**Project Plan for Data Jobs Salary**

1. Scoping the project

Define business challenges

Develop hypotheses

Define data metrics

Propose solution

Mock up solution

Propose pilot design to measure impact

Problem statement and create dashboard

1. Acquire Data

Identify data source and gain access

**Scope - plan the project**

**Design - create a blueprint of the project**

**Prepare - clean the data**

**Test - run data validation checks on the data**

**Visualize - create graphs and charts with the data**

**Analyse - explore the insights from the data**

**Recommend - provide options to the user and back them up with data**

**Dirty data:**

1. **Duplicate records**
2. **Inconstant format**
3. **Missing values on nulls**
4. **Inaccurate information**

I’m struggling to come up with a data project out if this data set. Help me simulate a problem statement and create a dashboard out it.

**Scoping the project**

‘*Our organisation lacks a comprehensive understanding of salary disparities and the factors influencing them, such as job title, experience level, remote work, geographical location, and company size. We need an interactive dashboard to visualise these trends and provide actionable insights for making informed salary-related decisions.’*

This dataset offers comprehensive insights into data salaries and employment attributes from 2020 to 2024. It includes key information such as salary, job title, experience level, employment type, employee residence, remote work ratio, company location, and company size. Using this dataset, I aim to explore the following questions:

1. **Analysing Salary Trends by Job Title and Experience Level**

* Investigate how salaries differ across job titles and experience levels.
* Identify which job titles and experience levels have the highest and lowest average salaries.

-- Summary Statistics for Salary by Job Title

WITH RankedSalaries AS (

SELECT

job\_title,

salary\_in\_usd,

ROW\_NUMBER() OVER (PARTITION BY job\_title ORDER BY salary\_in\_usd) AS rn,

COUNT(\*) OVER (PARTITION BY job\_title) AS cnt

FROM

data\_salaries

)

SELECT

job\_title,

COUNT(job\_title),

ROUND(AVG(salary\_in\_usd), 2) AS median\_salary,

MIN(salary\_in\_usd) AS min\_salary,

MAX(salary\_in\_usd) AS max\_salary

FROM

RankedSalaries

WHERE

-- Select the middle value(s) depending on odd/even count of salaries

rn IN ((cnt + 1) / 2, (cnt + 2) / 2)

GROUP BY

job\_title

ORDER BY

median\_salary DESC

LIMIT 10;

From the results, we observe that the analytical engineering manager has the highest median salary at $399,880.00, which also represents both the minimum and maximum salary for this role. This could suggest that this position might be relatively new in the data field, potentially resulting in a uniform salary due to limited data. Typically, emerging fields or specialised roles attract higher salaries due to a lack of expertise. Additionally, managerial roles like this one are often held by one or two individuals who occupy positions of authority and responsibility, involving leadership, organisation, and oversight of a group or project within a business which may count for fewer individuals in the survey.

The next highest salary is for a data science tech lead, with a median of $375,000.00, followed by the head of machine learning at $330,000.00. The top 10 roles are predominantly managerial, leadership, or architectural positions, which is logical given that these roles require greater responsibility, advanced technical skills and a deeper understanding of the field.

On the other hand, the lowest median salary is for the principal data architect, at $38,154.00, which is surprisingly low for a senior level position. This difference could be due to the role location possibly being outside the U.S., where salary standards may differ. Following this, the CRM data analyst role has a median salary of $40,000.00. The bottom 10 roles are primarily analyst positions, which suggests offers lower salaries compared to other data related roles.

-- Summary Statistics for Salary by Experience Level

WITH RankedSalaries AS (

SELECT

experience\_level,

salary\_in\_usd,

ROW\_NUMBER() OVER (PARTITION BY experience\_level ORDER BY salary\_in\_usd) AS rn,

COUNT(\*) OVER (PARTITION BY experience\_level) AS cnt

FROM

data\_salaries

)

SELECT

experience\_level,

ROUND(AVG(salary\_in\_usd), 2) AS median\_salary,

MIN(salary\_in\_usd) AS min\_salary,

MAX(salary\_in\_usd) AS max\_salary

FROM

RankedSalaries

WHERE

-- Select the middle value(s) depending on odd/even count of salaries

rn IN ((cnt + 1) / 2, (cnt + 2) / 2)

GROUP BY

experience\_level

ORDER BY

median\_salary DESC;

Examining salaries by experience level, executives (EX) earn the highest median salary at $192000.00. This is then followed by seniors (SE), with a median salary of $154380.00. Mid-level professionals (MI) earn a median salary of $112000.00, while entry-level positions (EN) have the lowest median salary at $80769.00. This progression aligns with expectations, as salaries typically increase with experience and career advancement.

1. **Impact of Remote Work on Salaries**

* Examine the relationship between remote work ratio and salary.
* Determine if employees working remotely tend to have higher or lower salaries compared to those who work on-site.

WITH RankedSalaries AS (

SELECT

remote\_ratio,

salary\_in\_usd,

ROW\_NUMBER() OVER (PARTITION BY remote\_ratio ORDER BY salary\_in\_usd) AS rn,

COUNT(\*) OVER (PARTITION BY remote\_ratio) AS cnt

FROM

data\_salaries

)

SELECT

remote\_ratio,

ROUND(AVG(salary\_in\_usd), 2) AS median\_salary,

MIN(salary\_in\_usd) AS min\_salary,

MAX(salary\_in\_usd) AS max\_salary

FROM

RankedSalaries

WHERE

-- Select the middle value(s) depending on odd/even count of salaries

rn IN ((cnt + 1) / 2, (cnt + 2) / 2)

GROUP BY

remote\_ratio

ORDER BY

median\_salary DESC;

Office workers had the highest median salary at $ 140100.00 likely due to the traditionally higher pay associated with on-site roles, which often include more senior or specialised positions. Remote workers earned a median of $137785.50, reflecting competitive pay in flexible work arrangements that still offer substantial compensation. Hybrid workers, who split their time between the office and remote work, had the lowest median salary at $66022.00. This lower average might be attributed to hybrid roles potentially encompassing a wider range of positions, including those with less specialisation or fewer responsibilities compared to full-time office roles.

1. **Geographical Salary Analysis**

* Compare salaries across different employee residences and company locations.
* Identify regions with the highest and lowest salaries.

WITH salary\_stats AS (

SELECT

employee\_residence,

salary\_in\_usd,

COUNT(\*) OVER (PARTITION BY employee\_residence) AS num\_employees,

ROUND(AVG(salary\_in\_usd) OVER (PARTITION BY employee\_residence), 2) AS avg\_salary,

MIN(salary\_in\_usd) OVER (PARTITION BY employee\_residence) AS min\_salary,

MAX(salary\_in\_usd) OVER (PARTITION BY employee\_residence) AS max\_salary,

ROW\_NUMBER() OVER (PARTITION BY employee\_residence ORDER BY salary\_in\_usd) AS rn,

COUNT(\*) OVER (PARTITION BY employee\_residence) AS total\_count

FROM data\_salaries

WHERE employee\_residence = company\_location

),

median\_calc AS (

SELECT

employee\_residence,

num\_employees,

avg\_salary,

min\_salary,

max\_salary,

ROUND(

CASE

WHEN total\_count % 2 = 0 THEN

(SELECT AVG(salary\_in\_usd)

FROM salary\_stats

WHERE rn IN (total\_count / 2, total\_count / 2 + 1)

AND employee\_residence = s.employee\_residence)

ELSE

(SELECT salary\_in\_usd

FROM salary\_stats

WHERE rn = (total\_count + 1) / 2

AND employee\_residence = s.employee\_residence)

END, 2

) AS median\_salary

FROM salary\_stats s

GROUP BY employee\_residence, num\_employees, avg\_salary, min\_salary, max\_salary, total\_count

),

top\_10\_residences AS (

SELECT

employee\_residence,

num\_employees,

avg\_salary,

min\_salary,

max\_salary,

median\_salary

FROM median\_calc

ORDER BY num\_employees DESC

LIMIT 10

)

SELECT

employee\_residence,

num\_employees,

avg\_salary,

min\_salary,

max\_salary,

median\_salary

FROM top\_10\_residences

ORDER BY median\_salary DESC;

This analysis was challenging due to the presence of extreme outliers. Initially, I found that Qatar had the highest salary by country, but this result was based on only one participant, making it unreliable to conclude that Qatar truly had the highest salary. To address this, I focused on the 10 countries with the most participants in the dataset. I then calculated the median salary for each of these countries and ranked them in descending order.

From this refined analysis, several key insights emerged:

U.S. Dominates in Median Salary: At the top of the list is the U.S., with a median salary of $147,500.00. This suggests that the U.S. remains a highly lucrative market for professionals, likely due to the strong demand for skilled workers, the presence of global tech companies, and the higher cost of living in major cities.

Canada as a Competitive Market: Canada follows closely with a median salary of $137,665.00, indicating that it is also a competitive market. This may be driven by its growing tech sector, favorable immigration policies for skilled workers, and a high standard of living that attracts talent.

Australia and Germany on Par: Australia has a median salary of $105,600.00, while Germany's median salary is $79,425.00. This difference suggests that while both countries offer attractive compensation, the market dynamics might vary. Australia's rapidly growing tech sector may command higher salaries, while Germany's strong economy and focus on engineering provide solid opportunities, though with relatively lower median pay.

Impact of Outliers and Data Distribution: The initial finding that Qatar had the highest salary underscores the importance of considering sample size in data analysis. With only one participant from Qatar, it became clear that outliers can distort perceptions. By focusing on countries with more participants, a more accurate and reliable picture of the salary landscape emerged.

Global Disparities in Pay: The significant differences in median salaries across countries highlight the global disparities in pay, even for similar roles. Factors such as cost of living, the strength of local economies, demand for specific skills, and tax policies likely contribute to these disparities.

Economic and Industry Implications: Countries with higher median salaries, like the U.S. and Canada, may attract more global talent, reinforcing their competitive advantage in industries like technology and finance. Conversely, countries with lower median salaries might face challenges in retaining top talent unless they offer other incentives, such as quality of life, work-life balance, or career growth opportunities.

1. **Company Size and Salary Correlation**

* Analyse how company size affects salaries.
* Determine if larger companies pay more than smaller companies.

WITH RankedSalaries AS (

SELECT

company\_size,

salary\_in\_usd,

ROW\_NUMBER() OVER (PARTITION BY company\_size ORDER BY salary\_in\_usd) AS rn,

COUNT(\*) OVER (PARTITION BY company\_size) AS cnt

FROM

data\_salaries

)

SELECT

company\_size,

ROUND(AVG(salary\_in\_usd), 2) AS median\_salary,

MIN(salary\_in\_usd) AS min\_salary,

MAX(salary\_in\_usd) AS max\_salary

FROM

RankedSalaries

WHERE

-- Select the middle value(s) depending on odd/even count of salaries

rn IN ((cnt + 1) / 2, (cnt + 2) / 2)

GROUP BY

company\_size

ORDER BY

median\_salary DESC;

The data shows that medium sized companies offer the highest median salary of $140250.00, possibly due to a balance between resources and flexibility, allowing them to attract and retain top talent. Large companies follow with an average salary of $108000.00, benefiting from economies of scale but often having more standardised pay structures and rigid work environment. Small companies, while offering more intimate work environments and potential for rapid growth, provide a lower average salary of $72000.00, reflecting their typically limited financial resources.

1. **Employment type**

* Analyse how company size affects salaries.

WITH RankedSalaries AS (

SELECT

employment\_type,

salary\_in\_usd,

ROW\_NUMBER() OVER (PARTITION BY employment\_type ORDER BY salary\_in\_usd) AS rn,

COUNT(\*) OVER (PARTITION BY employment\_type) AS cnt

FROM

data\_salaries

)

SELECT

employment\_type,

ROUND(AVG(salary\_in\_usd), 2) AS median\_salary,

MIN(salary\_in\_usd) AS min\_salary,

MAX(salary\_in\_usd) AS max\_salary

FROM

RankedSalaries

WHERE

-- Select the middle value(s) depending on odd/even count of salaries

rn IN ((cnt + 1) / 2, (cnt + 2) / 2)

GROUP BY

employment\_type

ORDER BY

median\_salary DESC;

The data shows that full-time employees earn the highest median salary at $139152.00

, likely due to the stability and benefits associated with full-time positions. Contract workers follow with a median of $93856.00, reflecting higher pay for specialised skills but often without the same benefits. Part-time employees earn a median of $66451.50, as they typically work fewer hours and may not receive full benefits. Freelancers have the lowest median earnings at $47777.50, likely due to the variability of work and the absence of long-term contracts. It is important to note that in Europe, most contract workers out earn full-time employees. This is because healthcare, a significant benefit in the U.S, is either free or low-cost in Europe, which likely influences the data.

**Data Metrics**

1. **Average Salary:**
   * By job title
   * By experience level
   * By employment type
   * By remote work ratio
   * By geographical location (employee residence and company location)
   * By company size
2. **Salary Distribution:**
   * Across different job titles and experience levels
3. **Remote Work Impact:**
   * Comparison of average salaries based on remote work ratio
4. **Geographical Salary Trends:**
   * Mapping average salaries by employee residence and company location

**Mock-Up Solution**

1. **Main Dashboard Components**:
   * **Job Title and Experience Level Analysis**:
     + Bar charts or box plots for average salaries by job title and experience level.
   * **Remote Work Impact**:
     + Bar chart comparing average salaries based on remote work ratio (0%, 50%, 100%).
   * **Geographical Analysis**:
     + Map visualisation showing average salaries by employee residence and company location.
   * **Company Size Impact**:
     + Bar chart showing average salaries across different company sizes (small, medium, large).
2. **Interactive Features**:
   * Filters for job title, experience level, employment type, and company size.
   * Drill-down capabilities to view detailed salary distributions.
3. Mastery by Robert Greene   
   2. Do Hard Things by Alex Harris and Brett Harris   
   3. Thinking, Fast and Slow by Daniel Kahneman

Why I chose median rather than mean:

**Mean (Average):**

* **Sensitive to Outliers**: The mean is affected by extremely high or low salaries. If your dataset contains outliers (e.g., a few very high salaries that are much higher than the rest), the mean might be skewed and not representative of the typical salary for a job title.
* **Use Case**: The mean is useful when you want to consider the total sum of salaries and get an overall average, particularly if your data is normally distributed without significant outliers.

**Median:**

* **Resistant to Outliers**: The median represents the middle value of your data when sorted, meaning it is not affected by outliers. It gives a better sense of what a typical employee earns within a job title, especially in datasets with significant salary disparities.
* **Use Case**: The median is more appropriate when you want to understand the typical salary for a job title and your data has outliers or is skewed.

**Which to Use for Salary by Job Title:**

* **If your salary data contains outliers or is skewed** (e.g., a few employees with significantly higher or lower salaries), it's better to use the **median** to represent the central tendency more accurately.
* **If your salary data is relatively normally distributed** without extreme outliers, the **mean** can be used, and it will likely provide a similar result to the median.

Given that salary data often contains outliers (e.g., high executive salaries or very low entry-level positions), using the **median** is generally recommended for a more accurate reflection of typical salaries by job title.