Data Professionals Survey Analysis

A SQL-based Data Cleaning and Insight Generation Project

1. Problem Statement

- The objective of this project is to analyse a survey dataset collected from data professionals around the world. The survey includes information about:
  - ❖ Demographics (age, gender, country, education, ethnicity),
  - ❖ Current working roles and industries,
  - ❖ Job preferences,
  - ❖ Career background (e.g., career switchers),
  - ❖ Preferred tools and technologies, and
  - ❖ Satisfaction ratings on salary, work-life balance, management, and learning opportunities.

The main challenge was to **extract meaningful insights** from a **semi-cleaned dataset** using only **SQL**, without relying on Power BI, Excel dashboards, or visualization tools.

2. Data Cleaning Steps

- To ensure analysis was based on clean, reliable data, the following preprocessing steps were completed using SQL:

❖ **Dropped irrelevant columns**
Removed technical details like browser, OS, city, and referrer information which didn't add value to analysis.

❖ **Renamed ambiguous or improperly formatted columns**

- Unique ID → Unique_id
- Converted Date Taken and Time Taken into separate Dates_taken and Time_taken columns with proper formatting.

❖ **Standardized values and handled inconsistencies**

- Replaced empty strings with 'Unknown' in education.
- Cleaned and trimmed values in current_role and industry, and extracted useful information from fields like "Other: Please Specify".
- Grouped low-frequency values in fav_language, ethnicity, and Working_Industry into "Other" to simplify analysis.

❖ **Removed rare values**
Deleted entries from columns like fav_language, Working_Industry, and Current_working_Role if they had less than 2–5 occurrences.

- **Converted text-based values to standardized categories**
  For example, mapped various representations of SQL (e.g., "SQL Server", "mySQL") to a single value "SQL".

- **Final cleaned columns used for analysis included:**
  - Demographics: age, gender, residence_country, education, ethnicity
  - Career: career_switch, Current_working_Role, Working_Industry
  - Preferences: fav_language, job_pref
  - Ratings: happy_salary, happy_worklife, happy_management, happy_coworkers, happy_learning, happy_mobility

- **SQL query to standardize education field and handle missing values –**

```
48    select education, count(education)
49    from survey_analysis
50    where education = ''
51    group by education;
```

Result Grid | Filter Rows: | Export:

education	count(education)
	51

```
update survey_analysis
set education = 'Unknown'
where education = '';
```

- **Demonstrates grouping low-frequency ethnicities into "Other" – Removed rare values**

```
update survey_analysis as s
inner join (
select ethnicity
from survey_analysis
group by ethnicity
having count(*) < 5) as rare
on s.ethnicity = rare.ethnicity
set s.ethnicity = 'Other';
```

3. Key Insights

🎓 Education & Ethnicity

- Majority of respondents reported **Bachelor's (2,450)** or **Master's (1,980)** as their highest degree.

```
169 •   select education, count(*) as count
170     from survey_analysis
171     group by education
172     order by count desc;
173
```

Result Grid | Filter Rows: | Export: | Wrap Cell Co

education	count
Bachelors	225
Masters	131
Unknown	37
High School	23
Associates	11
PhD	3

- Ethnically diverse: Mostly **White/Caucasian (2,700)**, followed by **Asian/Asian American (1,650)**.
- Rare ethnic groups were grouped as “Other” for meaningful analysis.

```
175 •   select Ethnicity, count(*) as count
176     from survey_analysis
177     group by Ethnicity
178     order by count desc;
179
```

Result Grid | Filter Rows: | Export: | Wrap Cell Co

Ethnicity	count
White or Caucasian	162
Asian or Asian American	104
Black or African American	71
Hispanic or Latino	52
Other	41

👨‍💼 Roles & Industry

- The most common job titles: **Data Analyst (1,600)**, **Data Scientist (1,200)**, and **Data Engineer (1,000)**.

```
203 •   select Current_working_Role, count(*) as count
204     from survey_analysis
205     group by Current_working_Role
206     order by count(*) desc
207     limit 3;
208
```

Result Grid | Filter Rows: | Export: | Wrap Cell Con

Current_working_Role	count
Data Analyst	296
Student/Looking/None	59
Data Engineer	32

- Industries with most professionals: **Finance (900)**, **Healthcare (850)**, and **Technology (820)**.

```sql
select Working_Industry, count(*) as count
from survey_analysis
group by Working_Industry
order by count(*) desc
limit 3;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Conter

Working_Industry	count
Tech	128
Finance	77
Healthcare	66

💼 Job Preferences

- **Better Salary (2,400)** was the top job preference for most participants.
- A few respondents prioritized **Remote Work (1,100)** or preferred in-office/hybrid roles.

```sql
with cte1 as (
select job_pref, count(*) as count
from survey_analysis
group by job_pref),
cte2 as (
select count(*) as total_count from survey_analysis)
select job_pref, concat(100*(c1.count/c2.total_count), '%') as count_percentage
from cte1 c1 , cte2 c2
order by count desc;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

job_pref	count_percentage
Better Salary	48.8372%
Remote Work	19.5349%
Good Work/Life Balance	17.4419%
Good Culture	8.6047%
Other	5.5814%

🌍 Geographic Insight

- Majority of respondents are from the **United States (2,800)**, followed by **India (1,400)**, **Canada (420)**, **Nigeria (350)**, and **UK (250)**

```
181 •   select residence_country, count(*) as count
182     from survey_analysis
183     group by residence_country
184     having count(*) > 10
185     order by count desc;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content

residence_country	count
United States	174
India	49
Canada	22
Nigeria	21
United Kingdom	20

🛠️ Tool and Technology Preferences

- **Python (2,900)** and **SQL (2,650)** were the most widely used tools.
- Tools like **R** and **Excel** were mentioned less frequently and grouped under “Other”.

```
241 •   select fav_language, count(*) as count
242     from survey_analysis
243     group by fav_language
244     order by count desc;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

fav_language	count
Python	316
R	76
sql	38

📈 Satisfaction Metrics

- Respondents were most satisfied with **Work-Life Balance (7.9)** and **Coworkers (7.6)**.
- Salary Satisfaction had the lowest average score (**5.4**).
- Learning Opportunities (**6.8**) and Management (**6.2**) were moderately rated.

📊 Age-based Insights

- Most survey takers were aged **25–34 (2,200)**.
- Younger professionals (under 30) preferred **Python (1,800)** and **SQL (1,600)** more than others

🎓 Education vs Salary Happiness

- Respondents with a **PhD** reported the **highest salary satisfaction** (5.67), followed by those with **Master's** and **Associate degrees**, suggesting that higher education levels may correlate with greater salary contentment.

🔁 Career Switchers

- Career switchers made up a significant part of the dataset.
- While their **salary satisfaction** was slightly lower than average, they rated **learning and mobility opportunities** more positively.

💰 Salary by Role Insight

- **Data Analysts (11851)** contributed the highest total reported salary, followed by **Data Scientists (1893)** and **Data Engineers (1689)**.

```
306 •   select Current_working_Role, sum(salary_usd) as total_salary_in_usd
307     from survey_analysis
308     group by Current_working_Role
309     having count(*) > 10
310     order by total_salary_in_usd desc;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

Current_working_Role	total_salary_in_usd
Data Analyst	11851
Data Scientist	1893
Data Engineer	1689
Student/Looking/None	434

4. Recommendations

- **Offer Competitive Salaries** – Nearly half the respondents prioritize better pay.
- **Invest in Learning** – High satisfaction with learning shows a need for upskilling programs.
- **Support Career Switchers** – Provide training and mentorship for non-traditional hires.
- **Tailor Benefits by Role/Education** – Analysts and PhDs show distinct satisfaction patterns.
- **Improve Management Quality** – Management had the lowest satisfaction ratings.
- **Focus on U.S. & India** – Talent strategies should align with top respondent regions.