French Load Consumption, Phase 1

John Chidiac, Carlo Oueiss

2024-10-24

1. The Dataset

The dataset is built for energy demand forecasting, specifically in france, where the goal is to predict the load based on various factors, which are primarily weather-related. The dataset also contains temporal information such as day of the week, holidays, and whether it's a weekend, which allow us to form a more precise study. We obtain the French Load Consumption dataset from OpenML. The dataset ranges from 2017 to 2022, spanning 105k rows, and the data is recorded each 30 minutes.

We use farff to import the dataset as it is in ARFF format.

```
library(farff)
data <- readARFF("./french_load.arff")
```

2. Preprocessing and Visualizing

As a first step, we take a look at the structure of our data:

```
str(data)
```

```
## 'data.frame':    105168 obs. of  17 variables:
##  $ id              : num  0 1 2 3 4 5 6 7 8 9 ...
##  $ date            : chr  "2017-01-01T00:00:00Z" "2017-01-01T00:30:00Z" "2017-01-01T01:00:00Z" "2017-
##  $ temp            : num  -1.94 -1.97 -2 -2.04 -2.09 ...
##  $ wind            : num  1.97 1.98 1.99 1.99 1.99 ...
##  $ sun             : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Load            : num  15052349 17580789 19854004 18741368 17859804 ...
##  $ Instant         : num  2 3 4 5 6 7 8 9 10 11 ...
##  $ Posan           : num  0.000114 0.000171 0.000228 0.000285 0.000342 ...
##  $ JourSemaine     : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ JourFerie       : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ offset          : num  20 20 20 20 20 20 20 20 20 20 ...
##  $ DayType         : num  30 30 30 30 30 30 30 30 30 30 ...
##  $ Weekend         : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ temp_liss_fort  : num  -0.75 -0.784 -0.818 -0.853 -0.887 ...
##  $ temp_liss_faible: num  3.38 3.36 3.34 3.33 3.31 ...
##  $ tempMax         : num  2.81 2.81 2.81 2.81 2.81 ...
##  $ tempMin         : num  -2.12 -2.12 -2.12 -2.12 -2.12 ...
```

We can see that the `id` column is useless, and that we have a number of columns presenting categorical data, such as `JourSemaine`, `JourFerie`, `DayType`, `Weekend`, and `Instant`. Given that the given information about the dataset is limited, we choose to omit `Instant`, `DayType`, and `offset`. We will also omit rows with N/A values. We will transform `date` into a viable format. The column `JourSemaine` describes what day of the week where we can assume `0` stands for Monday, in accordance with the French system. `JourFerie` is binary data stating whether the day is a holiday or not, and `Weekend` states whether it is a weekend or not. We transform these threes to factors so they are not confused with numerical data.

```r
data <- data[ , !(names(data) %in% c("offset", "Instant", "DayType", "id"))]
data$date <- as.POSIXct(data$date, format="%Y-%m-%dT%H:%M:%SZ")
data$JourSemaine <- as.factor(data$JourSemaine)
data$JourFerie <- as.factor(data$JourFerie)
data$Weekend <- as.factor(data$Weekend)
str(data)
```

```
## 'data.frame':    105168 obs. of  13 variables:
##  $ date            : POSIXct, format: "2017-01-01 00:00:00" "2017-01-01 00:30:00" ...
##  $ temp            : num  -1.94 -1.97 -2 -2.04 -2.09 ...
##  $ wind            : num  1.97 1.98 1.99 1.99 1.99 ...
##  $ sun             : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Load            : num  15052349 17580789 19854004 18741368 17859804 ...
##  $ Posan           : num  0.000114 0.000171 0.000228 0.000285 0.000342 ...
##  $ JourSemaine     : Factor w/ 7 levels "0","1","2","3",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ JourFerie       : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...
##  $ Weekend         : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...
##  $ temp_liss_fort  : num  -0.75 -0.784 -0.818 -0.853 -0.887 ...
##  $ temp_liss_faible: num  3.38 3.36 3.34 3.33 3.31 ...
##  $ tempMax         : num  2.81 2.81 2.81 2.81 2.81 ...
##  $ tempMin         : num  -2.12 -2.12 -2.12 -2.12 -2.12 ...
```

To gain a general view of the dataset, we will take a look at a summary of the data.

Now, we will explore the variables.

```r
summary(data)
```

```
##       date                          temp             wind
##  Min.   :2017-01-01 00:00:00.00   Min.   :-7.763   Min.   :0.8218
##  1st Qu.:2018-07-02 17:52:30.00   1st Qu.: 7.326   1st Qu.:2.3833
##  Median :2020-01-01 11:45:00.00   Median :12.104   Median :3.0636
##  Mean   :2020-01-01 11:09:44.11   Mean   :12.625   Mean   :3.3302
##  3rd Qu.:2021-07-02 05:37:30.00   3rd Qu.:17.586   3rd Qu.:4.0259
##  Max.   :2022-12-31 23:30:00.00   Max.   :36.163   Max.   :9.7061
##
##       sun              Load              Posan        JourSemaine JourFerie
##  Min.   :  0.0   Min.   : 1691236   Min.   :0.0000   0:15024     0:101424
##  1st Qu.:  0.0   1st Qu.: 4557927   1st Qu.:0.2501   1:15024     1:  3744
##  Median :  0.0   Median : 8093514   Median :0.5001   2:15024
##  Mean   :136.1   Mean   :10143159   Mean   :0.5001   3:15024
##  3rd Qu.:231.2   3rd Qu.:15116449   3rd Qu.:0.7501   4:15024
##  Max.   :884.1   Max.   :32458947   Max.   :1.0000   5:15024
##                                                      6:15024
##  Weekend   temp_liss_fort   temp_liss_faible     tempMax          tempMin
##  0:75120   Min.   :-5.051   Min.   :-0.1201   Min.   :-1.406   Min.   :-7.763
##  1:30048   1st Qu.: 7.620   1st Qu.: 7.5386   1st Qu.:10.849   1st Qu.: 4.561
##            Median :12.162   Median :12.0502   Median :16.344   Median : 8.607
##            Mean   :12.621   Mean   :12.6069   Mean   :16.612   Mean   : 8.804
##            3rd Qu.:17.945   3rd Qu.:18.2048   3rd Qu.:22.374   3rd Qu.:13.425
##            Max.   :30.020   Max.   :24.3297   Max.   :36.163   Max.   :21.412
##
```

We can see that the dates range from January 1, 2017 until December 31, 2022. Looking at the `Load`, it is clear there is a large difference between the minimum and maximum, so our task will be investigating the variables that impact `Load`.

To give a general overview of the temperature trends, we will smooth them using a moving average. We speculate that the `sun` variable either denotes the visibility of the sun, or the impact of the sun in a given day. We can find evidence of this by observing that the plots for temperature and sun follow the same pattern. We assume the opposite trend for wind, and this is confirmed by the plot.

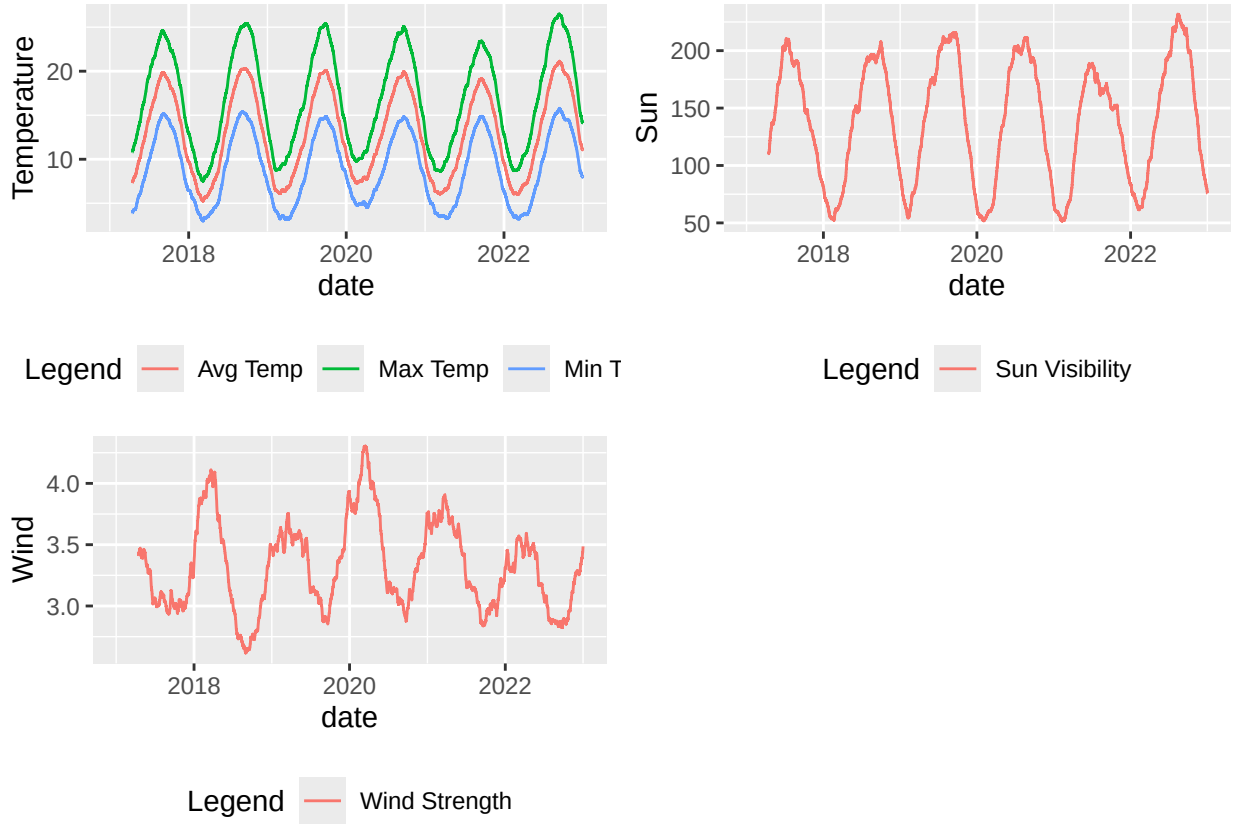

Figure 1: Temperature (left), sun visibility (middle), and wind strength (right) through time (2017-2022)

We hypothesize that electricity consumption remains, to an extent, the same between weekends and weekdays, and decreases significantly during week-day holidays and even more when it is a weekend and holiday.

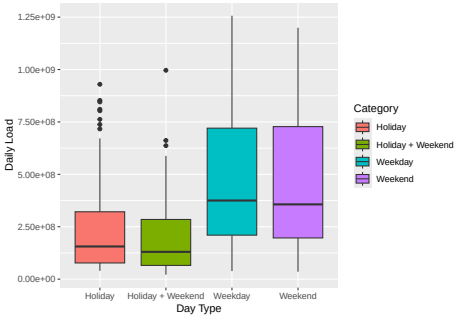

Figure 2: Electricity consumption on weekdays, weekends, and holidays

Regression

In this section, we will start with simple linear regression, then slowly progress into exploring more complex relationships between predictors through more complex techniques such as multiple linear regression with interaction.

Simple Linear Regression

Having taken a look at the temperature trends, we observe they are close to identical each year, so we will currently only concern ourselves with tasks during 2017. We will also disregard changes throughout the day, calculating daily `Load` as the sum of `Load` throughout the day, and the `temp`, `sun`, and `wind` as their respective means. For easier interpretation of linear regression values, we will scale the load values between 0 and 100. As a simple first experiment, we try to establish a relationship between `temp` and `Load`.

```
##
## Call:
## lm(formula = DailyLoad_scaled ~ temp, data = jan_to_june)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -67.738  -6.578   1.309   9.433  28.871
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   78.586      1.938   40.55   <2e-16 ***
## temp          -3.452      0.154  -22.42   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.54 on 190 degrees of freedom
## Multiple R-squared:  0.7257, Adjusted R-squared:  0.7242
## F-statistic: 502.6 on 1 and 190 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = DailyLoad_scaled ~ temp, data = july_to_dec)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -48.457  -5.243  -0.656   7.556  20.004
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  67.8668     2.6330   25.77   <2e-16 ***
## temp         -2.8514     0.1635  -17.45   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.09 on 159 degrees of freedom
## Multiple R-squared:  0.6568, Adjusted R-squared:  0.6547
## F-statistic: 304.3 on 1 and 159 DF,  p-value: < 2.2e-16
```

We split the 2017 data into two parts, since it is evident that the inverted bell shape presented by the temperature throughout time will result in a very poor linear model. This split allows for a more accurate model. In the linear model for January to June, the multiple R-squared is 0.7257, meaning that about 72.57% of the variance in the scaled daily load is explained by the temperature, the RSE is 13.54, meaning the model is making reasonable predictions, and the p-value is less than 0.001, indicating the relationship between temperature and daily load is significant. In the model for July to December the multiple R-squared is 0.6568, meaning that about 65.68% of the variance in the scaled daily load is explained by the temperature. We also obtain a RSE of 11.09, meaning on average, we may be around 11.09 the true value, which is reasonable to some extent, for this model, the p-value is also less than 0.001. In both cases we observe a large F-statistic, and we can conclude that temperature is a useful predictor of scaled daily load.

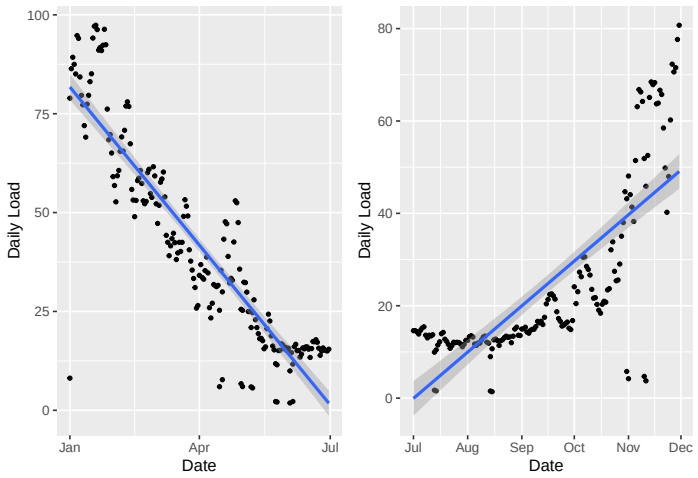

Figure 3: Electricity consumption between the months of January to June, and July to December

Next, we try to establish a simple linear relationship between `sun` and `Load`, as well as `wind` and `Load`. We expect a strong relationship between former, and a weaker one between the latter. We will also scale both `sun` and `wind` variables between 0 and 100 for easier interpretability of results. We will take into consideration the original data, rather than the cropped and aggregated daily data.

```
##
## Call:
## lm(formula = Load_scaled ~ sun_scaled, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -31.195 -16.978  -4.546  14.858  70.215
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 31.195095   0.074965  416.13   <2e-16 ***
## sun_scaled  -0.241885   0.002713  -89.15   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.18 on 105166 degrees of freedom
## Multiple R-squared:  0.07026,    Adjusted R-squared:  0.07026
## F-statistic:  7948 on 1 and 105166 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = Load_scaled ~ wind_scaled, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -36.339 -16.937  -7.068  14.963  74.194
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 19.759143   0.142818  138.35   <2e-16 ***
## wind_scaled  0.273118   0.004532   60.27   <2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.58 on 105166 degrees of freedom
## Multiple R-squared:  0.03339,    Adjusted R-squared:  0.03338
## F-statistic:  3632 on 1 and 105166 DF,  p-value: < 2.2e-16
```

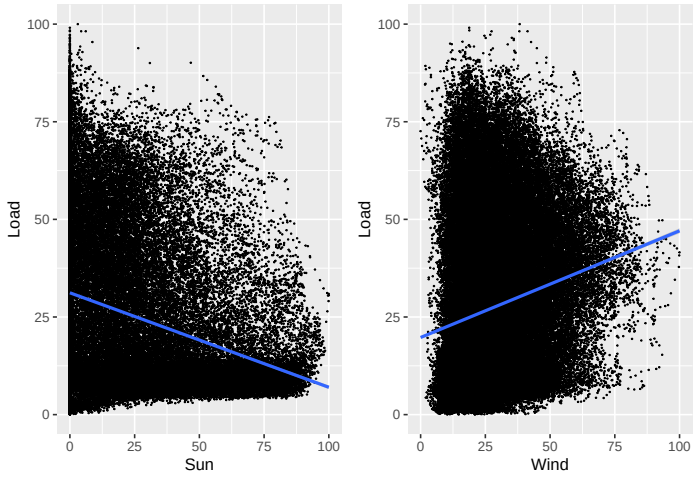

Figure 4: Electricity Consumption vs Sun Visibility (left) and Electricity Consumption vs Wind Strength (right)

We find that around 7% of the variability in the data can be explained by the sun conditions, while only 3% can be explained by the wind.

Multiple Linear Regression

From the results above, we hypothesize that using both the `sun` conditions and `temp` as predictors would give us a more accurate model. Thus, we perform multiple linear regression with these two predictors.

```
lm_temp_sun <- lm(Load_scaled ~ sun_scaled + temp, data)
summary(lm_temp_sun)
```

```
##
## Call:
## lm(formula = Load_scaled ~ sun_scaled + temp, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -31.281  -7.880  -1.062   7.929  42.414
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 60.063361   0.070921   846.9   <2e-16 ***
## sun_scaled   0.198629   0.001724   115.2   <2e-16 ***
## temp        -2.823847   0.005668  -498.2   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.01 on 105165 degrees of freedom
## Multiple R-squared:  0.7233, Adjusted R-squared:  0.7233
## F-statistic: 1.375e+05 on 2 and 105165 DF,  p-value: < 2.2e-16
```

The multiple linear regression model that includes both temperature and sunlight performs similarly to the temperature-only model, but adding sunlight doesn't provide a significant improvement in terms of R-squared, it has a slightly lower residual standard error (11.01). Finally, the F-statistic is massively large, indicating that this model is overall very statistically significant.