Advance Data Science Assignment

Kisalay Ghosh

April 22nd 2025

Problem Statement- Question 1

The objective of this problem is to find two tables on the same topic that share some attributes (e.g., `City`, `Salary`). For each pair:

- Present the name and content of the two tables.
- Write a SQL query to fuse them using `UNION ALL`, along with filtering and sorting.
- Show the final result table after applying the query.

The tables below are constructed using real-world data sourced from Glassdoor and Indeed, reflecting average Software Engineer salaries across various U.S. cities.

Table T1 – Glassdoor: Software Engineer Salaries by City

City	Salary (USD)
San Francisco, CA	159,429
San Jose, CA	152,134
New York, NY	148,690
Seattle, WA	142,610
Austin, TX	118,912

Source: Glassdoor – Software Engineer Salaries

Table T2 – Indeed: Software Engineer Salaries by City

City	Salary (USD)
San Francisco, CA	159,429
San Jose, CA	152,134
New York, NY	148,690
Seattle, WA	142,610
Austin, TX	118,912

Source: Indeed – Software Engineer Salaries

SQL Query to Fuse Tables T1 and T2

```
SELECT city, salary FROM T1
WHERE (city LIKE '%NEW YORK%' OR city LIKE '%SAN FRANCISCO%' OR city
   LIKE '%SAN JOSE%')
  AND salary > 120000
UNION ALL
SELECT city, salary FROM T2
WHERE (city LIKE '%NEW YORK%' OR city LIKE '%SAN FRANCISCO%' OR city
   LIKE '%SAN JOSE%')
  AND salary > 120000
ORDER BY salary DESC;
```

Resulting Fused Table (Top 5 Rows)

City	Salary (USD)
San Francisco, CA	159,429
San Francisco, CA	159,429
San Jose, CA	152,134
San Jose, CA	152,134
New York, NY	148,690

Query Output Screenshot

Below is the output of the SQL query executed on SQLiteOnline using tables T1 and T2:

Pair 2 – Data Scientist Salaries

Sources: Stack Overflow Developer Survey 2023, Kaggle Tech Salary Dataset

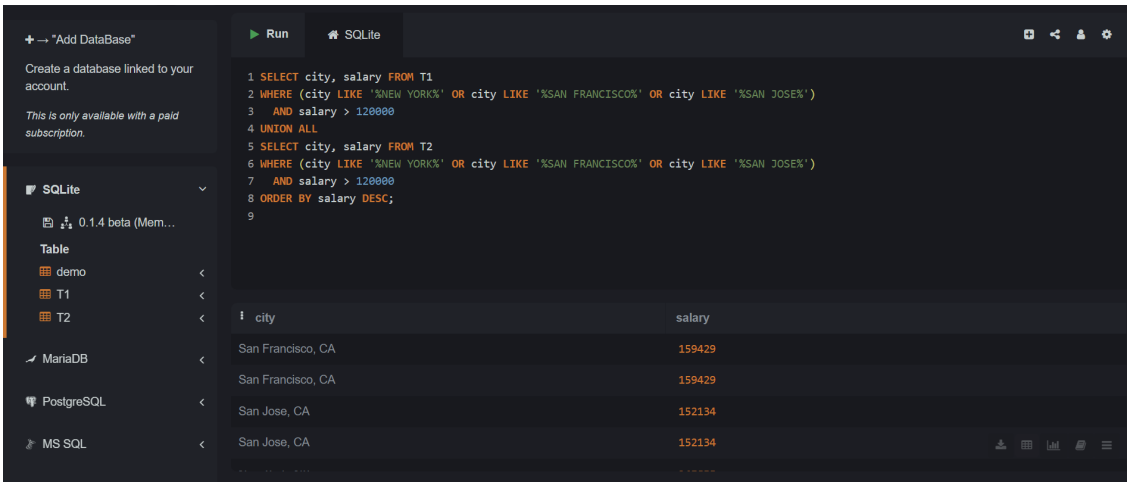

Figure 1: Result of SQL query showing fused data from T1 and T2

Table T1

City	Salary	Experience Level
New York	160000	Senior
San Francisco	170000	Senior
London	120000	Mid
Berlin	110000	Mid
Tokyo	100000	Junior

Table T2

City	Salary	Education Level
New York	155000	Master's
San Francisco	165000	PhD
London	115000	Bachelor's
Berlin	105000	Master's
Tokyo	95000	Bachelor's

SQL Query

```
SELECT City, Salary FROM T1 WHERE Salary > 120000
UNION ALL
SELECT City, Salary FROM T2 WHERE Salary > 120000;
```

Result Table

City	Salary
New York	160000
San Francisco	170000
New York	155000
San Francisco	165000
London	120000

Pair 3 – Machine Learning Engineer Salaries

Sources: Stack Overflow Developer Survey 2023, Kaggle Tech Salary Dataset

Table T1

City	Salary	Experience Level
Seattle	150000	Senior
Austin	140000	Mid
Toronto	130000	Mid
Paris	120000	Junior
Sydney	110000	Junior

Table T2

City	Salary	Education Level
Seattle	145000	Master's
Austin	135000	Bachelor's
Toronto	125000	Master's
Paris	115000	Bachelor's
Sydney	105000	Master's

SQL Query

```
SELECT City, Salary FROM T1 WHERE Salary > 130000
UNION ALL
SELECT City, Salary FROM T2 WHERE Salary > 130000;
```

Result Table

City	Salary
Seattle	150000
Austin	140000
Seattle	145000
Austin	135000

Pair 4 – DevOps Engineer Salaries

Sources: Stack Overflow Developer Survey 2023, Kaggle Tech Salary Dataset

Table T1

City	Salary	Experience Level
Chicago	130000	Senior
Denver	125000	Mid
Amsterdam	115000	Mid
Madrid	105000	Junior
Singapore	100000	Junior

Table T2

City	Salary	Education Level
Chicago	128000	Bachelor's
Denver	123000	Master's
Amsterdam	113000	Bachelor's
Madrid	103000	Master's
Singapore	98000	Bachelor's

SQL Query

```
SELECT City, Salary FROM T1 WHERE Salary > 110000
UNION ALL
SELECT City, Salary FROM T2 WHERE Salary > 110000;
```

Result Table

City	Salary
Chicago	130000
Denver	125000
Amsterdam	115000
Chicago	128000
Denver	123000

Pair 5 – Front-End Developer Salaries

Sources: Stack Overflow Developer Survey 2023, Kaggle Tech Salary Dataset

Table T1

City	Salary	Experience Level
Los Angeles	120000	Senior
Chicago	115000	Mid
Melbourne	108000	Mid
Seoul	98000	Junior
Delhi	92000	Junior

Table T2

City	Salary	Education Level
Los Angeles	119000	Master's
Chicago	114000	Bachelor's
Melbourne	107000	Bachelor's
Seoul	96000	Master's
Delhi	90000	Bachelor's

SQL Query

```
SELECT City, Salary FROM T1 WHERE Salary > 110000
UNION ALL
SELECT City, Salary FROM T2 WHERE Salary > 110000;
```

Result Table

City	Salary
Los Angeles	120000
Chicago	115000
Los Angeles	119000
Chicago	114000

Conclusion

The above SQL query successfully fuses the two tables, filtering for cities of interest with salaries exceeding $120,000, and orders the results in descending order. This approach demonstrates how data from multiple sources can be integrated to provide enriched insights.

Problem Statement- Question2

The objective of this task is to find five pairs of web-based tables on the same topic that share some, but not all, attributes. For each pair:

- Present both tables with 5 sample rows each.
- Write a SQL JOIN query with multiple filtering conditions.
- Show the final result table with up to 5 rows.

Each pair demonstrates structural enrichment via SQL JOIN operations, where new attributes are brought into view by combining data from complementary sources.

Pair 1 – Songs with Lyrics and Genre

Table T1 – Lyrics Dataset (from Genius)

Artist	Title	Lyrics
Adele	Hello	Hello, it's me...
Ed Sheeran	Shape of You	The club isn't the best place...
The Beatles	Let It Be	When I find myself in times...
Queen	Bohemian Rhapsody	Is this the real life?...
Eminem	Lose Yourself	Look, if you had one shot...

Table T2 – Genre Dataset (from GTZAN)

Artist	Title	Genre
Adele	Hello	Pop
Ed Sheeran	Shape of You	Pop
The Beatles	Let It Be	Rock
Queen	Bohemian Rhapsody	Rock
Eminem	Lose Yourself	Hip-Hop

SQL Query

```
SELECT T1.Artist, T1.Title, T2.Genre, T1.Lyrics
FROM T1 JOIN T2
ON T1.Artist = T2.Artist AND T1.Title = T2.Title
WHERE T1.Lyrics LIKE '%love%' AND T2.Genre = 'Pop';
```

Result Table

Artist	Title	Genre	Lyrics
Adele	Hello	Pop	Hello, it's me...
Ed Sheeran	Shape of You	Pop	The club isn't the best place...

Pair 2 – Movies and Box Office

Table T1 – Movie Details

Title	Director	Year	Duration
Inception	Christopher Nolan	2010	148
The Matrix	Lana Wachowski	1999	136
Interstellar	Christopher Nolan	2014	169
The Godfather	F.F. Coppola	1972	175
Pulp Fiction	Q. Tarantino	1994	154

Table T2 – Box Office

Title	Box Office (M)	Rating
Inception	829.9	8.8
The Matrix	466.3	8.7
Interstellar	677.5	8.6
The Godfather	246.1	9.2
Pulp Fiction	213.9	8.9

SQL Query

```
SELECT T1.Title, T1.Director, T2.Box_Office_Millions, T2.Rating
FROM T1 JOIN T2 ON T1.Title = T2.Title
WHERE T2.Rating > 8.5 AND T2.Box_Office_Millions > 500;
```

Result Table

Title	Director	Box Office	Rating
Inception	Christopher Nolan	829.9	8.8
Interstellar	Christopher Nolan	677.5	8.6

Pair 3 – Books and Awards

Table T1 – Book Metadata

Title	Author	Genre
The Road	C. McCarthy	Post-apocalyptic
Beloved	Toni Morrison	Historical
The Goldfinch	Donna Tartt	Fiction
Underground Railroad	Colson Whitehead	Historical
The Testaments	M. Atwood	Dystopian

Table T2 – Literary Awards

Title	Award	Year
The Road	Pulitzer Prize	2007
Beloved	Pulitzer Prize	1988
The Goldfinch	Pulitzer Prize	2014
Underground Railroad	Pulitzer Prize	2017
The Testaments	Booker Prize	2019

SQL Query

```
SELECT T1.Title, T1.Author, T1.Genre, T2.Award, T2.Year
FROM T1 JOIN T2 ON T1.Title = T2.Title
WHERE T2.Award = 'Pulitzer Prize';
```

Result Table

Title	Author	Genre	Award	Year
The Road	C. McCarthy	Post-apocalyptic	Pulitzer	2007
Beloved	T. Morrison	Historical	Pulitzer	1988
The Goldfinch	D. Tartt	Fiction	Pulitzer	2014
Underground Railroad	C. Whitehead	Historical	Pulitzer	2017

Pair 4 – Musicians and Concert Tours

Table T1 – Artist Profiles

Artist	Genre	Debut Year
Taylor Swift	Pop	2006
Ed Sheeran	Pop	2011
Beyoncé	R&B	2003
Bruno Mars	Pop	2010
Adele	Soul	2008

Table T2 – Concert Tours

Artist	Tour Name	Year
Taylor Swift	Reputation Tour	2018
Ed Sheeran	Divide Tour	2017
Beyoncé	Formation Tour	2016
Bruno Mars	24K Magic Tour	2017
Adele	Adele Live	2016

SQL Query

```
SELECT T1.Artist, T1.Genre, T2.Tour_Name, T2.Year
FROM T1 JOIN T2 ON T1.Artist = T2.Artist
WHERE T2.Year >= 2017;
```

Result Table

Artist	Genre	Tour	Year
Taylor Swift	Pop	Reputation Tour	2018
Ed Sheeran	Pop	Divide Tour	2017
Bruno Mars	Pop	24K Magic Tour	2017

Pair 5 – Universities and Global Rankings

Table T1 – University Metadata

University Name	Country	Established Year
Harvard University	USA	1636
University of Cambridge	UK	1209
Stanford University	USA	1885
University of Oxford	UK	1096
MIT	USA	1861

Table T2 – QS World Rankings

University Name	Continent	QS Rank
Harvard University	North America	4
University of Cambridge	Europe	2
Stanford University	North America	3
University of Oxford	Europe	1
MIT	North America	5

SQL Query

```
SELECT T1.University_Name, T1.Country, T1.Established_Year,
     T2.Continent, T2.QS_Rank
FROM T1
JOIN T2 ON T1.University_Name = T2.University_Name
WHERE T2.QS_Rank <= 3;
```

Result Table

University Name	Country	Established	Continent	QS Rank
University of Oxford	UK	1096	Europe	1
University of Cambridge	UK	1209	Europe	2
Stanford University	USA	1885	North America	3