2023

JOBS AND SALARIES IN DATA SCIENCE

Sijie ZHU

TABLE OF CONTENTS

INTRODUCTION OF DATASET

Before start this report, please let me introduce this dataset from three aspects : source, content and size.

SOURCE

This data set is found from Kaggle, a website where we often can find some data to practice. Actually, Kaggle is a data science competition_platform and online community of data scientists and machine learning practitioners under Google LLC. Kaggle enables users to find and publish datasets, explore and build models in a web-based data science environment, work with other data scientists and machine learning engineers, and enter competitions to solve data science challenges

CONTENT

This dataset is about jobs and salary in data field all over the world from year 2020 to year 2023. It includes the basic information about each job, like job category, salary, employee residence, employment type, company size, company location... All this dimensions give us a different way to analyze the current job situation in this sector.

SIZE

In total we have 9355 observations and 12 columns, among which 3 are numerical columns and 9 are categorical columns. We collected information form 83 countries, 10 job categories in the data secteur like Machine Learning , Data Analysis, Data Engineering etc.

Bellow are the basic information about the dataset

- **work_year**: The year in which the data was recorded. This field indicates the temporal context of the data, important for understanding salary trends over time.

- **job_title**: The specific title of the job role, like 'Data Scientist', 'Data Engineer', or 'Data Analyst'. This column is crucial for understanding the salary distribution across various specialized roles within the data field.

- **job_category**: A classification of the job role into broader categories for easier analysis. This might include areas like 'Data Analysis', 'Machine Learning', 'Data Engineering', etc.

- **salary_currency**: The currency in which the salary is paid, such as USD, EUR, etc. This is important for currency conversion and understanding the actual value of the salary in a global context.

- **salary**: The annual gross salary of the role in the local currency. This raw salary figure is key for direct regional salary comparisons.

- **salary_in_usd**: The annual gross salary converted to United States Dollars (USD). This uniform currency conversion aids in global salary comparisons and analyses.

- **employee_residence**: The country of residence of the employee. This data point can be used to explore geographical salary differences and cost-of-living variations.

- **experience_level**: Classifies the professional experience level of the employee. Common categories might include 'Entry-level', 'Mid-level', 'Senior', and 'Executive', providing insight into how experience influences salary in data-related roles.

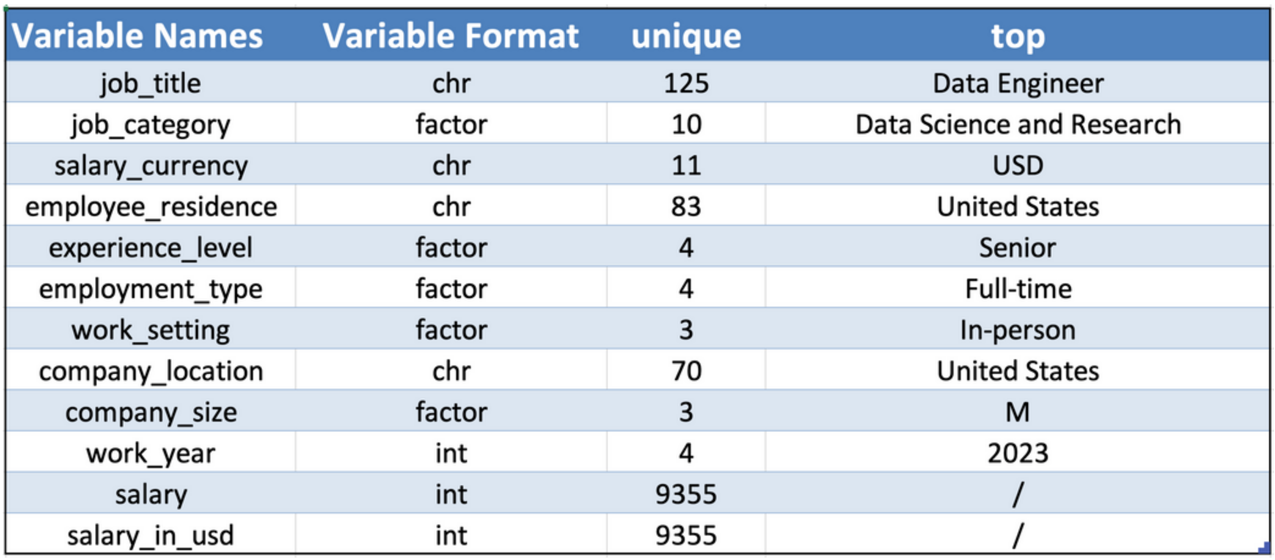
- **employment_type**: Specifies the type of employment, such as 'Full-time', 'Part-time', 'Contract', etc. This helps in analyzing how different employment arrangements affect salary structures.

- **work_setting**: The work setting or environment, like 'Remote', 'In-person', or 'Hybrid'. This column reflects the impact of work settings on salary levels in the data industry.

- company_location: The country where the company is located. It helps in analyzing how the location of the company affects salary structures.

- **company_size**: The size of the employer company, often categorized into small (S), medium (M), and large (L) sizes. This allows for analysis of how company size influences salary.

Basic information about the dataset



Variable Names	Variable Format	unique	top
job_title	chr	125	Data Engineer
job_category	factor	10	Data Science and Research
salary_currency	chr	11	USD
employee_residence	chr	83	United States
experience_level	factor	4	Senior
employment_type	factor	4	Full-time
work_setting	factor	3	In-person
company_location	chr	70	United States
company_size	factor	3	M
work_year	int	4	2023
salary	int	9355	/
salary_in_usd	int	9355	/

WHY THIS DATASET?

As a Digital Marketing and Data Science student, I always want to know what kind of job I can apply for and what are the salaries for the different jobs.

The R project gives me a good opportunity to find a date set in this field and I would like to have a basic insight for my future carrer choice.

In a nutshell, after this exploration, I would like to find out the different job opportunities provided by different countries. what kind of companies are hiring data people? Where are they located and what are their size. Also, I would like to know the changes of jobs over year.

My goal:

In a nutshell, after this exploration, I would like to have a basic insight of job offers and salaries in data field to facilitate my career in the near future.

DATA CLEAN

01

Explore dataset

Explore basic information, for example, number of rows, columns, datatype of each columns, number of missing values in each columns etc.

02

Change columns type

All the dimensions like job title, job category are stored in char type, but in R they need to changed to factor types.

03

Remove duplicates

Check duplicated rows and remove them, the original dataset is even bigger, but however there are many duplicates.

04

Remove unrelated columns

Since I want to comparing salaries of different jobs, so I will only keep the salary in usd and remove the salary currency and salary columns.

05

Deal with outliers and anormal values

Use scatter plot to check salaries, I found out some salary are extremely high and low, there are even negative values. This may impact our analysis, so I removed this part of values.

06

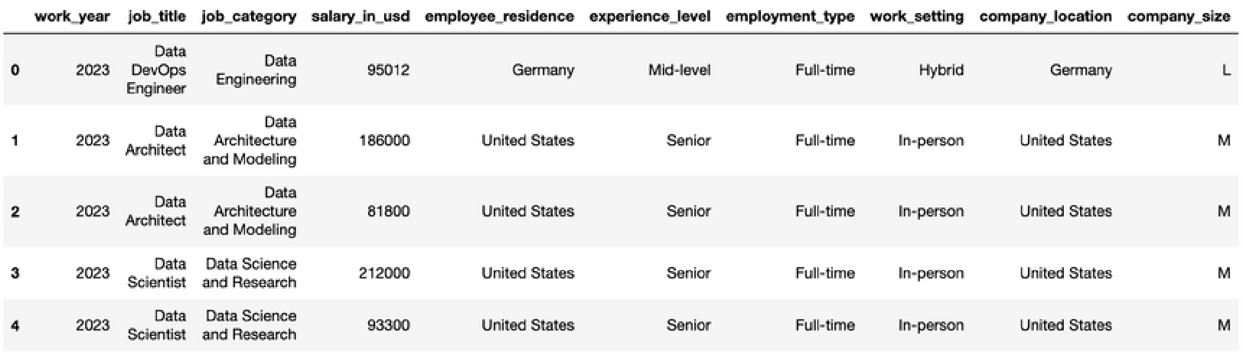
Deal with numerical missing values

Salary is a key indicator for this analysis, so in order to keep as much as data possible, I will not just remove all the missing values. Instead I will calculate the average salary of each job title and use the average salary to fill out all the missing salaries.

07

Remove categorical missing values

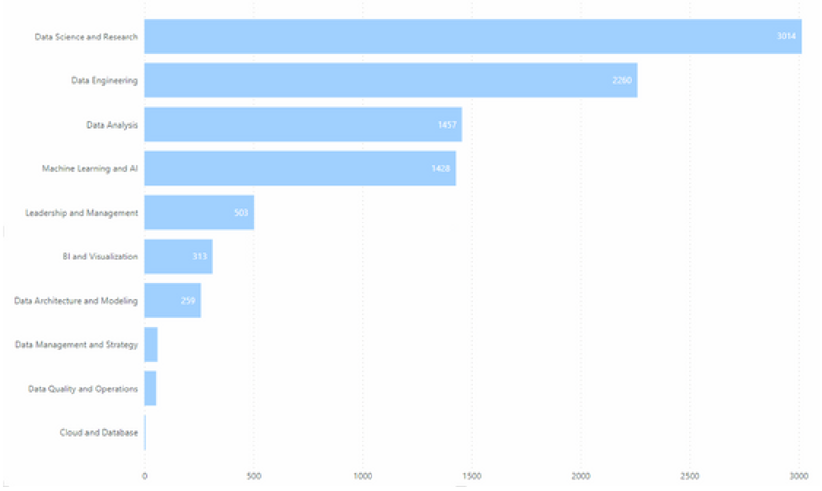
For instance, I cannot find a good way to deal with categorical missing values, so I will just remove them



	work_year	job_title	job_category	salary_in_usd	employee_residence	experience_level	employment_type	work_setting	company_location	company_size
0	2023	Data DevOps Engineer	Data Engineering	95012	Germany	Mid-level	Full-time	Hybrid	Germany	L
1	2023	Data Architect	Data Architecture and Modeling	186000	United States	Senior	Full-time	In-person	United States	M
2	2023	Data Architect	Data Architecture and Modeling	81800	United States	Senior	Full-time	In-person	United States	M
3	2023	Data Scientist	Data Science and Research	212000	United States	Senior	Full-time	In-person	United States	M
4	2023	Data Scientist	Data Science and Research	93300	United States	Senior	Full-time	In-person	United States	M

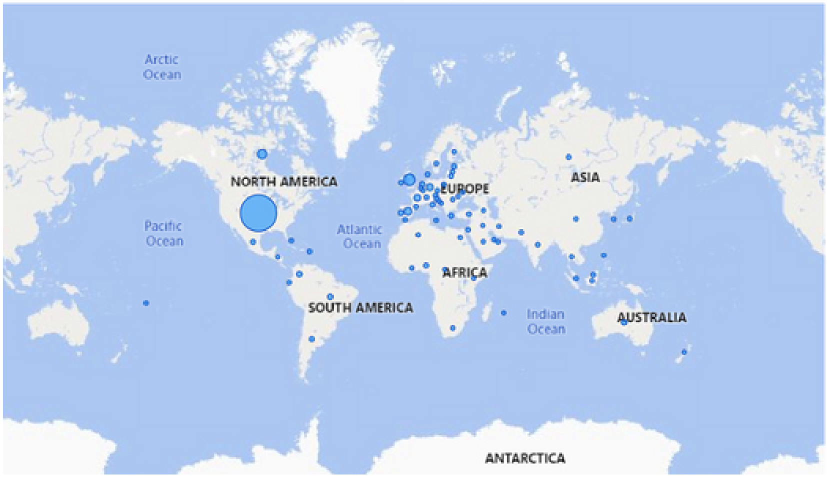
ANALYSIS AND GRAPHS

Job numbers by job category



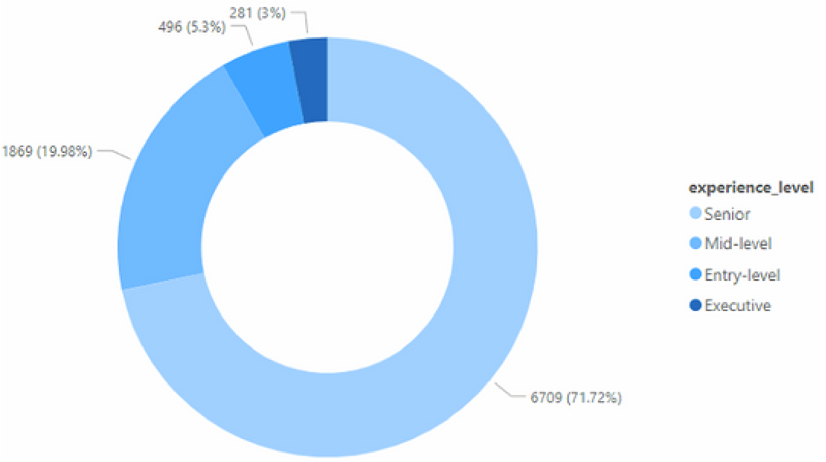
According the result, Data Science and Research provides more job offers than other categories. Every year this will change, so we will add a year slicer while making the website

Job numbers by job position



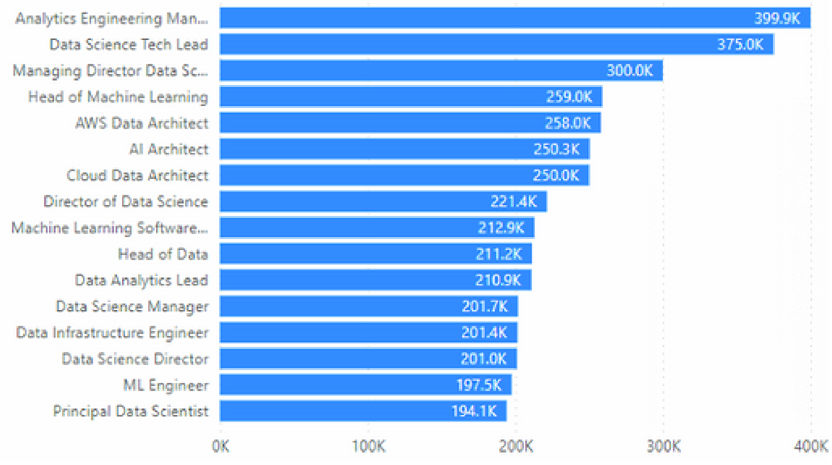
America provides way more job offers than other countries. We can find out that, most of these jobs focus on North America and Europe.

Job numbers by experience level



Working experience has a big impact on job numbers. According to the graph, companies often need more seniors than other levels. It accounts more than 71% of the jobs

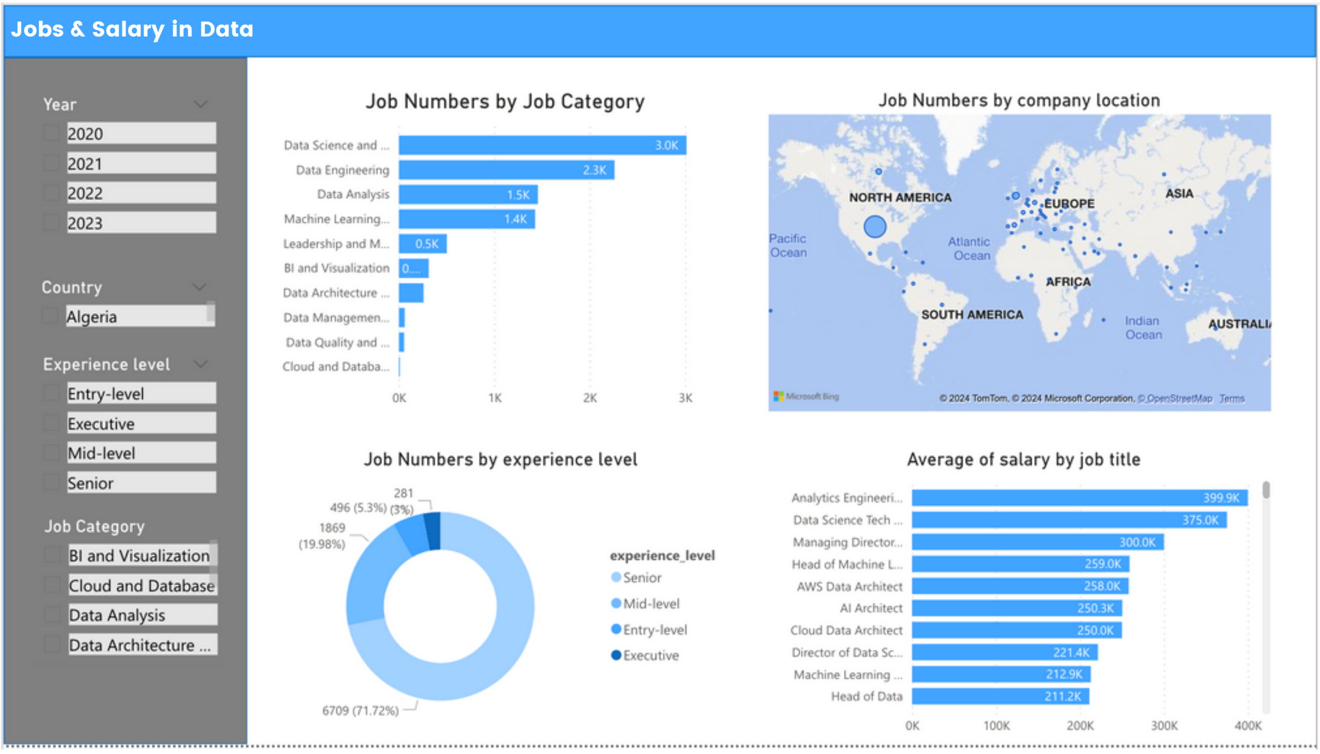
Average Salary by job title



Salaries change a lot according to different job titles. The tops are management positions, like Analytics Engineering Management.

APPLICATION DRAFT

Below is the outlook sketch of the App interface . On the left side is the widgets, where we can filter the year , country , job category and working experience. On the body part, there will be more buttons, where we can see more details, like table and more charts.


Jobs & Salary in Data
Year
2020
2021
2022
2023
Country
Algeria
Experience level
Entry-level
Executive
Mid-level
Senior
Job Category
BI and Visualization
Cloud and Database
Data Analysis
Data Architecture ...
Job Numbers by Job Category
Data Science and ...
Data Engineering
Data Analysis
Machine Learning...
Leadership and M...
BI and Visualization
Data Architecture ...
Data Managemen...
Data Quality and ...
Cloud and Databa...
3.0K
2.3K
1.5K
1.4K
0.5K
0....
0K
1K
2K
3K
Job Numbers by company location
NORTH AMERICA
EUROPE
ASIA
Pacific Ocean
Atlantic Ocean
AFRICA
SOUTH AMERICA
Indian Ocean
AUSTRALIA
Microsoft Bing
© 2024 TomTom, © 2024 Microsoft Corporation, © OpenStreetMap Terms
Job Numbers by experience level
281
496 (5.3%) (3%)
1869
(19.98%)
6709 (71.72%)
experience_level
● Senior
● Mid-level
● Entry-level
● Executive
Average of salary by job title
Analytics Engineeri...
Data Science Tech ...
Managing Director...
Head of Machine L...
AWS Data Architect
AI Architect
Cloud Data Architect
Director of Data Sc...
Machine Learning ...
Head of Data
399.9K
375.0K
300.0K
259.0K
258.0K
250.3K
250.0K
221.4K
212.9K
211.2K
0K
100K
200K
300K
400K