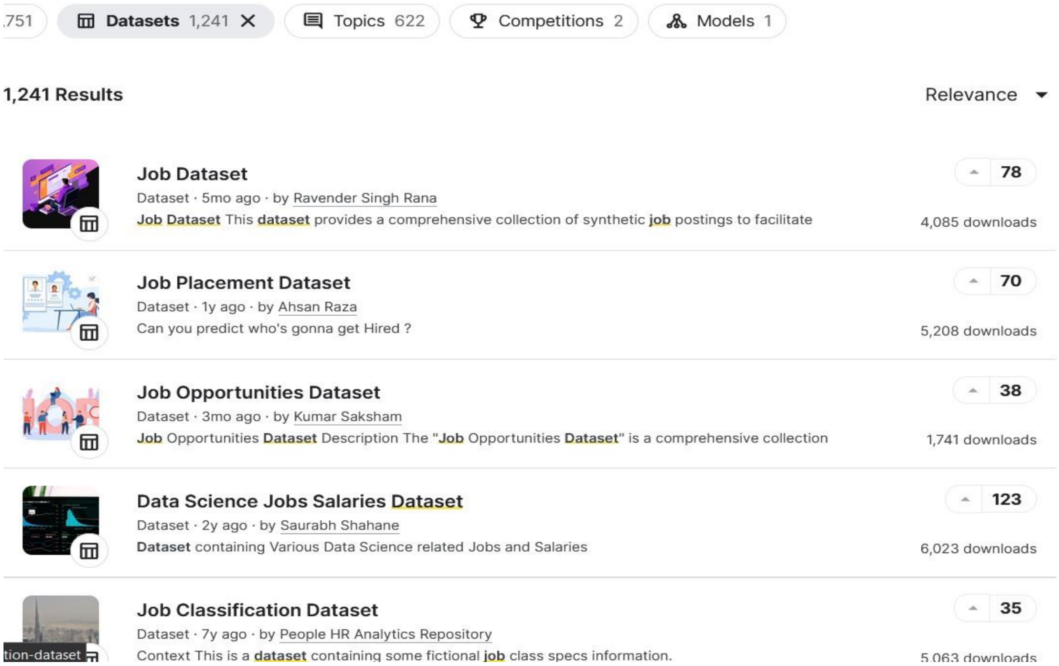
Data collection and completion of the project proposal are already done. On 2 May, it is planned to complete the data visualization project. Analysis and visualizations will be done within that time.

Literature review

Introduction to the Dataset Used

For this project, we overlooked several data samples that can be used. Instead of creating primary data through our data collection procedure, we have opted to use a publicly available secondary data source.

As our selected topic was Job salaries, it was convenient to use many available secondary data sources which were available. For the data source, we choose Kaggle as one of the best public data repositories for data enthusiasts. We sampled numerous public datasets available in Kaggle and these are some of the few datasets about jobs in Kaggle.



751 | Datasets 1,241 | Topics 622 | Competitions 2 | Models 1

1,241 Results

Relevance

Job Dataset
Dataset · 5mo ago · by Ravender Singh Rana
Job Dataset This dataset provides a comprehensive collection of synthetic job postings to facilitate
78
4,085 downloads

Job Placement Dataset
Dataset · 1y ago · by Ahsan Raza
Can you predict who's gonna get Hired ?
70
5,208 downloads

Job Opportunities Dataset
Dataset · 3mo ago · by Kumar Saksham
Job Opportunities Dataset Description The "Job Opportunities Dataset" is a comprehensive collection
38
1,741 downloads

Data Science Jobs Salaries Dataset
Dataset · 2y ago · by Saurabh Shahane
Dataset containing Various Data Science related Jobs and Salaries
123
6,023 downloads

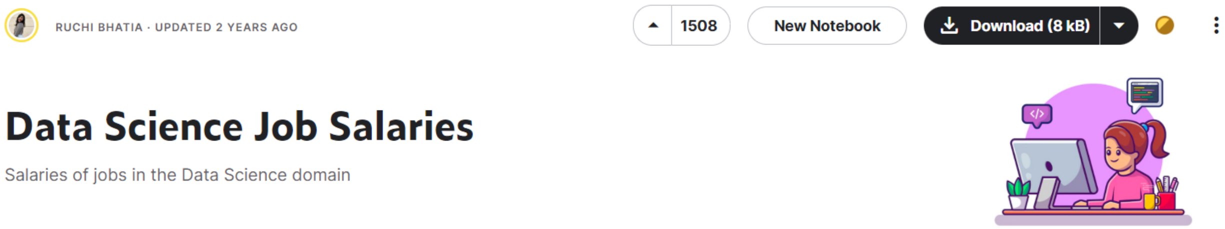
Job Classification Dataset
Dataset · 7y ago · by People HR Analytics Repository
Context This is a dataset containing some fictional job class specs information.
35
5,063 downloads

tion-dataset

There were 1241 datasets in Kaggle related to jobs. However, after considering the volume, parameters, and variables available for analysis, we selected a few datasets that were good candidate datasets that we could use.

Further looking into their accuracy and the amount of preprocessing needed to get the data ready for our purposes, we looked for the most comprehensive data source that would allow us to get the greatest number of insights.

Therefore, we chose the following dataset for this project.



RUCHI BHATIA · UPDATED 2 YEARS AGO

1508 | New Notebook | Download (8 kB)

Data Science Job Salaries

Salaries of jobs in the Data Science domain

This dataset can be downloaded through this Kaggle link:

https://www.kaggle.com/datasets/ruchi798/data-science-job-salaries

Overview of the data set

Kaggle is the data retrieved source for our project. This data set is a comprehensive job data set for data science, research, and analysis. The dataset was created by Ruchi Bhata for educational research purposes. This dataset has been created from a web scraping of the jobs available at https://ai-jobs.net/ It includes job insights which are essential for understanding the roles and salaries associated with various positions. This data set includes a diverse set of jobs within the data science domain.

The data set contains 11 variables. They are,

1. Work year: the year the salary was paid.
2. Experience level: The level of experience for a particular job title. EN: Enty-level / MI: Mid-level / SE: Senior-level / EX: Executive-level
3. Employment type: The type of employee. PT for part-time, FT for full-time, CT for contract, and FL for freelance.
4. Job Title: The role worked in the year.
5. Salary: Total salary paid for the role.
6. Salary currency: the currency of the paid as an ISO 4217 currency code.
7. Salary in usd: the salary in USD
8. Employee residence: Employee's country of residence in during the work year.
9. Remote Ratio: the overall amount of work done remotely.
10. Company location: The country of the employee's main office or contracting branch.
11. Company size: The average number of employees worked for the company during this year.

The dataset is small, and it contains about 608 individual data rows. Because of that reason, we can get an overall idea of how the various job titles in data science domain salaries are distributed. Most of the data contains categorical data types. However, there are some numerical variables in the data set. Because of that reason, we identify that we can perform some statistical analysis as well.

Preprocessing methodologies

Preprocessing refers to the steps or techniques applied to data before it is used for analysis or modeling. Preprocessing includes data cleaning, data transformation, feature engineering, and dimensionality reduction.

Data Blending & Integration

The dataset that is chosen for this project is a secondary one. Because of that reason, the data blending and integration are already done. Plus, all the data was in a singular tabular format. Thus, we were free from data modeling tasks and blending of several queries together.

Data cleaning

Since this assignment is based mostly on descriptive statistics, we only needed to work on data cleaning and data transformation. When choosing the dataset, we opted for a dataset that was clean and needed the least amount of preprocessing of data. Thus, the amount of cleaning needed will be minimal. When it comes to data transformation, the data columns such as salary are given as ranges.

Data Transformation & Reduction

Though the cleaning steps in the dataset were minimal, we still had to replace certain values for the ease of visualizing. The column values and values in the categorical data columns were mostly in shortened forms. Thus, we had to replace them with much simpler names for better clarity. The following transformations were done in the data columns and values.

Renamed the empty column of the index as “job id”

Replacing values in the Experience column

EN -Entry-level
MI -Mid-level
SE -Senior-level
EX -Executive-level

Replacing values in the employment type column

PT -Part-time
FT -Full-time
CT- Contract
FL –Freelance

Replacing values in the Company size column

S- small
M- medium
L- Large

Methodology

Introduction

As the dataset we used was a job salaries dataset with over 600 rows of data, we wanted to choose a tool that could make our analysis easier to handle. We wanted to choose a tool that makes every step of this project easier and well-polished in the outcome. For the preprocessing of data, we had the option of using programming libraries such as pandas in Python or dplyr in R. We also have the option of using power query builder in Power BI or prep builder in a tableau that can make the preprocessing much more intuitive.

After the preprocessing of data, we assume that acquired new numerical data can give us an interactive and more clear solution to our problem statement.

Type of Data to be collected and data sources.

The data collection process for our project involved sourcing a comprehensive dataset from Kaggle, a prominent platform for data enthusiasts and professionals. Rather than conducting our primary data collection, we opted for a secondary data source due to the convenience and richness of publicly available datasets related to job listing on Kaggle. With a pool of 1241 datasets about jobs, we carefully evaluated their volume, parameters, and suitability for analysis and visualization.

The chosen dataset, accessible through the Kaggle link provided above, was created by Ruchi Bhatia with the help of ai-jobs.net. Diverse job salaries across industries serve as the foundation for our exploration into the dynamic landscape of the job market.

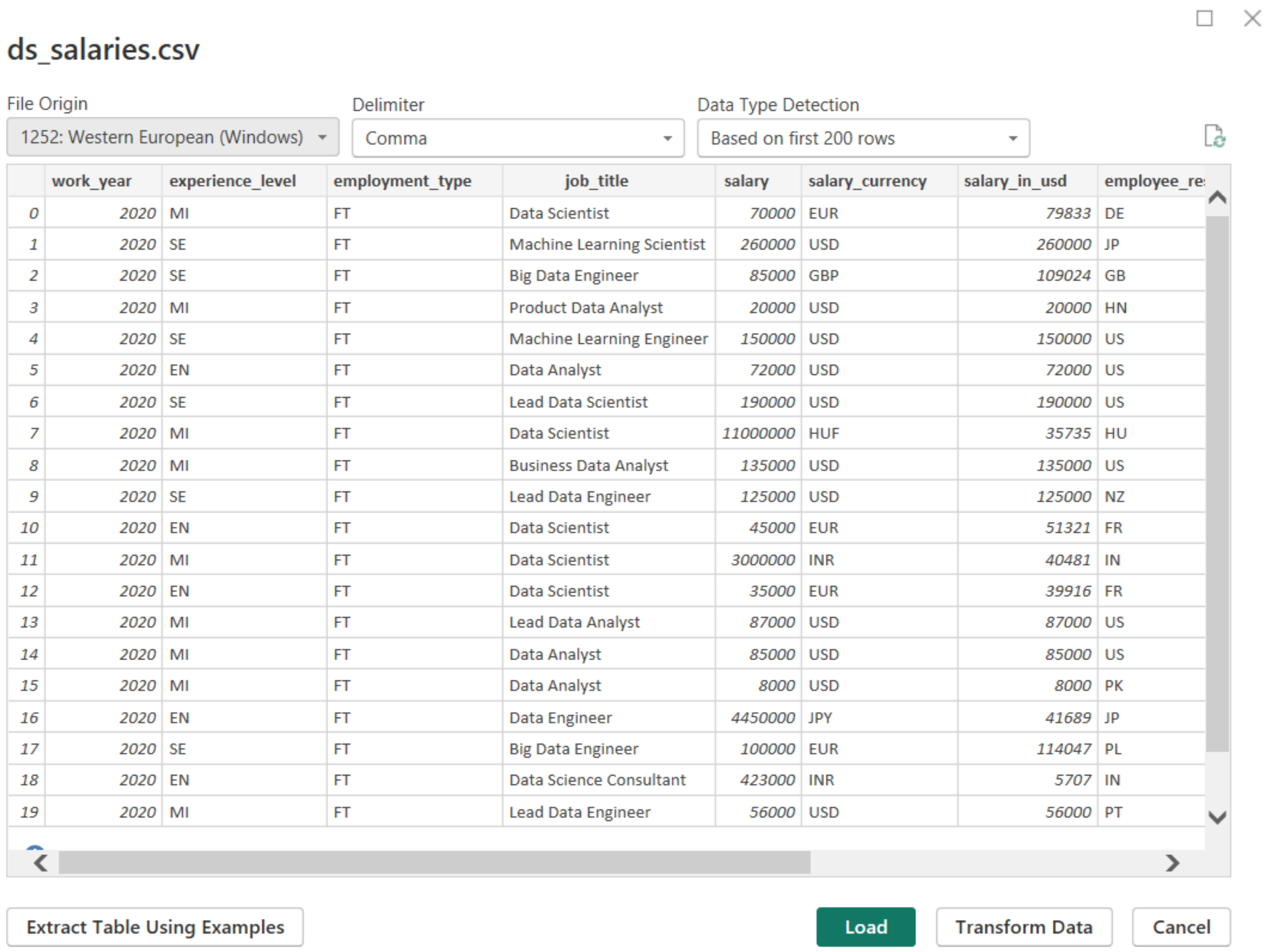
Methods techniques and tools used to visualize the Data

We have decided to have Power BI as the dashboarding tool of our choice for this project. Our previous familiarity with the application plus its wide usage in the industry made us choose Power BI with the integration of Python for further analysis.

For creating interactive dashboards and visualizations, we'll turn to Power BI, a tool developed by Microsoft. Power BI lets us build engaging dashboards that showcase key insights from the job listing data. Its user-friendly interface makes it simple to explore data, create visuals, and share findings. Plus, Power BI offers seamless integration with various data sources, ensuring our dashboards are always up to date.

By combining Python / R and Power BI, we'll conduct thorough analysis and visualization of job salary data, providing valuable insights for informed decision-making.

The following is an image of how the data is represented when loaded into Power BI.



ds_salaries.csv

File Origin: 1252: Western European (Windows)

Delimiter: Comma

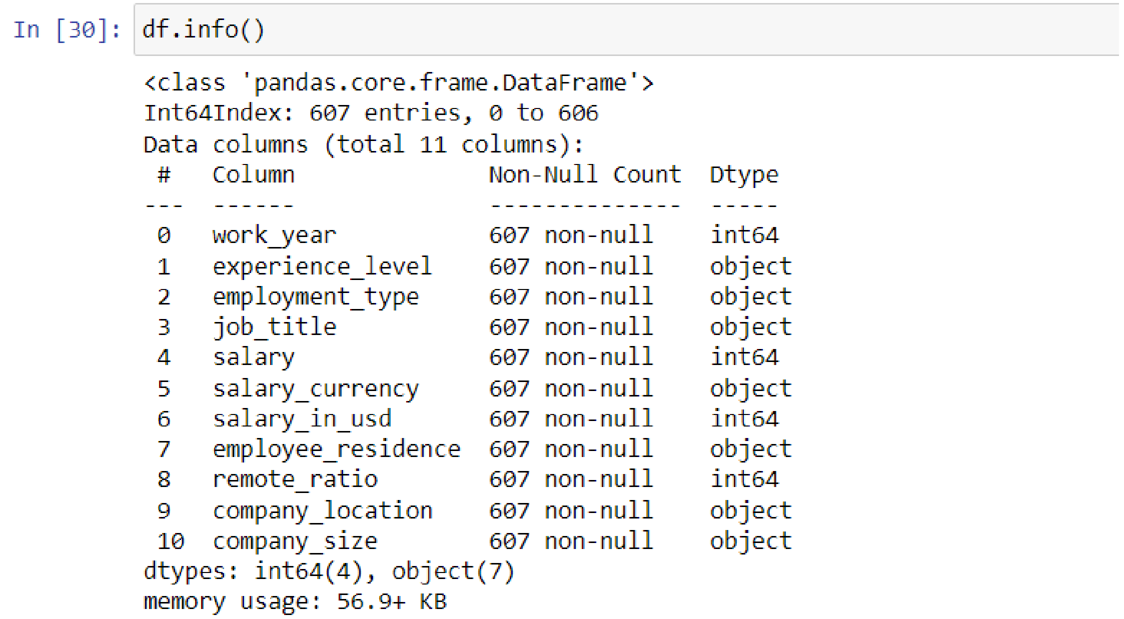
Data Type Detection: Based on first 200 rows

	work_year	experience_level	employment_type	job_title	salary	salary_currency	salary_in_usd	employee_re
0	2020	MI	FT	Data Scientist	70000	EUR	79833	DE
1	2020	SE	FT	Machine Learning Scientist	260000	USD	260000	JP
2	2020	SE	FT	Big Data Engineer	85000	GBP	109024	GB
3	2020	MI	FT	Product Data Analyst	20000	USD	20000	HN
4	2020	SE	FT	Machine Learning Engineer	150000	USD	150000	US
5	2020	EN	FT	Data Analyst	72000	USD	72000	US
6	2020	SE	FT	Lead Data Scientist	190000	USD	190000	US
7	2020	MI	FT	Data Scientist	11000000	HUF	35735	HU
8	2020	MI	FT	Business Data Analyst	135000	USD	135000	US
9	2020	SE	FT	Lead Data Engineer	125000	USD	125000	NZ
10	2020	EN	FT	Data Scientist	45000	EUR	51321	FR
11	2020	MI	FT	Data Scientist	3000000	INR	40481	IN
12	2020	EN	FT	Data Scientist	35000	EUR	39916	FR
13	2020	MI	FT	Lead Data Analyst	87000	USD	87000	US
14	2020	MI	FT	Data Analyst	85000	USD	85000	US
15	2020	MI	FT	Data Analyst	8000	USD	8000	PK
16	2020	EN	FT	Data Engineer	4450000	JPY	41689	JP
17	2020	SE	FT	Big Data Engineer	100000	EUR	114047	PL
18	2020	EN	FT	Data Science Consultant	423000	INR	5707	IN
19	2020	MI	FT	Lead Data Engineer	56000	USD	56000	PT

Data Analysis

Descriptive Analysis

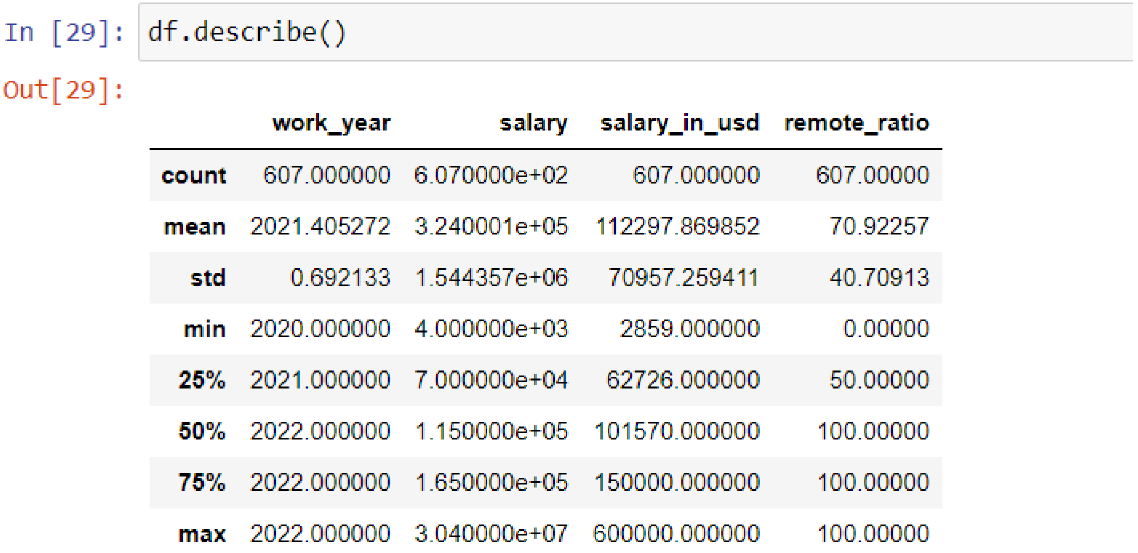
Descriptive analysis is a key area to consider when it comes to analysis. Therefore, we decided to use Python to interpret the descriptive analysis of our data science job salaries dataset. We use some functions to perform that task. They are info () and describe () functions. Also, we describe the object variables as well.



```
In [30]: df.info()

         <class 'pandas.core.frame.DataFrame'>
         Int64Index: 607 entries, 0 to 606
         Data columns (total 11 columns):
          #   Column              Non-Null Count  Dtype 
         ---  ------              --------------  ----- 
          0   work_year           607 non-null    int64 
          1   experience_level    607 non-null    object
          2   employment_type     607 non-null    object
          3   job_title           607 non-null    object
          4   salary              607 non-null    int64 
          5   salary_currency     607 non-null    object
          6   salary_in_usd       607 non-null    int64 
          7   employee_residence  607 non-null    object
          8   remote_ratio        607 non-null    int64 
          9   company_location    607 non-null    object
          10  company_size        607 non-null    object
         dtypes: int64(4), object(7)
         memory usage: 56.9+ KB
```

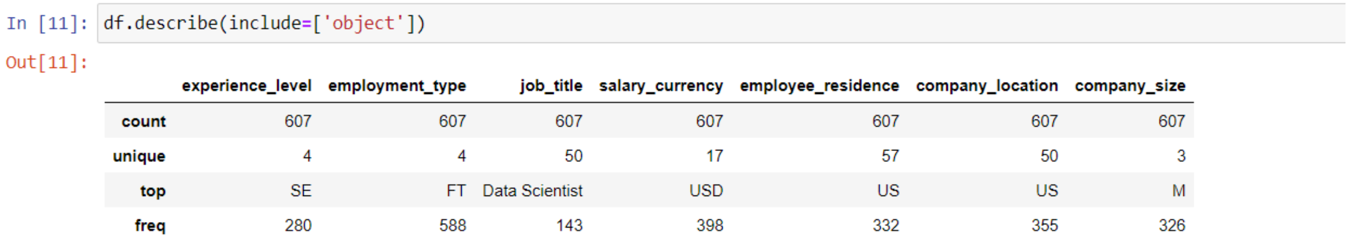
When we observe the dataset, we can see that there are no null values spread among the variables. There are 4 integer data types, 7 object data types, and no float or any other data types. It has a total number of 607 rows.



```
In [29]: df.describe()
Out[29]:
```

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

According to the description function, we can get a brief idea of the integer variables. There we can see count, mean, standard deviation, minimum and maximum values, and interquartile ranges. We consider about work year we can see the distribution is from 2020 to 2022. In the salary column, individual salaries are represented according to their currencies which are used by countries. Therefore, we consider salary in the USD column for visualization. The mean salary is 112297 USD. The maximum salary is 600000 USD, and the minimum salary is 2589 USD. Most of the salary lies between +70957 and – 70957 from the mean value. The average remote ratio is 70. This means the majority of jobs are allowed remotely.



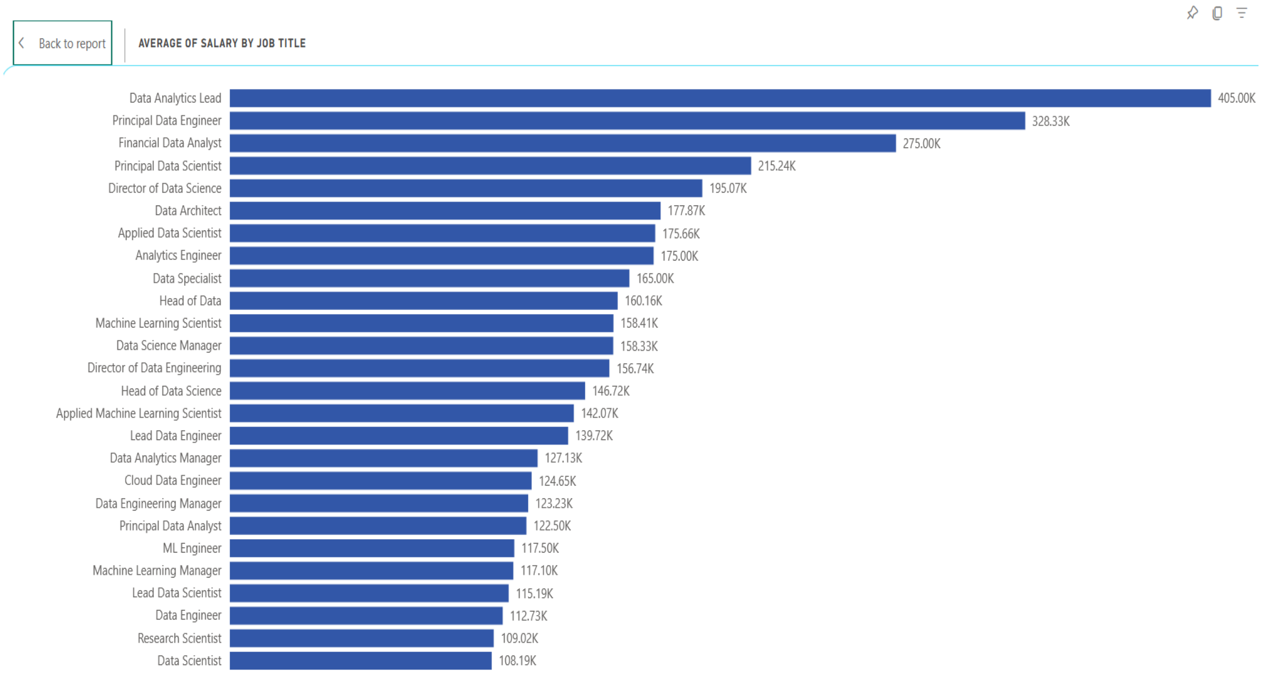
```
In [11]: df.describe(include=['object'])
Out[11]:
```

	experience_level	employment_type	job_title	salary_currency	employee_residence	company_location	company_size
count	607	607	607	607	607	607	607
unique	4	4	50	17	57	50	3
top	SE	FT	Data Scientist	USD	US	US	M
freq	280	588	143	398	332	355	326

In this, we get the summary of object data types. There are 7 in total. At the experience level, there are unique 4 values and the SE is the most frequent item in the column, which means, most of the time they asked for senior-level experience. The full-time employment type has the highest frequency rather than other employment types like part-time, contract, and freelance. From the job titles Data science is the most demanding role in the data science domain. According to the employee residences, most of the jobs are available in the US and the number of companies situated in the US. When comes to company sizes, most companies are medium-sized companies.

Data Analysis with Data Visualizations

Average of salary by Job title

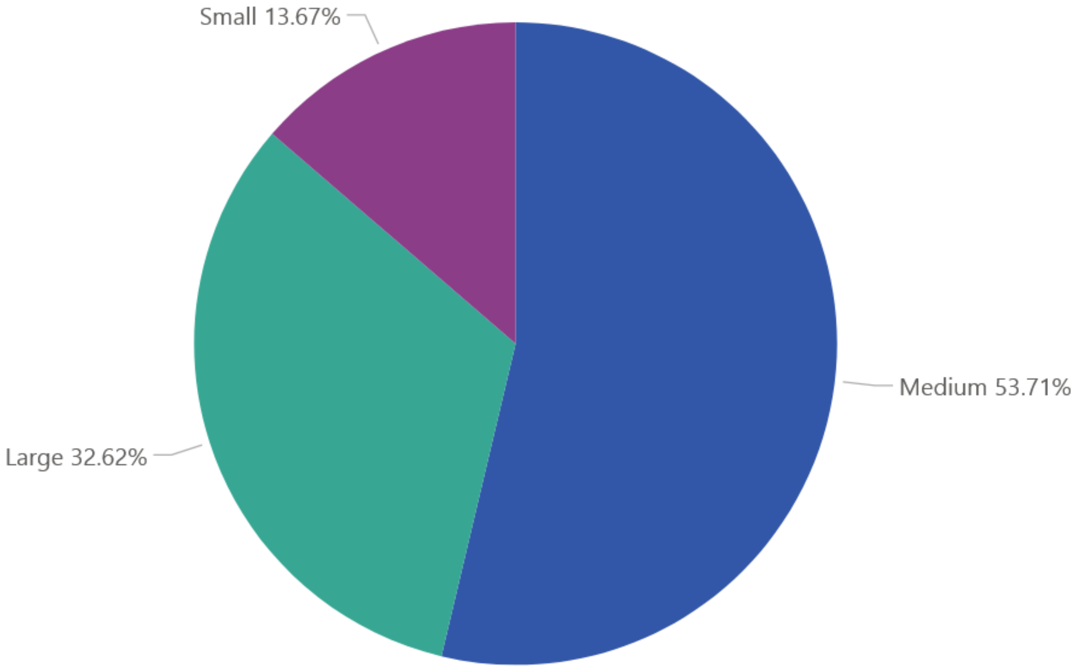


The dashboard represents a historical analysis of average salaries within data science job roles for the years 2020, 2021, and 2022. Bar chart visualizations offer a comparative overview of average salaries across various roles in making informed decisions. By providing separate bar charts for each year, the dashboard enables users to delve into year-wise salary trends, gaining valuable insights into the evolving earning potential over time.

The displayed salaries vary widely from $405,000 for a Data Analyst and $5409 for a 3D Computer Vision Researcher. This reflects how different skills, responsibilities, and demand levels influence earning potential in data science roles. The highest average salary is attributed to the role of Data Analytic Lead, showcasing the potential of leadership positions. Principal Data Engineers also earn significantly with an average salary of $328,330, indicating the value of leadership and technical expertise.

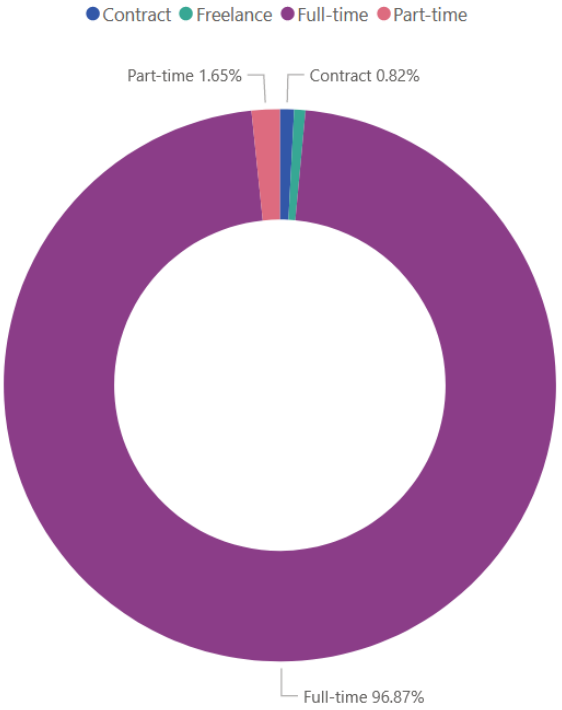
The dashboard represents a diverse array of job titles, ranging from specialized roles like financial Data Analyst and Applied Data Scientist to leadership positions such as Director of Data Science and Head of Data. This diversity reflects the multifaceted nature of careers in data science. Based on the data, aiming for roles like Data Analytics Lead or Principal Data Engineer could lead to higher earnings.

Company size



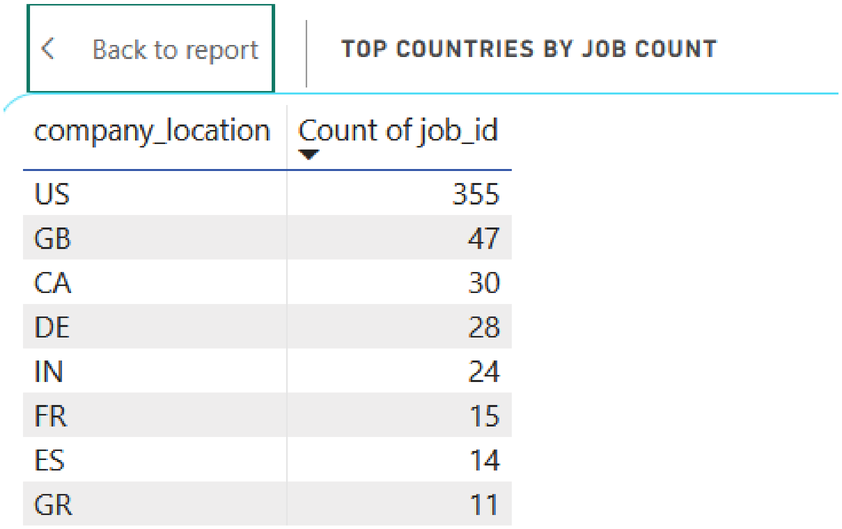
Mainly there are three types of companies. They are small companies, medium companies, and large companies. The small company has less than 50 employees, the medium-sized company has 50 to 250 employees, and the large company has more than 250 employees. We consider overall job offerings, 13% of companies are small, 32% of large companies, and 53% of companies are large-scale companies. Therefore, we can see that most data science domain jobs rely on large-scale companies. Also, we can see that in 2020- and 2021-years large-scale companies are leads in offering jobs. But in 2022 medium companies are leading job offers in the job market. Therefore, we can see that most of the time data science domain degree holders can have a job role in large or medium-scale companies.

Count of Jobs by Employment Type



This chart depicts the distribution of jobs across four employment types: Contract, Freelance, Full-time, and Part-time. Freelance employment shows the minimum count of 4, while Full-time positions top the chart at 588. Part-time positions amount to 10, surpassing the count of Contract roles, which stands at 5. The dataset reveals that the majority of jobs are Full-time positions, constituting 98.87%, with the others forming smaller fractions. Part-time positions stand at 1.65%, slightly surpassing the counts of Contracts and Freelance roles. Additionally, there's a notable trend from 2020 to 2021 where Part-time and other small fraction counts increase, followed by a decrease from 2021 to 2022. It's evident that many holders of degrees in the data science domain are transitioning towards full-time employment.

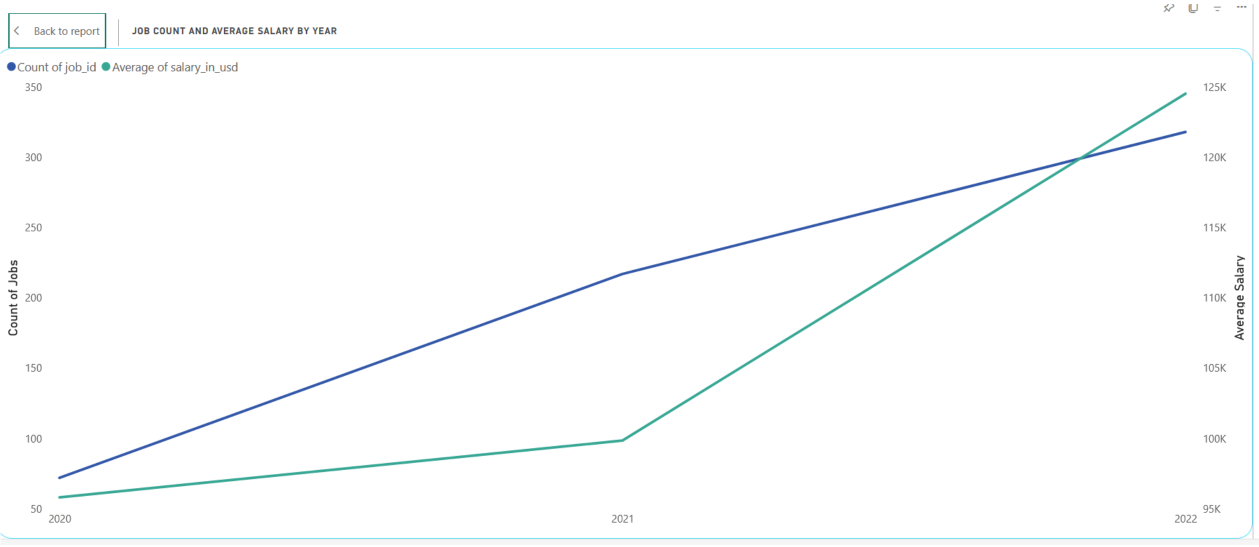
Top countries by job count



company_location	Count of job_id
US	355
GB	47
CA	30
DE	28
IN	24
FR	15
ES	14
GR	11

Most of the jobs in our dataset are from the United States over the years. The United States accounts for more than half of all the job postings with a count of 355. The next spot is taken by the United Kingdom (GB) with 47 jobs listed. Canada (CA), Germany (DE), and India (IN) are the next top countries by the number of jobs posted. Since there were a large number of jobs posted in the year 2022 compared to both previous 2 years, the count of jobs in the year 2022 has had the most impact. For example, in both the years 2020 and 2021 Germany has accounted to be the second highest count of jobs. Therefore, a data science domain degree holder has many opportunities to work abroad in countries like America, the United Kingdom, Canada, etc.

Number of jobs and average salary by year

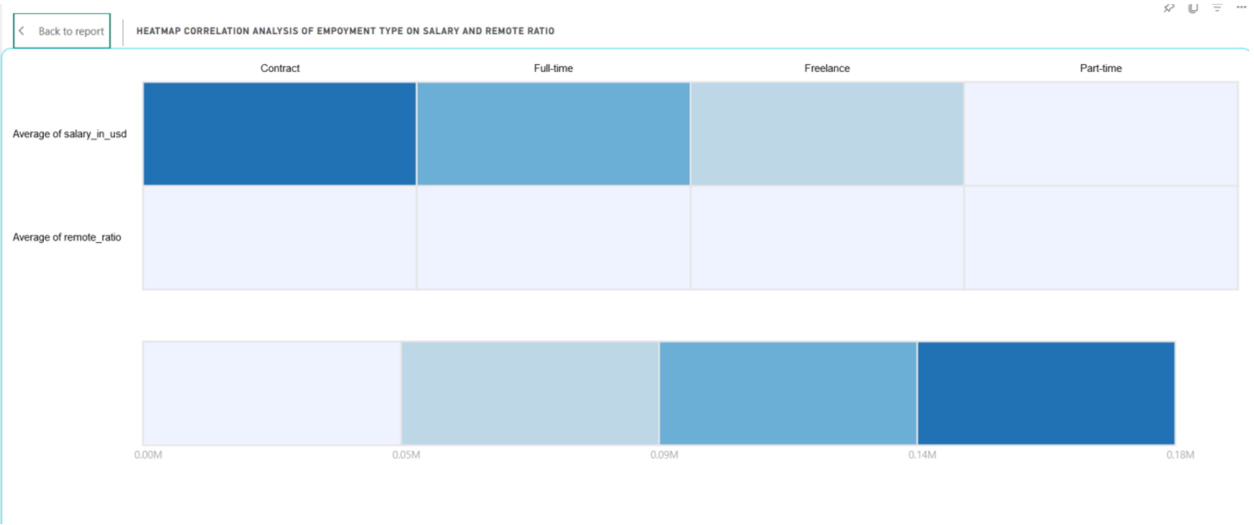


This line plot visualizes the overall trend of salary and number of jobs by year. As it is seen, the overall trend of job postings is increasing every year. But when it comes to the average salary, it shows a dip in the average salary in the year 2021 before exponentially increasing again in 2022. One interesting finding is that there has been a steady increase in salary in large companies regardless of the year. This adds to the fact that salaries in large companies are more consistent. For medium-scale companies, which account for over 50% of all, the number of jobs posted, and salaries show a similar trend.

Statistical Analysis

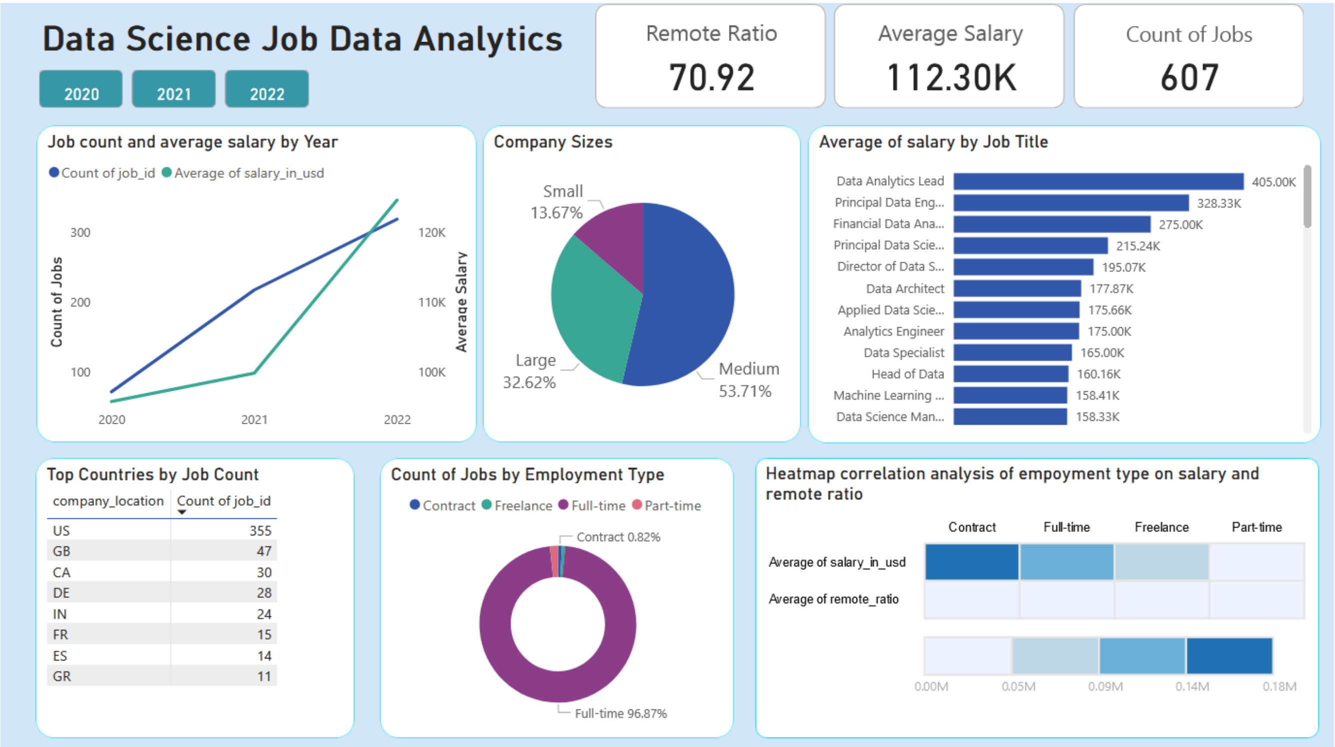
Heatmap correlation analysis involved us in mapping the correlation between variables and how much of an effect each variable had on the other. An interesting finding we found was that though the number of full-time jobs was high, contract jobs were the jobs that had the highest correlation with the average salary. Part-time and freelance work jobs displayed a low correlation with the salary. The remote ratio has little to no correlation with the employment type.

Employment type on salary and remote ratio



The listed employment types are contract, freelance, full-time time, and part-time. As stated above, 96% of all job listings are full-time. While accounting for a greater number, we wanted to see the relationship between employment type, salary, and remote ratio. To our surprise, there was little to no relationship between the remote ratio and employment type. This means, working hybrid, remote and on-site are equally distributed across all employment types.

Final dashboard


Data Science Job Data Analytics
2020
2021
2022
Remote Ratio
70.92
Average Salary
112.30K
Count of Jobs
607
Job count and average salary by Year
● Count of job_id ● Average of salary_in_usd
Count of Jobs
300
200
100
Average Salary
120K
110K
100K
2020
2021
2022
Company Sizes
Small 13.67%
Large 32.62%
Medium 53.71%
Average of salary by Job Title
Data Analytics Lead 405.00K
Principal Data Eng... 328.33K
Financial Data Ana... 275.00K
Principal Data Scie... 215.24K
Director of Data S... 195.07K
Data Architect 177.87K
Applied Data Scie... 175.66K
Analytics Engineer 175.00K
Data Specialist 165.00K
Head of Data 160.16K
Machine Learning ... 158.41K
Data Science Man... 158.33K
Top Countries by Job Count
Count of Jobs by Employment Type
● Contract ● Freelance ● Full-time ● Part-time
Contract 0.82%
Full-time 96.87%
Heatmap correlation analysis of empoyment type on salary and remote ratio
Contract
Full-time
Freelance
Part-time
Average of salary_in_usd
Average of remote_ratio
0.00M
0.05M
0.09M
0.14M
0.18M

company_location	Count of job_id
US	355
GB	47
CA	30
DE	28
IN	24
FR	15
ES	14
GR	11

Conclusion

In conclusion, this thorough analysis of the data science jobs brings several insights about the landscape of jobs in data science. The average annual salary of a data science professional exceeds 100,000$. There is a growth in the trend of jobs and salaries in data science and most of the hires are done by medium and large-scale companies.

Senior/lead/principal positions yield more salaries. When it comes to the number of jobs listed, more than 50% of the jobs are from the US, and over 96% of the jobs are full-time. Though most of these jobs are full-time, almost all these jobs allow hybrid or fully remote work. More experienced senior professionals who work in contract and full-time positions have a higher correlation with a higher salary.

Attachments

Power bi link:https://app.powerbi.com/groups/me/reports/3f182181-f9bc-499a-bdee-e9a0d94a91ae/ReportSectioncbe73a82aa2851c7d773?experience=power-bi