



Edoardo Cavaliere

Analisi dei dati con SQL

Descrizione Progetto

In un laboratorio chimico è stato introdotto un nuovo macchinario per eseguire gli esperimenti su determinate molecole.

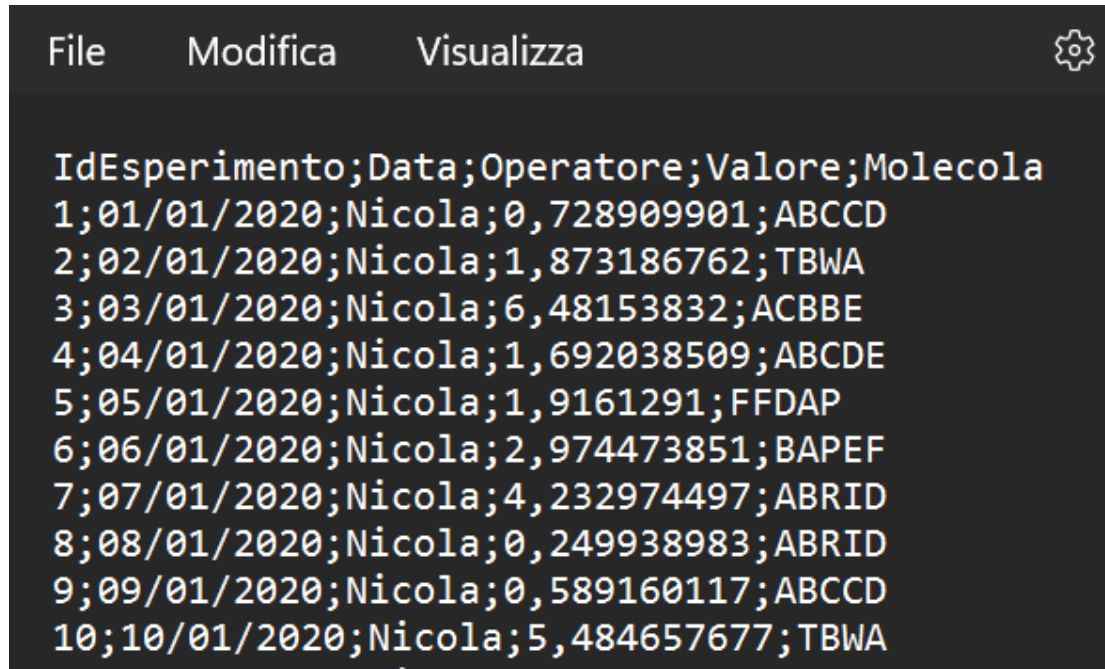
Sono interessate dal macchinario le molecole:

- il cui nome inizia con AB e finisce con D;
- il cui nome inizia con F e NON finisce con P.

Il macchinario è stato introdotto in data 1 maggio 2020.

Lo scopo dell'analisi è quello di analizzare per ogni operatore come è variato il **valore** ottenuto degli esperimenti su tali molecole, prima e dopo la data di cambio del macchinario.

Analisi del file di input



The screenshot shows a CSV file viewer with a menu bar (File, Modifica, Visualizza) and a settings icon. The data is as follows:

IdEsperimento	Data	Operatore	Valore	Molecola
1	01/01/2020	Nicola	0,728909901	ABCCD
2	02/01/2020	Nicola	1,873186762	TBWA
3	03/01/2020	Nicola	6,48153832	ACBBE
4	04/01/2020	Nicola	1,692038509	ABCDE
5	05/01/2020	Nicola	1,9161291	FFDAP
6	06/01/2020	Nicola	2,974473851	BAPEF
7	07/01/2020	Nicola	4,232974497	ABRID
8	08/01/2020	Nicola	0,249938983	ABRID
9	09/01/2020	Nicola	0,589160117	ABCCD
10	10/01/2020	Nicola	5,484657677	TBWA

- File con estensione CSV
- Carattere delimitatore ;
- Presenza di riga con l'intestazione
- Date in formato DD/MM/YYYY
- Numero con carattere , come separatore dei decimali
- 5 colonne, 321 righe + intestazione

Import dei dati in SQL Server – pt.1

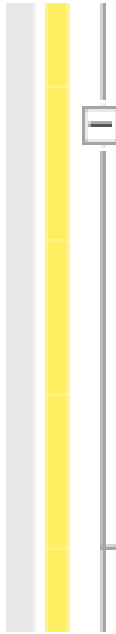
Carichiamo preliminarmente i dati in una tabella di Staging senza vincoli.

```
= CREATE TABLE dbo.EsperimentiStaging(  
  IdEsperimento VARCHAR (255),  
  Data VARCHAR (255),  
  Operatore VARCHAR (255),  
  Valore VARCHAR (255),  
  Molecola VARCHAR (255));
```

```
= BULK INSERT dbo.EsperimentiStaging  
  FROM 'Progetto_Esperimenti.csv'  
  WITH (  
    FIRSTROW = 2,  
    FIELDTERMINATOR = ';',  
    ROWTERMINATOR = '\n');
```

Import dei dati in SQL Server – pt.2

Trasferiamo i dati nella tabella target, con colonne tipizzate, vincoli NOT NULL e chiave primaria (PRIMARY KEY).

A vertical bar with a yellow highlight and a small icon at the top, resembling a code editor's line indicator.

```
CREATE TABLE dbo.Esperimenti (  
    IdEsperimento INT NOT NULL PRIMARY KEY,  
    Data DATE NOT NULL,  
    Operatore VARCHAR (255) NOT NULL,  
    Valore DECIMAL (18,10) NOT NULL,  
    Molecola VARCHAR (255) NOT NULL);
```

Import dei dati in SQL Server – pt.3

Utilizziamo le funzioni CAST, CONCAT, LEFT, SUBSTRING, RIGHT e REPLACE per trasformare i dati di input e inserirli nella tabella target.

```
INSERT INTO dbo.Esperimenti (IdEsperimento,
Data, Operatore, Valore, Molecola)
SELECT CAST (IdEsperimento AS INT) AS IdEsperimento,
        CAST(CONCAT(RIGHT(Data,4),
        '-',
        SUBSTRING(Data, 4,2),
        '-',
        LEFT(Data, 2)) AS DATE) AS Data,
        Operatore,
        CAST(REPLACE(REPLACE(Valore, '.', ''), ',', '.')) AS DECIMAL(18,10)),
        Molecola
FROM dbo.EsperimentiStaging;
```

Scrittura della query SQL

- eseguiamo il filtro sulle molecole con l'operatore LIKE

- raggruppiamo i dati per operatore

- utilizziamo la CASE WHEN all'interno della funzione AVG per calcolare le medie ristrette ai dati precedenti il 1 maggio 2020 e successivi al 1 maggio 2020



- combiniamo le colonne per calcolare la differenza e lo scostamento percentuale

- utilizziamo una CTE per render il codice più leggibile

```
with DatiPerOperatore AS (  
    SELECT Operatore,  
           AVG(CASE WHEN Data < '20200501' THEN Valore ELSE NULL END) AS MediaPre,  
           AVG(CASE WHEN Data >= '20200501' THEN Valore ELSE NULL END) AS MediaPost  
    FROM   dbo.Esperimenti  
    WHERE  (Molecola LIKE 'AB%'  
            AND Molecola LIKE '%D')  
           or  
           (Molecola LIKE 'F%'  
            AND Molecola NOT LIKE '%P')  
    )  
    GROUP BY Operatore)  
SELECT Operatore,  
       MediaPre,  
       MediaPost,  
       MediaPost-MediaPre as Differenza,  
       CASE WHEN MediaPre = 0 THEN NULL ELSE ( MediaPost-MediaPre)/MediaPre END AS ScostamentoPercentuale  
FROM   DatiPerOperatore;
```

Analisi dei risultati

L'output della query mostra che per tutti e tre gli operatori si è registrato un incremento nel valore degli esperimenti dal 1 maggio 2020 in poi. Per Alberto e Nicola l'incremento è stato del 41,9% e 46,5%, mentre per Giovanni del 10,6%

<div><div> Results</div><div> Messages</div></div>					
	Operatore	MediaPre	MediaPost	Differenza	ScostamentoPercentuale
1	Alberto	2.0342603905	2.8868315072	0.8525711167	0.419106
2	Giovanni	2.7079465561	2.9942716615	0.2863251054	0.105735
3	Nicola	2.1117036017	3.0945281702	0.9828245685	0.465417