ANALISI PRE/POST MACCHINARIO: SQL & MONGODB

DESCRIZIONE PROGETTO

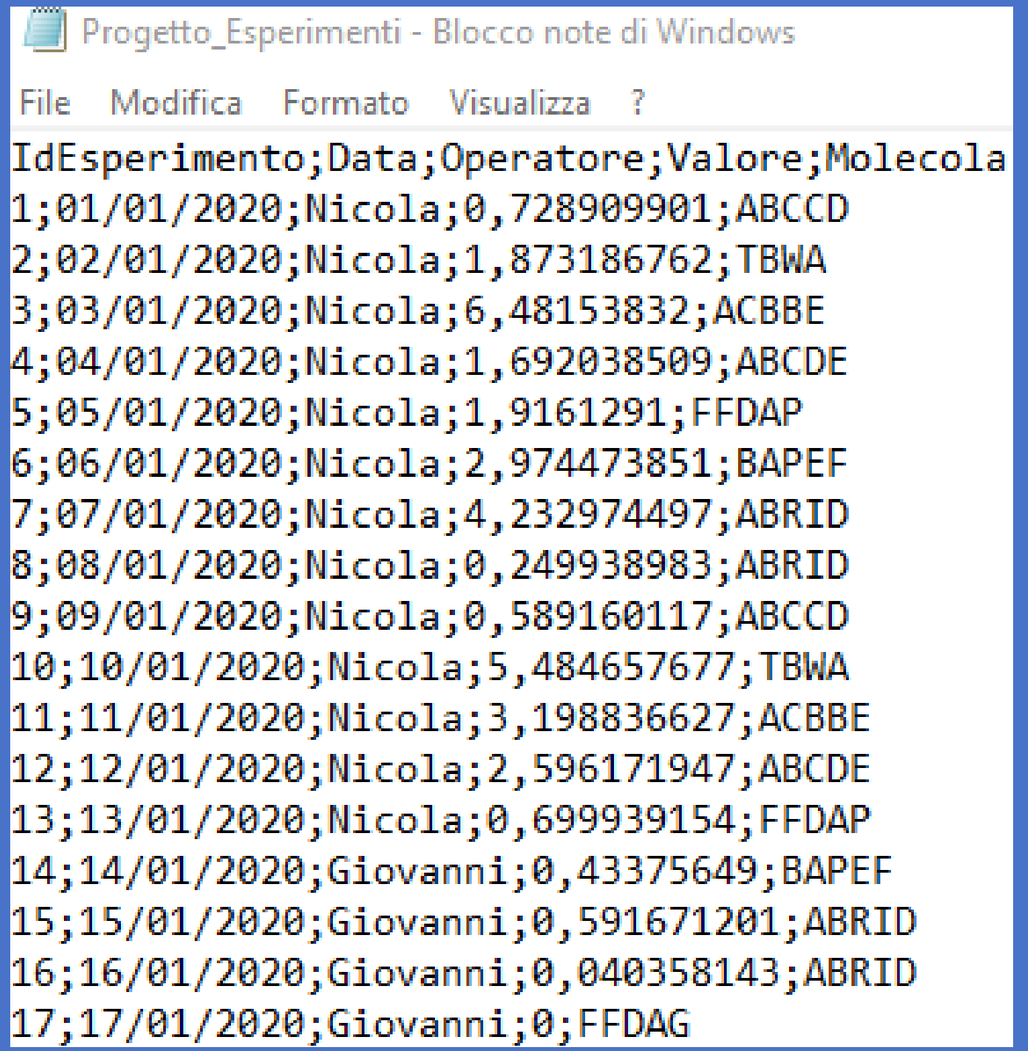
In un laboratorio chimico è stato introdotto un nuovo macchinario il 1° maggio 2020, che coinvolge le molecole:

- con nome che inizia per “AB” e finisce con “D”
- con nome che inizia per “F” e non finisce con “P”.

Obiettivo: analizzare come è cambiato il valore medio degli esperimenti per ciascun operatore prima e dopo l’introduzione del macchinario.

DETTAGLI FILE DI INPUT

- 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



DATI DI INPUT

SQL

- Dati strutturati in una tabella relazionale
- Ogni riga rappresenta un esperimento (ID esperimento, data, operatore, valore, molecola)

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

PREPARAZIONE DATI

```sql
ALTER TABLE dbo.progetto_esperimenti
ADD data_converted DATE;

UPDATE dbo.progetto_esperimenti
SET data_converted =
CONVERT(DATE, Data, 103);
```

Dopo l'importazione, ho convertito la colonna Data (formato stringa) in un formato DATE standard, creando la nuova colonna data_converted, necessaria per l'analisi temporale.

Il formato 103 è stato scelto per convertire le date da DD/MM/YYYY, permettendo una valutazione accurata tramite ordinamento, filtro e raggruppamento dei dati.

STUDIO COMPARATO

Ho elaborato una query SQL per confrontare, per ogni operatore, la media dei valori degli esperimenti su specifiche molecole, prima e dopo l'introduzione del macchinario.

Utilizzando CTE, ho creato due sottoinsiemi (prima e dopo), ripulito i dati da formati errati (es. 1.234,56), e calcolato le differenze assolute e percentuali tra le medie.

```
with datiprima as
(
select Operatore,AVG(CAST(REPLACE(REPLACE(Valore,'.',''),',','.')
AS DECIMAL(18,10))) as mediaprima
from dbo.progetto_esperimenti
where data_converted<'2020-05-01' AND   ((molecola LIKE 'AB%'
        AND molecola LIKE '%D')
        or
        (molecola LIKE 'F%'
        AND molecola NOT LIKE '%P'
        ))
group by Operatore),
datidopo as
(
select Operatore,AVG(CAST(REPLACE(REPLACE(Valore,'.',''),',','.')
AS DECIMAL(18,10))) as mediadopo
from dbo.progetto_esperimenti
where data_converted>='2020-05-01'  and   ((molecola LIKE 'AB%'
        AND molecola LIKE '%D')
        or
        (molecola LIKE 'F%'
        AND molecola NOT LIKE '%P'
        ))
group by Operatore)

SELECT
    dp.Operatore,
    dp.mediaprima,
    dd.mediadopo,
    (dd.mediadopo - dp.mediaprima) AS differenzaassoluta,
    ((dd.mediadopo - dp.mediaprima)/dp.mediaprima) as scostamento
FROM datiprima dp
INNER JOIN datidopo dd ON dp.operatore = dd.Operatore;
```

DATI DI INPUT MONGODB

- Dati strutturati come documenti JSON
- Ogni documento rappresenta un esperimento con attributi chiave (ID, data, operatore, valore, molecola)

```
{
  _id: ObjectId('685bde1dc41e71aba011dcbe'),
  IdEsperimento: 1,
  Data: '01/01/2020',
  Operatore: 'Nicola',
  Valore: '0,728909901',
  Molecola: 'ABCCD',
  molecola_prime_due: 'AB',
  molecola_primo: 'A',
  lunghezza_molecola: 5
}
{
  _id: ObjectId('685bde1dc41e71aba011dcbf'),
  IdEsperimento: 2,
  Data: '02/01/2020',
  Operatore: 'Nicola',
  Valore: '1,873186762',
```

ESTRAZIONE DATI

- Estrazione dei primi/ultimi caratteri da Molecola
- Parsing della data (gg/mm/aaaa → Date)
- Conversione di Valore: stringa → decimale
- Filtraggio su condizioni molecola (AB…D o F…(≠P))

```
db.esperimenti.aggregate([
  {$addFields: {
    primo: {$substrCP: ["$Molecola", 0, 1]},
    primi_due: {$substrCP: ["$Molecola", 0, 2]},
    ultimo: {$substrCP: ["$Molecola",
    {$subtract: [{$strLenCP: "$Molecola"}, 1]}, 1]},
    data_formattata: {$concat: [
      {$substrCP: ["$Data", 6, 4]}, "-",
      {$substrCP: ["$Data", 3, 2]}, "-",
      {$substrCP: ["$Data", 0, 2]]
    ]},
    valore: {$toDecimal: {$replaceAll: {input: "$Valore", find: ",", replacement: "."}}}
  }},
  {$match: {
    $or: [
      {primi_due: "AB", ultimo: "D"},
      {primo: "F", ultimo: {$ne: "P"}}
    ]
  }},
  {$project: {
    id_esperimento: "$IdEsperimento",
    data: {$toDate: "$data_formattata"},
    operatore: 1,
    valore: 1,
    molecola: 1
  }}
]);
```

PRE/POST

1 MAGGIO 2020

Aggiunta campi:
valore_pre (data < 01/05/2020),
valore_post (data ≥ 01/05/2020)

Raggruppamento:
media pre e post per operatore

Calcoli finali:
differenza e scostamento
percentuale tra media post e pre

```
db.esperimenti_transformed.aggregate([
  {
    $addFields: {
      valore_pre: {$cond: [{$lt: ["$data", ISODate("2020-05-01")]}, "$valore", null]},
      valore_post: {$cond: [{$gte: ["$data", ISODate("2020-05-01")]}, "$valore", null]}
    }
  },
  {
    $group: {
      _id: "$operatore",
      media_pre: {$avg: "$valore_pre"},
      media_post: {$avg: "$valore_post"}
    }
  },
  {
    $addFields: {
      differenza_post_pre: {$subtract: ["$media_post", "$media_pre"]},
      scostamento_percentuale: {
        $cond: [
          {$eq: ["$media_pre", 0]},
          null,
          {$divide: [{$subtract: ["$media_post", "$media_pre"]}, "$media_pre"]}
        ]
      }
    }
  }
]);
```

ANALISI DEI RISULTATI

L’output 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%

	Operatore	mediaprima	mediadopo	differenzaassoluta	scostamento
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


```
{
  mediaprima: 2.111703601727273,
  mediadopo: 3.0945281702727274,
  differenzaassoluta: 0.9828245685454546,
  scostamento: 0.46541785870969343,
  Operatore: 'Nicola'
}
{
  mediaprima: 2.707946556133334,
  mediadopo: 2.9942716615,
  differenzaassoluta: 0.2863251053666662,
  scostamento: 0.1057351389443626,
  Operatore: 'Giovanni'
}
{
  mediaprima: 2.0342603905384617,
  mediadopo: 2.886831507205128,
  differenzaassoluta: 0.8525711166666663,
  scostamento: 0.41910618750286616,
  Operatore: 'Alberto'
}
```