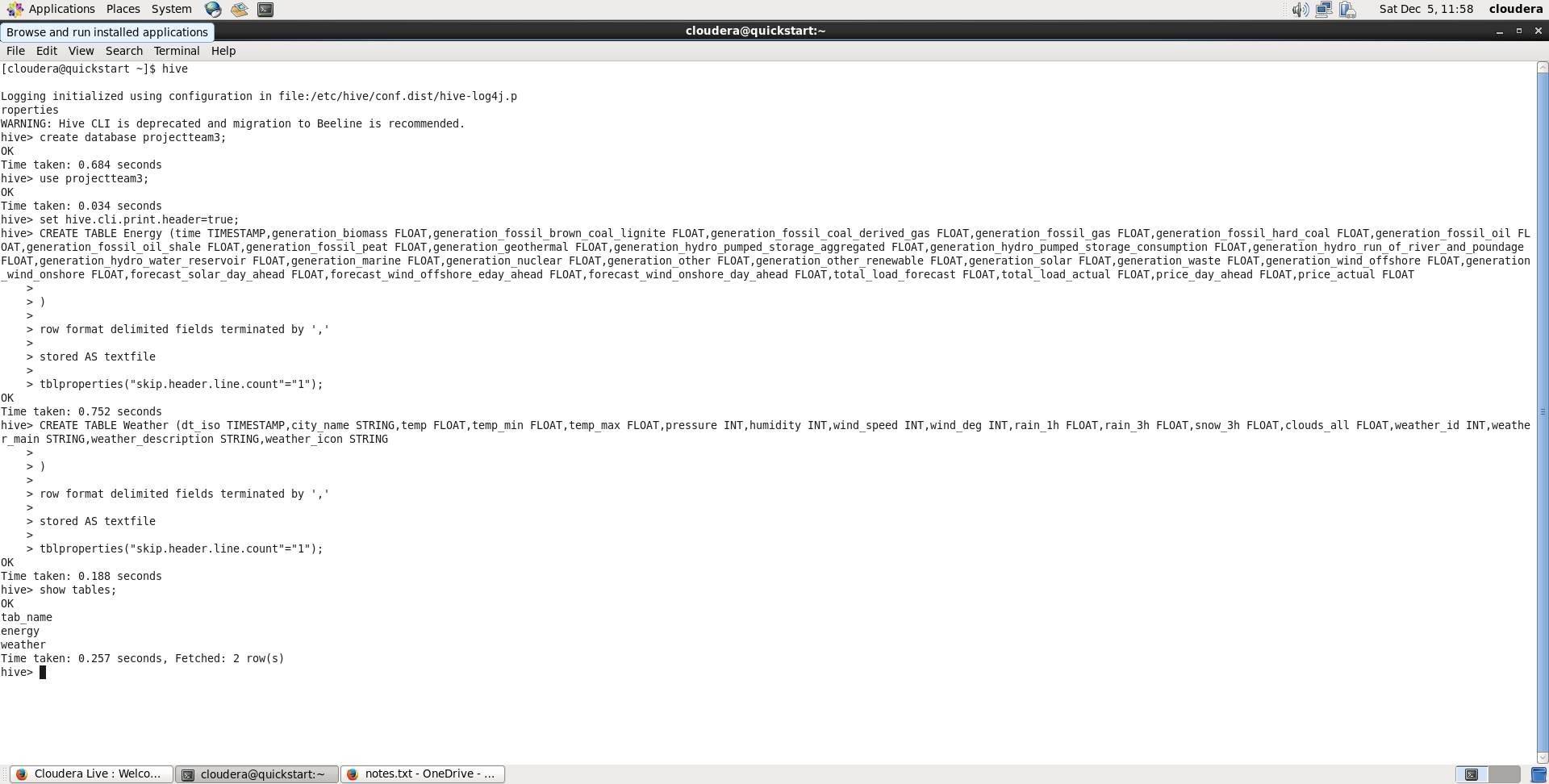
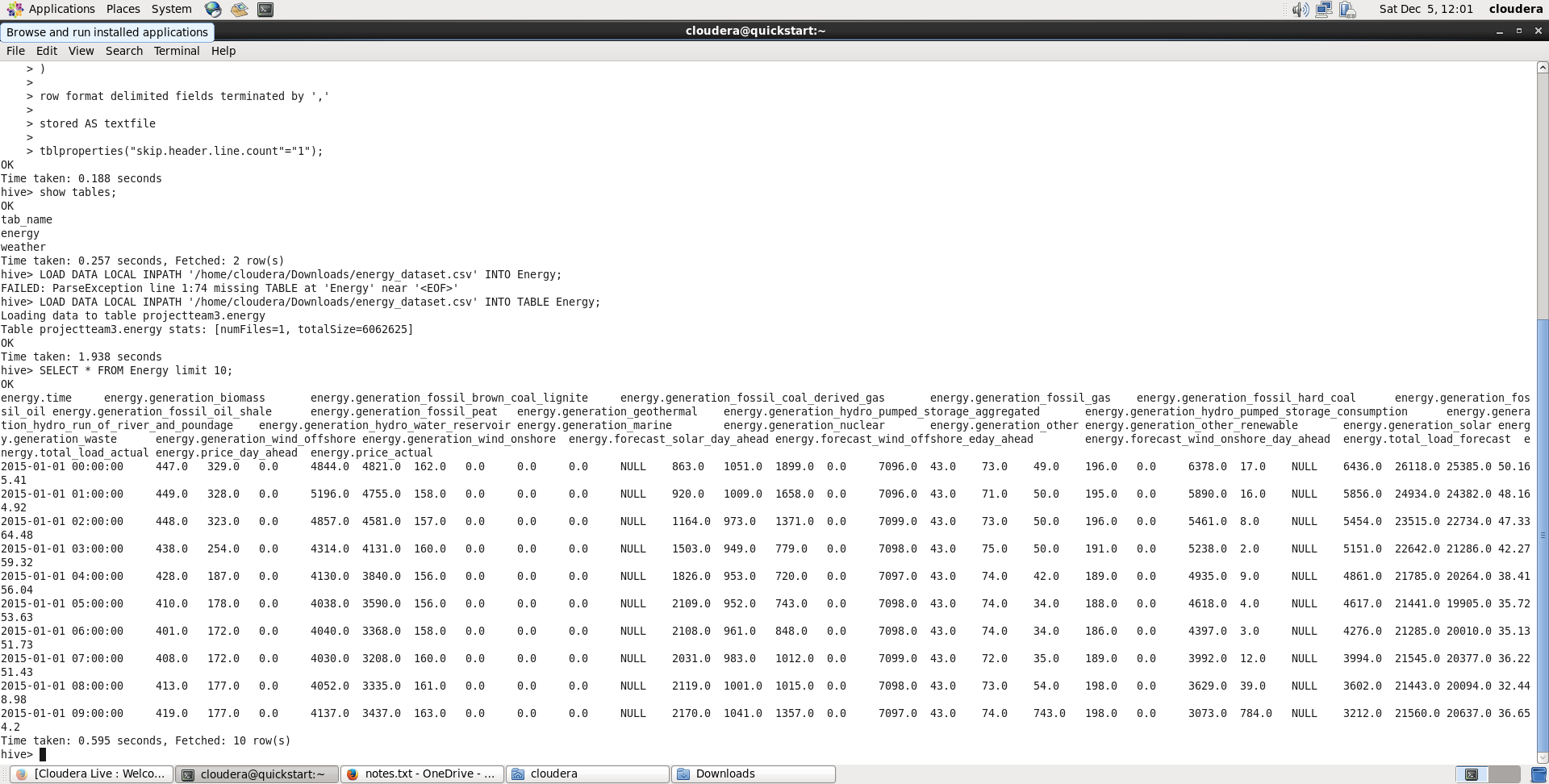
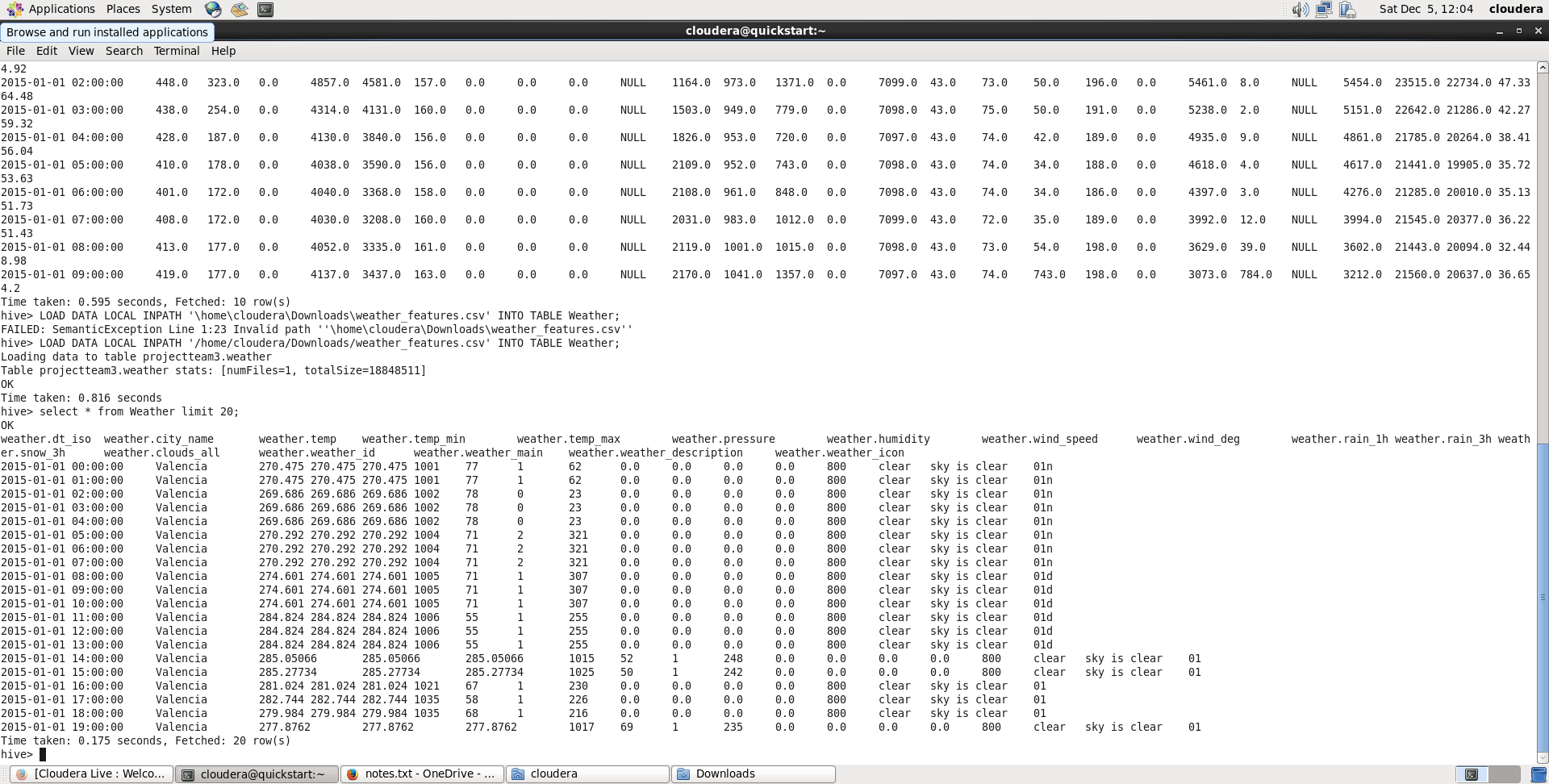
**PROJECT INCREMENT FINAL by CLAIRE**

We are working on two datasets for hourly energy demand generation and weather focusing on cities in spain. In this presentation, I will be importing both datasets into hive and then into mysql so as to run queries and see what we can predict or analyze from the data. The second part of my work, I will be creating a spark dataframe with the help of pyspark using google colabotary where I will do some visualization on the data and perform some aggregate functions like average, count, max etc

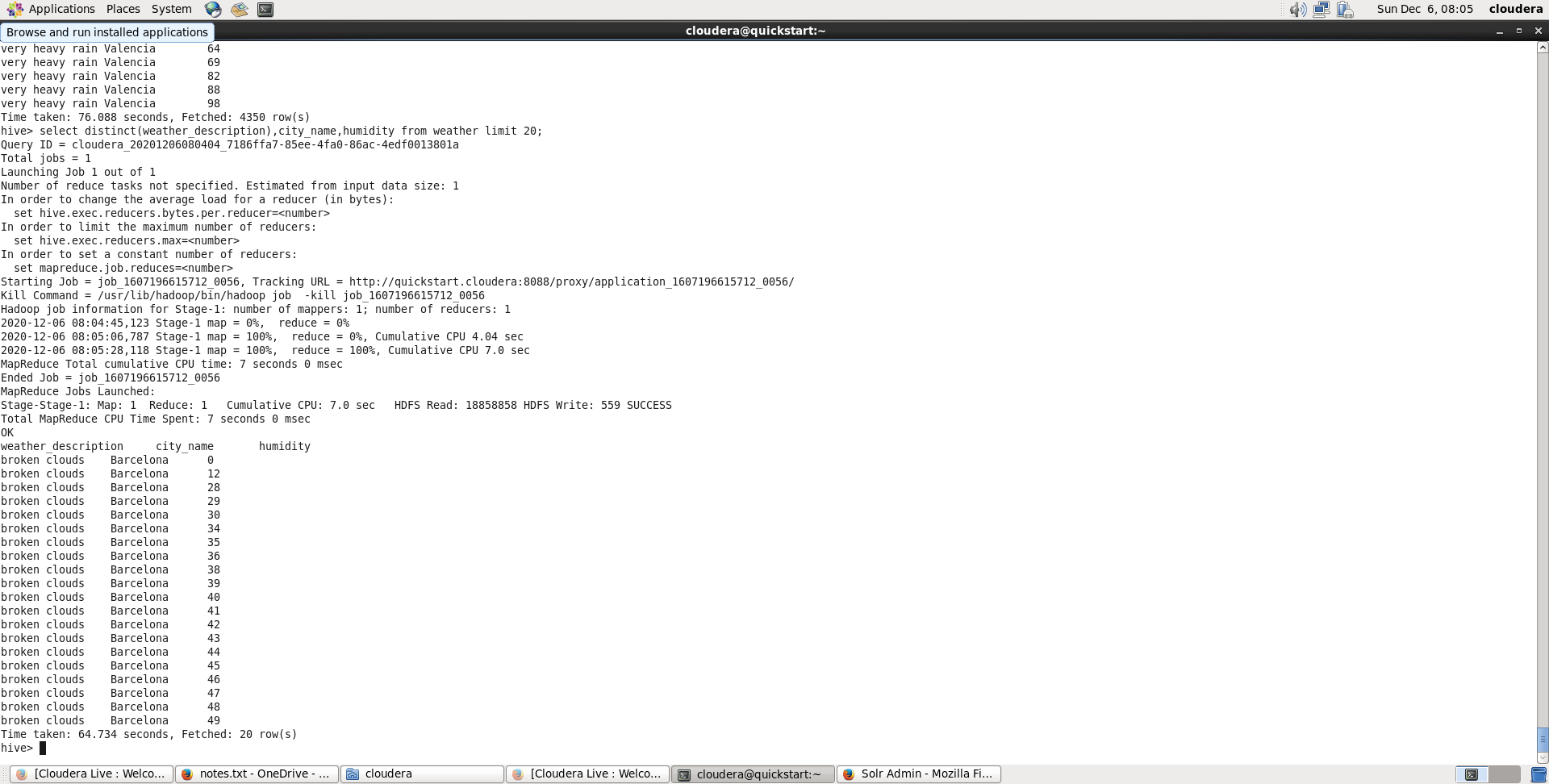
**HIVE**







With the dataset imported in Hive, I am able to query successfully see

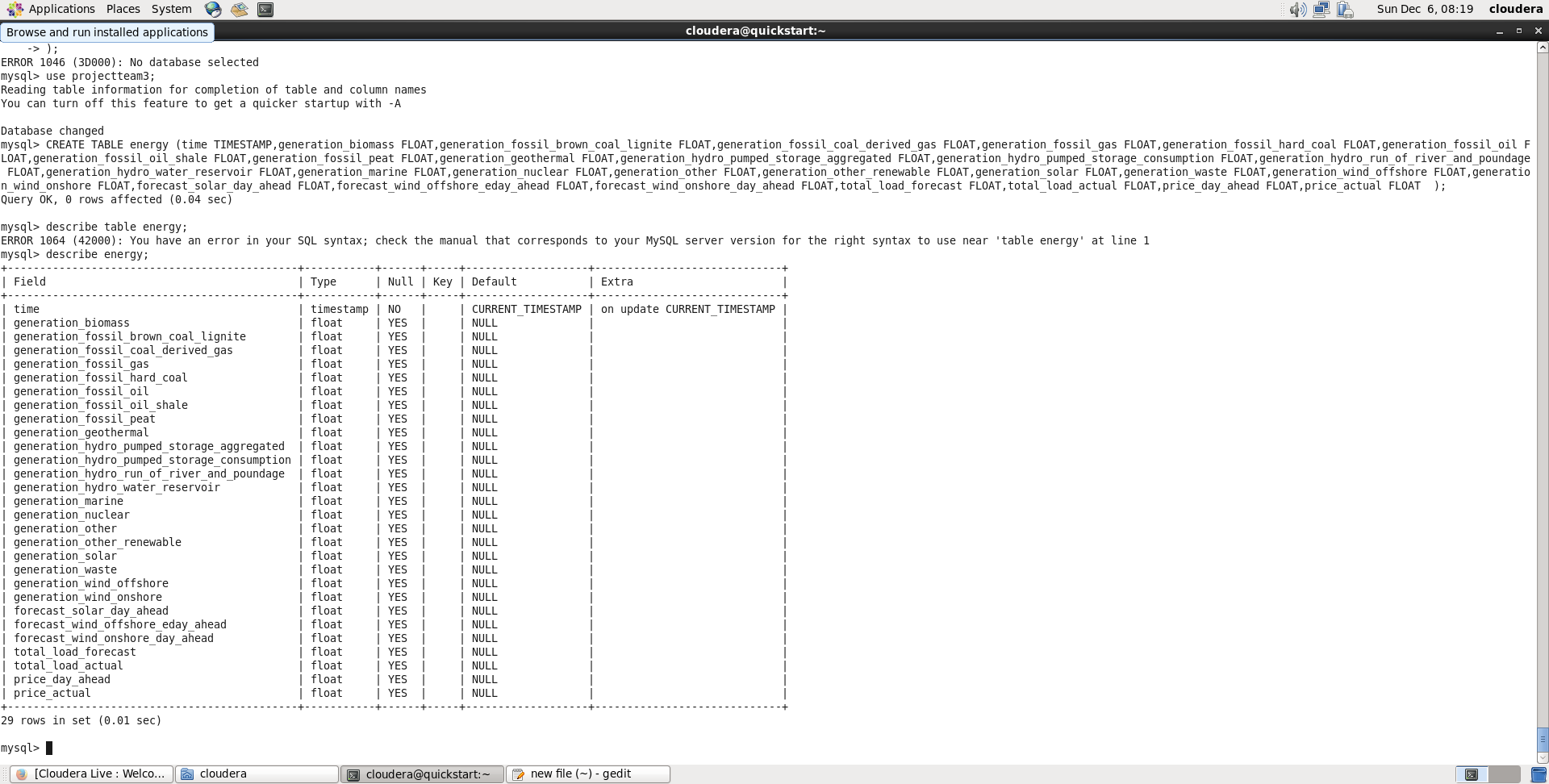


The next step was to move the data to mysql using sqoop export as we learned in class. Logging into mysql, I created a similar table like that of hive with datatypes with respect to mysql, see below;

First I created a database called projectteam3 which is a requirement to use mysql. Using this database I was able to create the two tables needed to perform the sqoop export

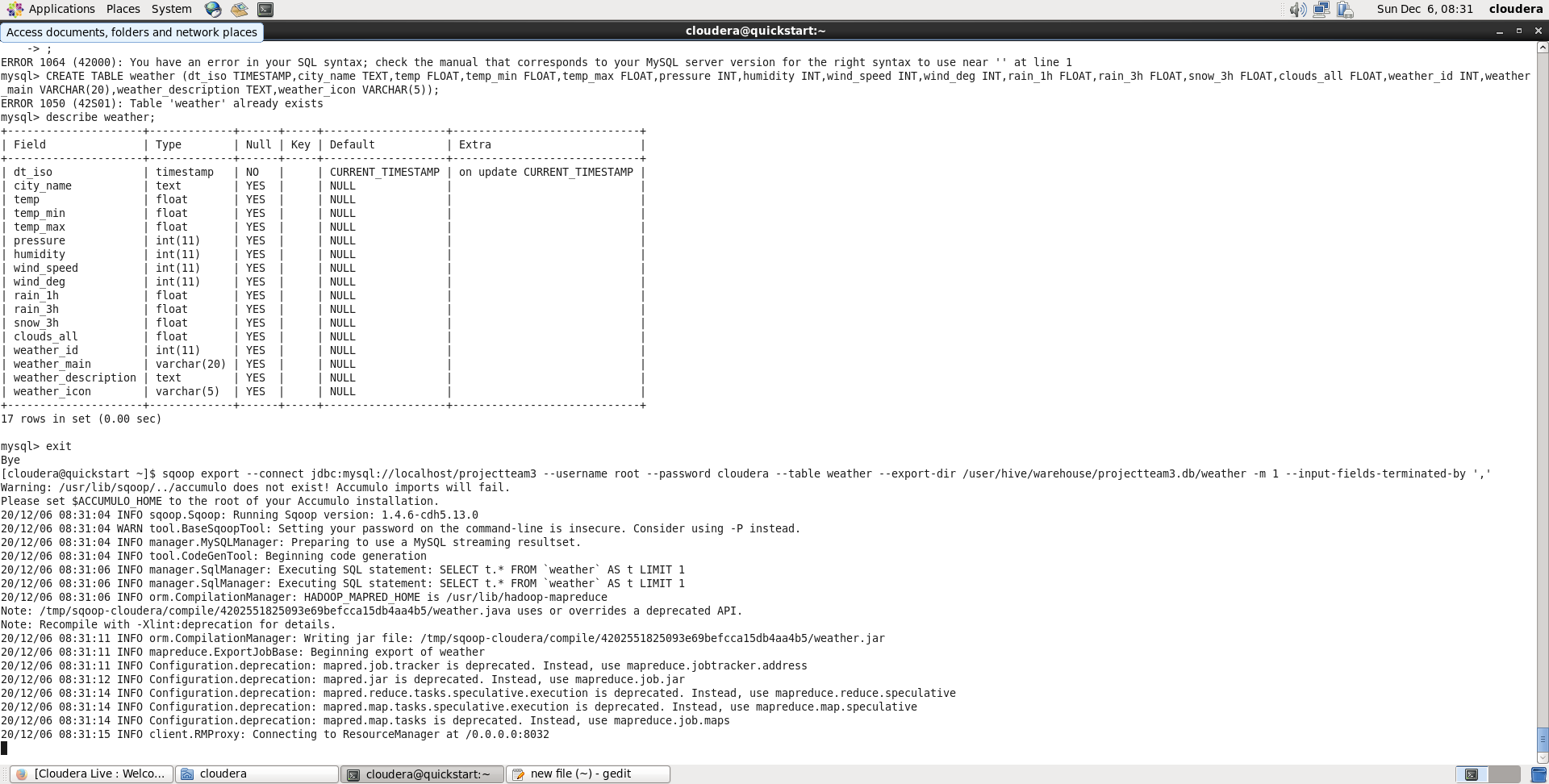
For energy table;

Table creation codes uploaded in hivemysql.txt file



Mysql accepts datatype of float and int as I have used for the creation of my table.

**Creating Weather table for the weather\_features dataset**

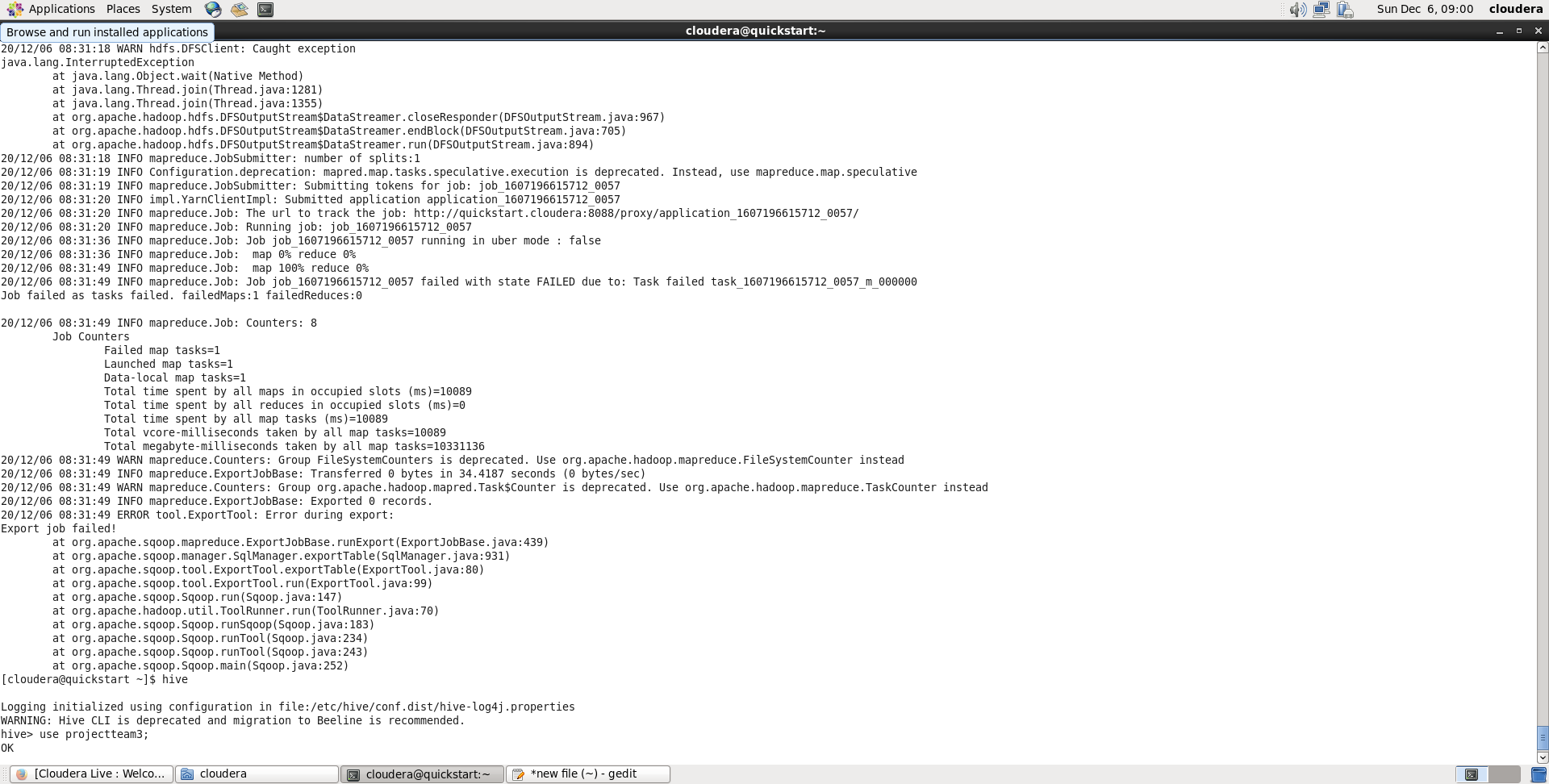


I have successfully created Energy table for the energy\_dataset.csv dataset I have and Weather table for the weather\_features.csv dataset in both hive and sql. From the screenshots you can see the table description of both tables in hive and mysql. Once I was certain that the table in hive matches that in mysql per column definitions I can now use sqoop export to move the data from hive to mysql;

**Using sqoop export**

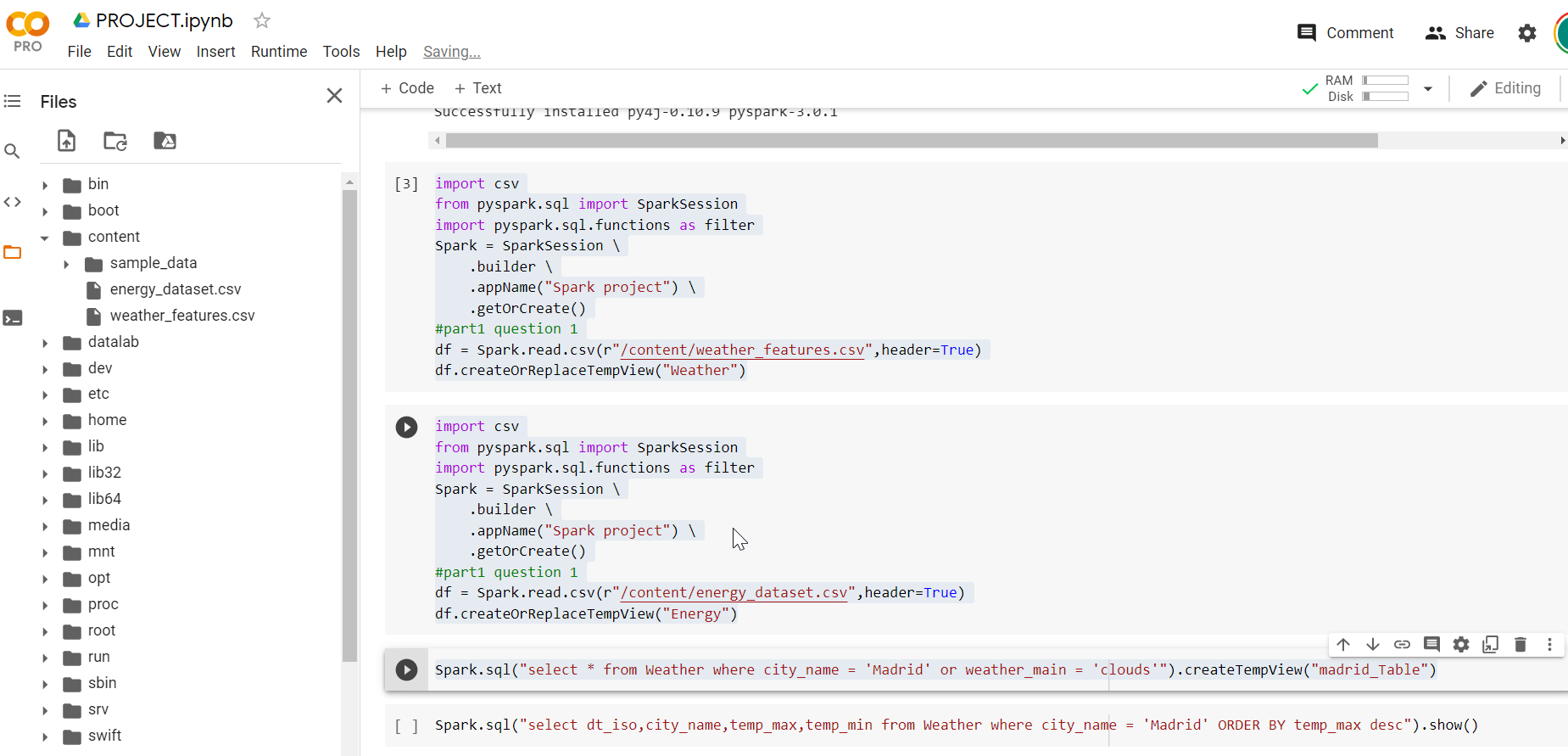
For increment 2 of the project, I faced two difficulties I had to improve on. One was on getting my data to load into hive without having Null values even though the csv file had values. I succeeded with that part as you can see from the screenshots above. However , the second difficulty I faced was with sqoop export where the export job will fail after mapping. I have attached the various codes I used for this part in a notepad file called hivemysql.txt. in this file you can see how the sqoop connects with jdbc successfully, reads username and password but reaching the export directory if fails at the level of the mapper. After debugging , my guess is the datatypes used in the table creation of mysql might be affecting the smooth transfer of data from hive. I changed the varchar datatypes to text and tried but had same error. JDBC connection looks ok as I created a dummy table and inserted values to do an export and it worked perfectly reason for my guess above.





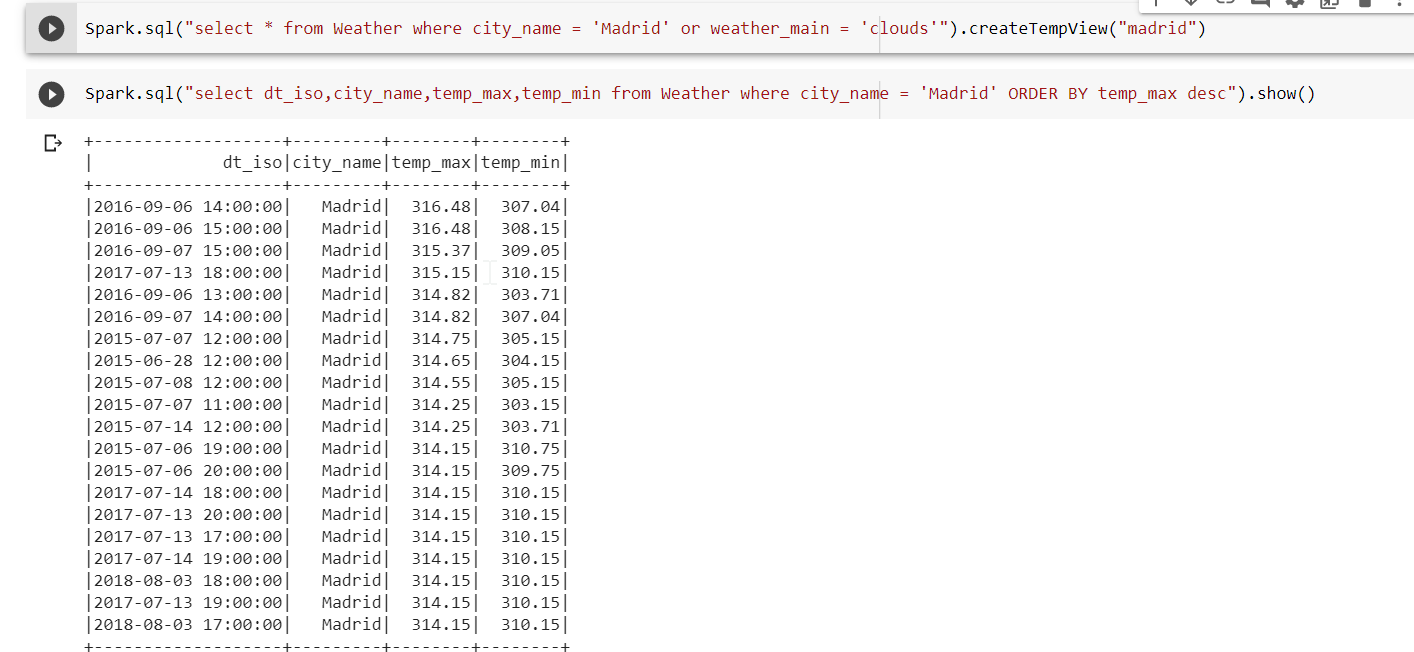
**SPARK.SQL**

Another interesting module we convered in class was Spark.sql where with the help of pyspark using dataframes I was able to create a view of the both datasets.

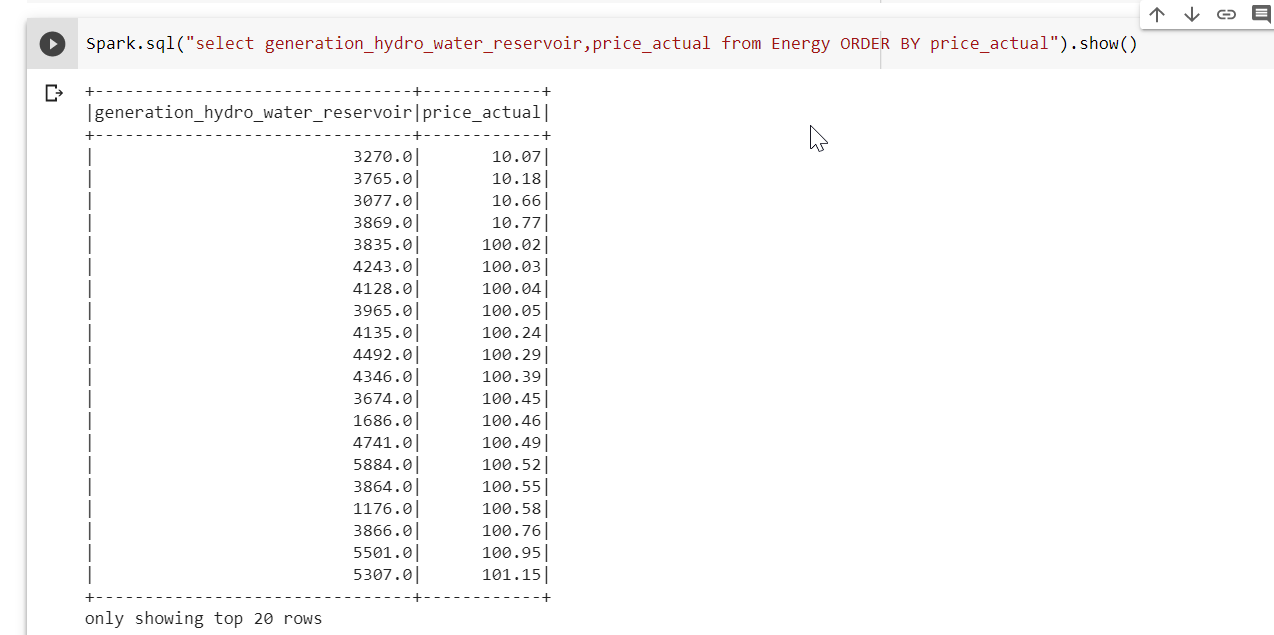
**Weather AND Energy table createion using Spark.sql** ****

With the successful upload of both tables, now I can see what to visualize. The Weather dataset has columns that a layman can easily understand what to visualize but as for the energy dataset, this will be more helpful for users in the electrical field as the terms would be familiar to them.

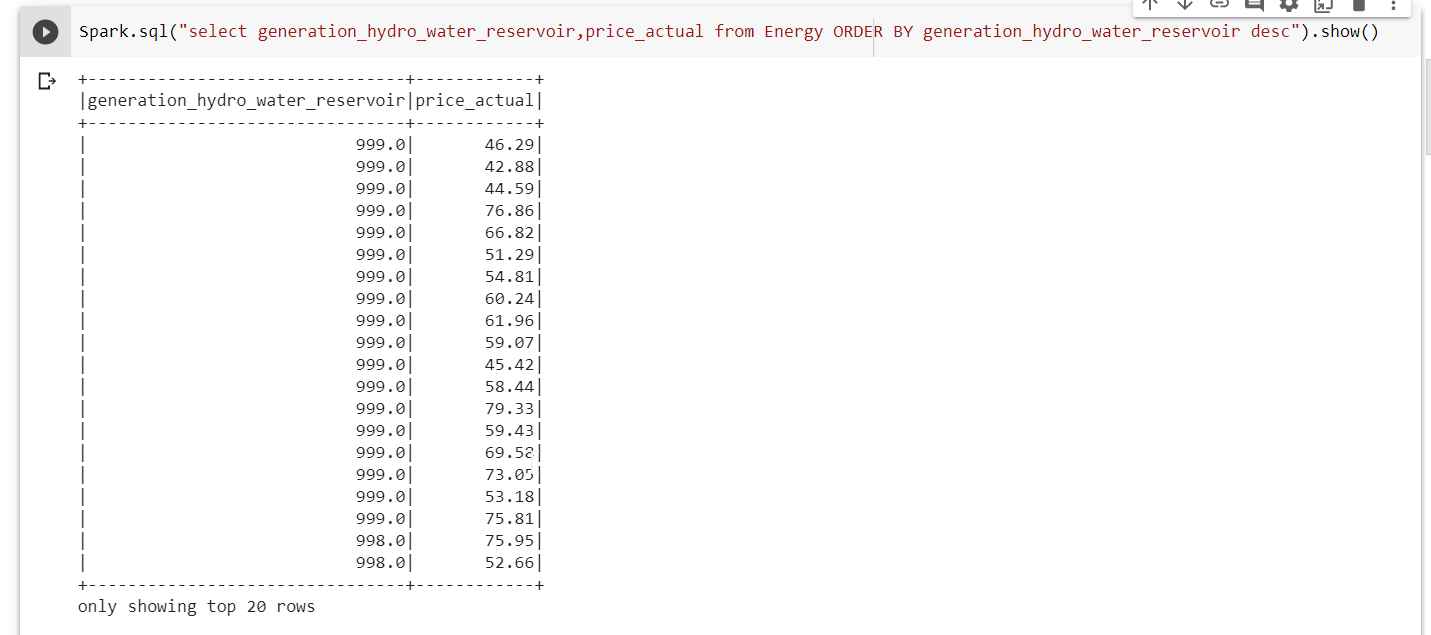
1. Viewing the maximum and minimum temperatiures in the city of Madrid this can be helpful for a weather channel to forecast how the temperatures might look like for the next few hours during a football match for example.

****

1. This next visualization, I want to see how the generation\_hydro\_water\_reservoir affects the price\_actual column

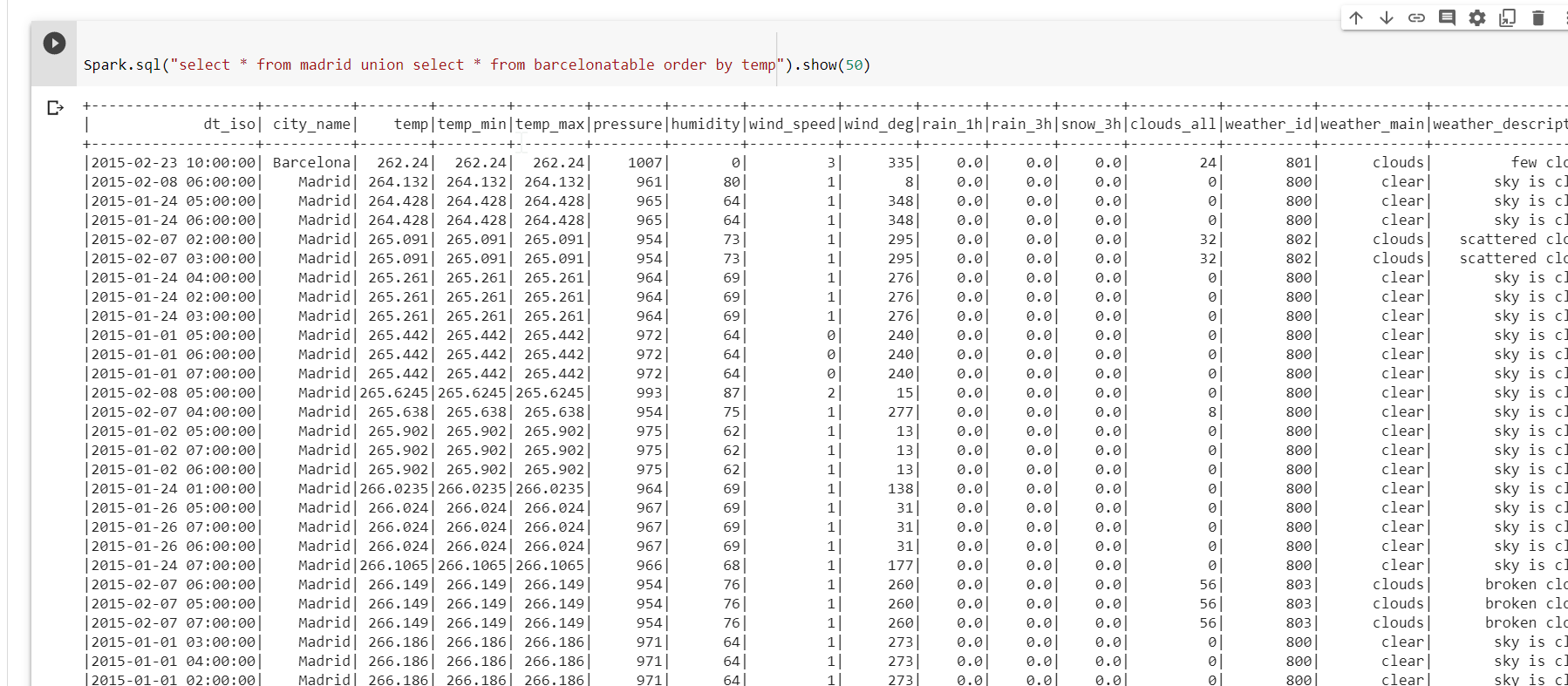
****

Looking at the outcome of this query, we notice that the generation\_hydro\_wter\_reservoir doesn’t really affect the price in the sense that, you will think that the higher the value, the higher the price bur I sorted my output in order of price ascending and we notice that a generation\_hydro\_water\_reservoir of 3270.0 incures a price of $10.07 but a value of 1176.0 which is obviously lower incures of price of up to 100.58.

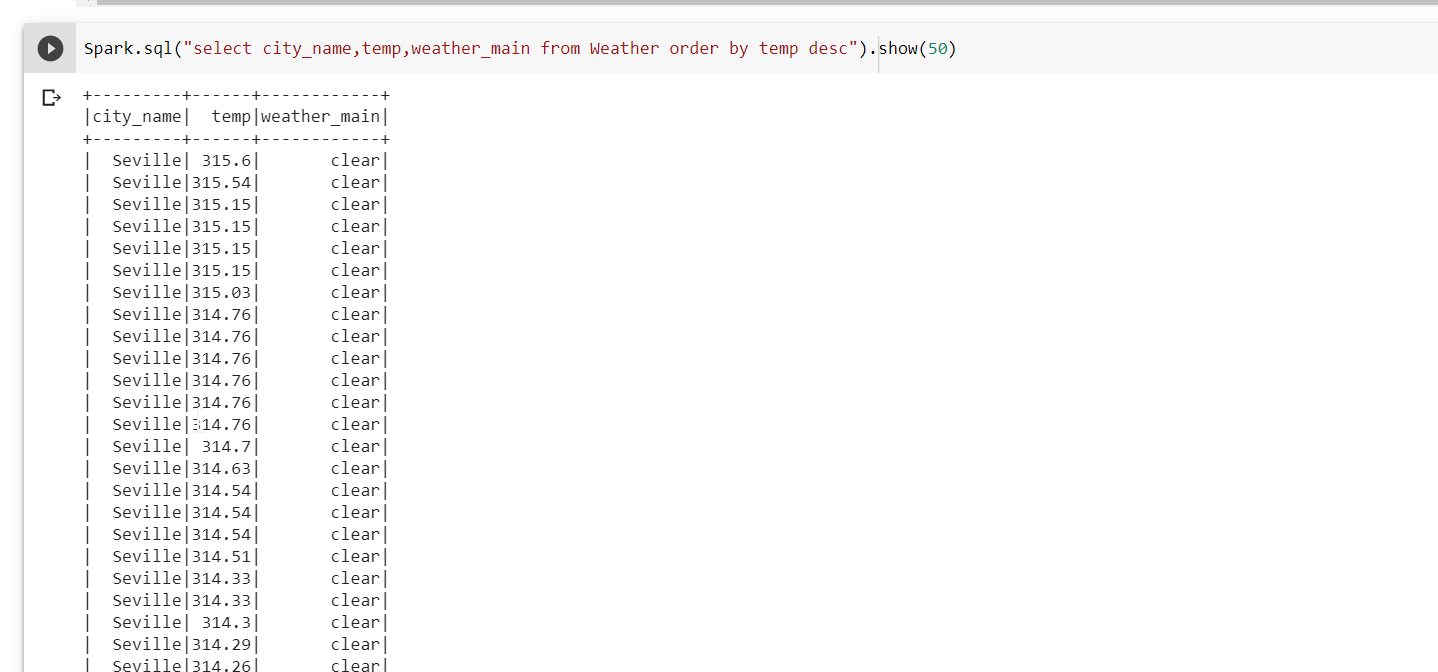
****

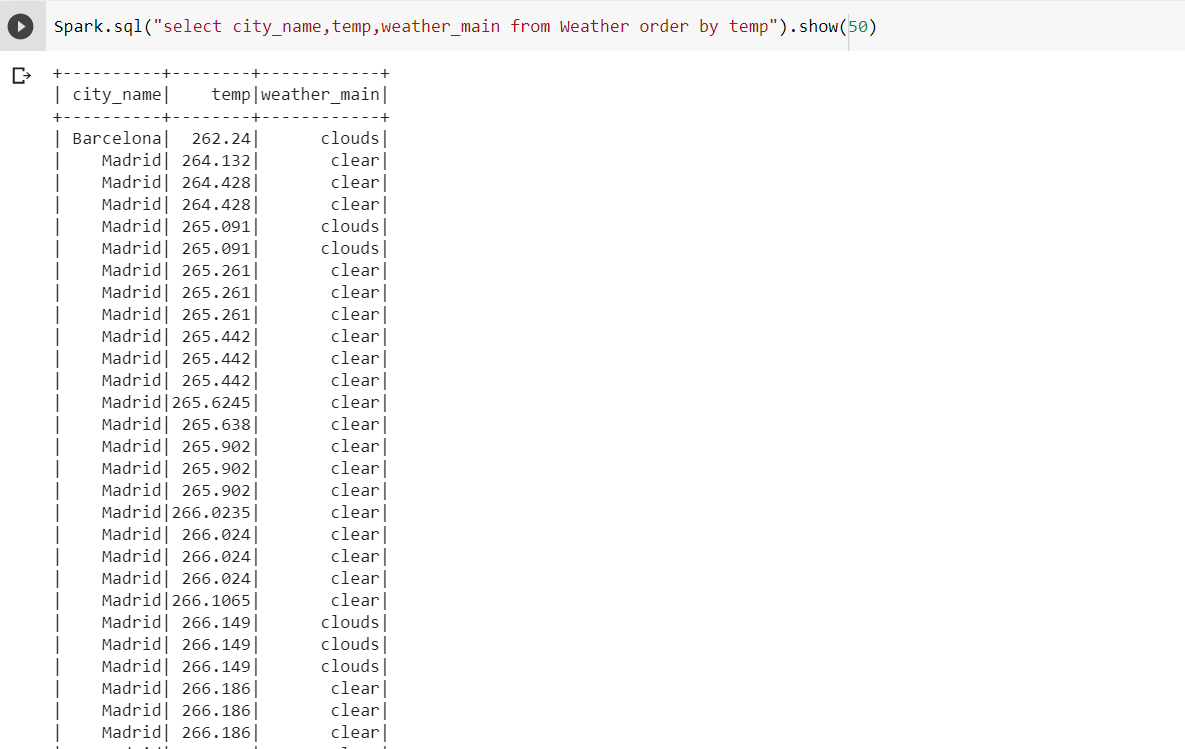
The screenshot above, I sorted my results in ascending order of the values of generation\_hydro\_water and we notice that even though the values are the same for most of the rows, the price\_actual is still not the same. Hence analyzing this I will say the generation\_hydro\_water\_reservoir does not have a high impact on the price\_actual column

1. You notice above that I created a tempview of madrid and Barcelona, this can be helpful when you want to filter out some cities and work with. For my case I am using madrid and Barcelona, I created a tempview for both cities and my first query will be merging these two tables together with the help of union operation and I am sorting it in ascending order of temperature.



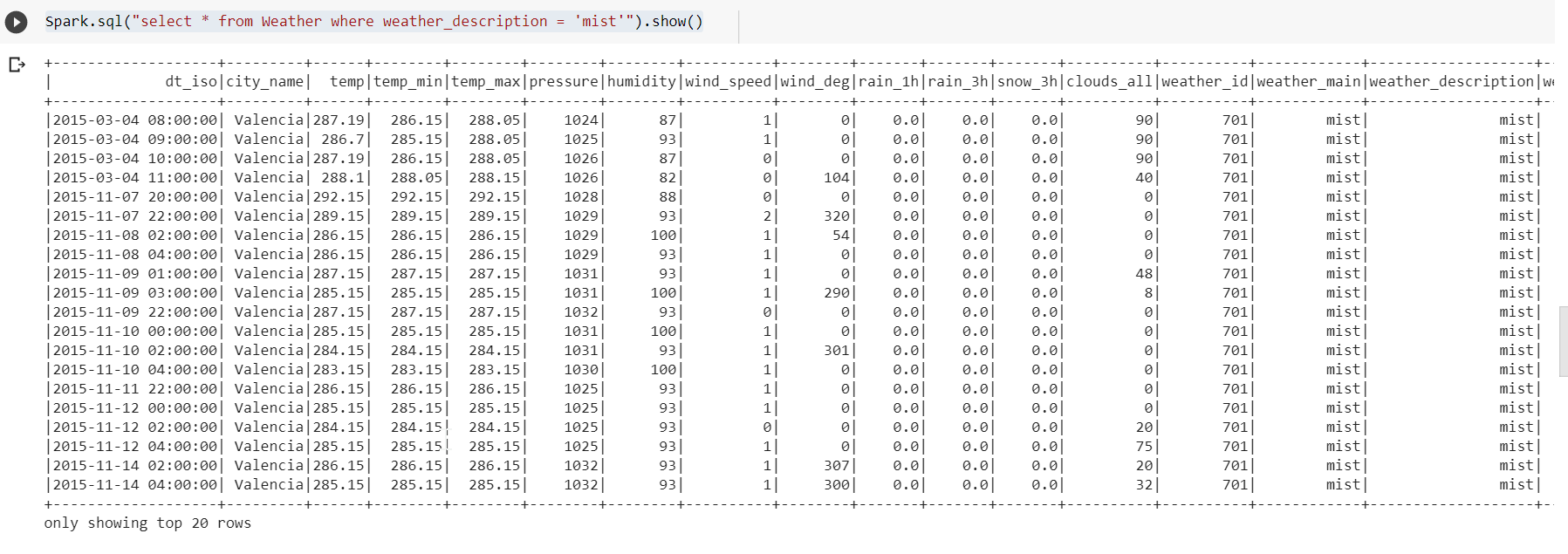
1. Our next visualization is one between the city, temperatiure and he weather. My idea here is to find out how the temperature is related to the weather\_main column which basically just tells you how the weather is.



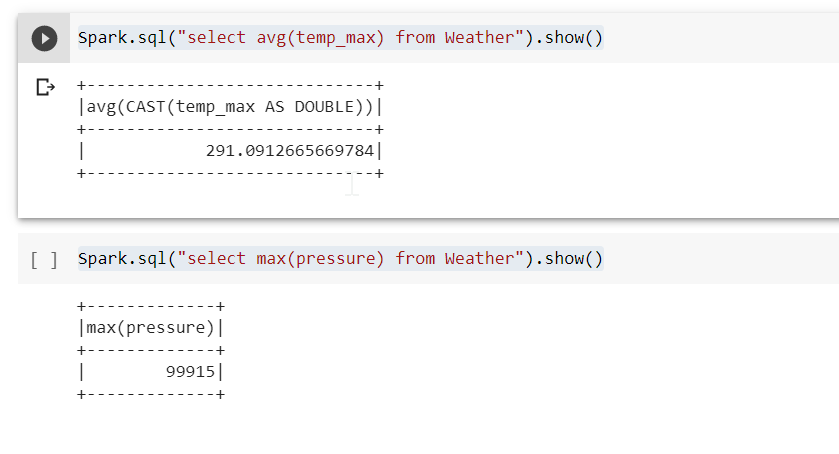


From the screenshot above I could come up with a conclusion that , when the temperature is high, the weather is clear and also the city of Seville has high temperatures compared to the other cities also Madrid city has lower temperature compared to the other cities. In order for me to make this conclusion, I viewed the data in ascending order of temp and descending order of temp.

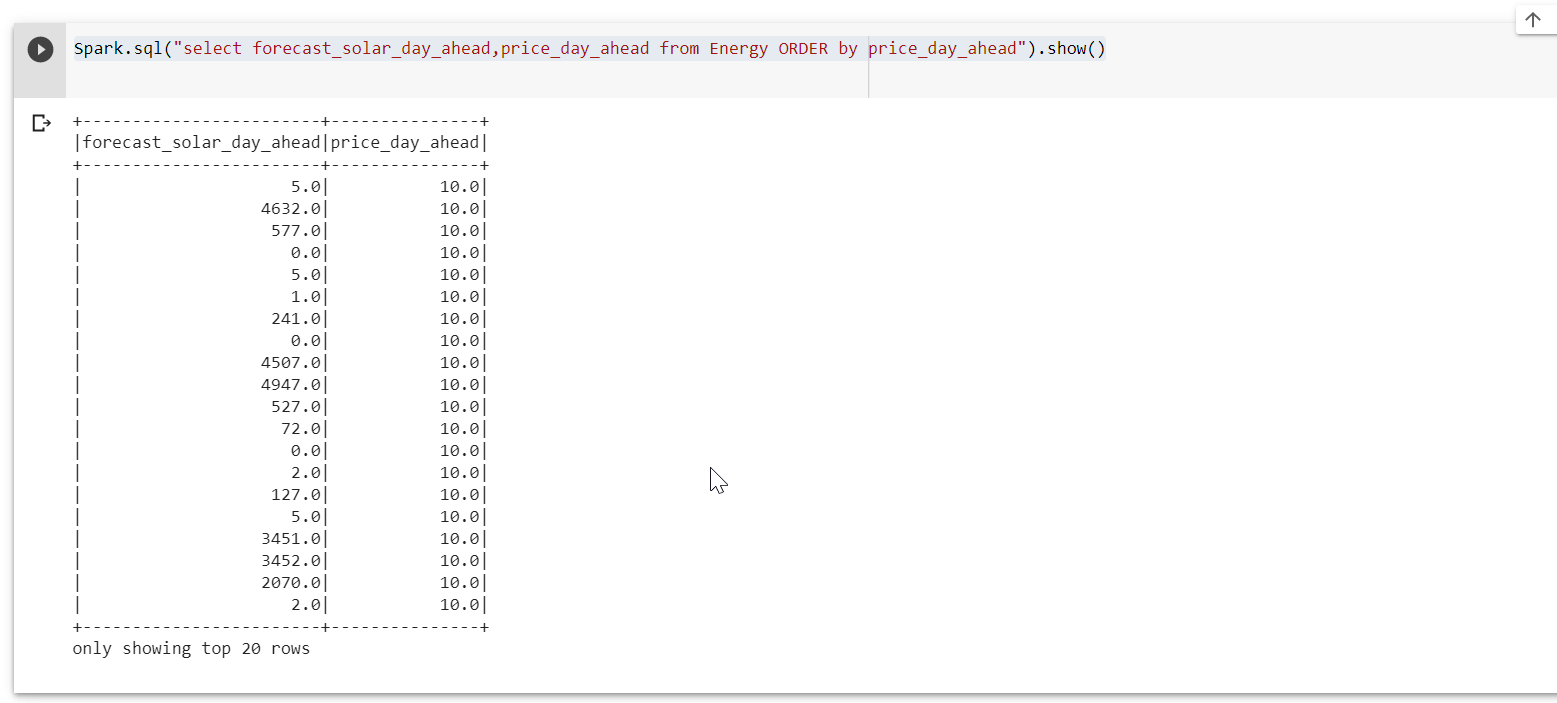
1. My next query is a simple one where I want to view those cities that had mist weather. This can be helpful when a user wants to know if a particular city has a history of this weather type



1. using avg and max function, I am able to get the average temperature and maximum temperature form my dataset and the maximu pressure. This kind of analysis can be useful to users who want to compare the difference in temperature values for the previous year and current year



e) with the Energy dataset, I could view solar focast for a day ahead and the price



From our result , we notice that the forecast\_solar\_day\_ahead doesn’t have a high impact in the price\_day\_ahead as the values of the forecast are quite different but the price is same

1. Using the distinct function as our dataset is very large and for this query i do not want repetitions, I can view cities , their maximum temperature and how the weather looks like with such temperatures

