CSEE5590/490: Big Data Programming

Increment 2

Project Title: Energy Demand Analysis in Spain

Team Members: Claire Ndofor, Wes McNall, Scott Howard, Shelby Mohar

Introduction:

Forecasting in energy markets is one exceedingly helpful tool in making the transition to a renewable-based electrical infrastructure (Rolnick et al, 2019). By improving forecasting, we can also increase the efficiency of a power grid and help reduce the usage of peak demand on power plants, which are generally less efficient than their counterparts. While the short-term results have the potential to improve 24-hour and hour-by-hour predictions, this work also has the potential predict energy prices for consumers.

Background:

The data is collected from the five largest cities in Spain: Madrid, Barcelona, Valencia, Seville, and Bilbao between the years of 2015 and 2019. This data has the ability to impact every community that uses an electrical grid. Not only is it advantageous at the individual level to be able to predict the cost of an electric bill, but it is also extremely helpful to be able to predict energy usage at a macro level as communities across the globe begin to make the transition to renewable energies in response to climate change. As stated in the introduction, forecasting in energy markets is an exceedingly helpful tool in making the transition to a renewable-based electrical infrastructure (Rolnick et al, 2019).

Goals and Objectives:

* Motivation:
  + Forecasting in energy markets is one exceedingly helpful tool in making the transition to a renewable-based electrical infrastructure, as stated in “Tackling Climate Change with Machine Learning” (see resources for link to paper). Our goal is to demonstrate this by leveraging Big Data analysis tools on a dataset that consists of energy usage and weather data for five large cities in Spain.
* Significance:
  + Predict energy usage to increase efficiency of electrical production
  + Predict energy price
  + Locate areas that would benefit from renewable energies
* Objectives:
  + Predict energy usage based on the weather
  + Predict energy prices by:
    - Time of day
    - Day of the week
    - Time of year
  + Analyze the factors that affect the fluctuations in energy usage, as well as the sources of energy
* Features:
  + dt\_iso (datetime index localized to CET)
  + generation biomass (in MW)
  + generation fossil brown coal/lignite (in MW)
  + generation fossil coal-derived gas (in MW)
  + generation fossil gas (in MW)
  + generation fossil hard coal (in MW)
  + generation fossil oil (in MW)
  + generation fossil oil shale (in MW)
  + generation fossil peat (in MW)
  + generation geothermal (in MW)
  + city\_name
  + temp (in kelvin)
  + temp\_min (in kelvin)
  + temp\_max (in kelvin)
  + pressure (in hPa)
  + humidity (in %)
  + wind\_speed (in m/s)
  + wind\_deg (wind direction)
  + rain\_1h (rain in last hour in mm)

Dataset

Our dataset is comprised of two .csv files:

* + weather\_features.csv – contains information about the weather
  + energy\_dataset.csv – contains information about the production, price, and variation of energy resources

The two files can be joined by a timestamp. The dataset can be found on Kaggle with the heading “Hourly energy demand generation and weather”. See resources for link.

Features Developed:

This section is dedicated to the features developed in this increment, and a guide to the files within the team repo.

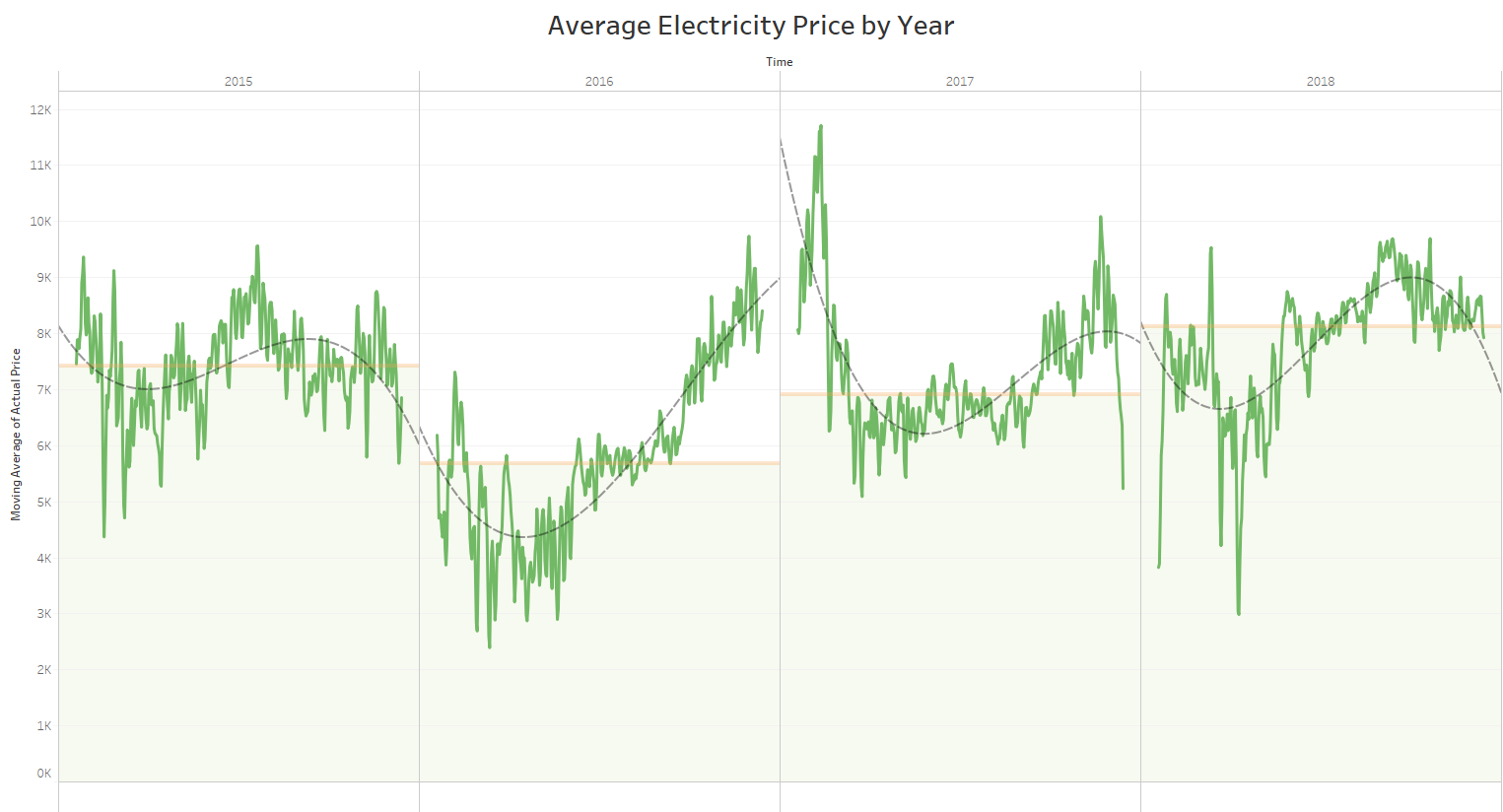
HiveQL:

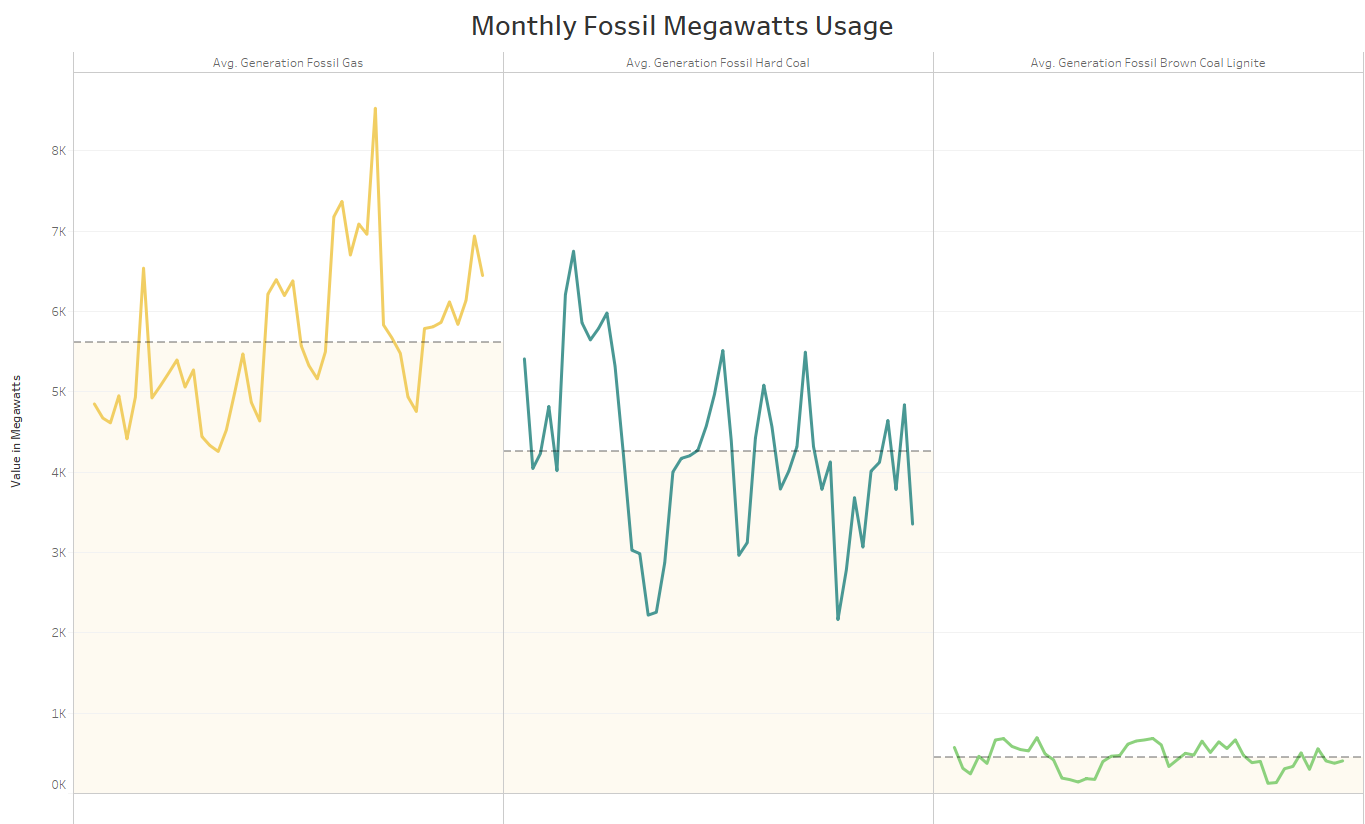
During a past class our professor mentioned using Graphs in this project and as soon as I heard that I knew I wanted to use Tableau to visualize some key aspects of the data, not only to learn more about it but to show key findings.

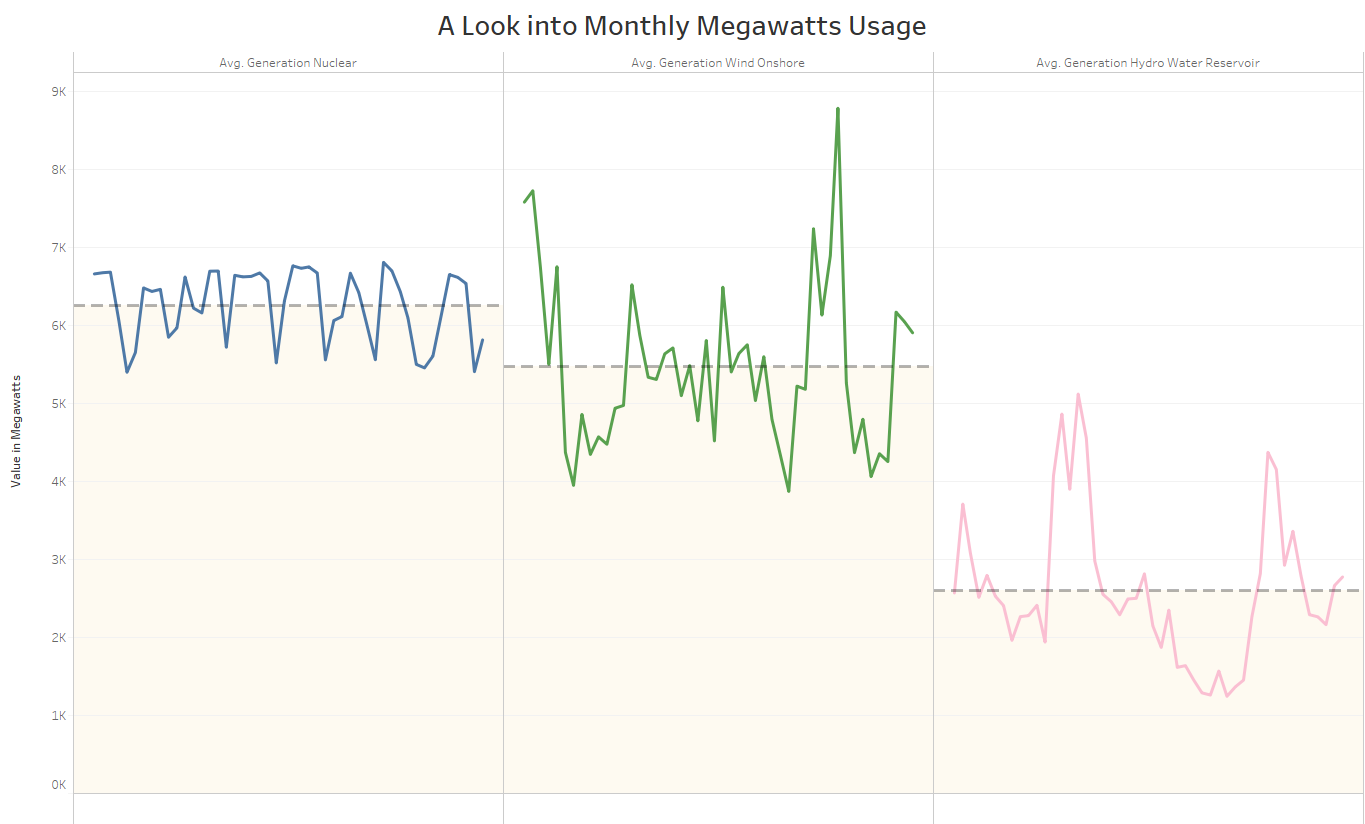
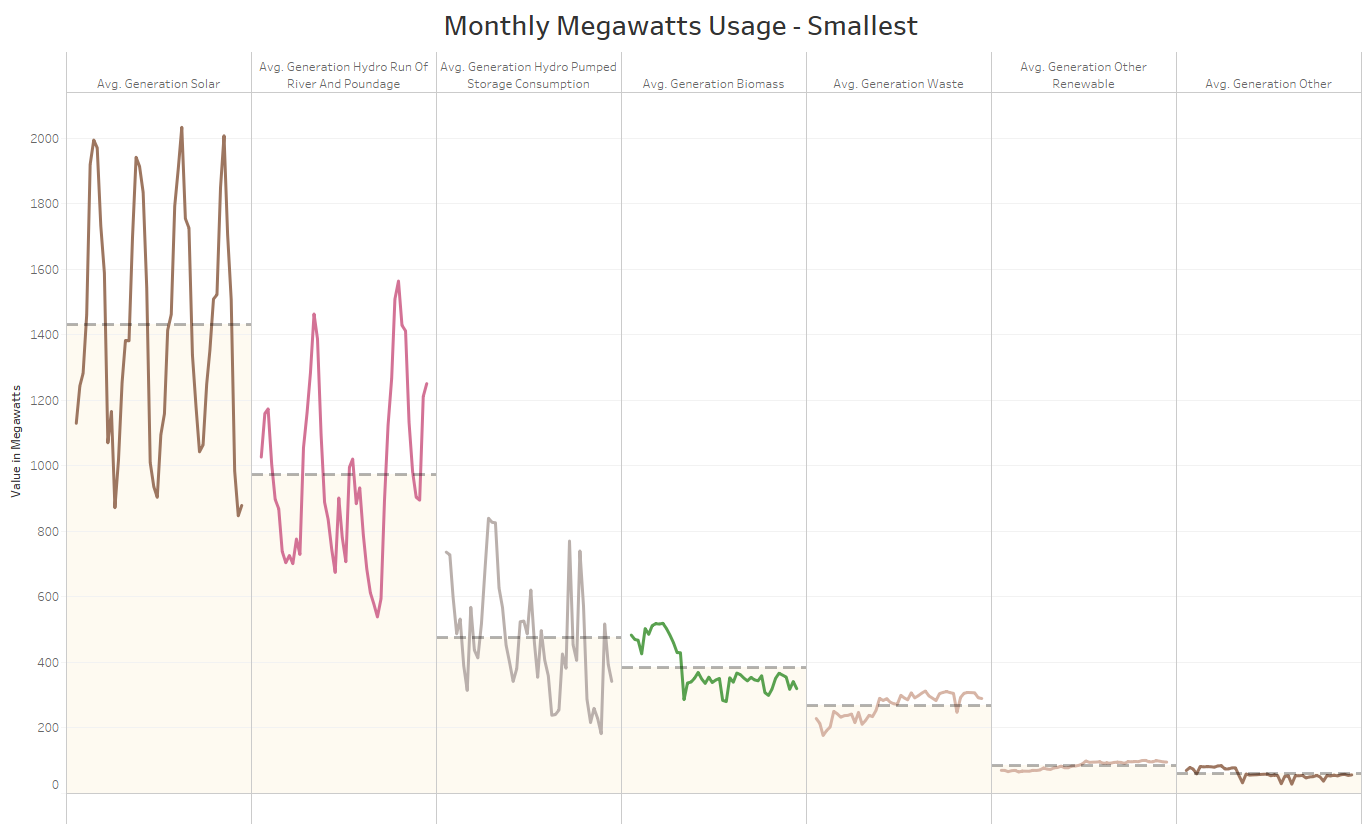
With the data already loaded from the previous Increment I took to asking some questions about the data and then visualizing the data to see what was interesting about it. Because we are dealing with trends of prices over time, that was a key aspect that I wanted to be able to visualize. Using Tableau I could add some extra visuals without having to calculate, such as trend lines for each particular year. Looking at the Average Electricity Price per Year Graph we can see there are clear lows and highs between the years, which will require some further investigation as to why those trends exist

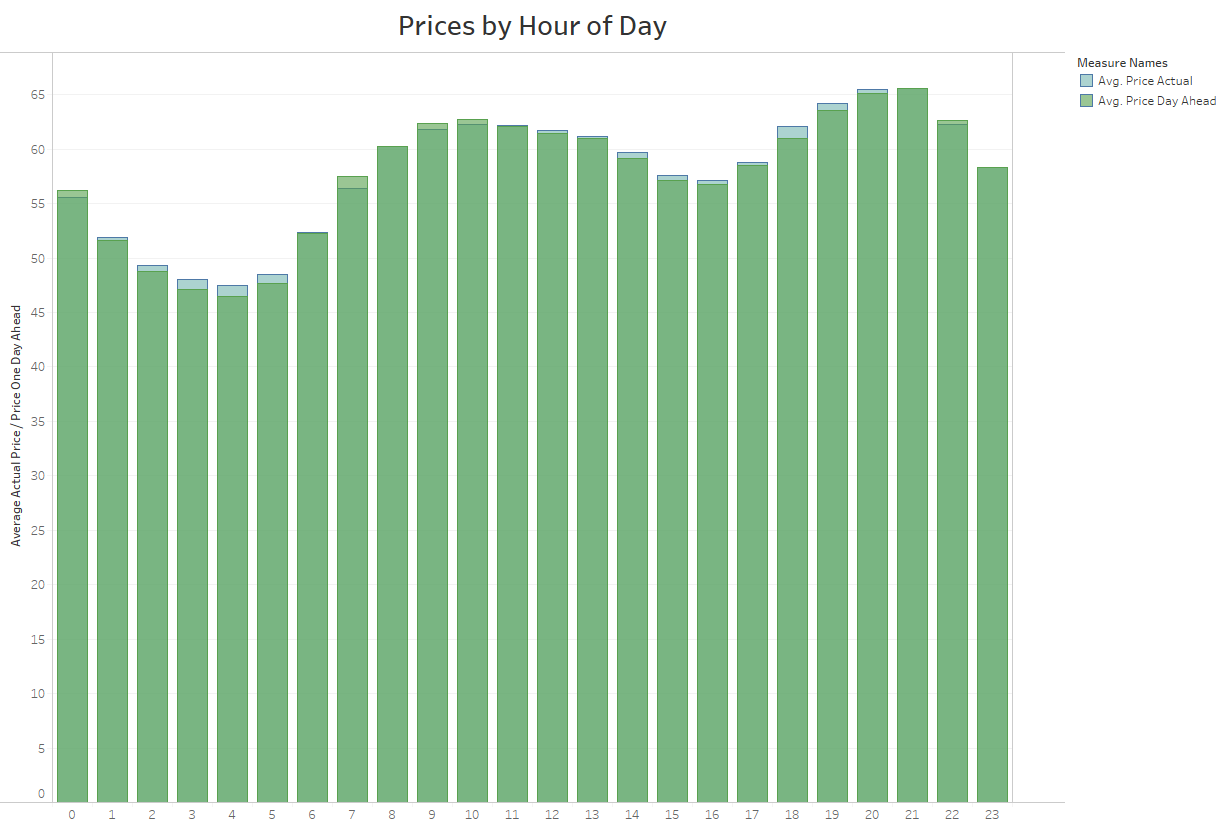
With a wide dataset, part of what I wanted to accomplish this Increment was to determine what columns had interesting and worthwhile data and what columns could be more or less ignored. By writing a large HiveQL query that included summary statistics over time, it would allow us to look for trends and determine which were worth investigating further. All separate Megawatts Usage graphs were trends I found interesting enough to highlight, and all others within the dataset I left ignored from the graphs and queries

Something I wanted to see was not only trends over months and years, but just over the course of a day. Specifically the average prices over different times of day. The Prices by Hour of Day Graph shows that there is a fluctuation of the cost throughout the course of the day. This is to be expected and the highs and lows also match times that make sense for what time most people will be working and most people will be sleeping

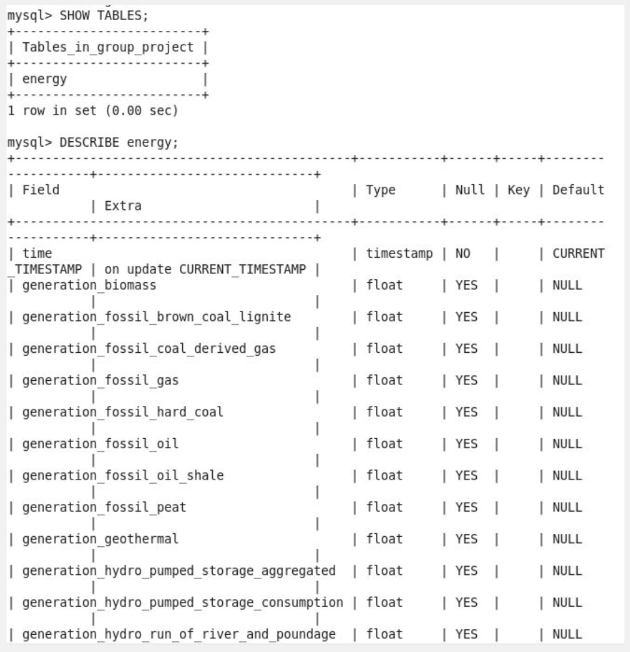
With the 5 different cities in the dataset I wanted to explore the quantitative difference between the locations and see if there was any interesting information that varied between them. Well, the answer was that there isn't, but this wasn't an unfortunate discovery it was a happy one! This means that these prices were being fairly priced between all of the different locations within the region, meaning that the pricing is independent of location which was a good thing to learn.







Sqoop:

Within Cloudera, we used Sqoop to transfer the merged dataset from Hive to mySQL.

mySQL:

MapReduce:

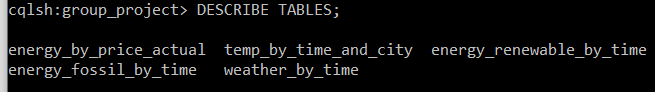
Cassandra:

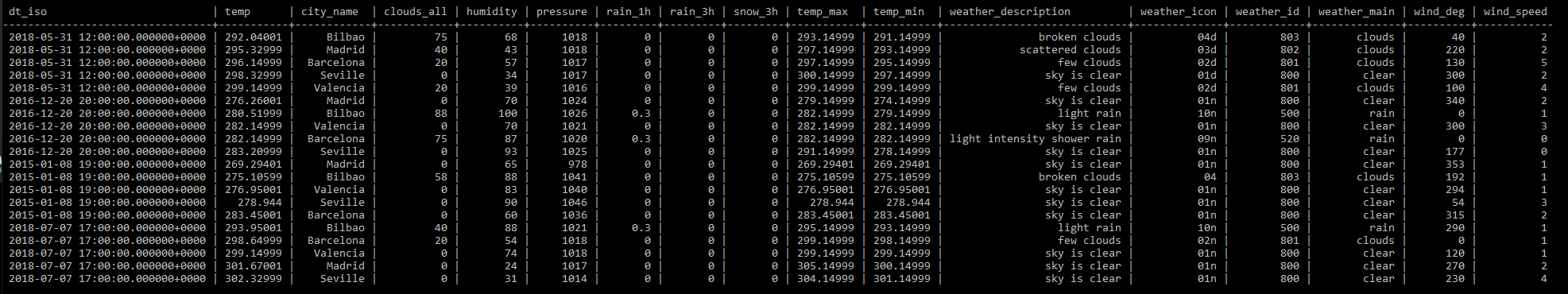
Because joins aren’t possible in Cassandra, it was necessary to keep the two tables separate. Furthermore, since Cassandra operates by a query-first approach, I created several tables within Cassandra such that each table was designed for a specific query. Though it did result it duplication of data, this design is good for high-load queries that usually happened in big data. The insights gleaned from these queries seemed rather unhelpful compared to the query capabilities of HQL and mySQL. Whereas HQL/mySQL can perform direct analysis on the data (such as calculating averages, join functions, etc.), it seems like there would have to be some secondary analysis step performed with any data returned from a Cassandra query.

‘Cassandra Tables Creation.cql’ – This file contains the script that was used to create and load data into five different Cassandra tables. Because the data is just text, the class used was SimpleStrategy. A replication factor of 3 was arbitrarily decided upon.

‘Cassandra Queries.cql’ – This file contains the queries used for each table. The result of the queries was stored into a unique txt file.

‘Cassandra Results” – This folder contains the results of the five .cql queries used for each of the Cassandra tables, as well as screenshots of the successfully created tables.





Project Management:

* Work completed:
  + Description: We have analyzed our dataset using Hive, mySQL, MapReduce, and Cassandra, as well as created some preliminary data visualizations.
  + Contributions:
    - Claire: mySQL queries
    - Wes: Hive table creation, HQL queries, visualizations
    - Scott: MapReduce queries
    - Shelby: Sqoop transfer from Hive to mySQL, Cassandra analysis, report composition
* Work to be completed:
  + Description: For our next increment, we will utilize Spark to gain more insights about our data.
  + Concerns: IntelliJ/Scala can be finicky, and some members of our group are using PySpark instead. It will be a challenge to coordinate our efforts when our setups are not the same.

Assignment 2 Questions:

* Who:
  + This dataset is about the people who use energy in Spain, whose energy production and grid was sampled for this dataset. There is no identifiable information on the individual level, meaning that there is little personal risk with this dataset.
* What:
  + The energy usage and sources of energy production of the people of Spain are what is being recorded by the data set. This addresses all of our questions in Assignment 1.
* When:
  + This data was collected between 2015 – 2019, meaning that the data is recent and therefore relevant. It is cross-sectional since the data was collected from several cities in Spain. This dataset contains real-time data.
* Where:
  + The data is collected from the five largest cities in Spain: Madrid, Barcelona, Valencia, Seville, and Bilbao. It could possible be extrapolated that the energy usage would be similar in the surrounding European countries with similar populations and weather as these five cities, and it is certainly possible that larger generalizations about predicting energy usage could be used for non-European locations.
* Why:
  + The data was collected by ENTSOE, a public portal for Transmission Service Operator (TSO) data and is publicly available.

References:

“Tackling Climate Change with Machine Learning”

<https://arxiv.org/abs/1906.05433>

“Hourly energy demand generation and weather – Electrical demand, generation by type, prices and weather in Space”

<https://www.kaggle.com/nicholasjhana/energy-consumption-generation-prices-and-weather?select=weather_features.csv>

“Chapter 4. The Cassandra Query Language”

<https://www.oreilly.com/library/view/cassandra-the-definitive/9781491933657/ch04.html>

“Defining Application Queries”

<https://cassandra.apache.org/doc/latest/data_modeling/data_modeling_queries.html>

“LanguageManual Select”

<https://cwiki.apache.org/confluence/display/Hive/LanguageManual+Select>