

Cleaning Data in SQL Server

Cleaning data fetched from various datasets.



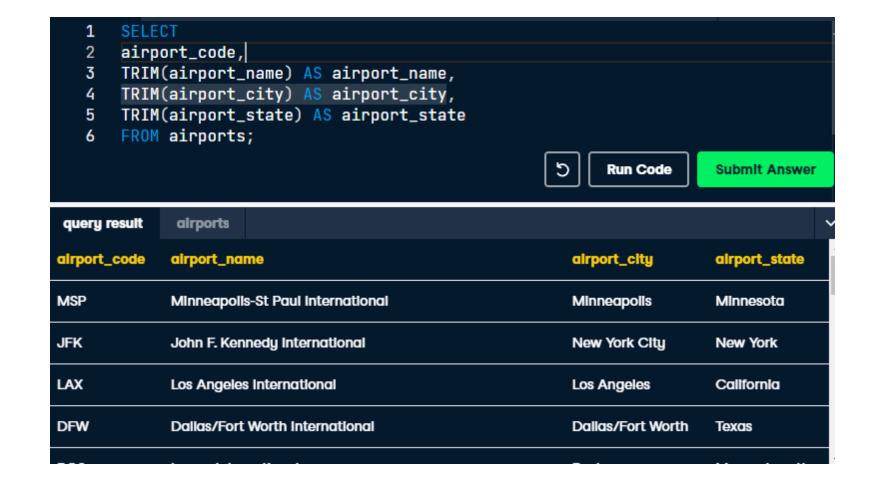
Unifying flight name formats using Replicate()

 You need to get every register with more than 100 delays from the flight_statistics table. In a unique column, you have to concatenate the carrier_code, registration_code, and airport_code, having a similar format to this one: "AA -000000119, JFK".

```
query.sql
       SELECT
       Concat(carrier_code,
       Replicate('0',9-len(registration_code)),
       registration_code,
       airport_code) AS registration_code
       FROM flight_statistics
  10
       WHERE delayed>100;
              flight_statistics
query result
registration_code
AA - 000000119, JFK
B6 - 000000120, JFK
AA - 000000127, JFK
```

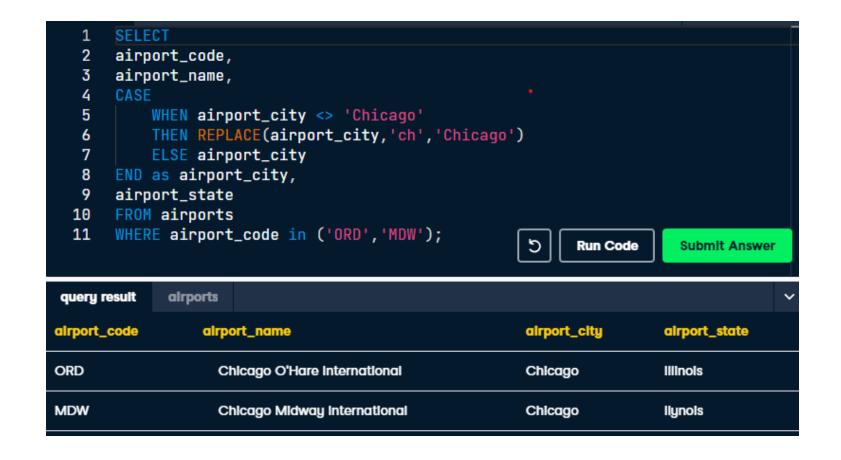
Trim() strings

 Examine the content of the airports table, and use the appropriate function to remove the leading and trailing spaces.



Unifying name formats with Case(), Replace()

 There are inconsistent values for 'Chicago' in the airport_city column, with values such as 'ch'. You will treat these inconsistent values by replacing them.



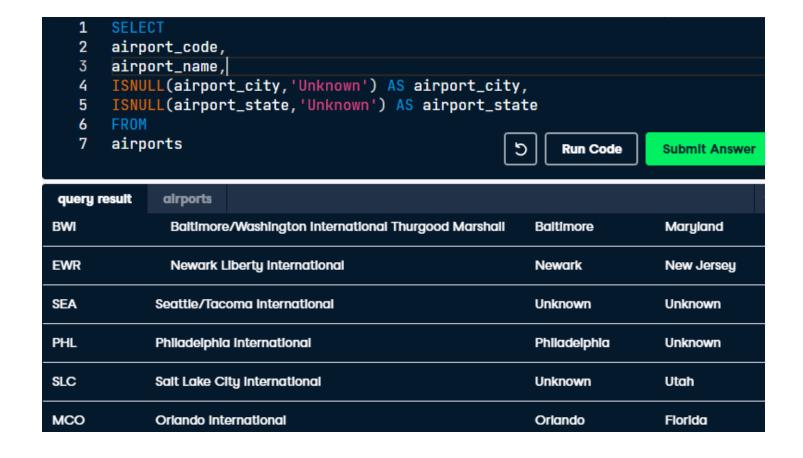
Comparing names with SOUNDEX()

• Some statistician_name and statistician_surname are written in a different way, such as Miriam Smith and Myriam Smyth. You think about comparing with SOUNDEX() the names of the statisticians. If the result of SOUNDEX() is the same, but the texts you are comparing are different, you will find the data you need to clean.

13	SELE	CT DISTINCT		
14	S1.statistician_name,S1.statistician_surname			
15	FROM flight_statistics S1			
16		INNER JOIN fli	ght_statistics S2	
17		ON SOUNDEX	(S1.statistician_name)=SOUNDEX(S2.statistician_name)	
18		AND SOUNDE	X(S1.statistician_surname)= <mark>SOUNDEX(</mark> S2.	
	statistician_surname)			
19	<pre>WHERE (S1.statistician_name <> S2.statistician_name)</pre>			
20		<pre>OR (S1.statistician_surname <> S2.statistician_surname);</pre>		
			Submit Answer	
query result		flight_statistics		
Brian			Page	
Bryan			Page	
Miriam			Smith	
Miriam			Smyth	

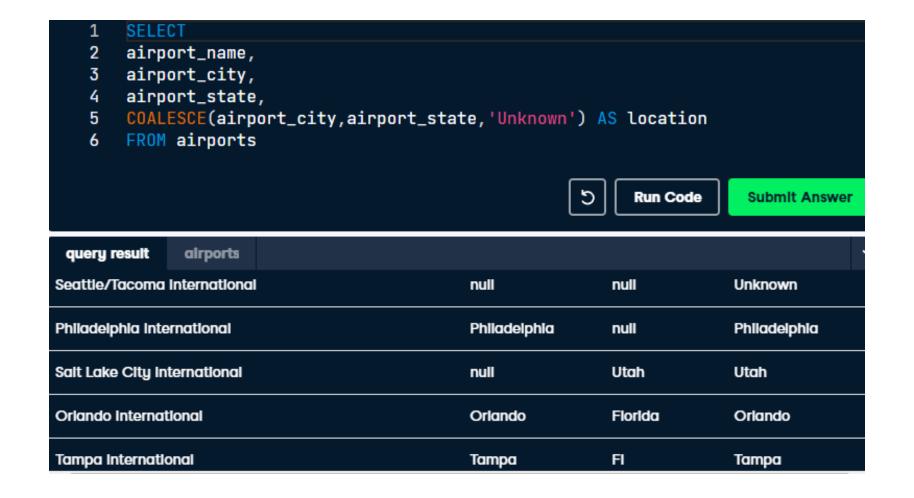
Filling missing values using ISNULL()

 Replace the missing values for airport_city and airport_state column with the 'Unknown' string.



Filling missing values using COALESCE()

 Now, you want to create a new column, location, that returns the values of the airport_city column, and in case it has NULL values, return the value of airport_state. Finally, if airport_state is also NULL, you want to return the string 'Unknown'.



Treating duplicates with Row_number()

• Get all the rows without duplicates. You consider that the repeating group for this table is formed by the columns airport_code, carrier_code, and registration_date.



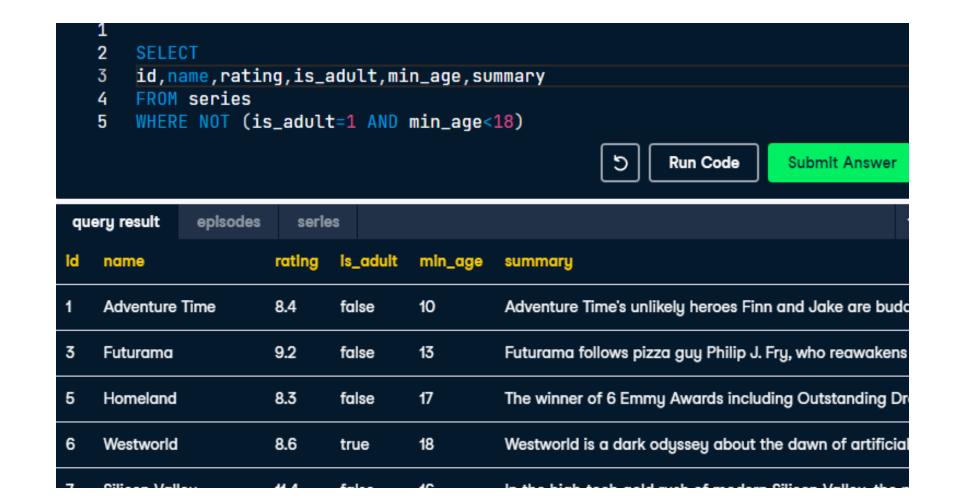
Convert() Dates Format

 The format of the registration_date column is yyyy-mm-dd, and you want to show the results in the format of mm/dd/yyyy, which is hardcoded as 101, using the CONVERT() function. Notice that the type of the registration_date column is VARCHAR(10) and not a date.



Excluding inaccurate data from a TV Series dataset

• In a TV series dataset there are contradictory values, some rows with a TRUE value in its is_adult column have a number smaller than 18 in its min_age column. Can you find these rows with inaccurate data?



Converting Data Types Using CAST() or CONVERT()

- Tables could store data with different types than you want. Sometimes you will need to convert these types to the correct ones to perform the operations you want.
- The num_ratings stores integer numbers, but this time it was designed as VARCHAR(5).



The TV series with most episodes

- In the episodes table, there is a column named number. It stores the number of each episode within a season for every series. This column was designed as VARCHAR(5), but it actually stores numbers.
- Can you guess which is the series with most episodes within a season?



Find Wrong URLs using Like()

 Prepare a script that checks every official_site value from the series table to analyze possible wrong URLs (not starting with www.).



Checking Phone Numbers using Like()

 prepare a script that checks every contact_number value from the series table to get those numbers which are not of the format 555-xxxxxxx.



Concatenating cities and states with Concat()

 Concatenate the names of the cities with the states using the CONCAT() function, while using a CASE statement that returns blank when state is NULL and performs a normal concatenation otherwise.



Split Column with Substring() and Charindex()

 You need to split this city_state column into two new columns, one for the city and the other one for the state.



Combine Different Parts of a Date: DATEFROMPARTS()

• Combine the columns, year_of_sale, month_of_sale, and day_of_sale, in the paper_shop_daily_sales table.



Turning rows into columns: Pivot()

 The structure of the table paper_shop_monthly_sales is not appropriate for the report. You want to generate a report with this appearance:

```
|year_of_sale|notebooks|pencils|crayons|
|------|
| 2018 | 150 | 150 | 80 |
| 2019 | 230 | 130 | 170 |
```

```
SELECT
       year_of_sale, notebooks, pencils, crayons
       FROM
       (SELECT
       SUBSTRING(product_name_units,1,charindex('-',product_name_units)-1)
   6
                    AS product_name,
       CAST(SUBSTRING(product_name_units, charindex('-', product_name_units)+1,
                        len(product_name_units)) AS INT) AS units,
   8
       year_of_sale
       FROM paper_shop_monthly_sales) sales
       PIVOT( SUM(units) FOR product_name IN (notebooks,pencils,crayons) )
  12
             AS paper_shop_pivot
                                                                          Submit Ans
                                                              Run Code
query result
             paper_shop_monthly_sales
year_of_sale
                            notebooks
                                                   pencils
                                                                    crayons
2018
                            150
                                                   150
                                                                    80
2019
                            230
                                                   130
                                                                    170
```

Turning columns into rows: Unpivot()

 Suppose you stored the result from the previous exercise in a new table called pivot_sales, and now you want to turn the columns notebooks, pencils, and crayons into row values.

