**CSEE 5590/490: Big Data Programming**

**Project Report**

**Due Date: Monday, May 10, 2021**

**Project Title:** Analysis for determination of a relationship between energy demand and weather.

**Team Members:** Joe Goldsich, Anna Johnson, Kyle Son, and Bill Yerkes

**Introduction:** Information overload can make it difficult to understand issues and solve problems. The appearance of excessive quantity of information prevents the understanding of the problem and the ability to construct a solution to the problem. Big Data technologies takes this paradigm and turns it inside out. The students working on this project are going to attempt to construct a model which will be able to take weather and energy related data to be able to forecast energy demand.

**Background**:

There are numerous articles and papers about leveraging weather data, weather forecast, to be able to forecast energy demand and energy production via renewable sources. An article in Wired talks about how this relates to energy pricing and how the technology has moved from using spreadsheets to the using of machine learning.1 The company BAXENERGY has a web page about the benefits of this field.2 Hugo Ferreira wrote a paper about predicting wind and solar generation using machine learning in 2018.3 They all talk about the vast amount of data which gets generated and stored. How this data can be leveraged to help improve the energy sector, and in turn help improve society.

**Goals and Objectives:** Utilize the tools and technologies learned from CSEE 5590 to be able to analyze collected data so that it will be possible to determine if there is a relationship between weather and energy consumption and if a relationship exists determine the possibilities of using that relationship to predict future energy needs.

**Motivation:** The global population continues to increase, and the weather patterns seem to be getting more extreme, from extending periods of both above and below normal temperatures in various parts of the world and in the United States. The demand and consumption of energy increases with the population and with the extreme weather, the need for air conditioning in the summer and for heating in the winter. The recent crisis in Texas has demonstrated what can happen if the energy providers are not able to meet the demands of the consumers. Being able to forecast accurately future demand and plan accordingly can help prevent or mitigate such crises in the future.

**Significance:** Better planning of resources for Utility Companies can result in reduced cost to the consumers and more reliable service. This also dips into the area of public safety, as loss of power during extreme weather with no warning can be dangerous for vulnerable groups.

**Objectives:**

What weather measurements and cities influence most the electrical demand, prices, generation capacity?

Forecast intraday price or electrical demand hour-by-hour.

What is the next generation source to be activated on the load curve?

Analyze the change of the electricity price yearly and hourly and get some insights

Create ML or DL model to predict the electricity price

Create Clustering model by the weather features

Could We perform Map Reduce on the datasets to get some refined data

Plotting some graphs for visualization with meaningful insights

**Approaches/Methods:**

The team’s initial form of communication was via Email. The discussion over email resulted in the team deciding to communicate via Discord, which has allowed for the team member to communicate on a regular basis. The team also meets periodically via Zoom to review items and discuss plans.

The team uses Google Docs and Google Slides to collaborate on documentation. The team also uses GitHub as the repository for source code and final version of documentation.

The team met to discuss several ideas on what topic to do the project on and came to a consensus to do the project on Energy prediction based on Weather. The team reviewed the information on the Kaggle site. The team decided to leverage the knowledge we had gained over the first half of the course to investigate both the technologies, Hadoop, MapReduce, Hive, Sqoop, Cassandra, Solr, that were covered and the data, Weather and Energy datasets.

Each team member worked on their own. They utilized the various tools to do analysis on the data for the project. This allowed the team members to get more practice with the various tools and to get familiar with the data. This approach led to discussions of common issues found and exchanges of ideas between team members. The team shared with each other their findings on various tools and data. The team also started discussion on what approach the team should take for the next iteration.

**Our Solution:**

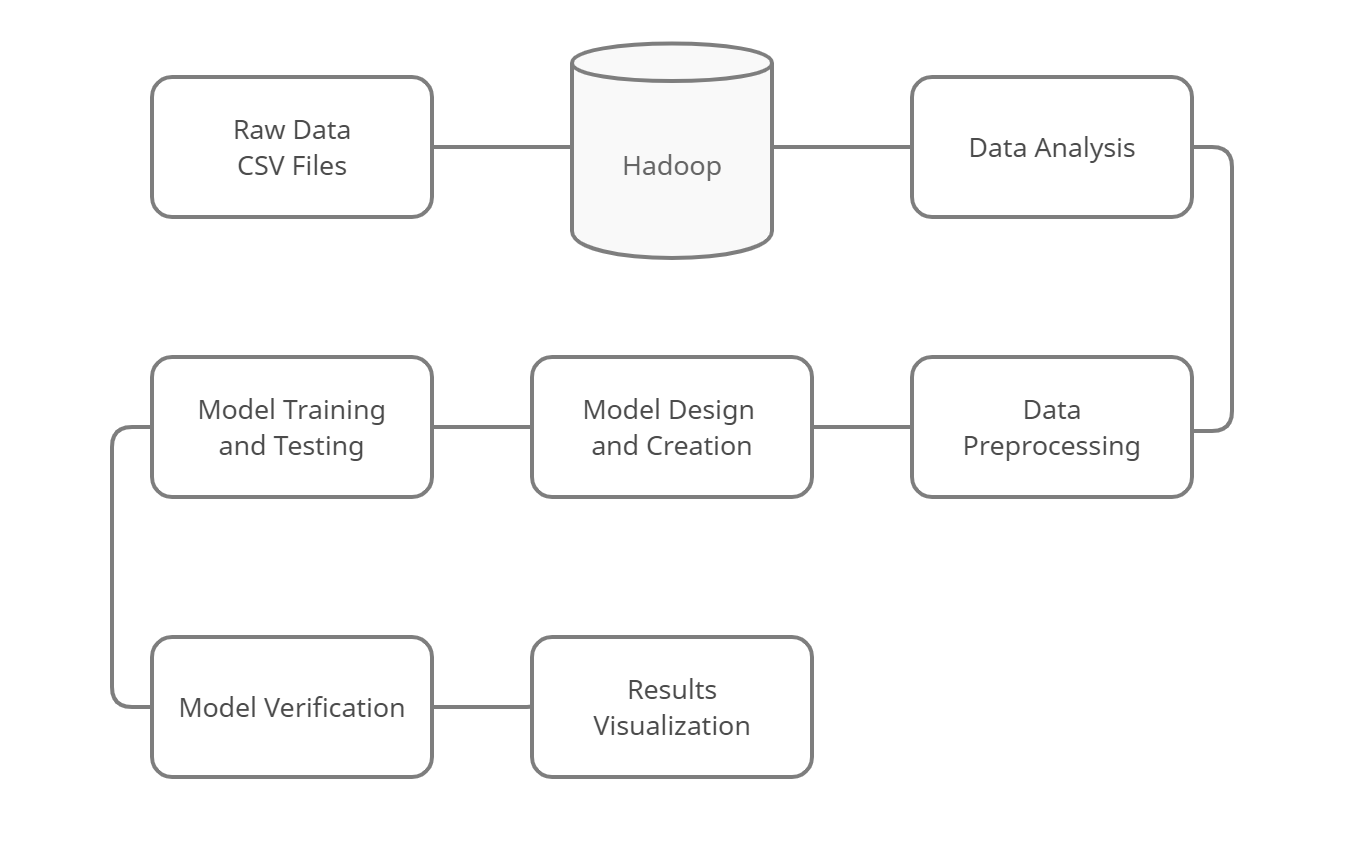
The processes of creating the solution for our project began with a set of CSV files. The data in the CSV files was stored in Hadoop. This process was done both in Cloudera, directly into Hadoop, and later via Spark.

The first half of the project effort was focused on utilizing the tools learned in the first half of the semester to perform data analysis of the data. This was done using Hive, MapReduce, Sqoop with MySQL and Hadoop. Insights into the data gave us an overview of the data that was to be utilized. This research highlighted an issue with the date time field and how it was configured, which we addressed in the second half of the project.

The second half of the project was focused on leveraging Spark. The first step which was taken was to “clean” the data so that it would be ready for analysis. Records with missing values were dropped from the datasets. The date time field, which included a time offset, had to be converted to a timestamp field. The two datasets were joined together based on the newly created timestamp field.

The Machine Learning module of the class was at the end of the semester, which required the team to investigate this topic on their own. The team created several models to predict energy demand based on features present in the weather dataset. These ranged from the straight forward linear regression to the complex Long Short Term Memory model. Ridge, Lasso, and Elastic regression models were created as well.

After a model was created the data was split into training and test sets. The model was trained on the training data and the accuracy of the model was checked on the test set. Finally the model was used to verify the prediction compared to the actual for the entire data set.

****

**Dataset**

[**Hourly energy demand generation and weather | Kaggle**](https://www.kaggle.com/nicholasjhana/energy-consumption-generation-prices-and-weather)

This dataset contains 4 years of electrical consumption, generation, pricing, and weather data for Spain. Consumption and generation data were retrieved from ENTSOE a public portal for Transmission Service Operator (TSO) data. Settlement prices were obtained from the Spanish TSO Red Electric España. Weather data was purchased as part of a personal project from the Open Weather API for the 5 largest cities in Spain and made public on Kaggle..

**Analysis of Data:**

Data preprocessing began in iteration two. Utilizing Hive, MapReduce, Sqoop with MySQL, and storing the data in Hadoop, the team performed various queries on the data. The preprocessing and analysis of the data continued into the first part of iteration three. Records and Columns with missing data were removed from the dataset. The issue concerning the date time field was resolved in the beginning of iteration three as well.

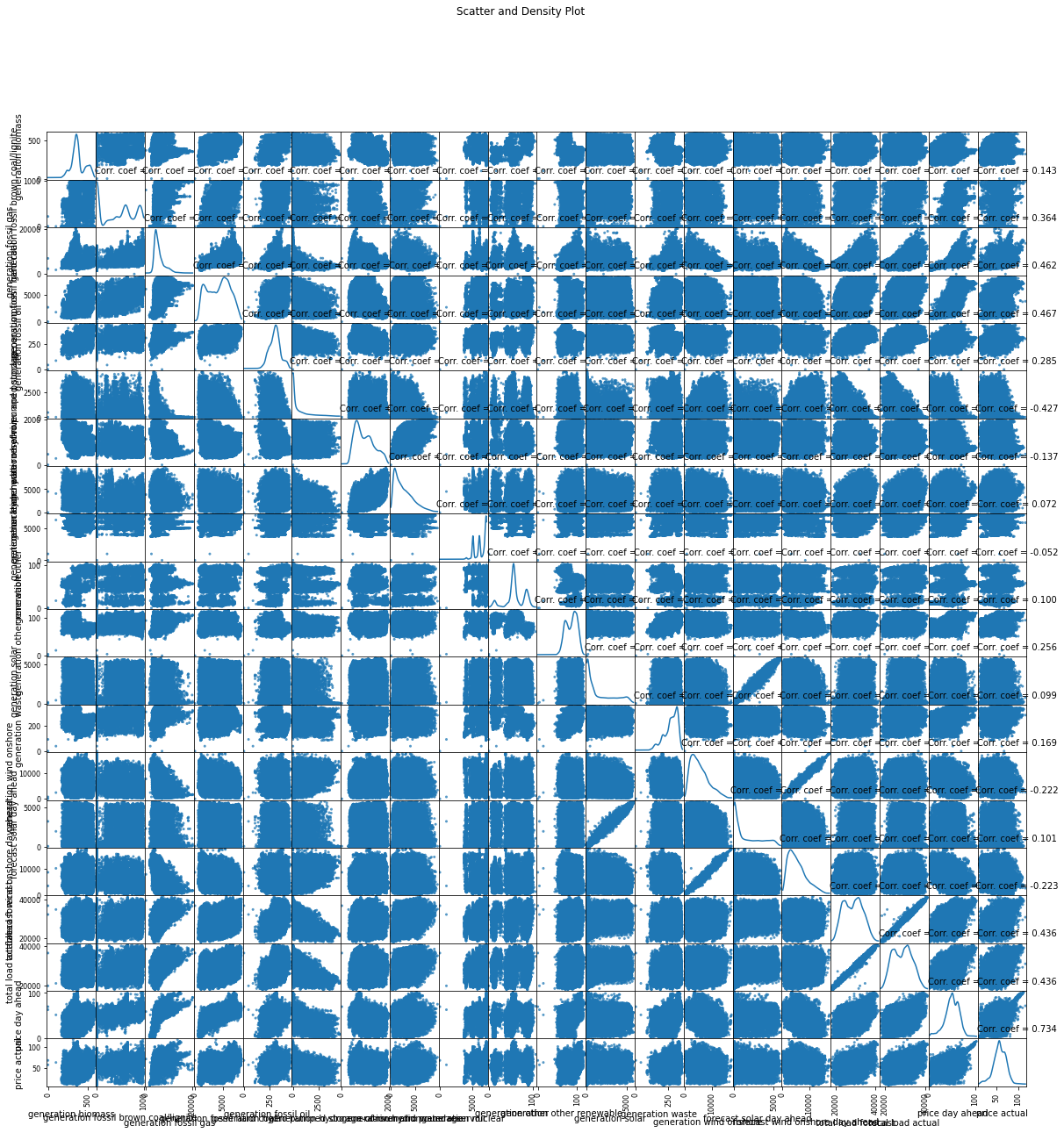
**Weather Data Summary**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **temp** | **temp\_min** | **temp\_max** | **pressure** | **humidity** |  |
| **Min.** | **262.2** | **262.2** | **262.2** | **0** | **0.00** |  |
| **1st Qu.** | **283.7** | **282.5** | **284.6** | **1013** | **53.00** |  |
| **Median** | **289.1** | **288.1** | **290.1** | **1018** | **72.00** |  |
| **Mean** | **289.6** | **288.3** | **291.1** | **1069** | **68.42** |  |
| **3rd Qu.** | **295.1** | **293.7** | **297.1** | **1022** | **87.00** |  |
| **Max.** | **315.6** | **315.1** | **321.1** | **1008371** | **100.00** |  |
|  |  |  |  |  |  |  |
|  | **wind\_speed** | **wind\_deg** | **rain\_1h** | **rain\_3h** | **snow\_3h** | **clouds\_all** |
| **Min.** | **0.00** | **0.00** | **0.00** | **0.00** | **0.00** | **0.00** |
| **1st Qu.** | **1.00** | **55.00** | **0.00** | **0.00** | **0.00** | **0.00** |
| **Median** | **2.00** | **177.00** | **0.00** | **0.00** | **0.00** | **20.00** |
| **Mean** | **2.47** | **166.60** | **0.08** | **0.00** | **0.00** | **25.07** |
| **3rd Qu.** | **4.00** | **270.00** | **0.00** | **0.00** | **0.00** | **40.00** |
| **Max.** | **133.00** | **360.00** | **12.00** | **2.32** | **21.50** | **100.00** |

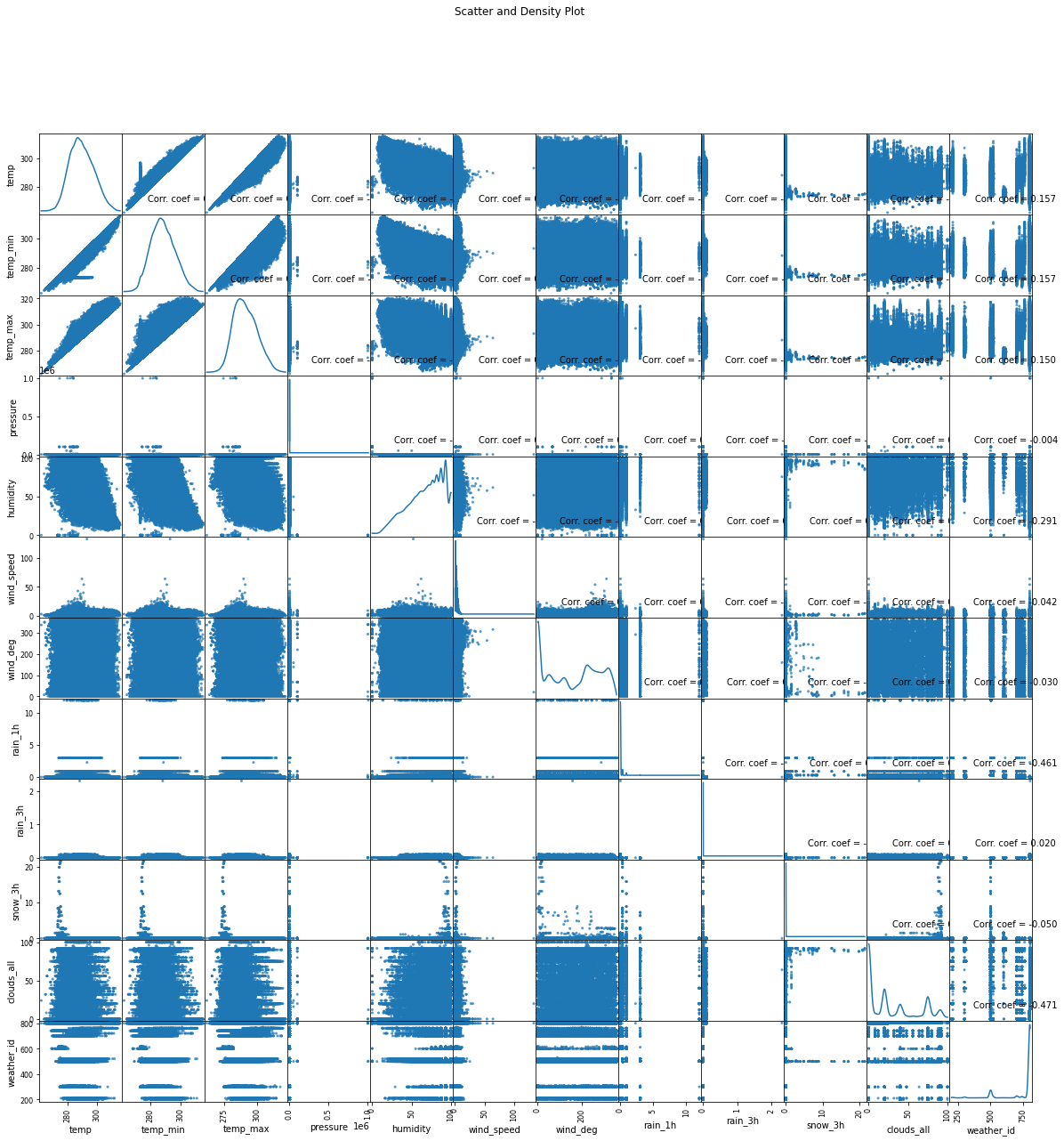
**Energy Data Summary**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Generation**  **biomass** | **Generation**  **Fossil hard coal** | **Generation hydro pumped storage consumption** | **Generation hydro water reservoir** | **Generation fossil brown coal lignite** |
| **Min.** | **0** | **0** | **0** | **0** | **0** |
| **1st Qu.** | **333** | **2527** | **0** | **1077** | **0** |
| **Median** | **367** | **4474** | **68** | **2164** | **509** |
| **Mean** | **383.5** | **4256** | **475.6** | **2605** | **448.1** |
| **3rd Qu.** | **433** | **5839** | **616** | **3757** | **757** |
| **Max.** | **592** | **8359** | **4523** | **9728** | **999** |
| **NA's** | **19** | **18** | **19** | **18** | **18** |
|  | **Generation waste** | **Forecast wind onshore day ahead** | **Generation wind onshore** | **Generation other** | **Generation fossil oil** |
| **Min.** | **0** | **237** |  | **0** | **0** |
| **1st Qu.** | **240** | **2979** | **0** | **53** | **263** |
| **Median** | **279** | **4855** | **2933** | **57** | **300** |
| **Mean** | **269.5** | **5471** | **4849** | **60.23** | **298.3** |
| **3rd Qu.** | **310** | **7353** | **5464** | **80** | **330** |
| **Max.** | **357** | **17430** | **7398** | **106** | **449** |
| **NA's** | **19** |  | **17436** | **18** | **19** |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  | **Forecast solar day ahead** | **Generation fossil gas** | **Generation other renewable** | **Generation nuclear** | **Generation hydro run of river and poundage** |
| **Min.** | **0** | **0** | **0** | **0** | **0** |
| **1st Qu.** | **69** | **4126** | **73** | **5760** | **637** |
| **Median** | **576** | **4969** | **88** | **6566** | **906** |
| **Mean** | **1439** | **5623** | **85.64** | **6264** | **972.1** |
| **3rd Qu.** | **2636** | **6429** | **97** | **7025** | **1250** |
| **Max.** | **5836** | **20034** | **119** | **7117** | **2000** |
| **NA's** |  | **18** | **18** | **17** | **19** |
|  | **Total load forecast** | **Total load actual** | **Price day ahead** | **Price actual** |  |
| **Min.** | **18105** | **18041** | **2.06** | **9.33** |  |
| **1st Qu.** | **24794** | **24808** | **41.49** | **49.35** |  |
| **Median** | **28906** | **28901** | **50.52** | **58.02** |  |
| **Mean** | **28712** | **28697** | **49.87** | **57.88** |  |
| **3rd Qu.** | **32263** | **32192** | **60.53** | **68.01** |  |
| **Max.** | **41390** | **41015** | **101.99** | **116.8** |  |
| **NA's** |  | **36** |  |  |  |

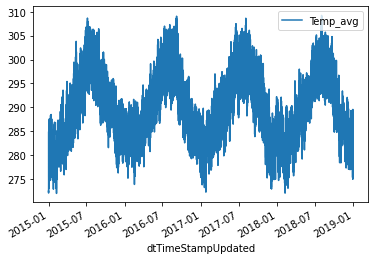
Relationship of the Energy Dataset Features



Relationship of the WeatherDataset Features

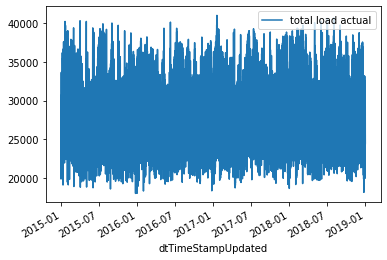


Average Temperature over the five cities:



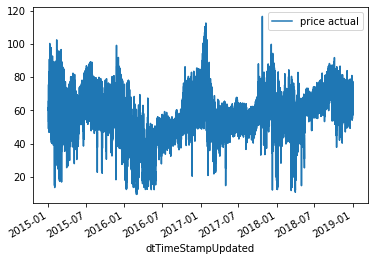
The previous graph shows the average temperature of the five cities over the time frame the data covers. It follows a known pattern of the temperatures being hotter during the summer seasons and colder during the winter seasons. The ten to fifteen degree range of temperature is consistent and reflects the rise and fall of the temperature during a twenty four hour period.

Total Actual Load:



The previous graph shows the Total Actual Load of energy over the time frame the data covers. The data appears very chaotic. It does not appear to follow a pattern. The load appears to fluctuate by a degree of twenty thousand units almost daily. There is some consistency in the data in that the range of Total Actual Load appears to be bounded and does not appear to be either steadily increasing or decreasing over time..

Actual Price:

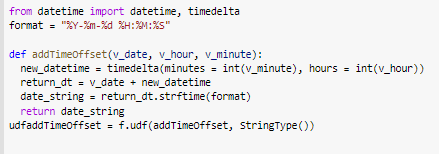


The previous graph shows the price of energy over the time frame the data covers. It does not appear to be following a pattern. There are several spikes in the data, both on the high and low ends of the range. There does appear to be a trend of the price on average increasing from about May of 2016. The daily range in values also narrows staring at this time as well. There are other economical factors which effect the price of energy over time, such as global economic boom and bust cycles.

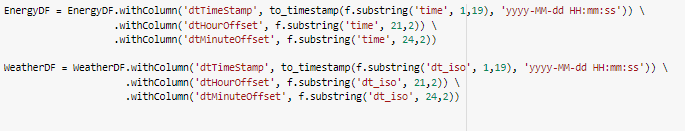
**Features Implementation**

**Resolving the issue with the Date Field**

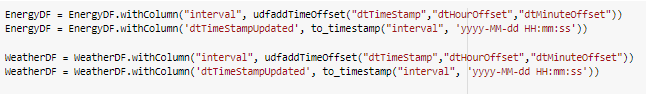
**Code to Add the time offset into the datefield:**

****

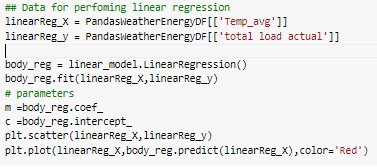
**Code to convert string to timestamp:**

****

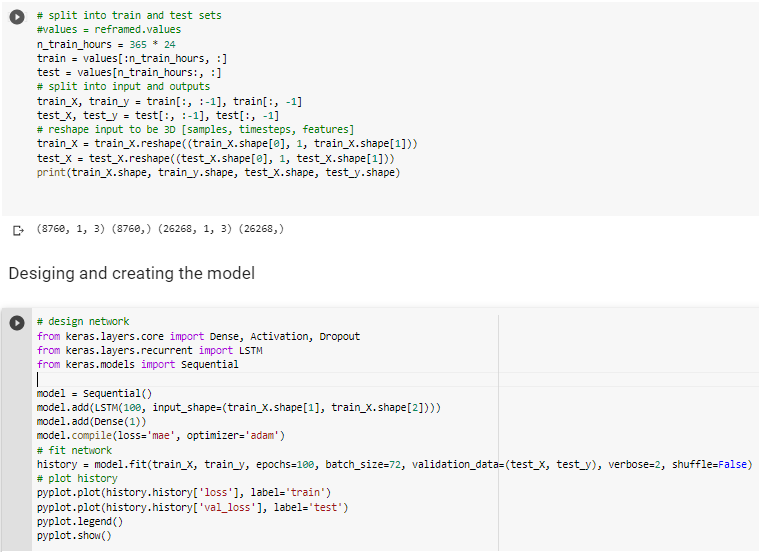
**Apply the time offset to the timestamp fields**

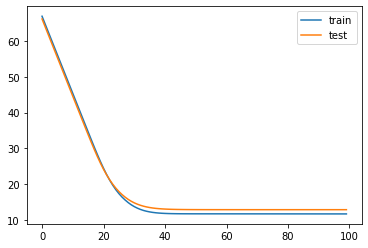
****

**Linear Regression**

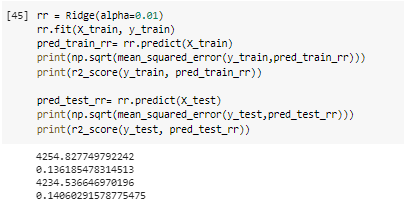
****

**Long Short Term Memory Model:**

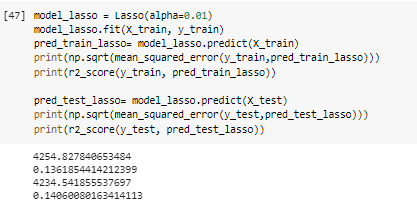
****

****

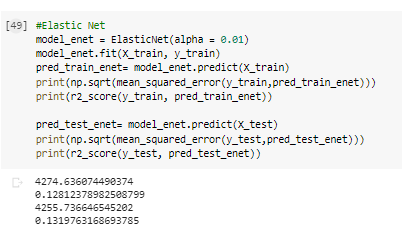
**Ridge Model:**

****

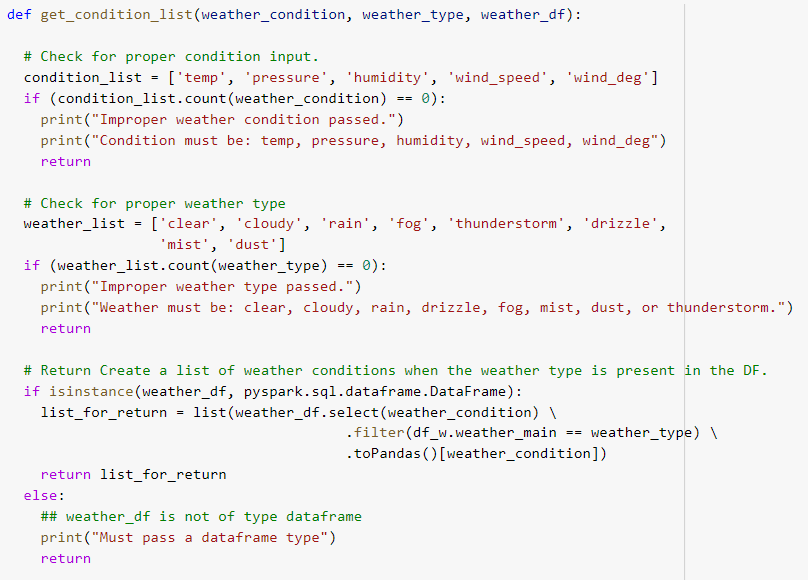
**Lasso Model:**

****

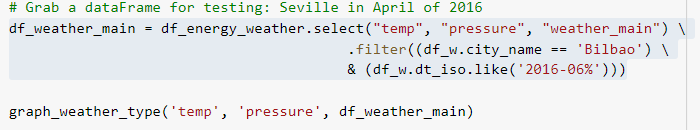
**Elastic Model:**

****

**Plot Weather Conditions for 2 Weather Variables:**

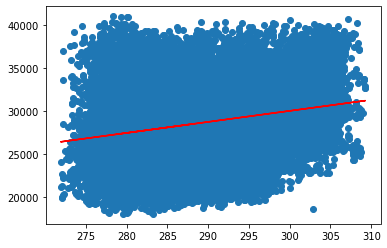
****

****

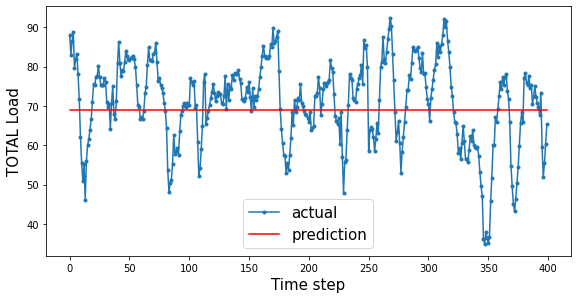
****

**Results :**

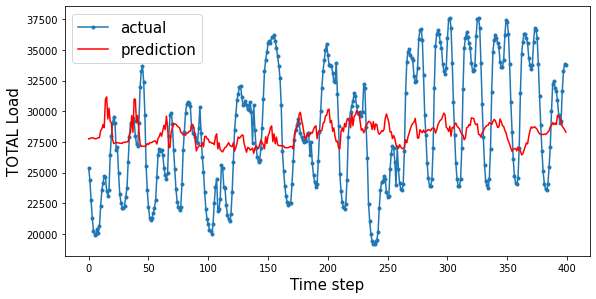
**Linear Regression Temperature ~ Total Actual Load:**

****

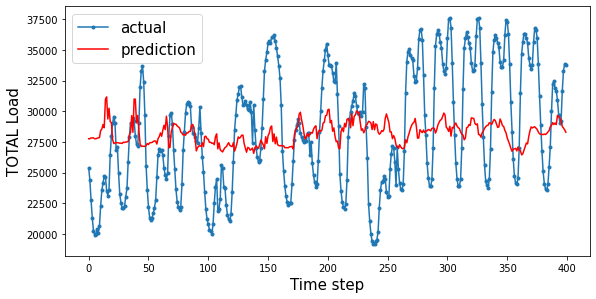
**LSTM [Temperature,Pressure,Humidity~Total Actual Load]:**

****

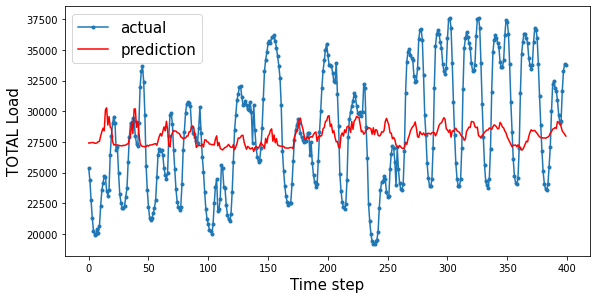
**Ridge Regression [Temperature,Pressure,Humidity~Total Actual Load]:**

****

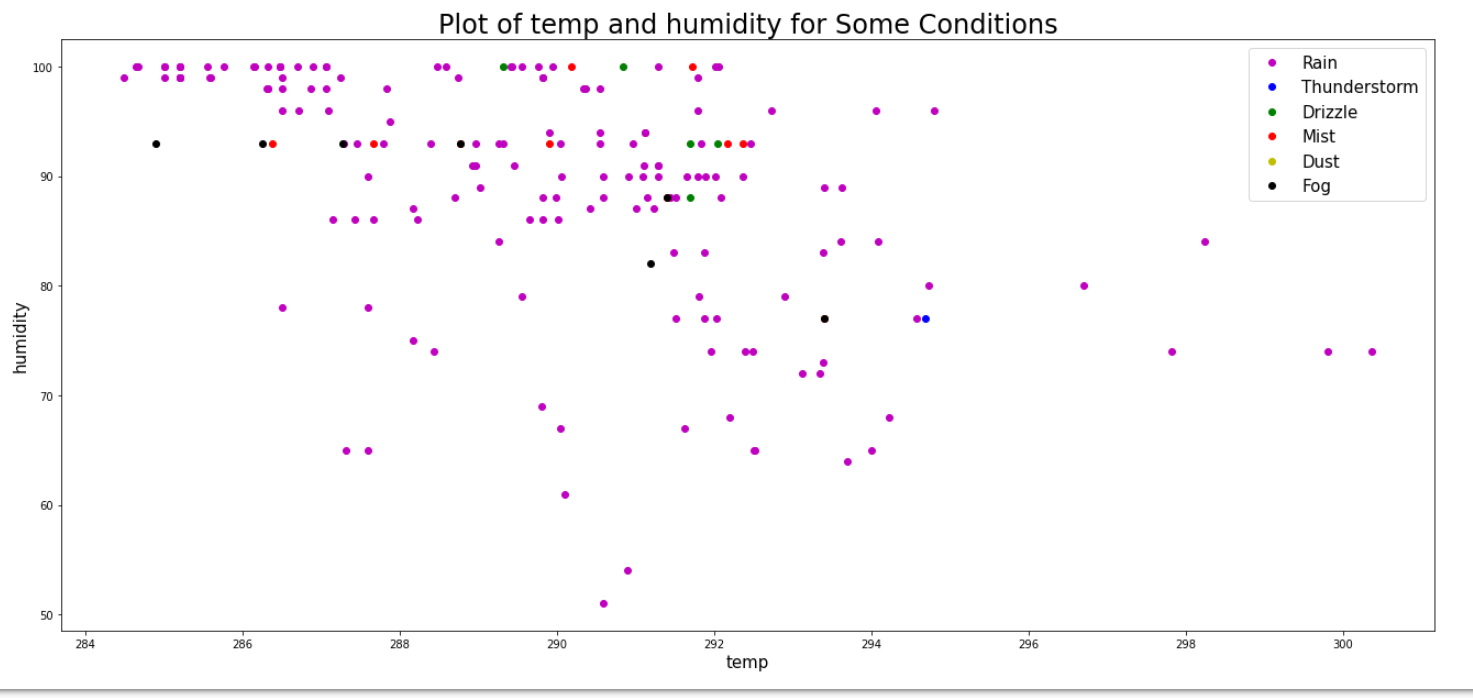
**Lasso Regression [Temperature,Pressure,Humidity~Total Actual Load]:**

****

**Elastic Regression [Temperature,Pressure,Humidity~Total Actual Load]:**

****

**Weather Type Plot:**

****

**Future Work**

This project afforded the team an excellent opportunity to explore various Big Data tools and Technologies. The concepts of how to store and query the data from the first half of the semester were very straightforward and were relatable to previous knowledge we had gained. The second half of the semester with Spark and the various components of Spark were more complex. The area of Machine Learning and the various types of prediction models is definitely an area the team plans to explore and learn more about. The models we were able to create reflect the “novice” level of understanding of the subject, which is understandable given the amount of exposure the team had to it.

Future work and plans would include to expand on the level of ML knowledge, both via coursework and independent learning. Improve or even create new and different prediction models. Locate other sources of data to be able to build better and more robust models. Explore concepts of semily dissimilar areas being able to predict the behavior of other areas, such as weather and pollution, or movie preferences based upon consumer spending patterns.

**Story Telling:**

**Life**

1. **Who** are the people or communities in need of help?

Consumers of energy utility companies (gas and electricity) are one portion of people who need help, as the recent crisis in Texas has demonstrated. The other people who need help are the producer of the energy being consumed (Utility Companies) and their suppliers. Being able to accurately forecast demand can help them better plan on how much to produce, how much raw materials to keep in stock, and how to better plan to use their resources, including labor, to meet their client’s demands. In the case where it is not possible for utility companies to keep up with demand, adequate warning could be given to energy consumers so they can plan for loss of power.

2. **What** problem happened to them?

Recently in Texas with the severe cold weather, the demand for electricity far exceeded the capacity the Utility Companies were able to provide. In addition, the cold weather caused some producers of gas to halt production. If the Utility Companies could have foreseen the spike in demand, it may have been possible for them to take actions to mitigate the negative impacts which were caused by the shortages in energy as compared to the demand that was required to keep people warm.

3. **When** did the problem take place?

This recent issue occurred in February 2021. There were also rolling blackouts in Kansas City Missouri in February 2021 because of the cold. There have been rolling blackouts in California because of the heat over numerous years.

4. **Where** means two things:

a. The environment and settings that the people or the community is living in, and

Due to a rapidly changing climate, these problems can affect a wide range of environments. Communities that are not accustomed to extreme cold or hot weather are particularly vulnerable to energy shortfalls when hit with unexpected demand. Notably in the case of Texas, homes there were designed to shed heat, so when they were hit with uncharacteristically cold weather, the infrastructure was especially unequipped to deal with power outages.

b. The place/location where the problem takes place.

The problem of energy shortages can occur anywhere, in any country. The recent cases have been in Kansas City Missouri and in Texas. They have occurred in California. They can occur anywhere.

5. **Why** means the possible causes and/origin of the problem.

The inability of the Utility Companies to accurately predict demand and a lack of excess capacity to handle a reduction in production capacity by outside forces results in a failure to meet the demands of the consumer. IT companies have HA and redundancy built into their server farms to help prevent outages of their services. Utility companies need to have the same HA and redundancy built into their systems, and they need to understand and be able to reasonably predict how much demand there will be from their consumers.

6. **How**: If you would like, you can add a dimension of how. How did it happen? Sometimes, the answer to how can be covered by what, when, and where.

In the case of Texas, lawmakers were warned years in advance of the possibility of this situation happening. They ignored recommendations that they winterize their energy infrastructure, and that contributed greatly to the energy shortages as some systems failed. Analysis of energy consumption vs. weather patterns could provide the necessary information to utility companies and lawmakers in advance of disastrous situations and could help hold them accountable for their lack of preparedness.

**Workflow:**

Gain knowledge and experience with Big Data Programming tools.

Select the problem the team wishes to solve.

Research aspects of the problem, data, tools, why, benefits, issues.

Select the tools the team wishes to use to solve the problem.

[Repeat following steps as necessary]

Refine the data associated with the problem.

Design Model for solving the problem.

Construct Model for solving the problem.

Test/validate Model for solving the problem.

Create a presentation of the solution the team has created.

**Project Management:**

**Implementation Status Report**

|  |  |  |
| --- | --- | --- |
| **Task Name** | **Assigned To** | **Iteration** |
| Store Weather data set in Hadoop | Joe Goldsich | 2 |
| Analyze Weather data set using Hive | Joe Goldsich | 2 |
| Analyze Weather data set using MapReduce | Joe Goldsich | 2 |
| Store data sets in Hadoop | Anna Johnson | 2 |
| Export Data out of Hadoop and into MySQL using Sqoop | Anna Johnson | 2 |
| Export Data out of MySQL and into Hadoop using Squoop | Anna Johnson | 2 |
| Analyze the data in MySQL | Anna Johnson | 2 |
| Store data sets in Hadoop | Kyle Son | 2 |
| Analyze Merged data using Hive | Kyle Son | 2 |
| Analyze Merged data using Cassandra | Kyle Son | 2 |
| Store Energy data set in Hadoop | Bill Yerkes | 2 |
| Analyze Energy data set using Hive | Bill Yerkes | 2 |
| Analyze Energy data set using MapReduce | Bill Yerkes | 2 |
| Convert string type to date-type in datasets | Joe Goldsich | 3 |
| Plotting Graphs of Query results on datasets | Joe Goldsich | 3 |
| Graphing Madrid's January 2015 actual and predicted energy loads | Joe Goldsich | 3 |
| Plot Sevilles' April 2016 weather\_main relative to humidity and pressure | Joe Goldsich | 3 |
| Altering previous code into more repeatable functions | Joe Goldsich | 3 |
| Create KNN classification for Various Column Combinations | Joe Goldsich | 3 |

**Implementation Status Report Continued**

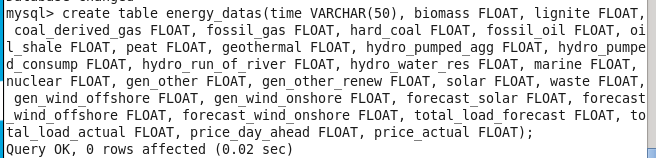
|  |  |  |
| --- | --- | --- |
| **Task Name** | **Assigned To** | **Iteration** |
| Plot the average temp of the city | Kyle Son | 3 |
| Plot the count of the city and avg price by city | Kyle Son | 3 |
| Plot the graph contain relationship between weather status and average\_price | Kyle Son | 3 |
| Hourly analysis of average price | Kyle Son | 3 |
| Hourly analysis of average load | Kyle Son | 3 |
| Hourly analysis of wind speed and humidity | Kyle Son | 3 |
| Add Code from Kaggle to Project to review and learn from | Bill Yerkes | 3 |
| Figure out how to read Data in Colab directly from GitHub | Bill Yerkes | 3 |
| Convert column with string of time to timestamp | Bill Yerkes | 3 |
| Create function to Add x hours to timestamp field | Bill Yerkes | 3 |
| Create function to add time offset back to timestamp field | Bill Yerkes | 3 |
| Implement an LSTM model | Bill Yerkes | 3 |
| Implement a Linear regression | Bill Yerkes | 3 |
| Implement a Lasso regression | Bill Yerkes | 3 |
| Implement a Ridge regression | Bill Yerkes | 3 |
| Create Kmeans clustering model and 3D visual | Bill Yerkes | 3 |
| Knn analysis of merged data set for energy and weather features | Anna Johnson | 3 |
| MapReduce functions for both energy and weather features dataset using RDD’s and Pyspark’s built-in MapReduce functionality | Anna Johnson | 3 |
| Analysis of datasets in Pyspark | Anna Johnson | 3 |
| Visualization of data in Pyspark | Anna Johnson | 3 |
| Querying energy and weather features datasets in Pyspark to pull useful information to inform our modeling | Anna Johnson | 3 |
| Merging the datasets in Pyspark so they could be queried together | Anna Johnson | 3 |
| Generalize Graphing weather\_main based on two numeric weather conditions with functions | Joe Goldsich | 3 |

**Working screens from project:**

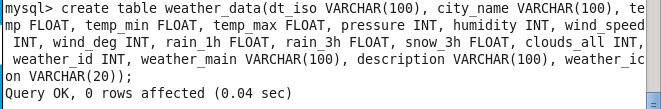
**Anna Johnson:**

Create tables in MySQL:

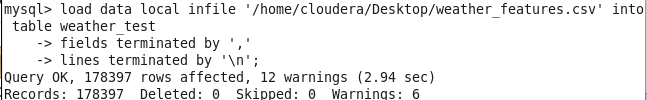
Energy dataset:



Weather dataset:

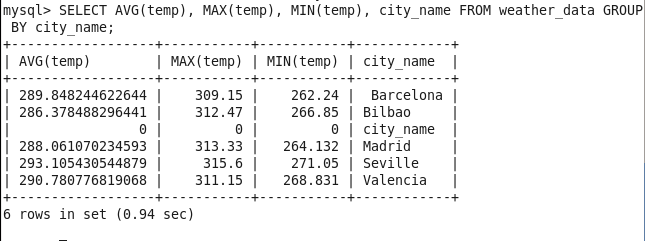


Load Data into MySQL tables from csv files:



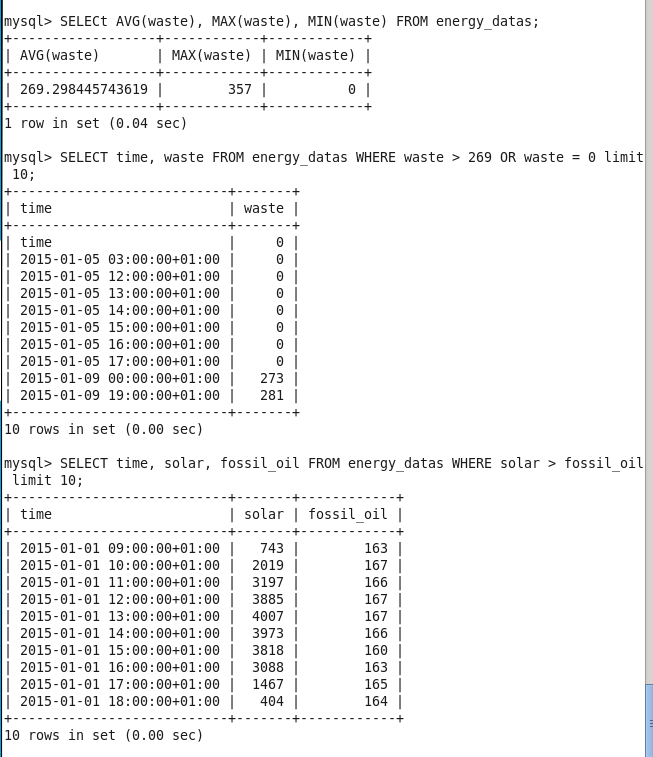
Run Queries on data in MySQL:

Query on weather dataset to get statistics on weather for each city:



Queries on energy dataset:

1. Find the average, min, and max for the ‘waste’ column
2. Find times where there was above average waste, or no waste at all
3. Find times where more energy was generated from solar sources than from fossil oil sources

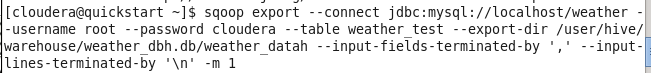


These queries helped us to gain a better insight into our datasets. We can apply this knowledge in the next steps of our project to identify important features to compare between the energy and weather datasets to determine the relationship between them.

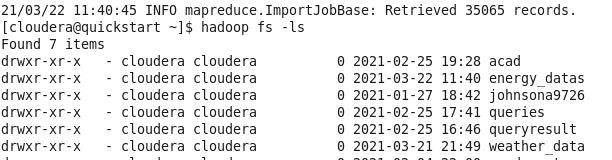
Lastly, it’s important that we are able to move these datasets between Hadoop and MySQL so that we can utilize the different functionality of both the Hive and MySQL databases. This can be done using Sqoop.

Exporting and importing tables between Hadoop, Hive, and MySQL:



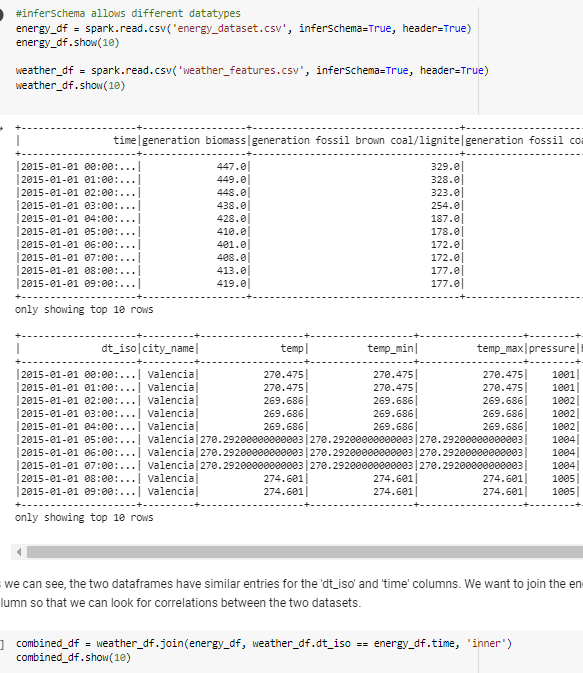


The MySQL tables in HDFS after importing using sqoop:



KNN Analysis in Pyspark:

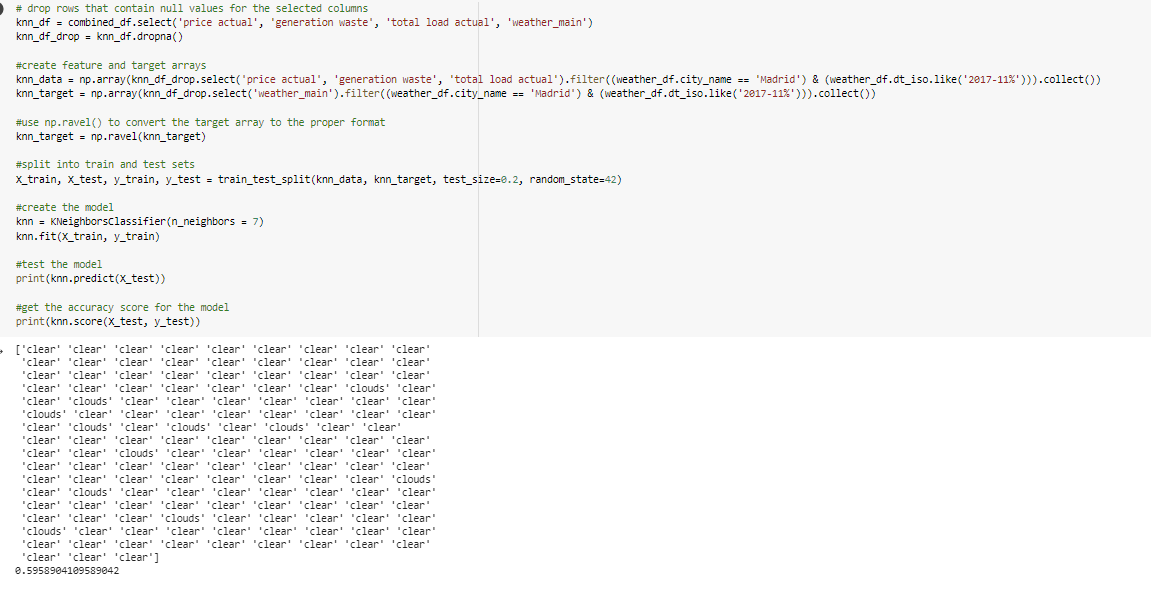
I created the dataframes in Pyspark so that they could be used for KNN modeling.



Then I used KNN analysis to determine the relationships between various features in the weather and energy datasets. Higher accuracy scores from the model correspond to a stronger correlation between the features.

An example of one of the models:



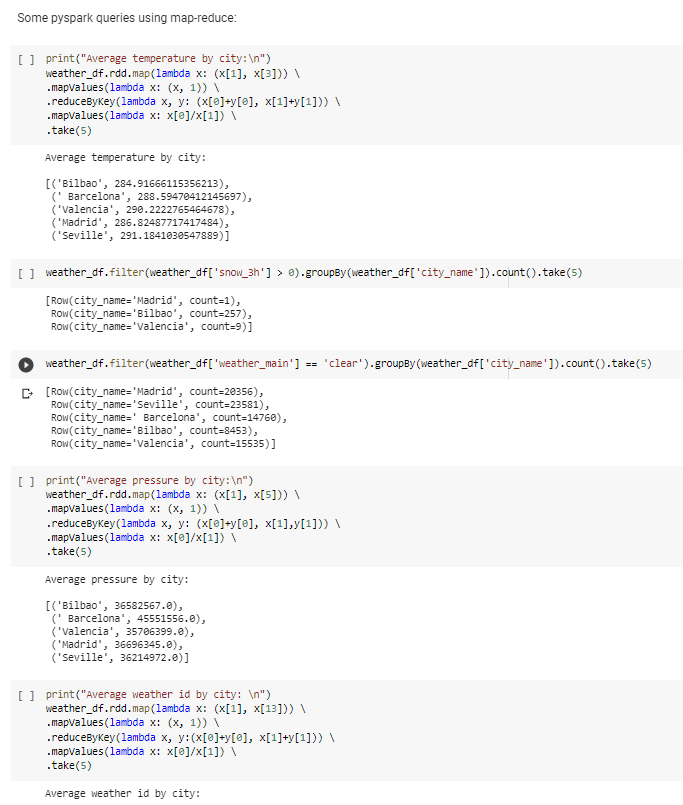


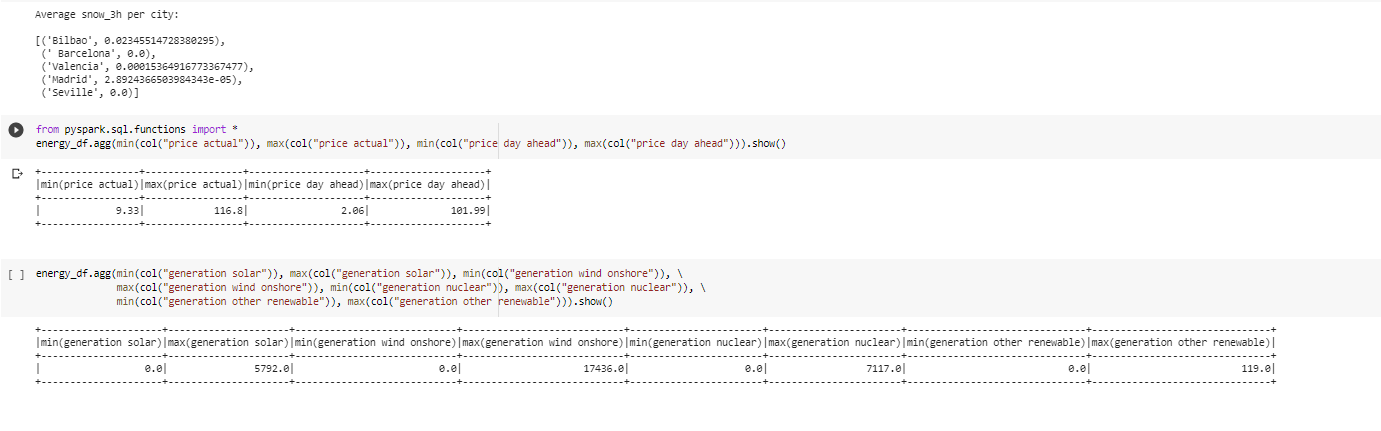
The KNN analysis showed that there was in fact a correlation between some of the columns in these two datasets. The code and output for the rest of the models can be found in the github repository for the project.

MapReduce:

I wanted to utilize the power of MapReduce functions to run several queries on our large datasets. I ran similar queries to the ones that I worked on in MySQL for increment 2, however I was able to work with larger datasets at much faster speeds by making the dataframe into RDD’s and utilizing Pyspark’s built-in MapReduce functionality.

Some examples of queries:

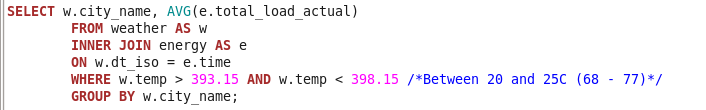




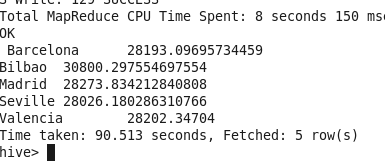
This was useful as it allowed us to get a better idea of what variables we should use in our modeling, as well as what kind of models to build.

**Joe Goldsich:**

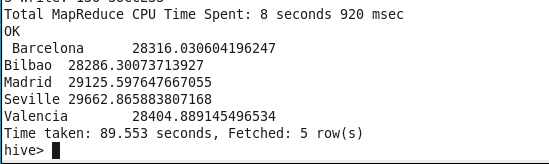
An example of a simple query on a join of the two data sets using HiveQL:



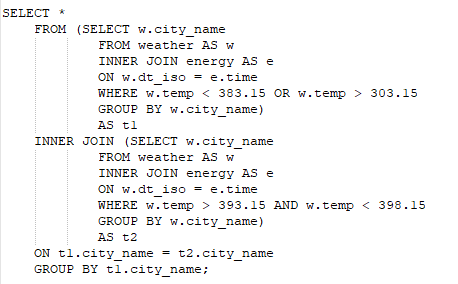
And the results of this simple query:



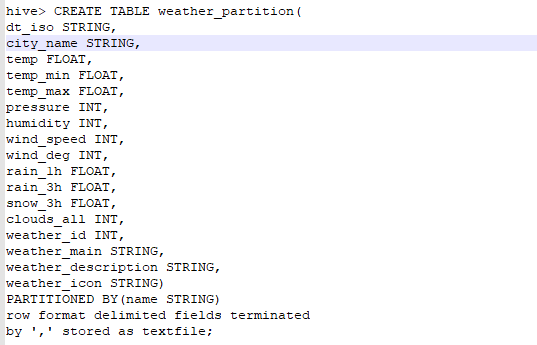
With the exception of Bilbao (which has a very temperate climate) most of the cities experienced an increase in their energy usage at more extreme temperatures (below 10C and over 30C):

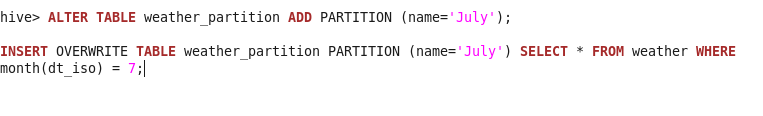


A slightly more complicated Select query was very slow to run (10 minutes slow).

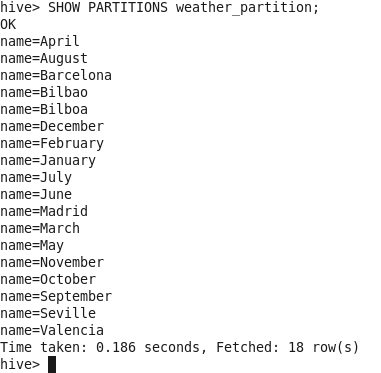


Creating and loading partitions into a new table for the weather dataset to look for performance gains:

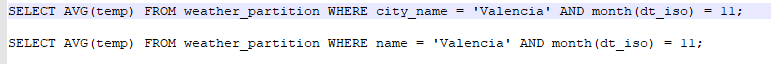




The result of all the partitions:

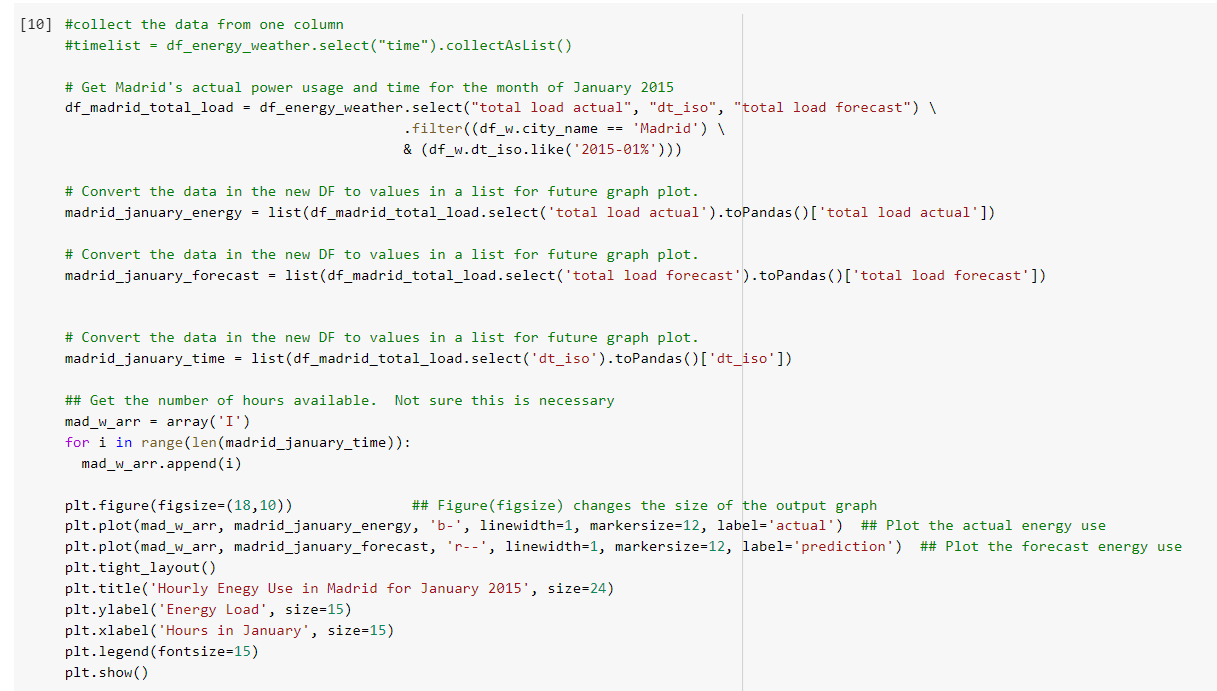


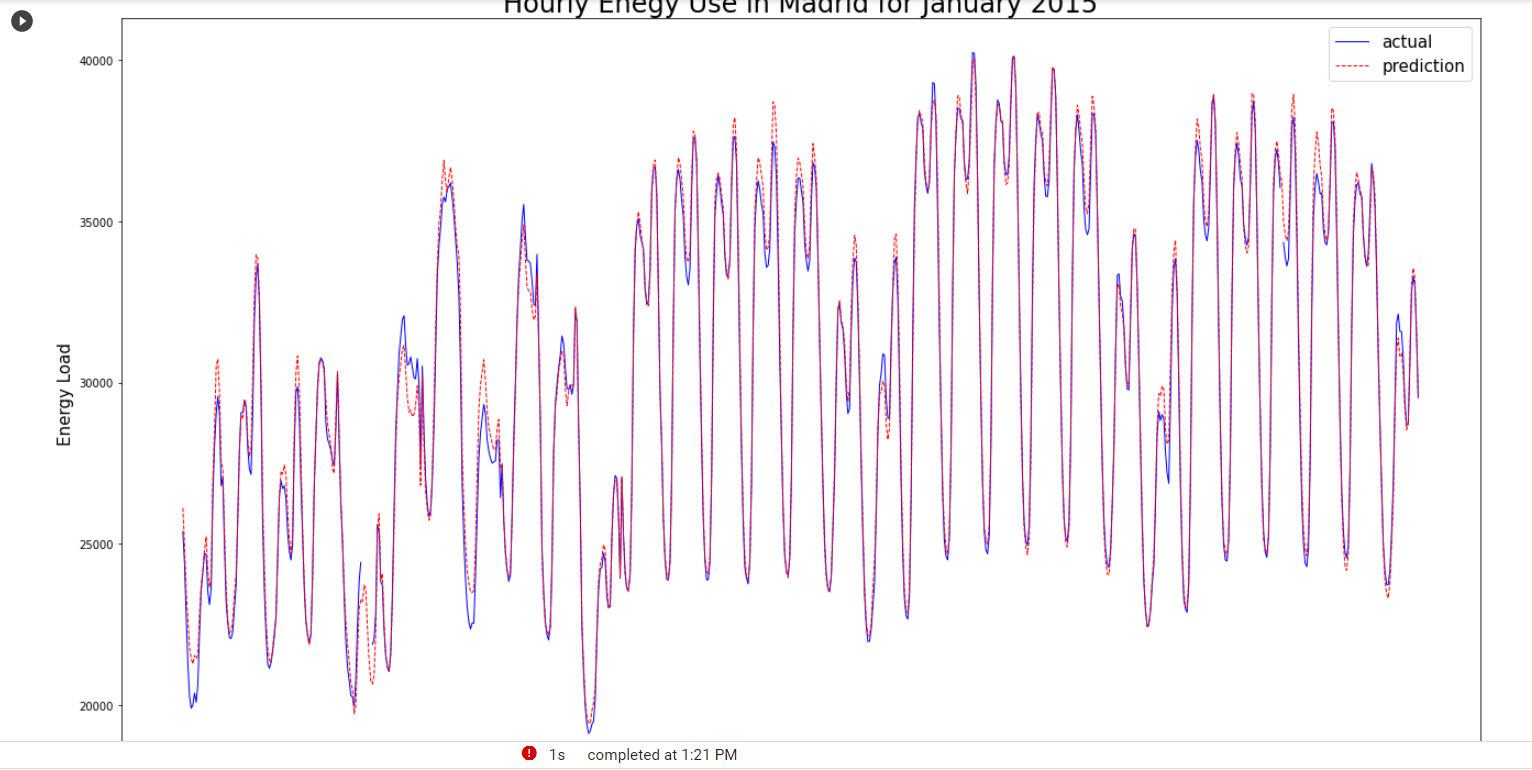
And some queries using the partitions and not using them. This did result in significantly quicker query times (~58seconds down from ~99 seconds).



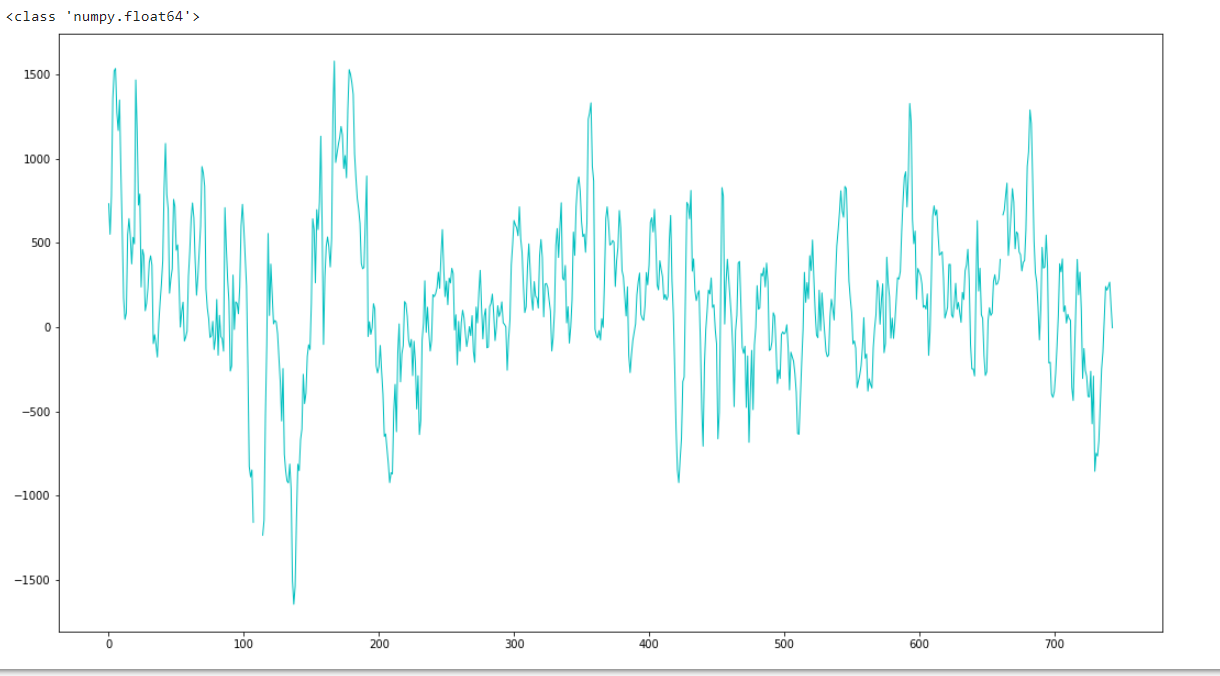
Once we moved onto the portion of the course that taught us Spark. I found using Dataframes was the most efficient way to manipulate the data that we had. The data was visualized with the use of the ‘matplotlib.pyplot’ library. I would primarily utilize Google Colabs for the rest of the work.

For our goal of trying to better predict energy usage based on weather information, I thought it would be good to get a baseline from which to compare the efficacy of our predictions. I created a subset of the total joined dataset for Madrid in January of 2015 (This period was chosen randomly). I then charted the actual energy load with the forecasted energy load.



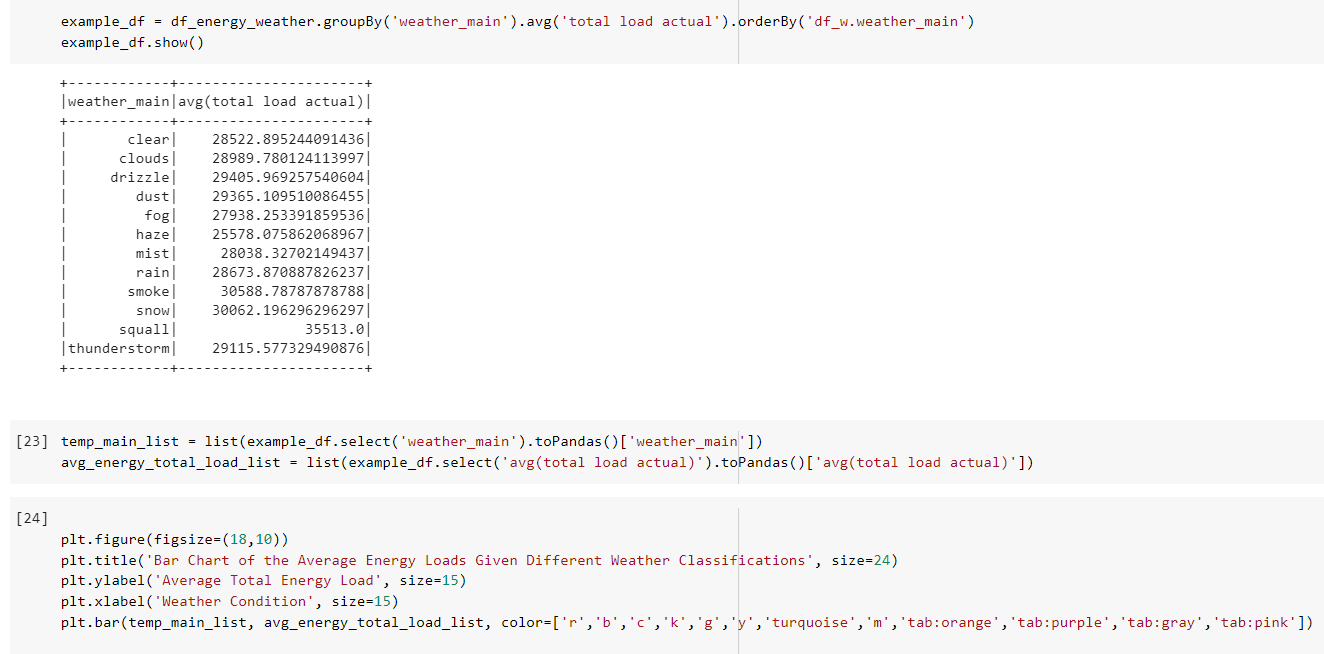


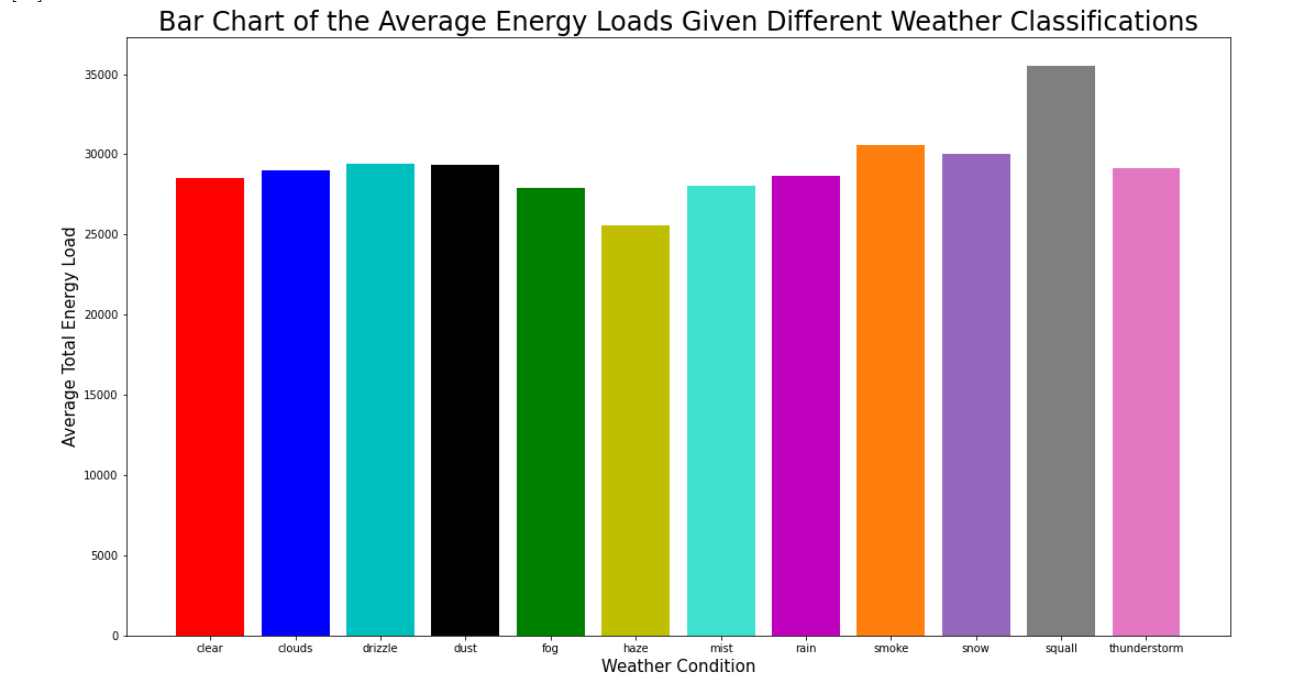
Then I charted the difference.



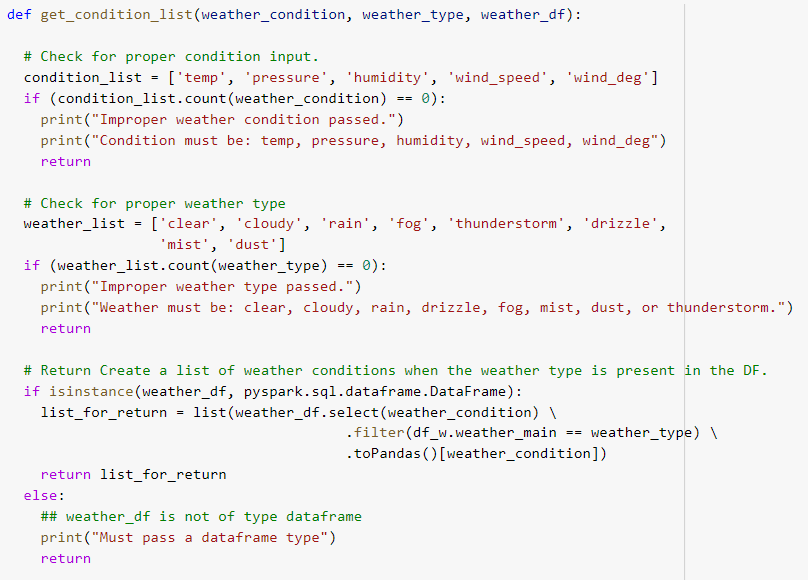
Ideally, we would want to flatten that line.

I compared total energy load with weather conditions. The column called ‘weather\_main’ in the data used a verbal description to classify the weather each hour. The possibilities were: clear, rain, cloudy, thunderstorm, drizzle, mist, dust, and fog. I created a bar graph to show how each condition related to energy load.

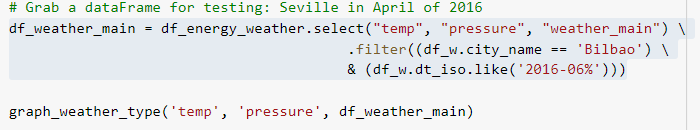


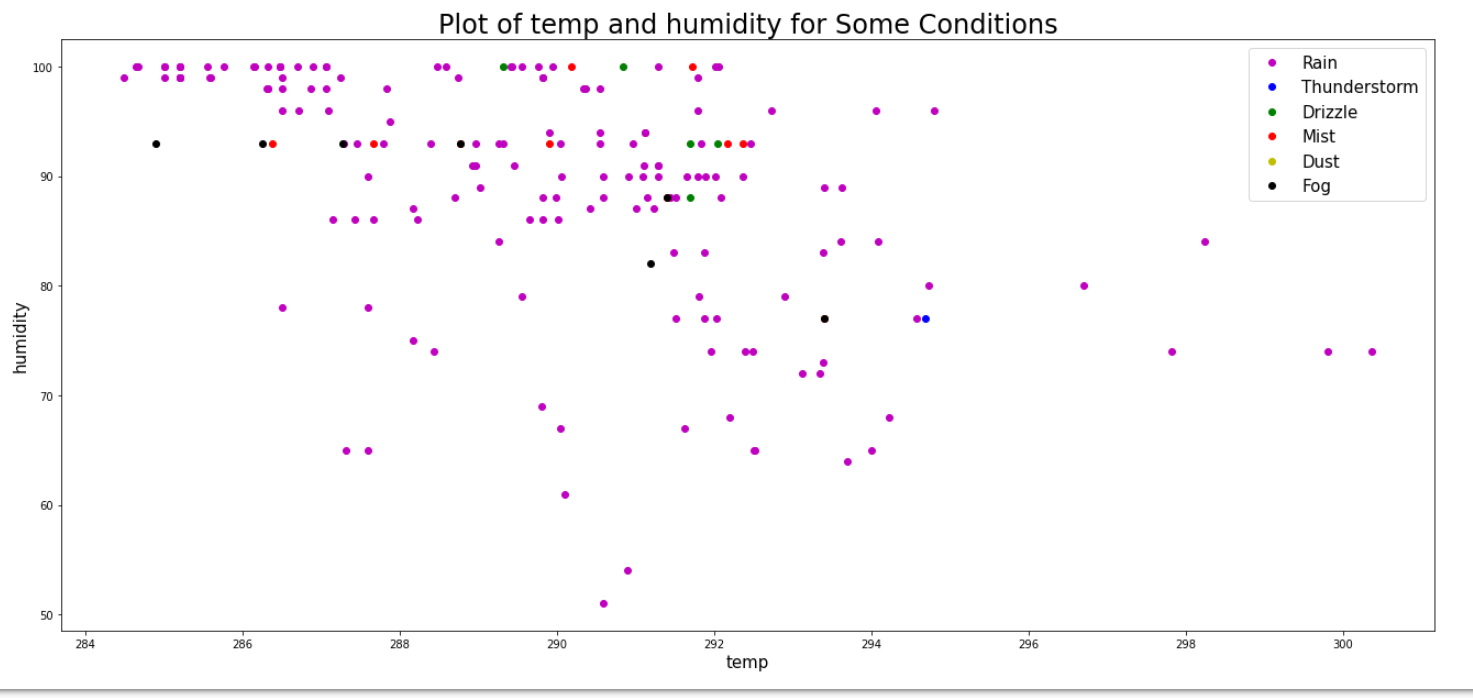


I thought these results indicated that this seemed a promising avenue to pursue. So, next I wanted to see if there was any kind of correlation between the numerically measured weather variables in our data set. So I wrote a couple of functions that together would allow us to print out a scatter plot of the weather condition (from ‘weather\_main’) based on two of the numerical columns in the set (‘pressure’, ‘temp’, ‘humidity’, ‘wind\_speed’, ‘wind\_deg’).

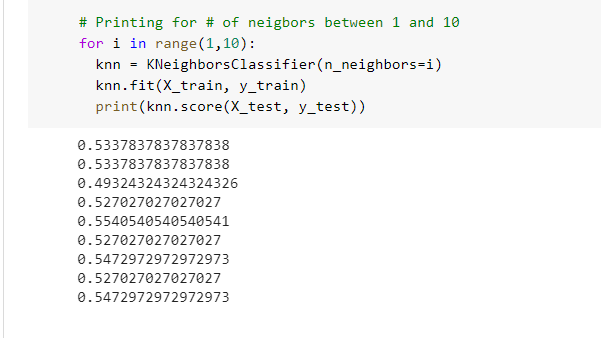


These two allow for the following:





While this looked promising, a more thorough analysis with knn functions indicated these were poor predictors of the weather condition.



Given more time I would like to explore some of the newer machine learning algorithms that we learned near the end of the course. I think that a classification algorithm like Naive Bayes may be better at predicting the weather, and then in turn make for better energy use predictions.

**Kyle Son**:

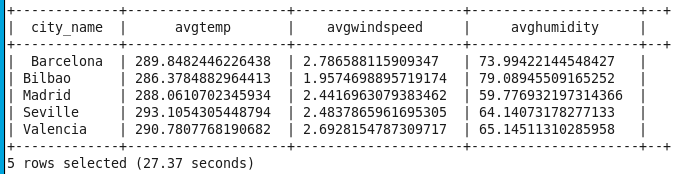
**Hive(Beeline)**

In hive, I used beeline command for the better visualization of the table then, I merged two tables to named merged

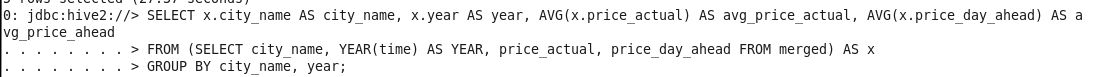
****

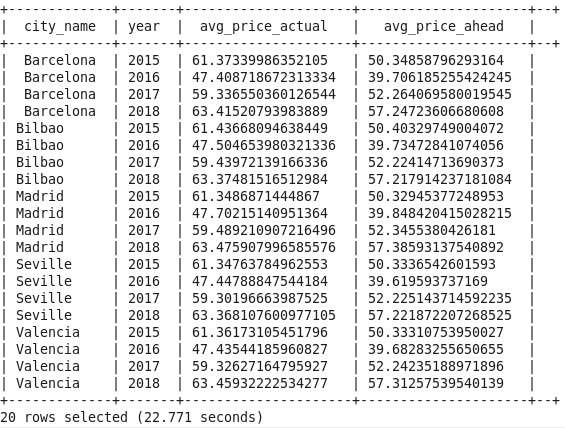
Displays avg temp, windspeed and humidity group by city name

****

****

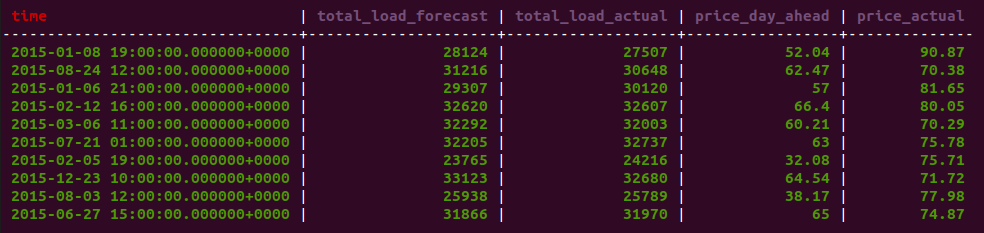
Show the chronological change of the price group by city name

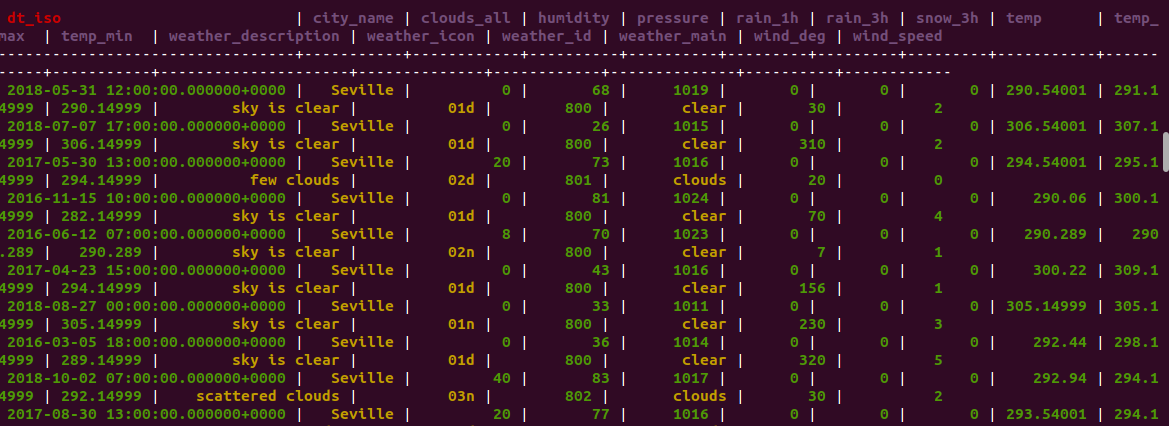
****

****

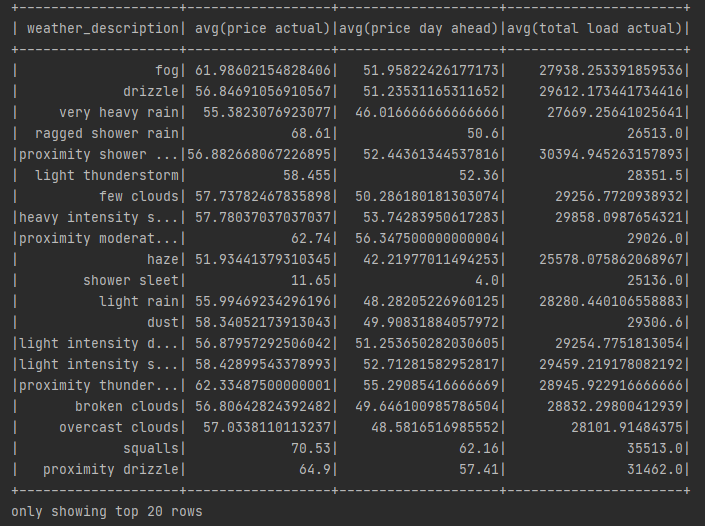
**Cassandra**

Perform some queries in cassandra, There is a limitation for our project because joining the table is not possible in cassandra

****

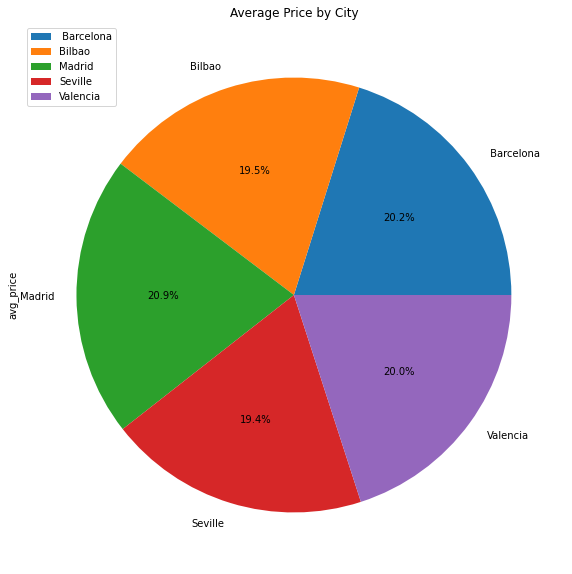
****

**Spark(Scala)**

Load the csv file in the spark in the intellij, I merged two table by using spark sql join function and then do some queries based on joined table. **** ****

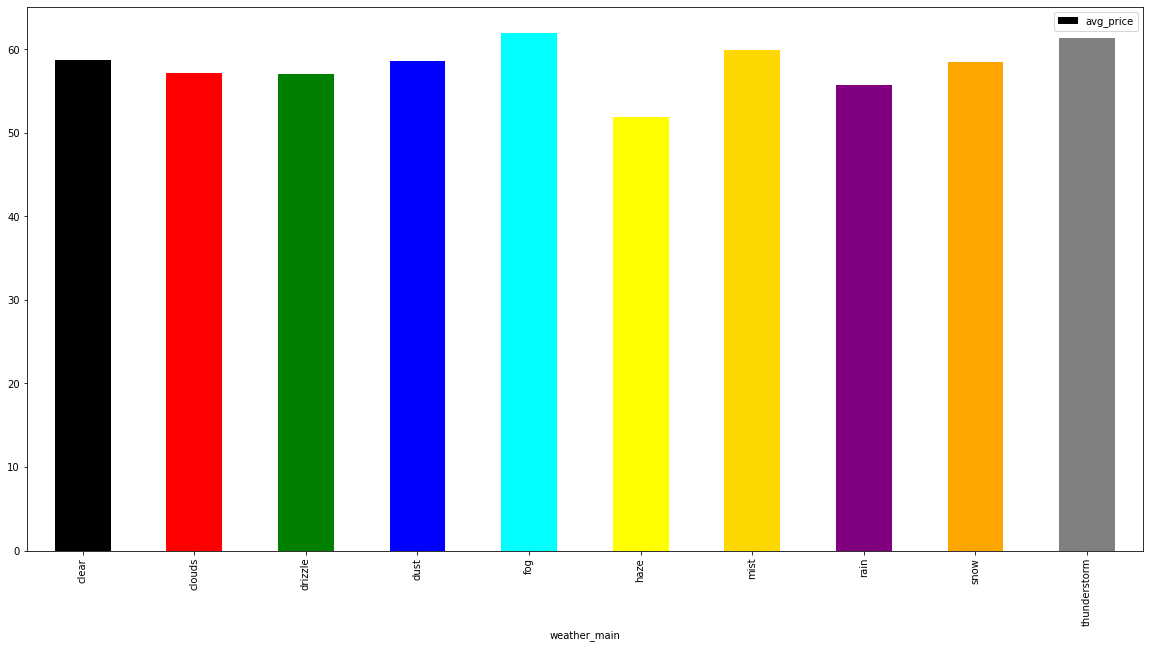
**Pyspark Graphs**

**Average price by city**

****

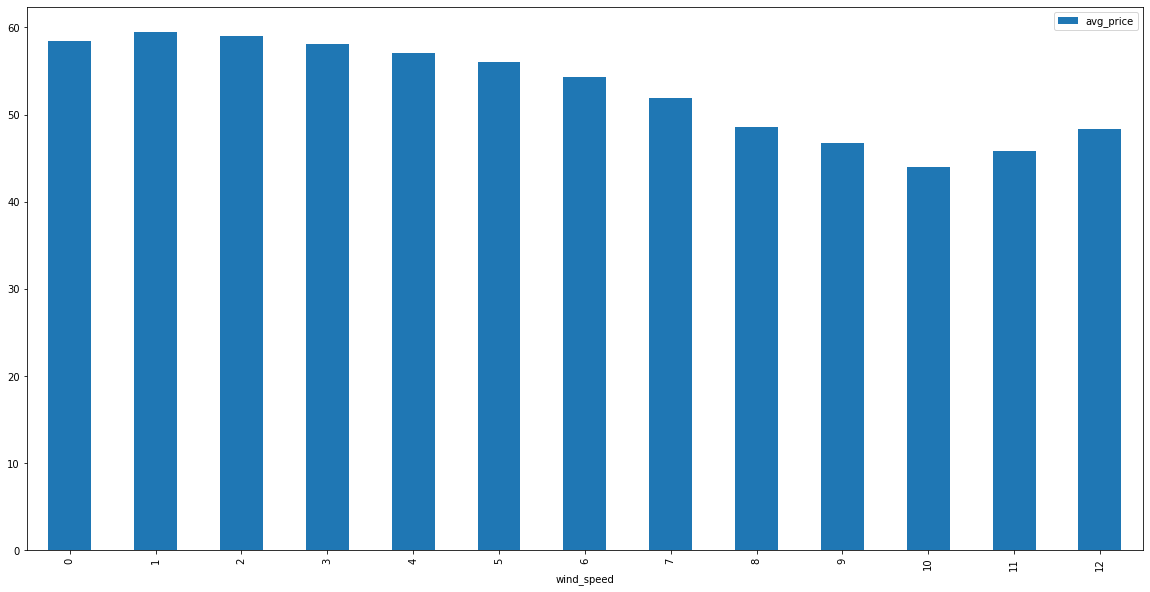
Average price of each city is almost similar to each city. But most highest one is Madrid

**Average Price by weather condition**

****

From the graph, We can see the average price by weather status. We can figure out the relationship between weather and price. Fog and the thunderstorm are most relevant weather condition which raise a electricity price

**Average wind speed and Average price**

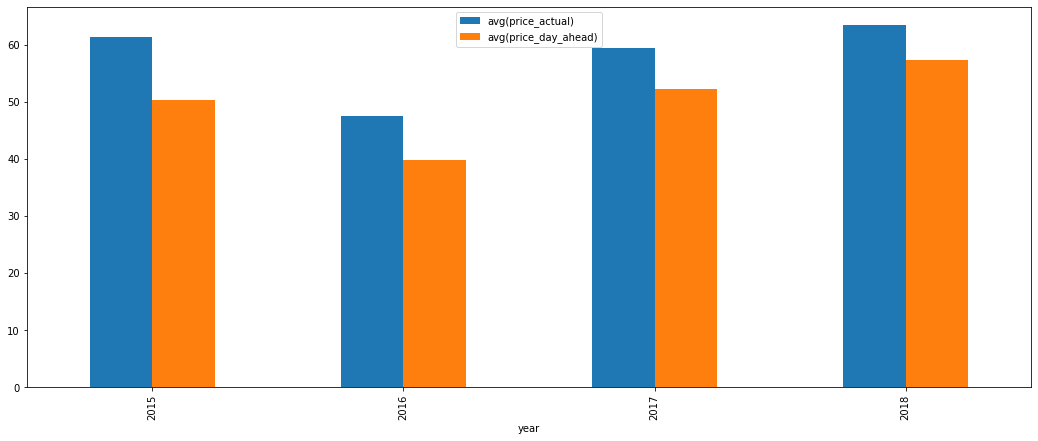
****

Average price seems to increase when average wind speed is at low level. We can infer that wind is supplied for the electricity, which makes the average price low.

**Yearly Analysis**

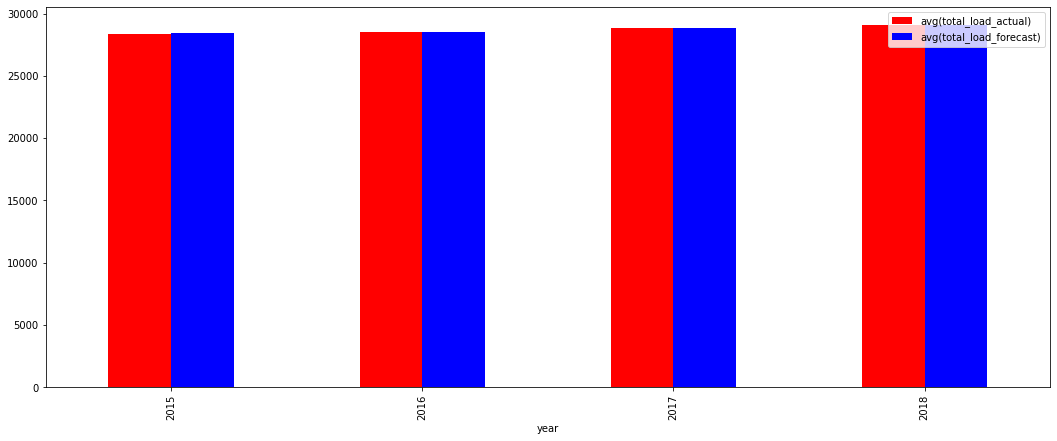
****

**Average Price and prediction by Year**

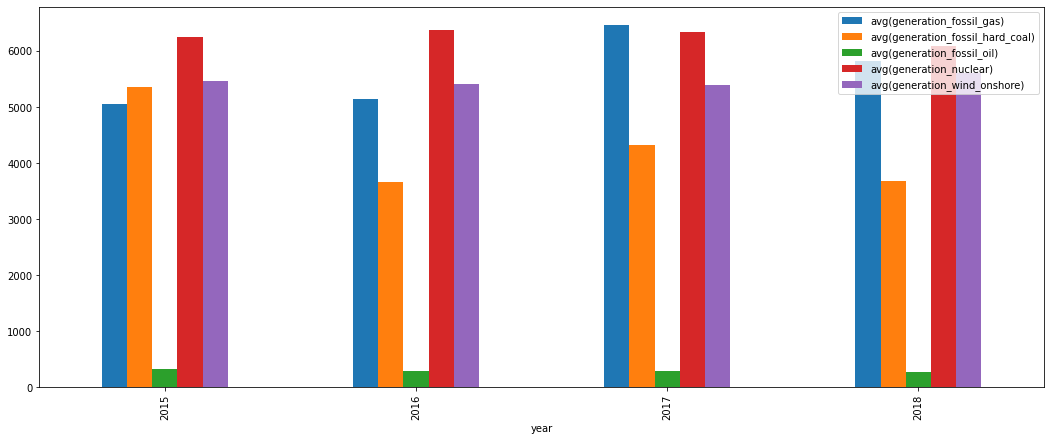
****

It shows each year’s average price and average price prediction. There is some gap between them

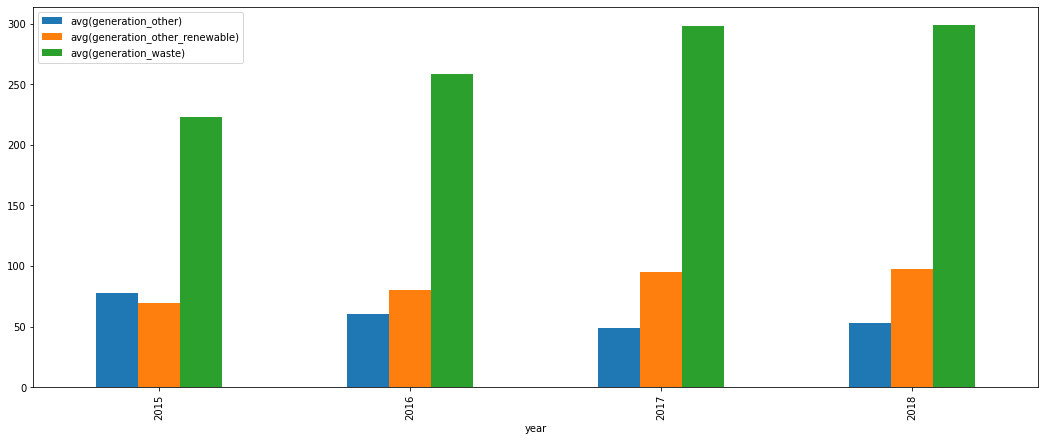
**Average Load and prediction by Year**



**Main Resources generation By Year**

****

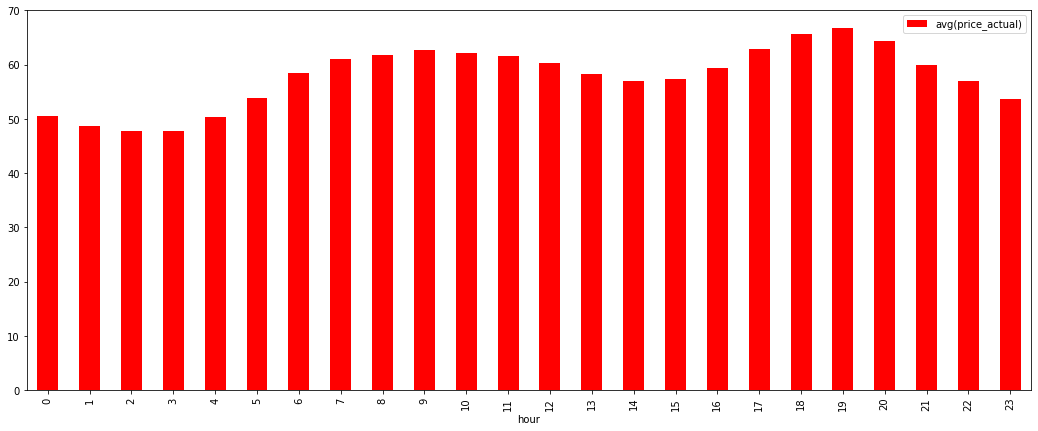
**Other Resources generation By Year**

****

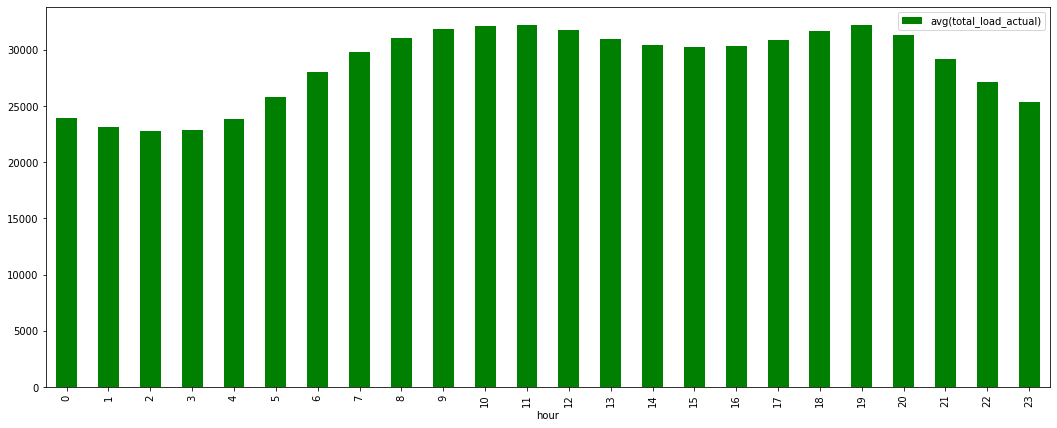
**Hourly Analysis**

****

**Average price by specific hour**

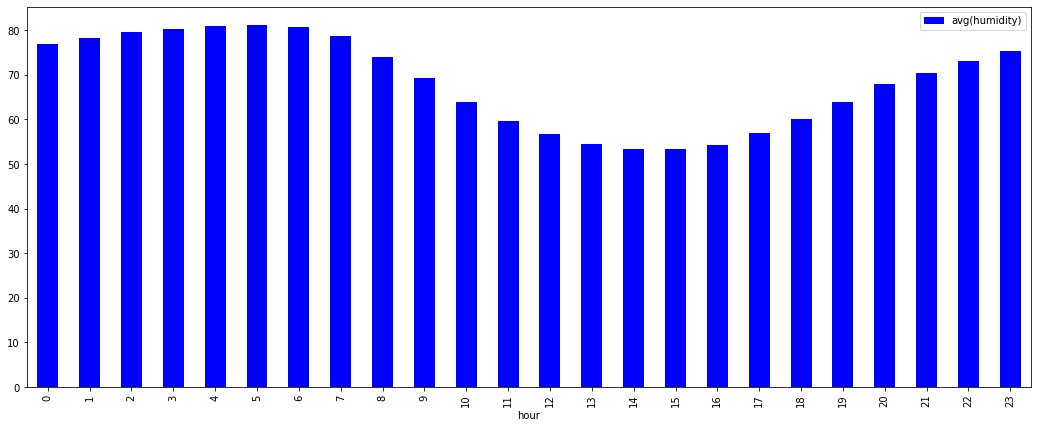
****

**Average Total load by specific hour**

****

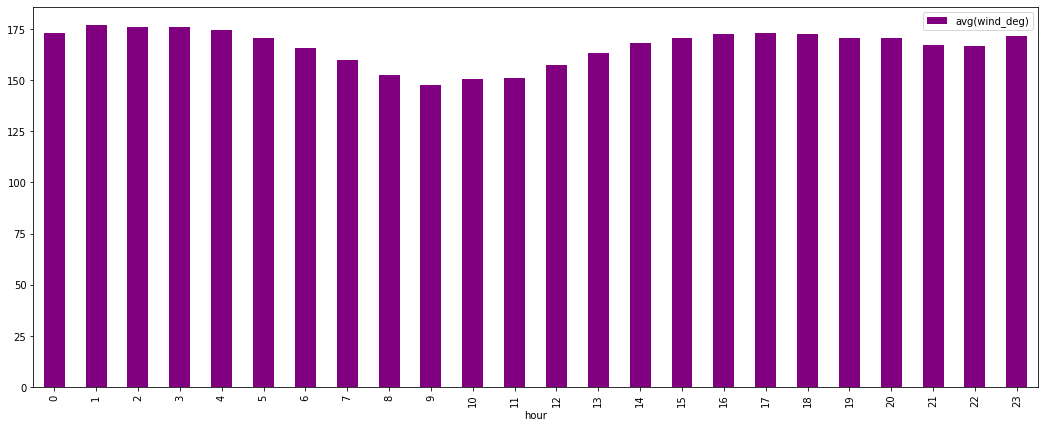
Load seems to increase in after noon time and decrease in morning. It is almost the same with the change of the average price graph.

**Avg humidity by hour**

****

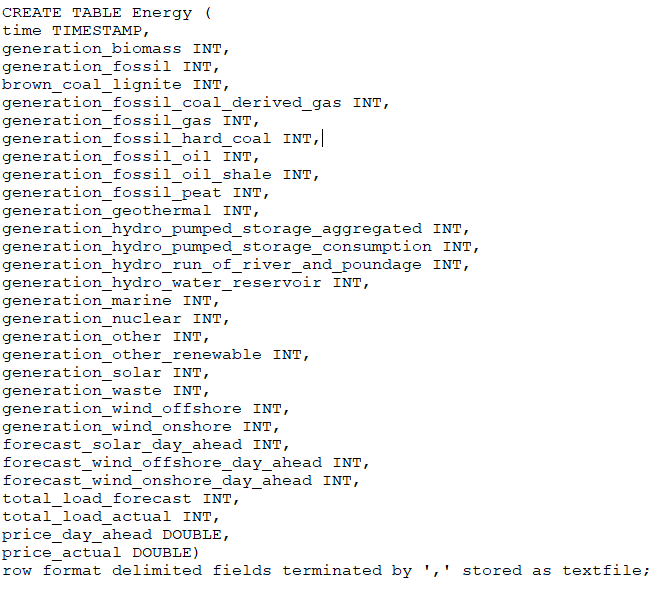
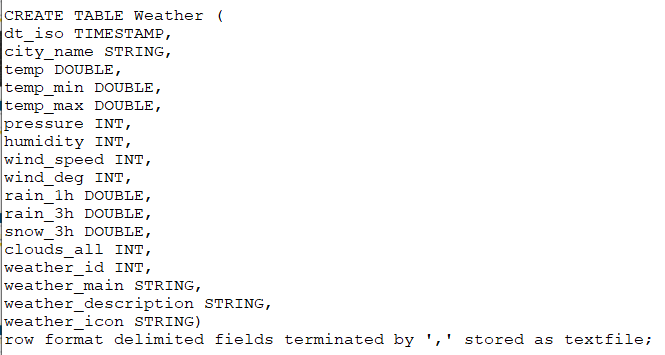
When average humidity is going up, the average price is decreased. So We can say that humidity and price is in opposite relation

**Average wind degree by hour**

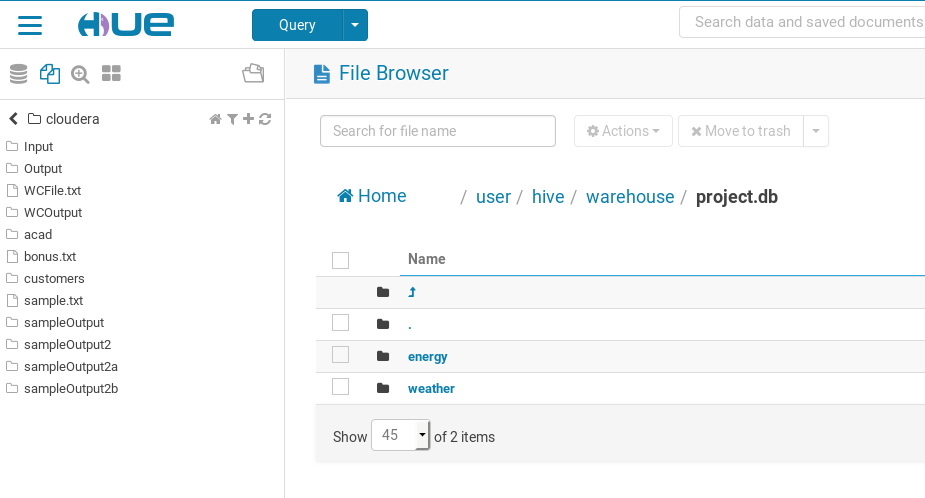
****Similar to the average humidity. opposite relation with the average price

**Bill Yerkes**:

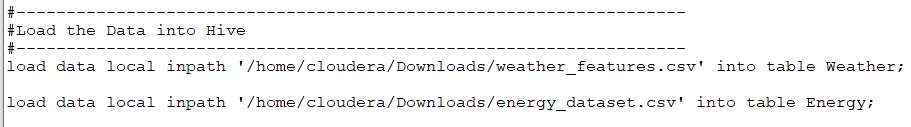
Create Tables in Hive:



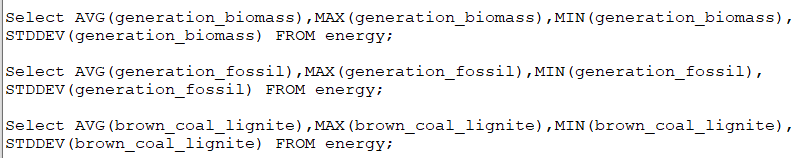
Hive Tables in Hadoop:

****

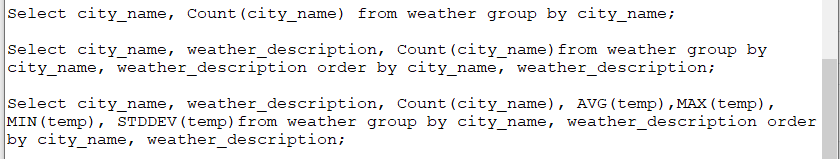
**Load the Data into Hive:**

****

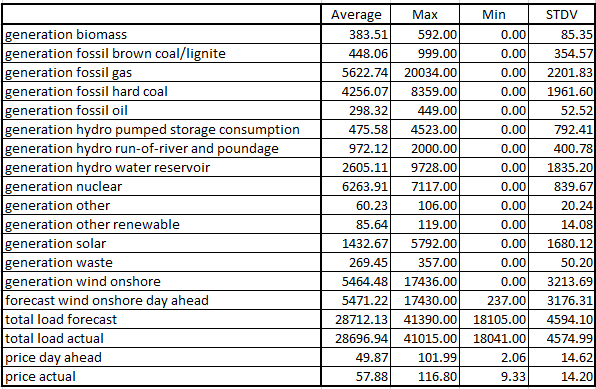
**Some of the queries on Energy Data Set:**

****

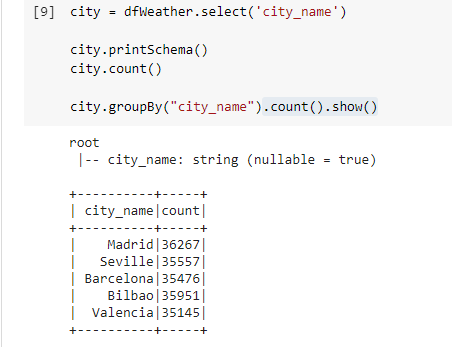
**Some of the queries on Weather Data Set:**

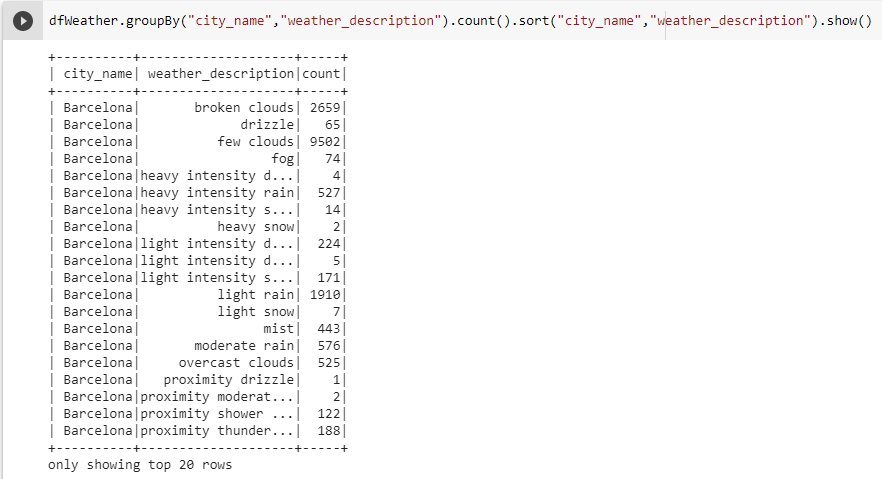
****

**Some of the metrics on the Data**

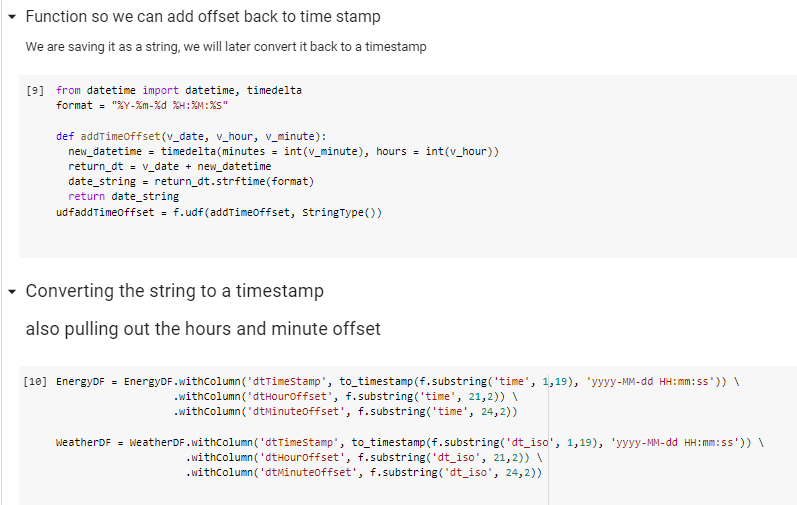
****

**Query and Results in Pyspark**

****

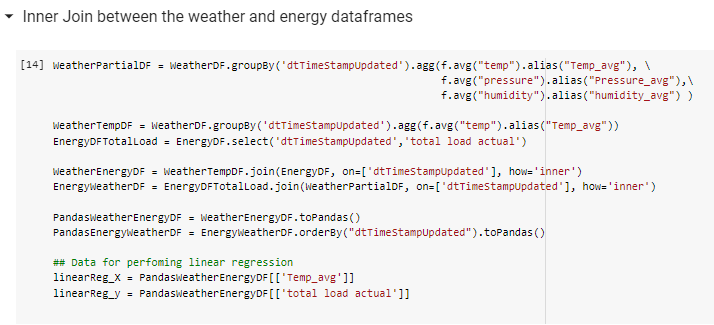
****

**Corrected issue with timestamp, converted from string and added offset**

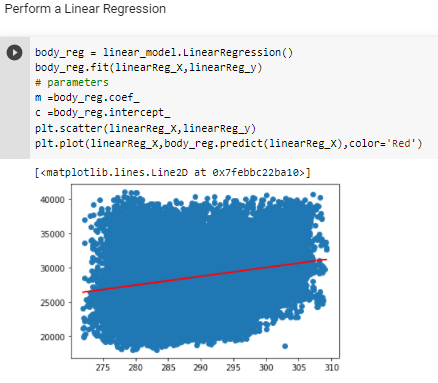
****

****

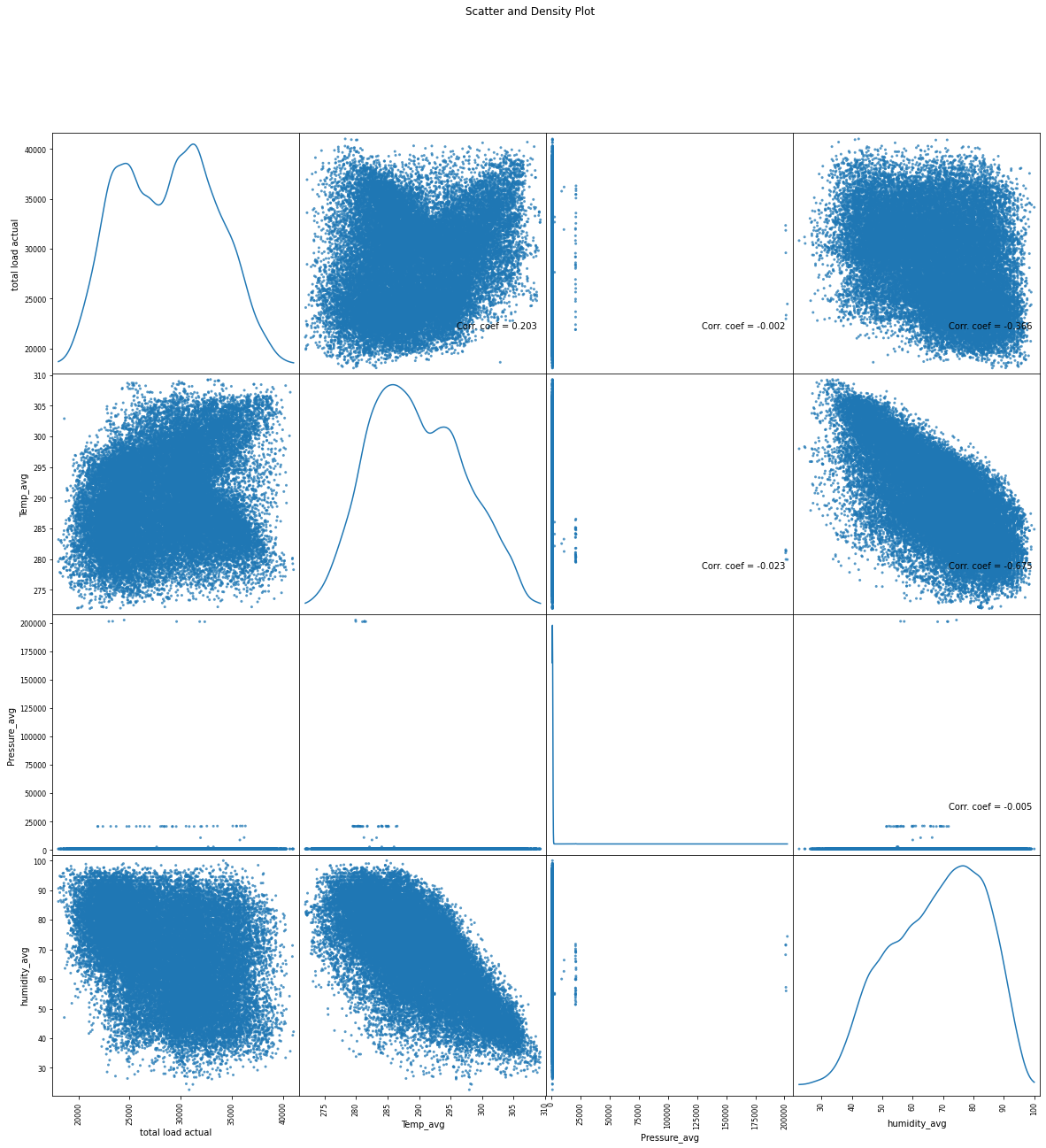
**Joining Weather and Energy Datasets**

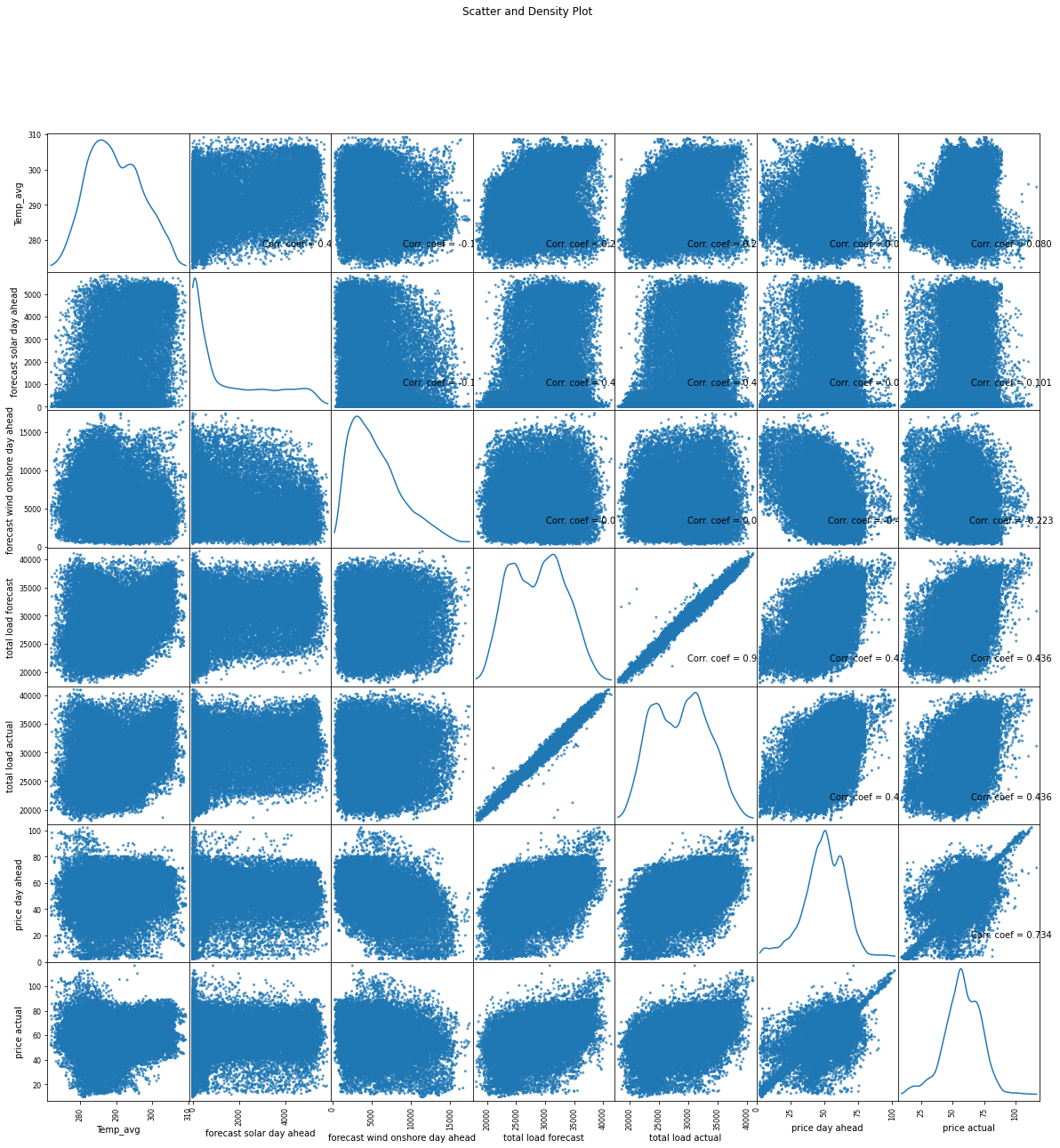
****

**Linear Regression between temperature and total energy load:**

****

**Relationship between: Total energy load, Temperature, Pressure, and Humidity:**

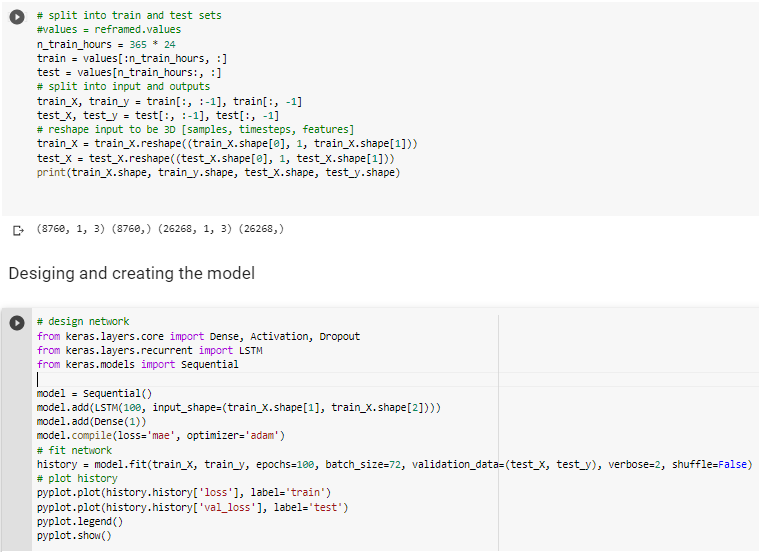
****

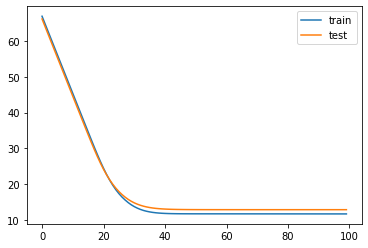
****

**View of Temperature, Total Load, and Price over time**

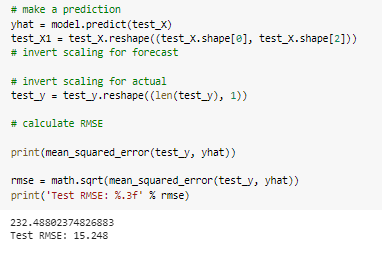
****

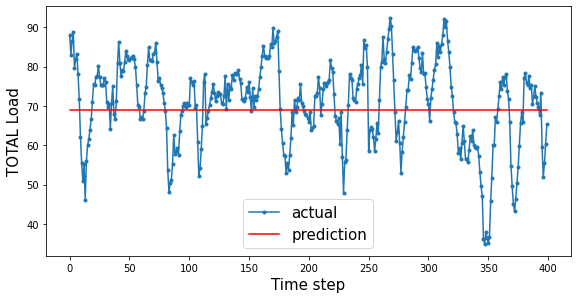
**Long Short Term Memory Model:**

****

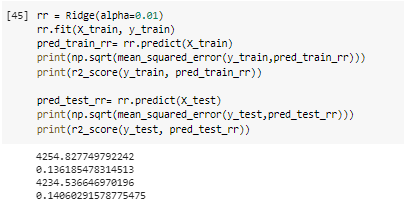
****

**LSTM Results:**

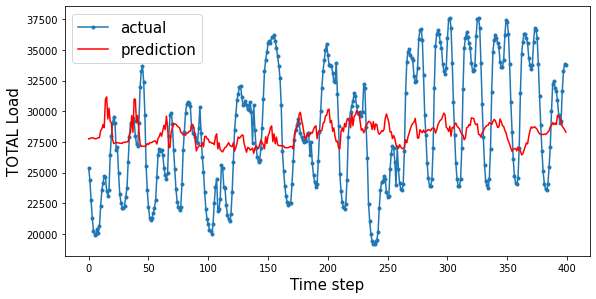
****

****

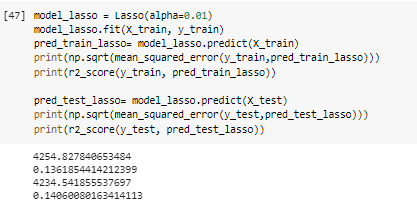
**Ridge Model:**

****

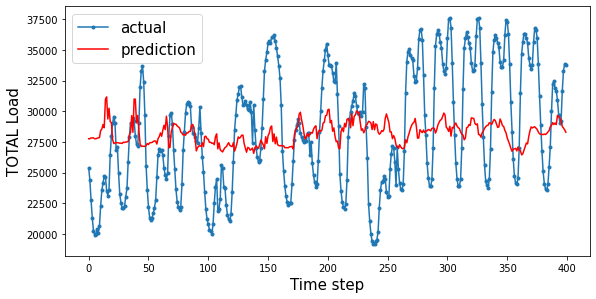
**Results:**

****

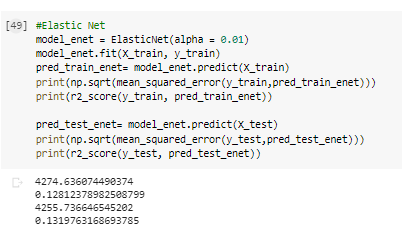
**Lasso Model:**

****

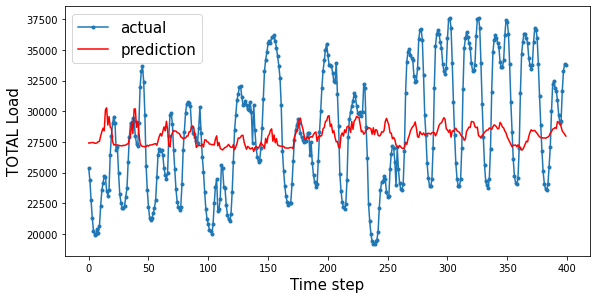
**Results:**

****

**Elastic Model:**

****

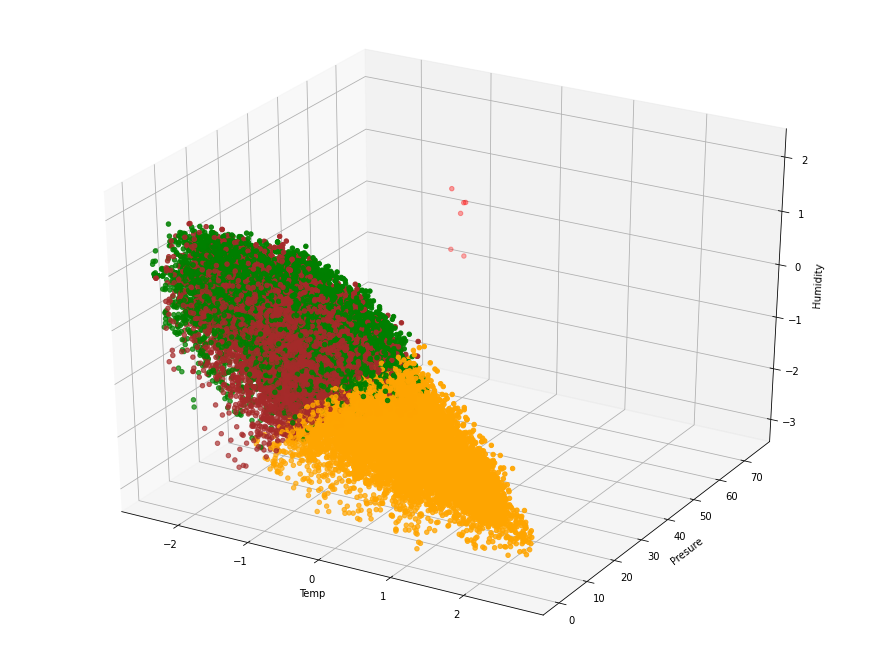
**Results:**

****

**KMeans Cluster**

****

**Visualization:**

****

**Data and Parameters:**

[**Hourly energy demand generation and weather | Kaggle**](https://www.kaggle.com/nicholasjhana/energy-consumption-generation-prices-and-weather)

This dataset contains 4 years of electrical consumption, generation, pricing, and weather data for Spain. Consumption and generation data were retrieved from ENTSOE a public portal for Transmission Service Operator (TSO) data. Settlement prices were obtained from the Spanish TSO Red Electric España. Weather data was purchased as part of a personal project from the Open Weather API for the 5 largest cities in Spain and made public here.

**GitHub Link:**

[BillYerkes/CSEE5590\_GroupProject (github.com)](https://github.com/BillYerkes/CSEE5590_GroupProject)

**Issues or Concerns:**

Big Data Programming comprises a large area. Spending one or two weeks on a topic offered the team the ability to get exposure to the various components. The subject of Machine Learning is one area the team could have used more time on.

Working remotely from team members during a pandemic offers its own set of challenges and stress.

One aspect of Big Data is the volume of data available to perform analysis on, our weather data could have been a bit larger to help improve the development of models. Finding data sets that had both weather and energy information was a bit difficult, learn how to manipulate the data set to get information relevant to the problem we were trying to solve.

**Conclusion:**

The team has familiarized themselves with the tools learned over the semester. The team has also investigated the two datasets to be able to better understand how to construct our solution. The team has received exposure to how Big Data can be used to solve real world problems. The team learned how to work together remotely using different technologies to help facilitate communication and collaboration, such as Discord, Zoom, GitHub and Email..

**The scientist:**

UMKC Students / CSEE 5590 Big Data Programming. .

Anna Johnson, Joe Goldsich, Jongkook Son, and Bill Yerkes

**Users**

There are two main users for our application. Consumers of energy utility companies and producers of the energy being consumed.

**The Society**

Thanks to our application, the overall total utility/energy cost for our society would decrease because each subject would be able to act appropriately according to the prediction of the application. Producers would be able to expand or decrease their production line based on the weather forecast. Consumers of energy would be able to avoid huge amounts of electricity bills because of a more efficient system.

**Video Presentations:**

[**Iteration 2**](https://youtu.be/DXZD-4CaSlI)

[**Final Iteration**](https://youtu.be/dtww_w4Z0Pk)

**References:**

1. [**Betting on the Weather: Predicting Tomorrow’s Energy Prices Using Machine Learning**](https://www.wired.com/wiredinsider/2019/10/betting-on-the-weather-predicting-tomorrows-energy-prices-using-machine-learning/)
2. [**Predict the Unpredictable: Power Forecasting and Energy Trading (baxenergy.com)**](https://www.baxenergy.com/power-forecasting-energy-trading/)
3. [**Predicting wind and solar generation from weather data using Machine Learning | by Hugo Ferreira | Hugo Ferreira’s blog | Medium**](https://medium.com/hugo-ferreiras-blog/predicting-wind-and-solar-generation-from-weather-data-using-machine-learning-998d7db8415e)
4. [**https://www.kaggle.com/nicholasjhana/energy-consumption-generation-prices-and-weather**](https://www.kaggle.com/nicholasjhana/energy-consumption-generation-prices-and-weather)
5. [**Thousands caught off guard by rolling blackouts in Kansas City metro Monday (fox4kc.com)**](https://fox4kc.com/news/thousands-caught-off-guard-by-rolling-blackouts-in-kansas-city-monday-afternoon/)
6. **https://www.texastribune.org/2021/02/17/texas-power-grid-failures/**
7. [**https://www.forbes.com/sites/arielcohen/2021/02/19/texas-energy-crisis-is-an-epic-resilience-and-leadership-failure/?sh=46d08806eee8**](https://www.forbes.com/sites/arielcohen/2021/02/19/texas-energy-crisis-is-an-epic-resilience-and-leadership-failure/?sh=46d08806eee8)
8. [**California rolling blackouts during summer heat wave caused by 3 main factors, report says | Fox Business**](https://www.foxbusiness.com/energy/california-rolling-blackout-power-outage-summer-heat-wave-water-climate-change)