ANALYSIS OF JOBS AND SALARIES IN DATA SCIENCE USING SQL

- PRADEEP K I

Introduction

In the rapidly evolving landscape of the data industry, understanding salary trends is crucial for professionals, employers, and policymakers alike. This project aims to provide comprehensive insights into the factors influencing salaries across different roles within the data field on a global scale

The dataset at the core of this analysis contains diverse information, ranging from job titles and job categories to geographical and temporal contexts. Each data point is carefully curated to facilitate a nuanced exploration of the intricacies that contribute to salary variations in the data industry.

About dataset

- **ID:** A unique identifier assigned to each employee
- **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 analysing 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 analysing 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 **.**

Tech-Stack Used

- Dataset: Jobs and Salaries in Data Science (kaggle.com)
- Creating SQL Database
- Testing SQL Database
- Executing SQL query
- Perform Analysis
- Version – Postgresql v16.1

Questionnaire:

1. Find the average salary for each job category.
2. Identify the top 5 countries with the highest average salary.
3. Retrieve the job title with the highest salary in each country.
4. Identify the top 3 job categories with the highest total salary in USD.
5. Find the top 3 company location with the highest average salary for remote jobs.
6. Display a list of all job category and the total number of employees in each category in USA.
7. Find the highest salary job in each company size in United States.
8. Find the job title with the highest salary in each experience level category.
9. Find employee id who work in Data Analyst job and have a higher salary than average salary in different countries.
10. Identify the job category with the highest average salary-to-experience ratio in India.
11. List any 5 pairs of employees who have the same job category with a salary difference of less than $5,000.

Questionnaire:

12. What is the average salary in USD for each job title within the 'Data Analysis' job category, and how does it compare to the overall average salary in USD across all job titles?
13. Calculate the year-over-year percentage change in the average salary for each job category.
14. Determine the median salary for Data engineering within each country.
15. List the job categories with the highest and lowest salary standard deviation in UK.

1. Find the average salary for each job category

Input SQL Query

	job_category character varying (255)	average_salary numeric	average_salary_in_usd numeric
1	Data Quality and Operations	100554.36	100879.47
2	Data Management and Strategy	103085.95	103139.93
3	Data Analysis	107801.47	108505.72
4	BI and Visualization	135935.84	135092.10
5	Leadership and Management	144164.41	145476.02
6	Data Engineering	145588.37	146197.66
7	Cloud and Database	155000.00	155000.00
8	Data Architecture and Modeling	155343.89	156002.36
9	Data Science and Research	163401.93	163758.58
10	Machine Learning and AI	179338.91	178925.85

Output SQL Query

```
SELECT
      job_category,
       round(avg(salary),2) as average_salary,
       round(avg(salary_in_usd),2) as average_salary_in_usd
FROM
     job
group by
     job_category
order by
     average_salary,average_salary_in_usd ;
```

2. Identify the top 5 countries with the highest average salary.

Input SQL Query

```
SELECT
      company_location, salary_currency,
      round(avg(salary),2) as avg_salary ,
      round(avg(salary_in_usd),2) as avg_salary_in_usd
FROM
      job
group by
       company_location, salary_currency
order by
       avg_salary,avg_salary_in_usd
limit 5
```

Output SQL Query

	company_location character varying (255)	salary_currency character varying (3)	avg_salary numeric	avg_salary_in_usd numeric
1	Thailand	USD	15000.00	15000.00
2	Indonesia	USD	15000.00	15000.00
3	Russia	EUR	15662.00	16455.00
4	Ecuador	USD	16000.00	16000.00
5	Moldova	USD	18000.00	18000.00

3. Retrieve the job title with the highest salary in each country.

Input SQL Query

```
With t1 as (
            SELECT job_title,salary,company_location,
            ROW_NUMBER() OVER(partition by company_location order by salary DESC) as rank
            FROM job)
SELECT
       job_title,
       salary,
       company_location as country
FROM  t1
WHERE rank=1
LIMIT 10;
```

Output SQL Query

	job_title character varying (255)	salary numeric (10,2)	country character varying (255)
1	Data Scientist	100000.00	Algeria
2	Business Data Analyst	50000.00	American Samoa
3	Data Scientist	47000.00	Andorra
4	Data Engineer	80000.00	Argentina
5	Machine Learning Engineer	50000.00	Armenia
6	ML Engineer	300000.00	Australia
7	Data Scientist Lead	85000.00	Austria
8	Autonomous Vehicle Technician	45555.00	Bahamas
9	Data Scientist	80000.00	Belgium
10	AI Developer	120000.00	Bosnia and Herzegovina

4. **Identify the top 3 job categories with the highest total salary in USD.**

Input SQL Query

```
SELECT
	job_category,
	SUM(salary_in_usd) as Total_salary_usd
FROM
	job
GROUP BY
	job_category
ORDER BY
	Total_salary_usd desc
LIMIT 3;
```

Output SQL Query

	job_category character varying (255)	total_salary_usd numeric
1	Data Science and Research	493568348.00
2	Data Engineering	330406703.00
3	Machine Learning and AI	255506110.00

5. **Find the top 3 company location with the highest average salary for remote jobs.**

Input SQL Query

```
SELECT
	company_location,round(avg(salary),2) as avarage_salary
from
	job
where
	work_setting='Remote'
group by
	company_location
order by
	avarage_salary
limit 3
```

Output SQL Query

	company_location character varying (255)	avarage_salary numeric
1	Thailand	15000.00
2	Ghana	15000.00
3	Ecuador	16000.00

6. Display a list of all job category and the total number of employees in each category in USA.

Input SQL Query

```sql
SELECT
    job_category,
    count(*) as Total_no_employee
FROM
      job
WHERE
     company_location='United States'
Group by
     job_category
ORDER BY
     total_no_employee DESC
```

Output SQL Query

	job_category character varying (255)	total_no_employee bigint
1	Data Science and Research	2635
2	Data Engineering	1977
3	Data Analysis	1252
4	Machine Learning and AI	1190
5	Leadership and Management	442
6	BI and Visualization	288
7	Data Architecture and Modeling	237
8	Data Management and Strategy	56
9	Data Quality and Operations	50
10	Cloud and Database	5

7. Find the highest salary job in each company size in United States.

Input SQL Query

```
WITH t1 AS (
    SELECT job_title,salary_in_usd,
       case
       when company_size ='L' then 'Large'
       when company_size='M'  then 'Medium'
       else 'Small'
       END AS company_size
FROM job
WHERE company_location='United States'),
     t2 as (
           select job_title,company_size,salary_in_usd,
           DENSE_RANK() over(partition by company_size order by salary_in_usd) as rank
           from t1
           )
SELECT company_size,job_title,salary_in_usd
from t2
where rank=1
```

Output SQL Query

	company_size text	job_title character varying (255)	salary_in_usd numeric (10,2)
1	Large	Data Engineer	20000.00
2	Large	Applied Scientist	20000.00
3	Medium	Data Analyst	20000.00
4	Small	ML Engineer	25500.00

8. Find the job title with the highest salary in each experience level category.

Input SQL Query

```
with t1 as (
          SELECT job_title,salary_in_usd,experience_level,
          dense_rank() over( partition by experience_level order by salary_in_usd desc) as rank
          FROM job)
SELECT
      experience_level,job_title, salary_in_usd
from   t1
where rank=1
```

Output SQL Query

	experience_level character varying (20)	job_title character varying (255)	salary_in_usd numeric (10,2)
1	Entry-level	Applied Scientist	281700.00
2	Executive	Principal Data Scientist	416000.00
3	Mid-level	Research Scientist	450000.00
4	Senior	Data Scientist	412000.00

9. **Find employee id who work in Data Analyst job and have a higher salary than average salary in different countries.**

Input SQL Query

```
with t1 as (
           SELECT * FROM job
           WHERE job_title='Data Analyst'),
     t2 as (
          SELECT id as employee_id,salary,company_location,
          dense_rank()over(partition by company_location order by salary) as rank
          FROM t1
           WHERE salary > (select avg(salary) from t1))
SELECT
      employee_id,salary,company_location
FROM t2
WHERE rank = 1
```

Output SQL Query

	employee_id integer	salary numeric (10,2)	company_location character varying (255)
1	4820	142500.00	Australia
2	2245	112000.00	Canada
3	208	110000.00	Poland
4	5443	110000.00	United Kingdom
5	4035	110000.00	United Kingdom
6	8952	109280.00	United States
7	8756	109280.00	United States

10. Identify the job category with the highest average salary-to-experience ratio in India.

Input SQL Query

```
With t1 as(
      SELECT
          job_category,
          round(avg(salary_in_usd)/COUNT(DISTINCT experience_level ),2) AS avg_salary_to_experience_ratio
       FROM
          Job
       WHERE
          Company_location='India'
       Group By
          job_category )
SELECT
      job_category, avg_salary_to_experience_ratio
FROM t1
ORDER BY avg_salary_to_experience_ratio DESC;
```

Output SQL Query

	job_category character varying (255)	avg_salary_to_experience_ratio numeric
1	Machine Learning and AI	45000.00
2	Data Science and Research	29000.00
3	Data Analysis	24080.00
4	Data Engineering	18000.00

11. List any 5 pairs of employees who have the same job category with a salary difference of less than $5,000.

Input SQL Query

```
SELECT DISTINCT(j1.id,j2.id) as pairs_of_employee,
      j1.id as employee1_id,
      j1.job_category,
      j1.salary_in_usd as salary1,
      j2.id as employee2_id,
      j2.salary_in_usd as salary2
FROM
    job j1
JOIN
    job j2
ON
   j1.job_category=j2.job_category
   AND j1.id < j2.id
   AND ABS(j1.salary_in_usd-j2.salary_in_usd) < 5000
LIMIT 5
```

Output SQL Query

	pairs_of_employee record	employee1_id integer	job_category character varying (255)	salary1 numeric (10,2)	employee2_id integer	salary2 numeric (10,2)
1	(1,39)	1	Data Engineering	95012.00	39	94500.00
2	(1,141)	1	Data Engineering	95012.00	141	99000.00
3	(1,145)	1	Data Engineering	95012.00	145	99000.00
4	(1,249)	1	Data Engineering	95012.00	249	100000.00
5	(1,281)	1	Data Engineering	95012.00	281	94500.00

12. What is the average salary in USD for each job title within the 'Data Analysis' job category, and how does it compare to the overall average salary in USD across all job titles?

Input SQL Query

```sql
WITH DataAnalysisAvg AS (
    SELECT
        job_title,
        round(AVG(salary_in_usd),2) AS avg_salary_data_analysis
    FROM
        job
    WHERE
        job_category='Data Analysis'
    GROUP BY
        job_title),

OverallAvg AS (
    SELECT
        round(AVG(salary_in_usd),2) AS avg_salary_overall
    FROM
        job)
SELECT
    d.job_title,
    d.avg_salary_data_analysis,
    o.avg_salary_overall,
    round(d.avg_salary_data_analysis - o.avg_salary_overall,2) AS salary_difference
FROM
    DataAnalysisAvg d
CROSS JOIN
    OverallAvg o;
```

Output SQL Query

	job_title character varying (255)	avg_salary_data_analysis numeric	avg_salary_overall numeric	salary_difference numeric
1	Marketing Data Analyst	144327.00	150299.50	-5972.50
2	Financial Data Analyst	90375.00	150299.50	-59924.50
3	Insight Analyst	47673.88	150299.50	-102625.62
4	Compliance Data Analyst	45000.00	150299.50	-105299.50
5	BI Data Analyst	71074.50	150299.50	-79225.00
6	Sales Data Analyst	60000.00	150299.50	-90299.50
7	Product Data Analyst	80040.00	150299.50	-70259.50
8	Business Data Analyst	78600.22	150299.50	-71699.28
9	Data Analyst	109911.47	150299.50	-40388.03
10	Staff Data Analyst	79917.00	150299.50	-70382.50
11	Finance Data Analyst	141933.67	150299.50	-8365.83
12	Principal Data Analyst	122500.00	150299.50	-27799.50
13	Lead Data Analyst	108333.33	150299.50	-41966.17
14	Business Intelligence Data Analyst	83209.50	150299.50	-67090.00

13. Calculate the year-over-year percentage change in the average salary for each job category.

Input SQL Query

```sql
WITH t1 AS (
    SELECT
        work_year,
        job_category,
        round(AVG(salary_in_usd),2) AS avg_salary_in_usd,
        LAG(AVG(salary_in_usd)) OVER (PARTITION BY job_category ORDER BY work_year) AS prev_avg_salary
    FROM
        Job
    GROUP BY
        work_year, job_category
)
SELECT
    work_year,
    job_category,
    avg_salary_in_usd,
    CASE
        WHEN prev_avg_salary IS NULL THEN 0
        ELSE round(((avg_salary_in_usd - prev_avg_salary) / prev_avg_salary) * 100,2)
    END AS percentage_change
FROM
    t1
ORDER BY
    job_category, work_year
LIMIT 10;
```

Output SQL Query

	work_year integer	job_category character varying (255)	avg_salary_in_usd numeric	percentage_change numeric
1	2022	BI and Visualization	100875.00	0
2	2023	BI and Visualization	135989.60	34.81
3	2022	Cloud and Database	175000.00	0
4	2023	Cloud and Database	141666.67	-19.05
5	2020	Data Analysis	68014.67	0
6	2021	Data Analysis	87818.50	29.12
7	2022	Data Analysis	107384.22	22.28
8	2023	Data Analysis	109848.58	2.29
9	2021	Data Architecture and Modeling	169940.60	0
10	2022	Data Architecture and Modeling	166091.54	-2.26

14. Determine the median salary for Data engineering within each country.

Input SQL Query

```sql
SELECT
    job_category,
    employee_residence AS country,
    PERCENTILE_CONT(0.5) WITHIN GROUP (ORDER BY salary_in_usd) AS median_salary
FROM
    Job
WHERE
    job_category='Data Engineering'
GROUP BY
    job_category, country
ORDER BY
    job_category, country
LIMIT 10;
```

Output SQL Query

	job_category character varying (255)	country character varying (255)	median_salary double precision
1	Data Engineering	Argentina	66000
2	Data Engineering	Australia	75050
3	Data Engineering	Austria	74130
4	Data Engineering	Brazil	72000
5	Data Engineering	Canada	145000
6	Data Engineering	Colombia	80000
7	Data Engineering	Estonia	50529
8	Data Engineering	France	67640
9	Data Engineering	Germany	86374
10	Data Engineering	Greece	63040
11	Data Engineering	Hong Kong	66022
12	Data Engineering	India	18000
13	Data Engineering	Ireland	102569
14	Data Engineering	Italy	62433
15	Data Engineering	Lithuania	94812
16	Data Engineering	Malta	28369
17	Data Engineering	Mauritius	100000
18	Data Engineering	Mexico	66000

15. List the job categories with the highest and lowest salary standard deviation in UK.

Input SQL Query

```sql
(SELECT
    job_category,
    round(STDDEV(salary_in_usd),2) AS salary_std_dev
FROM
    Job
WHERE
     company_location = 'United Kingdom'
GROUP BY
    job_category
ORDER BY
    salary_std_dev DESC, job_category
LIMIT 1)

UNION ALL

(SELECT
    job_category,
    round(STDDEV(salary_in_usd),2) AS salary_std_dev
FROM
    Job
WHERE
     company_location = 'United Kingdom'
GROUP BY
    job_category
ORDER BY
    salary_std_dev ASC, job_category
LIMIT 1);
```

Output SQL Query

	job_category character varying (255)	salary_std_dev numeric
1	Leadership and Management	96759.00
2	BI and Visualization	24817.64