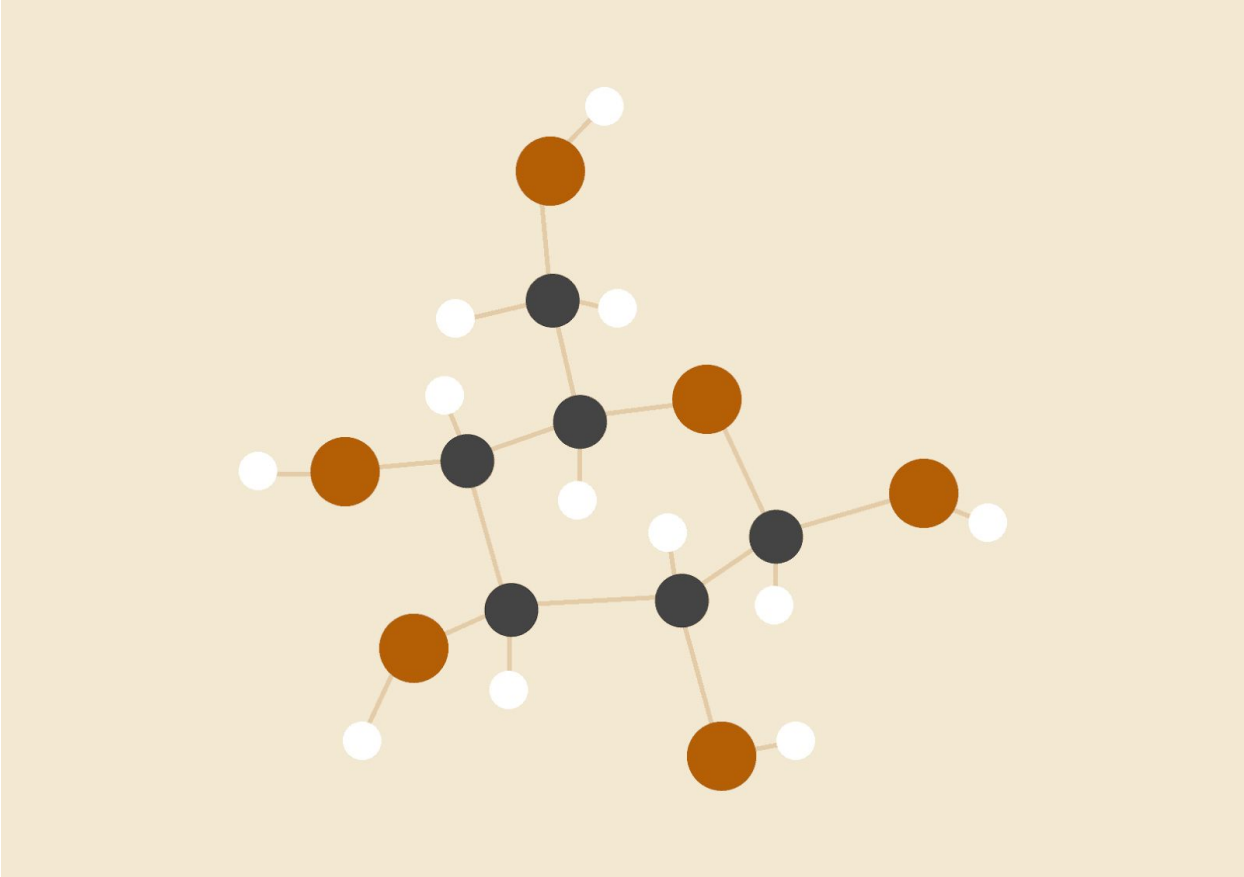
DATA ANALYTICS REPORT

Assignment 1: Modelling Electricity Consumption As a Function of Weather



CHRISTOPHER DARE

cdare

Andrew ID: cdare

Date: 27.01.2020

INTRODUCTION

The task at hand is to investigate the relationship between weather and electricity consumption in France. We will use corresponding 2017 data to achieve our tasks.

HYPOTHESIS

It's possible that weather can influence the demand and consumption of electricity. Should this be proved in the following procedure, it will be very helpful in informing power production corporations on how to manage demand and supply, and ultimately, price the price of electricity.

ANALYTICAL TOOLS

1. Jupyter notebook
2. Python and associated modules: pandas, sklearn, numpy, matplotlib

PROCEDURE

For the purpose of presentable narration, I assume the 3rd person in my writing.

Q.1 Importing The Weather Dataset And Preprocessing Data

The first thing to do is to import the data sets for 2017 weather and electricity consumption data respectively.

It's always important to preprocessing the data; transform it to exhibit some uniform and predictable characteristics. And so we do that. We first start ensure that both filesare in csv format and remove unnecessary headers for each files.

Next we import the datasets via custom load functions. Peeking the dataset for nan values, we see that the following columns have NaNs:

1. high Visibility (km)
2. avg Visibility (km)
3. low Visibility (km)
4. high Gust Wind (km/h)
5. Events

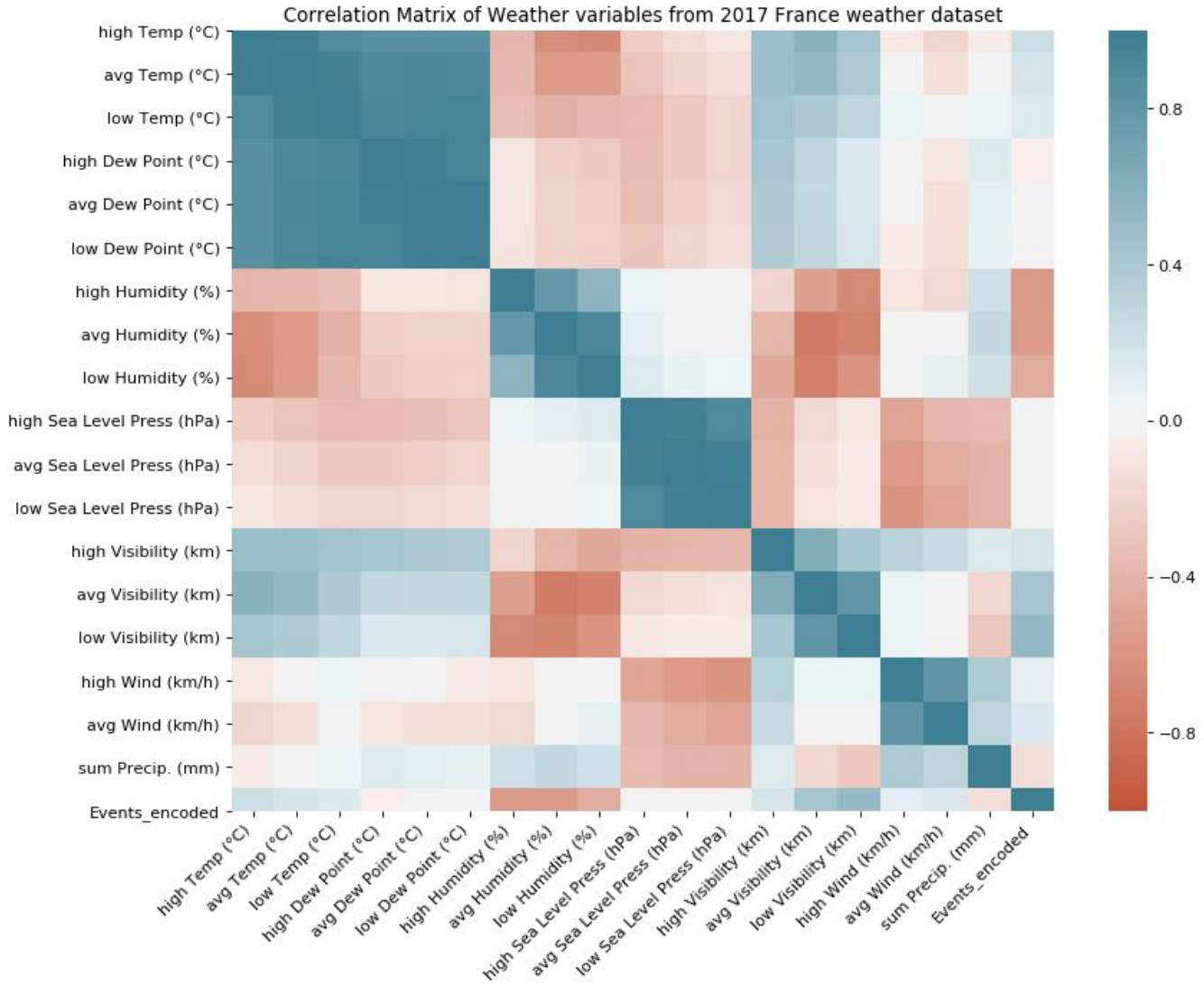
The first 3 are indicated by NaN. Missing data for “high Gust Wind (km/h)”, however is indicated with a hyphen (dash). And the “Events” data is provided as a string.

We can take quite a number of approaches, however we will take the following approach to handle missing values:

1. Use linear interpolation to fill traditional NaNs. For each of those 3 values, there are only 3 values.
2. Ignore “high Gust Wind (km/h)” in the predictive modelling techniques. (There are too many NaNs and creating values out of thin air - whether using the mean or past values - may not be so accurate since the dataset is small)
3. Encoding the “Events” data in numerical format - with a number to represent NaNs. NaNs will not be interpolated for the same reason as stated in 2. The new events data will be stored in “events_encoded” and the “Events” column disregarded

Q.2 Investigating Correlation

The next thing we do is to visualize and inspect the correlation between weather variables. We do this by plotting a correlation matrix of all weather variables as a heatmap. This is depicted in Fig. 1.1 below:



We see some very interesting patterns. There are some inverse relationships between certain variables such as:

- Temperature and humidity
- Sea level and visibility
- Precipitation and sea level

We also see some directly proportional but rather weak relationships between variables such as:

- Temperature/Dew point and visibility

Temperature and sea level (with their various variations) seem to be the most critical variables at play here. We will investigate their effect on energy consumption later. Now that we have seen this pattern it's time to load, inspect and model electricity consumption.

Q.3 Loading Electricity Consumption Data

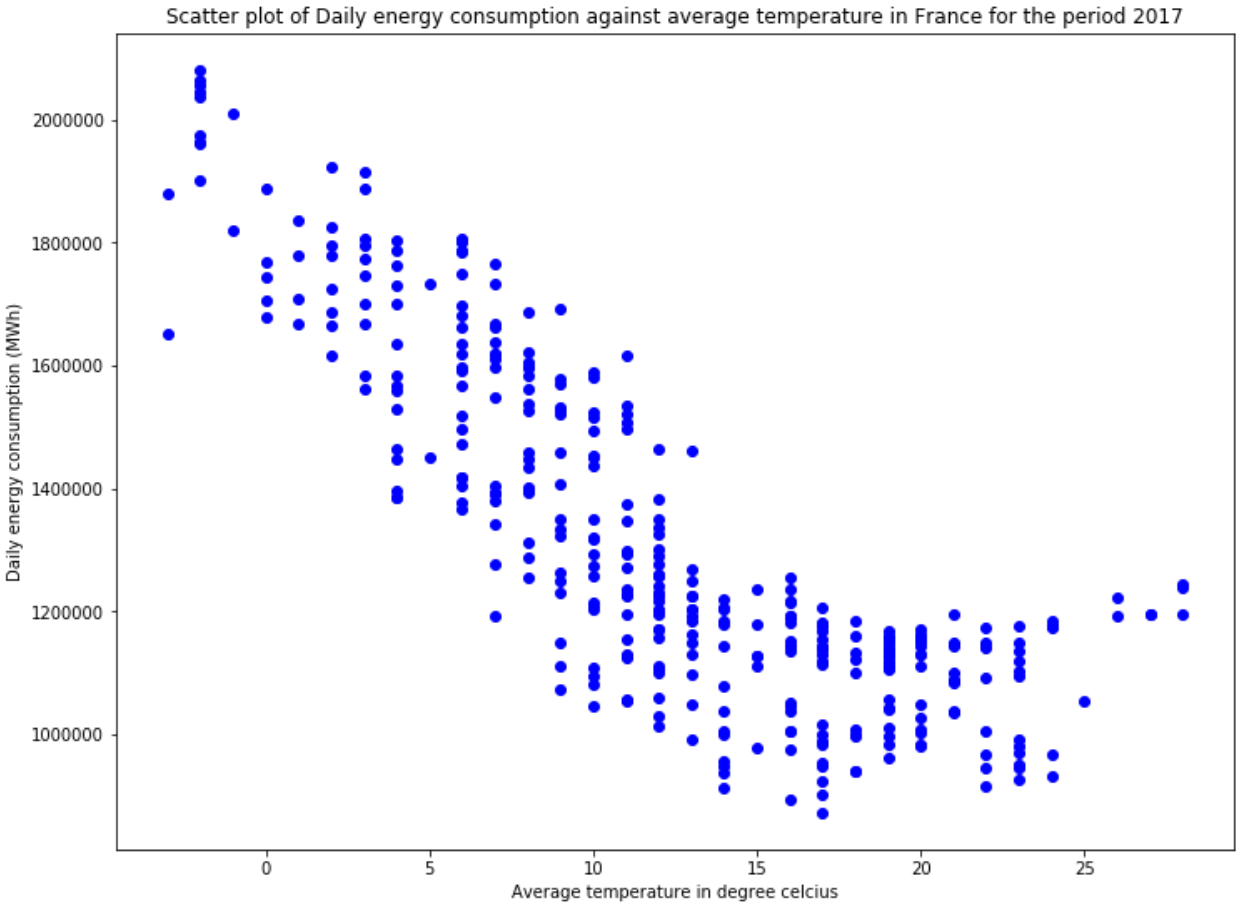
We repeat a similar process for loading the France's 2017 daily electricity consumption dataset. Peeking the dataset for NaNs, we see that there are no missing values for relevant columns that may require interpolation i.e electricity consumption by megawatts.

Q.4 Merging Weather And Electricity Consumption Datasets

In order to develop a predictive model of electricity consumption based on weather data, we need to merge both datasets. Since this is daily dataset for the period 2017, we will merge the datasets on the dates.

Q.5-6 Plotting Average Energy Against Temperature

Now that we have merged the dataset, let's visualize it (well a part of it) on a graph.

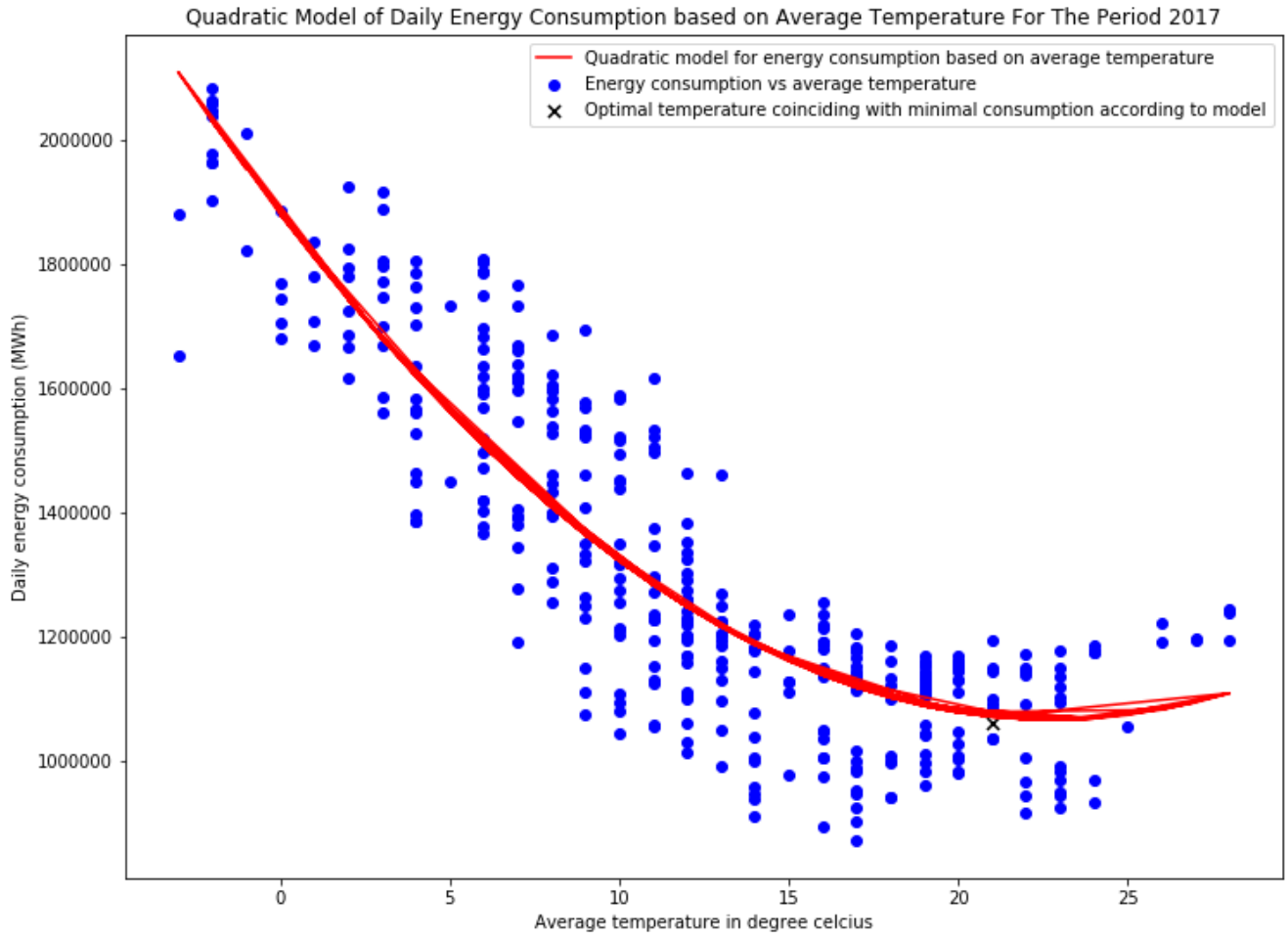


We use a scatter plot to show the average temperature and energy consumption. Looking at this graph, we see that there neither a linear or exponential relationship between both variables. A polynomial function is likely to model this relationship.

Modelling a Polynomial/Quadratic Function

In order to model a quadratic function we will use the numpy polyfit function. We fit this on average temperature and energy consumption to obtain the weights of each component of the function. Then we establish y as a function of this new equation.

Plotting the equation yields the curve shown in Fig 1.3 below:



The average temperature coinciding with the minimum energy consumption is 21 degrees celscius. This is shown by the x on the graph above.

Q.7 Stepwise Selection

Let's take a step back. We need to find out the most significant features to build a multivariate model. To do this, we will use a forward stepwise selection approach to select optimal features

Running the stepwise algorithm yields the following features:

1. high Temp (°C)
2. high Visibility (km)
3. low Temp (°C)
4. low Humidity (%
5. avg Dew Point (°C)
6. low Sea Level Press (hPa)

We also see that the co-efficient of determination for this model is 0.762

Q.8 Running a Second Reiteration

Can we improve this model? One way would be to introduce other variables that leverage the concept of mutual information. One such way would be to square independent variables. This would make the relationship between the original independent variables more significant for consideration since energy consumption depends on them more than their squared values.

When we perform a forward stepwise selection on this new set of features, we then obtain the following selected independent variables:

1. high Temp (°C)
2. high Temp (°C)_sq
3. low Humidity (%
4. avg Dew Point (°C)
5. high Wind (km/h)_sq

We also see that the coefficient of determination value is 0.785. That's a 3% increase from the previous and corresponds to an improvement.

Q.9 Running a Third Iteration With Dummy Variables

What about the days of the week? Could energy consumption be higher or lower depending on the day of the week? We can investigate that relationship as well. We do this by implementing one-hot encoding for each row by date in the dataset. (This is made possible by using dummy variables to get the day of the week. Then we implement one-hot encoding.)

After implementation, we pass this through out forward stepwise selection. The result is interesting:

1. high Temp (°C)
2. high Temp (°C) _sq
3. Sunday
4. Saturday
5. low Temp (°C)
6. low Humidity (%)
7. high Wind (km/h)_sq
8. Monday
9. low Temp (°C)_sq
10. avg Dew Point (°C)_sq
11. low Dew Point (°C)
12. sum Precip. (mm)

In plain English, this tells is that

1. Rainfall
2. Temperature
3. Humidity
4. Wind speed
5. The intensity of dew
6. The day being a Saturday, Sunday or Monday

...all have an effect on energy consumption!
By reason of gumption, we can probably understand why this is so: they often keep people indoors or in areas that use electricity - especially homes and places to hang out. Furthermore, Monday is a time when the working population uses appliances in getting ready for a work week.

So it makes sense that electricity consumption is dependent on these variables.

Q.10 Avoiding Over-fitting

The correlation matrix is a great way to check whether or not the model is overfitting. If we have features that are dependent on each other and found in the model, then we should have reason to suspect over-fitting. In this case however, we see that it is not a problem. Variables such as visibility are not accounted for in the predictive model - and well so because they depend on others such as humidity - which obviously as a higher statistical significance.(humidity in cold regions affects visibility)

Approaches to avoid over-fitting:

Our lecturer, Prof. McSharry, has taught us from the previous course on “Data, Inference and Applied Machine Learning”, some approaches to avoid over-fitting:

1. Cross-validating the model by splitting the training dataset into training and testing sets, then using the test data set to evaluate the accuracy of the model so as to fine tune the parameters.
2. Training with more data. The more the data, the better the model can generalize to other cases.

CONCLUSION

In this assignment, I have confirmed my hypothesis, practised techniques crucial to investigating and establishing relationships between variables, eliminating false relationships and developing predictive models for dependent variables.