Question #1:

Given:

Database: Oracle

Tablename: table1

Columns: attribute (varchar2(100)), timestamp (date), value (number)

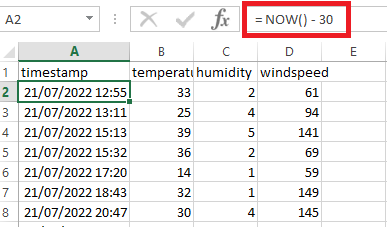
Example of attribute: temperature, humidity, wind speed etc.

Question:

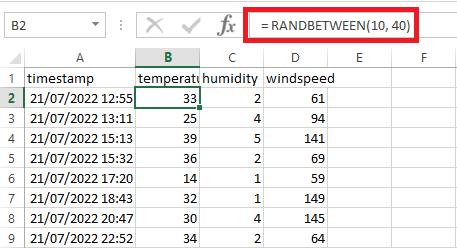
Write a SQL query to show the daily maximum of the values for each attribute over the month of July 2022, where the maximum value was greater than 90, ordered by attributes.

**Solution:**

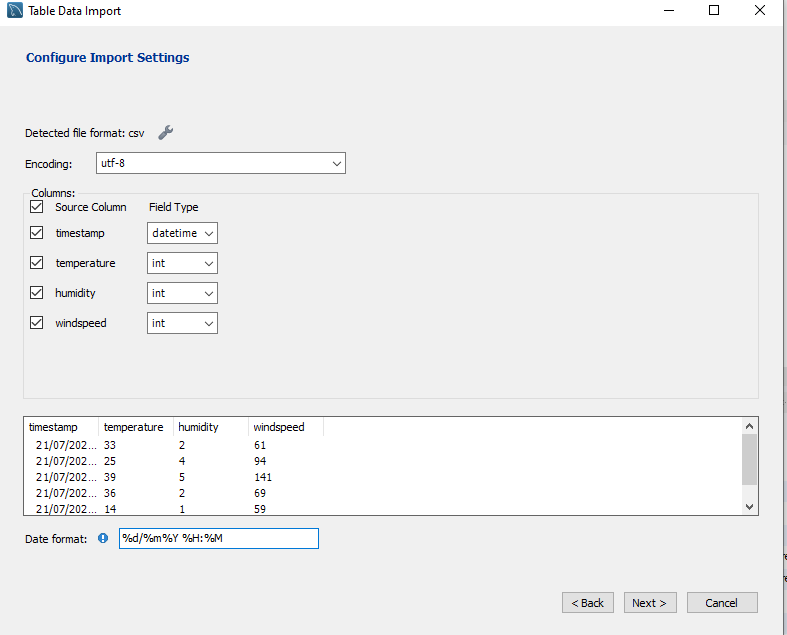
**The very first step, I have performed is to make dummy data in excel. I have subtracted 30 from current datetime, to get DATES of month JULY 2022.**



**Same I have done with other columns, I have generated random data between ranges.**



**To show working of queries, I have uploaded dummy data to my existing database using import wizard of MYSQL WORKBENCH. I have set the data types and date time format.**



**Note That:**

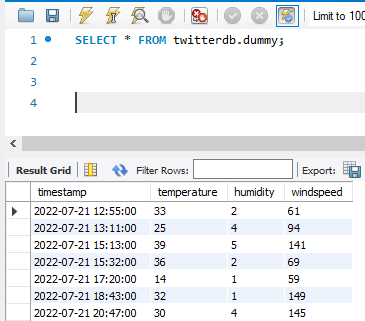
**I have used my existing database “twitterdb” and keep the table name same as file name “dummy”. Please consider twitterdb as Oracle and dummy as table name. We can easily create new database using following;**

|  |
| --- |
| CREATE DATABASE *databasename*; |

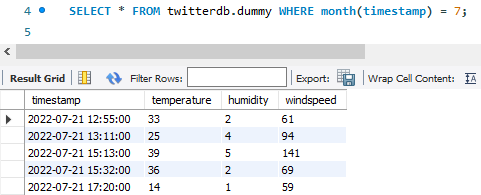
**And to use Database we need following**

|  |
| --- |
| use *databasename*; |

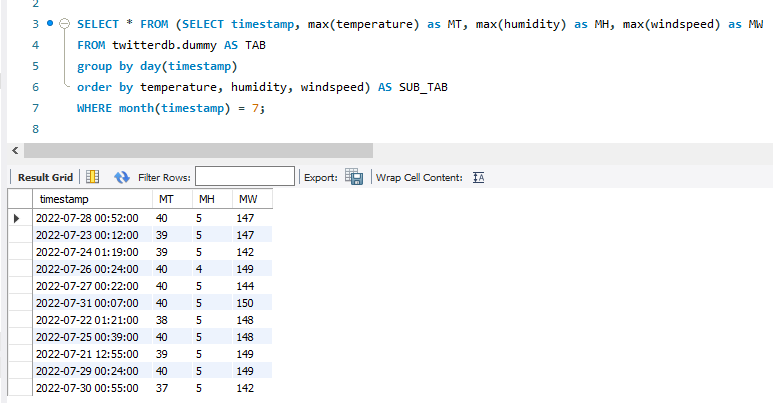
**Just to verify the database, I have select all records as shown;**



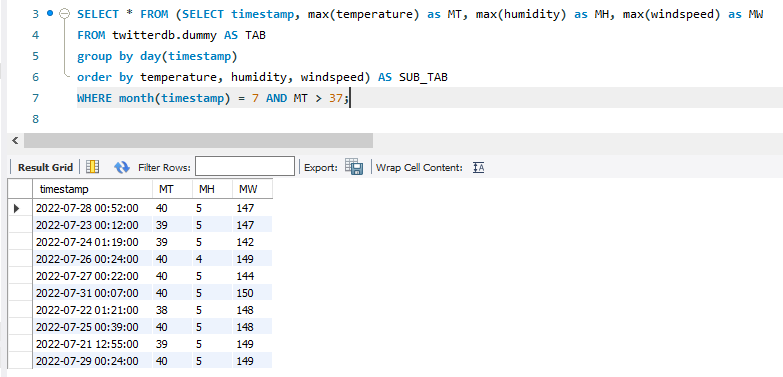
**I have used month() function to return month number form datetime. Number 7 means July month.**



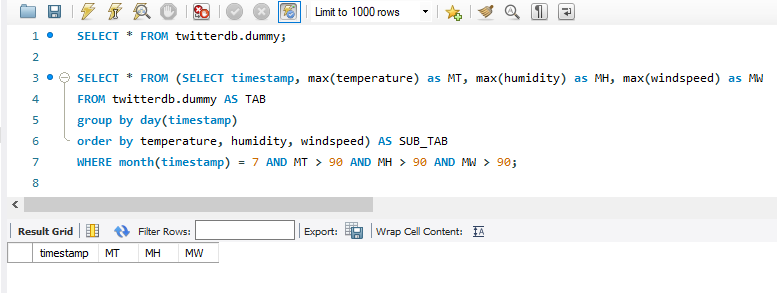
**Select Maximum values for attributes only for month of July;**



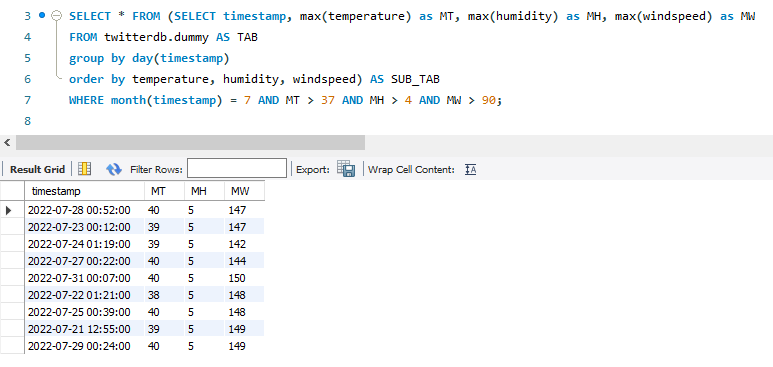
**I have select data using conditions such max(attribute) greater than VALUE. In given example it will return all values where maximum temperature is greater than 37 for month of July.**



**This is the required query as per question. You will observe that it return nothing because I my dummy data, I have values ranges for some attributes less than 90.**



**This is just to check working of above query with different conditions. See its working for other values.**



**Final SQL Query will be:**

|  |
| --- |
| SELECT \* FROM (SELECT timestamp, max(temperature) as MT, max(humidity) as MH, max(windspeed) as MW  FROM twitterdb.dummy AS TAB  group by day(timestamp)  order by temperature, humidity, windspeed) AS SUB\_TAB  WHERE month(timestamp) = 7 AND MT > 37 AND MH > 4 AND MW > 90; |

Question #2:

Given:

Database: Oracle

Tablename: table1

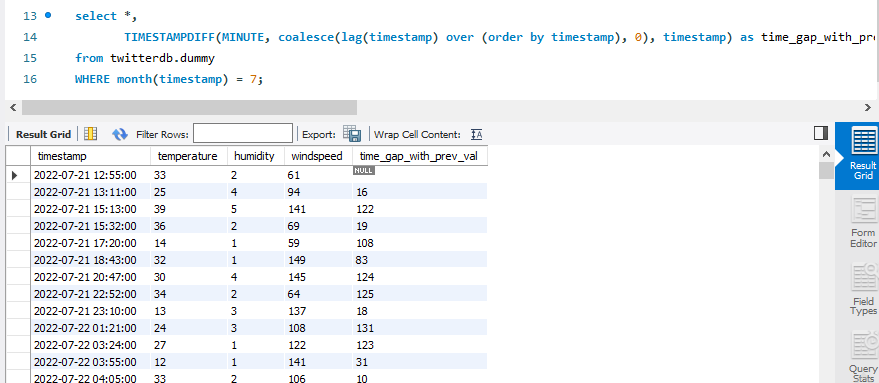
Columns: attribute (varchar2(100)), timestamp (date), value (number)

Example of attribute: temperature, humidity, wind speed etc.

Question:

Write a SQL query to present the data in the following structure: attribute, timestamp, value, time\_gap\_with\_prev\_val, where time\_gap\_with\_prev\_val is the duration between a particular value and its previous value, when ordered in ascending order of time, for a given attribute. Show data only for the month of 2022 July.

**Solution:**



|  |
| --- |
| select \*,  TIMESTAMPDIFF(MINUTE, coalesce(lag(timestamp) over (order by timestamp), 0), timestamp) as time\_gap\_with\_prev\_val  from twitterdb.dummy  WHERE month(timestamp) = 7; |

Question #3:

Given:

Database: Oracle

Tablename: table2

Columns: id (number), comment1 (varchar2(max)), comment2 (varchar2(max))

The varchar2 columns can have large amount of text data. We have identified 10 features that will need to be extracted based on the combination of columns comment1 and comment2.

Question:

What steps will need to be performed before creating the model?

**Solution:**

**Step1: Database Connection**

|  |
| --- |
| This step can either using SQL editor or using Python. You just need database credential where the data is available. For Python you have an open source python modules to work with Databases. (**as a programmer I will recommend this although we can do this using SQL Editor including step 2 and 3**) |

**Step 2: Data Extraction**

|  |
| --- |
| After connecting to database, now you can simply extract the required columns. It is recommended to query only required columns for fast and optimal execution of query. You can execute SQL queries in Pandas library or MYSQL library. If you are using PANDAS, you have the required table in Dataframe form; easy to manipulate and work with. This dataframe will be further used for analysis and modeling. |

**Step 3: Save Data to Local File (Optional)**

|  |
| --- |
| It is best to save data to local files, as it’s easy to read data from local files as compared to database (the execution time mainly depends on Internet speed). Also you can reuse this data again and again if needed. In case of secret/ confidential data, simply read from database, avoid files on local system. |

**Step 4: Data Pre-processing**

|  |
| --- |
| This step mainly depends on the dataset nature. For textual data you need following pre-processing steps (very from problem to problem, you may skip some steps).   * Convert text to lower case * Remove any Garbage Text * Remove Stopwords * Remove Symbols * Replace short forms by full forms * Do stemming * Do Lemmatization |

**Step 5: Feature Engineering**

|  |
| --- |
| The next step is most important and critical one. You will apply different feature engineering steps. This depends on which features are you using. In NLP we can use following features:-   * Term Frequency * Inverse Document Frequency * Term Frequency – Inverse Document Frequency * Bag of Words * Big-Grams * N-Grams etc., |

**Step 6: Machine Learning Modeling**

|  |
| --- |
| After having features, you will move to machine learning model. You can use any machine learning model, if labelled data use supervised machine learning otherwise use unsupervised machine learning model. |

**Step 7: Machine Learning Model Evaluation**

|  |
| --- |
| After having model, you have to evaluate the model on testing/ unseen data. Usually we split data into two parts, training and testing. You can use different parameters to evaluate models:-   * Classification   + Accuracy   + Precision   + Recall   + F1-Measure   + Loss * Regression   + Mean Squared Error   + Root Mean Squared Error   + Mean Absolute Error   + R2 Score   + Mean Absolute Percentage Error |

**Step 8: Model Deployment**

|  |
| --- |
| Once you have model on which you are confident that it is optimal, in the last stage you deploy model for execution/ usage. |