Data Analytics: Assignment 1

by Anthony Rizkallah

Table of contents

1 Python Packages

Chardet

Pandas

Sklearn

Matplotlib

Seaborn

Scipy

Numpy

Statsmodels

2 Exploratory Data Analysis

2.1 Correlation Analysis of Weather Variables

2.1.1 Steps

A heatmap expresses the collinearity between independent variables. Collinearity is the correlation that exists between independent variables (predictors). This means that the variables have a relationship with each other. In other words, multicollinearity is the term used to describe this dependence among multiple intercorrelated predictors. This is generally problematic because independent variables should be independent, similar to the linear algebra concept of linear dependence among elements in a matrix.

So, the effects of collinearity among predictors can have counterproductive effects on regression models. In principle, each predictor is associated with a $\widehat{B_i}$ value in a regression model. For example:

$$\hat{y} = \hat{B}_0 + \hat{B}_1 x_1 + \hat{B}_2 x_2 + \cdots \widehat{B}_i x_i + \epsilon$$

where $\widehat{B_0}$ is the intercept, $\widehat{B_i}$ is the estimated coefficient for the predictor x_i, and ϵ is the residual. If the variables are correlated, the estimated coefficients won't be distinct– each predictor should fit a separate piece of the dependent variable. As collinearity increases, it becomes more difficult to determine the effect of each variable on the estimated coefficients which undermines the clarity of the model.

2.1.2 Results

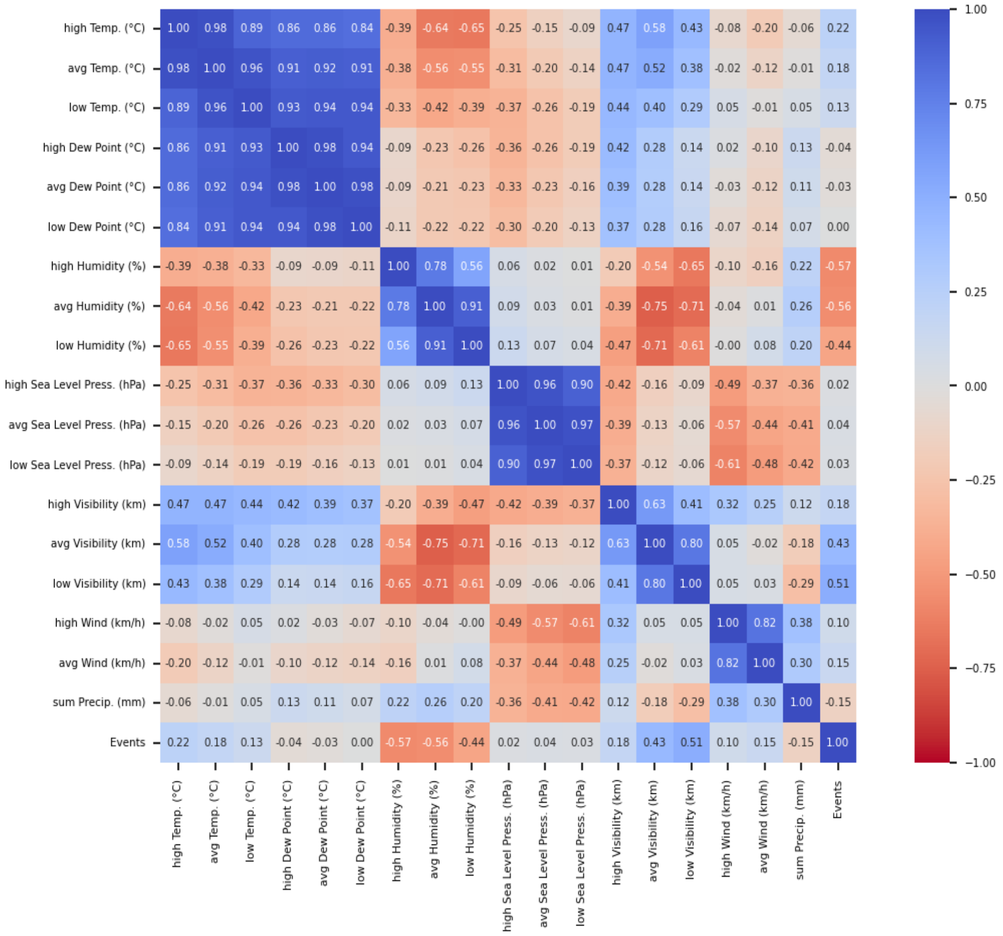

Figure 1. Correlation Heatmap for Weather Variables

2.1.3 Inference

In Figure 1, the color scale represents the correlation on a scale of [-1,1] where -1 is a perfect negative correlation (Red) and 1 is the perfect positive correlation (Blue). Trivially, the diagonal of the matrix is positively correlated, as those are the variables correlated with themselves. Moreover, the darker the color, the higher the correlation (either red or blue).

For example, the temperature variables "high Temp. (°C)", "avg Temp. (°C)", "low Temp (°C)", "High Dew Point (°C)", "avg Dew Point (°C)", and "low Dew Point (°C)" are strongly positively correlated, meaning, information about is a great predictor of the other. However, if one were to build a machine-learning model from the list of variables, choosing all of the aforementioned variables has at least a neutral effect on the model and at worst a counterproductive effect. They can all be expressed as a linear combination of each other. This is true for humidity variables, sea level pressure, visibility, and wind.

Putting it simply, the variables that are most likely to be chosen for a regression model will be the ones with the lowest correlation in the heatmap, as they would express the different components of the dependent variable. For example only, average humidity and average sea level pressure have a correlation of 0.03, implying they have low collinearity, possibly making them good variables to incorporate in a model or at least worth exploring.

2.2 Scatter Plot: Energy Consumption vs Mean Temperature

2.2.1 Steps

If one were to predict electricity consumption of France, there are trivial assumptions to be made. However, one can first ask the question, "What affects variation in electricity consumption?". Keeping all factors constant (manufacturing, office computers, military equipment, etc.), then one significant factor is space heating and cooling. According to the EIA, 40-55\% of residential consumption is space heating and cooling.

Hence, what affects space heating and cooling should be the temperature, where the lower/higher the temperature the higher the consumption for heating/cooling. Additionally, as temperature approaches moderate level, so should electricity consumption. So, to verify this idea, it is worth graphing electricity consumption vs temperature.

One thing that can be done before plotting is a two-sided t-test, where

H_0: Electricity consumption is not affected by temperature

H_1: Electricity consumption is affected by temperature

2.2.2 Results

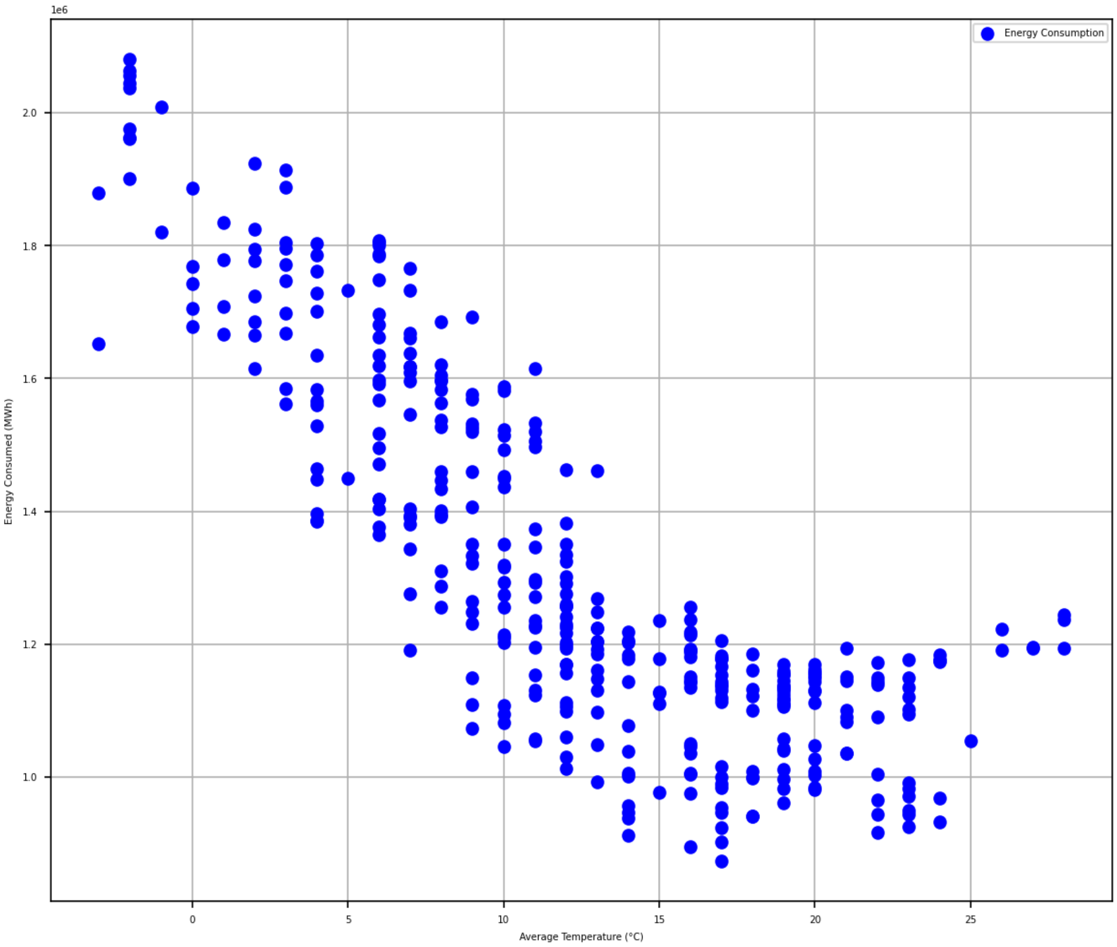

Figure 2. Electricity Consumption (MWh) vs Average Temperature (°C) in France

2.2.3 Inference

There is a clear relationship between electricity consumption and average temperature, where low temperatures have higher electricity consumption and when temperatures start exceeding 25°C, it is almost clear that the consumption begins sloping back up. In fact, the t-test yielded a p-value smaller than $1\times10^{-308} << 0.05$ (Python returns 0 for values smaller than 1×10^{-308}), meaning that the null hypothesis is rejected and that temperature is a very significant predictor of electricity consumption and the values are not correlated by chance.

2.3 Quadratic Fit: Energy Consumption vs Mean Temperature

2.3.1 Steps

From Figure 2, the Electricity Consumption vs Temperature graph has a minimum, similar to x^2. Generally, the equation that represents the a quadratic equation vs number of edges is

$$\text{Degree} = \text{Number}_{\text{edges}} + 1$$

Also from Figure 2, the number of edges (the minima in this case) is 1, yielding a second-degree quadratic equation.

2.3.2 Results

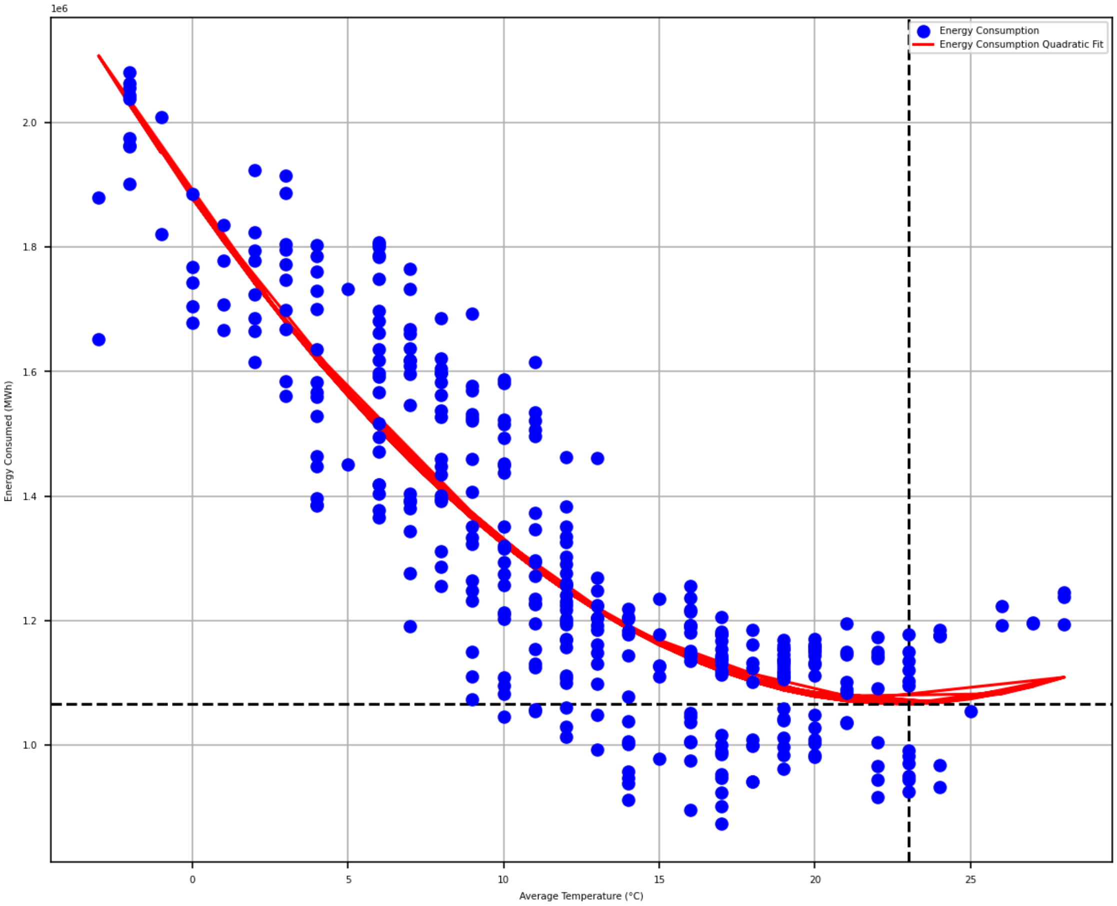

Figure 3. Quadratic Fit of Electricity Consumption (MWh) vs Average Temperature (°C) in France

2.3.3 Inference

In Figure 3, the quadratic fit of a second order fits the data properly, showcasing that as temperature decreases, electricity consumption decreases, reaching an inflection point at 23°C and then increases again.

There is not enough data from France to simulate the electricity consumption at higher temperatures, but one can expect the consumption to be symmetric to the left hand side (Again, similar to $a(x-h)^2+b$; centered at h, sloped by a, and lifted by b from x=0.)

3 Multivariate Regression Models

3.1 Selecting Variables: Stepwise Approach & Initial Model

3.1.1 Steps

1. Choose the variable (x_1) with the lowest p-value after calculating the correlation coefficient of all variables as a preliminary step (see previous section).

2. Regress the dependent variable on x_1 and x_2, x_1 and x_3 … x_1 and x_p. The variables in the model that have a p-value higher than the exit alpha $\alpha_{\text{exit}} = 0.15$ (chosen for this problem) are removed (Backward step).

3. Assuming that x_1 and x_2 remained, regress the independent variable onto the model containing both variables and repeat step 2 to evaluate the p-values.

4. Re-iterate the process until no addition nor removal changes the p-values.

3.1.2 Results

Variable	**P-Value**
high Humidity (%)	0.01569306059593367
avg Dew Point (°C)	$3.921849603027428 \times 10^{-67}$
low Humidity (%)	$8.995438236487256 \times 10^{-27}$
high Visibility (km)	0.00024963370079595667
avg Sea Level Press. (hPa)	0.02268404875892892

Table 1. Initial Model Selected Variables

Metric	**Value**
R^2	0.749
R^2_{adj}	0.746

Table 2. R^2 Values for Initial Model

3.1.3 Inference

The initial regression model selected five explanatory weather variables: high humidity, average dew point, low humidity, high visibility, and average sea level pressure, all of which have statistically significant p-values. The p-values for most variables are below α=0.05, indicating strong evidence against the null hypothesis, meaning these variables have a meaningful impact on electricity consumption. The highest p-value in the model, 0.0157 (for high humidity), is still well below the threshold, suggesting that all selected variables contribute significantly to predicting energy consumption

The model's coefficient of determination (R^2) is 0.749, and the adjusted coefficient of determination (R^2_{adj}) is 0.746, indicating that approximately 74.9% of the variance in daily electricity consumption is explained by the selected weather variables. The close proximity between R^2 and R^2_{adj} suggests that the model does not suffer from excessive inclusion of unnecessary variables. However, R^2_{adj} accounts for the number of predictors, meaning further model refinements—such as introducing quadratic terms or additional relevant variables—may still lead to improvement, as 74.9% is a good variability, but can certainly be improved for scientific applications.

3.2 Model with Squared Terms

3.2.1 Steps

By taking the squared of the independent variables in the dataset and running a stepwise regression to select new features (Section 3.1.1), one can identify new, non-linear, relationships between weather variables and electricity consumption.

3.2.2 Results

Variable	**P-Value**
avg Dew Point (°C)	$4.364537250811625 \times 10^{-14}$
low Humidity (%)	$4.102352681187931 \times 10^{-25}$
avg Sea Level Press. (hPa)	0.010362873705672244
avg Visibility (km)	0.08550952758860506
avg Temp.(°C)^2	0.0004844036044857471
low Temp. (°C)	0.0002820900798414514
low Sea Level Press. (hPa)	0.011512168504248044
low Temp.(°C)^2	0.022863767860225176
avg Sea Level Press.(hPa)^2	0.008629845879083681

Table 3. Model with Squared Terms Selected Variables

Metric	**Value**
R^2	0.806
R^2_{adj}	0.802

Table 4. R^2 Values for Model with Squared Terms

3.2.3 Inference

The inclusion of squared terms in the regression model has led to a notable improvement in performance, as indicated by the increase in R^2 from 0.749 to 0.806 and R^2_{adj} from 0.746 to 0.802. This suggests that accounting for non-linear relationships between weather variables and electricity consumption enhances the model's explanatory power. Compared to the initial model, which primarily included humidity, visibility, dew point, and sea level pressure, the new model introduces squared terms for temperature and sea level pressure, indicating that these variables have non-linear effects on electricity consumption.

3.3 Model with Day of Week

3.3.1 Steps

By selecting the date of the datapoint and transforming the date into a day-of-the-week boolean columns, one can find the day of the week that has strong statistical significance, which indicates an underlying behavioral aspect for predicting electricity consumption for that day. Again, turning dates into boolean columns and running stepwise regression again should yield new selected variables and an improvement in R^2.

3.3.2 Results

Variable	**P-Value**
avg Dew Point (°C)	$2.293004544316856\times10^{-11}$
low Humidity (%)	$6.948764789822583\times10^{-11}$
avg Sea Level Press. (hPa)	0.015617194193443635
avg Visibility (km)	0.06566795343708158
avg Temp.(°C)^2	$8.152858298510277\times10^{-7}$
low Temp. (°C)	$5.612414585968247\times10^{-7}$
low Sea Level Press. (hPa)	0.03656417003010709
low Temp.(°C)^2	0.010890741129090817
avg Sea Level Press.(hPa)^2	0.013822731061848123
Day_Saturday	$6.840041301367005\times10^{-21}$
Day_Sunday	$4.0606410151305245\times10^{-34}$
avg Wind (km/h)	0.0033434174242887096
sum Precip. (mm)	0.0017830393601520073
avg Humidity (%)^2	0.02935153162262643
Day_Monday	0.006581988646607499

Table 5. Model with Day of Week Selected Variables

Metric	**Value**
R^2	0.887
R^2_{adj}	0.882

Table 6. R^2 Values for Model with Day of Week

3.3.3 Inference

The inclusion of day-of-the-week effects in the regression model has significantly improved its performance, as reflected in the increase in R^2 from 0.806 to 0.887 and R^2_{adj} from 0.802 to 0.882 compared to the previous model, which included squared weather variables. This suggests that electricity consumption is not only influenced by weather conditions but also follows behavioral patterns based on the day of the week.

The stepwise selection process retained previously significant weather-related predictors, such as dew point, humidity, visibility, sea level pressure, and temperature (including squared terms), but also identified Saturday, Sunday, and Monday as significant categorical variables, indicating that consumption patterns differ on weekends and the start of the workweek. The p-values for these day variables are extremely low, confirming their statistical significance.

4 Preventing Overfitting

4.1 Cross-Validation

4.1.1 Explanation

One of the most effective ways to verify that a model is not overfitting is by performing cross-validation. Overfitting occurs when a model performs exceptionally well on training data but fails to generalize to unseen data. To test this, cross-validation, particularly KFold Cross Validation, splits the dataset into multiple subsets, training the model on some subsets while testing it on others. By evaluating the model's R^2 scores across different validation sets, its stability can be assessed. If the validation scores remains almost consistent, this indicates the model generalizes well and is not overfitting. However, if there is a significant drop in validation performance compared to training, it suggests overfitting, meaning the model is too finely tuned to the training data and does not capture broader trends.

4.1.2 Results

R_i^2	Value
$i=1$	0.8713245884056604
$i=2$	0.8987936760927481
$i=3$	0.9204623566366454
$i=4$	0.9327776313685638
$i=5$	0.8574249394795141
$i=6$	0.8413617282697086
$i=7$	0.8501561911001807
$i=8$	0.7934521437725884
$i=9$	0.8436590588356765
$i=10$	0.8893128319309561

Table 7. Scores of 10 KFold Cross Validation

4.1.3 Inference

The mean R^2 value is approximately 0.87, which is essentially very close to the R^2 of the model, which means that it is not overfitting.

4.2 Regularization using LASSO (L1) or Ridge (L2)

4.2.1 Explanation

Another approach to detecting and preventing overfitting is regularization. These techniques help by penalizing excessively large coefficients in the regression model, effectively reducing complexity and improving generalization. Ridge regression shrinks all coefficients proportionally, making the model more robust against noise, while Lasso regression forces some coefficients to zero, acting as an automatic feature selector. By comparing the performance of an unregularized OLS model with Ridge and Lasso models, overfitting can be determined. If regularization significantly improves validation R^2 while lowering training R^2, it indicates that the original model was overfitting. Conversely, if regularization has little effect, the model is already well-balanced and generalizes effectively to unseen data.

4.2.2 Results

Regularization	Value
Ridge R^2	0.8775892950369358
LASSO R^2	0.8765378000952952

Table 8. Regularization Scores

4.2.3 Inference

The Ridge and LASSO R^2 values are approximately 0.88, which is essentially very close to the R^2 of the model, which means that it is not overfitting.