**Linear Regression Model to predict the factors impacting the Temperature.**

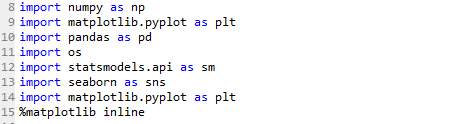
Our goal is to explore the factors which have an impact on the Temperature in Hungary. Using ML Technique in python we find out the factors having an impact on the Temperature.

The Dataset contains the following variables:

* time
* summary
* precipType
* temperature
* apparentTemperature
* humidity
* windspeed
* windBearing
* visibility
* loudcover
* pressure

Here the variable temperature is our dependent variable.

First, we import the libraries.



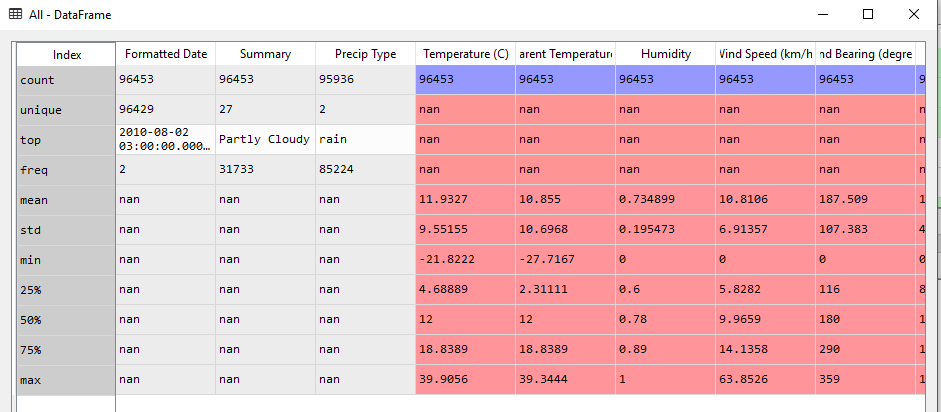
Then, we read the dataset in python.



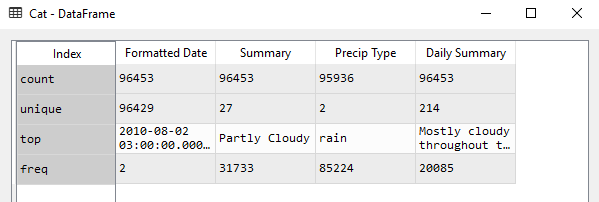


Then we see some statistical values of this data.





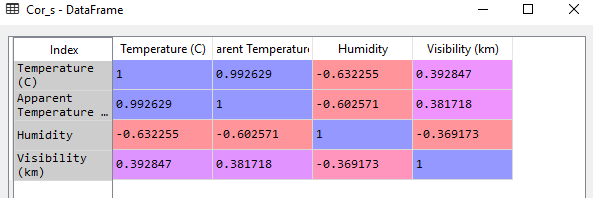




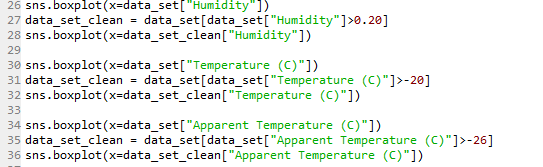


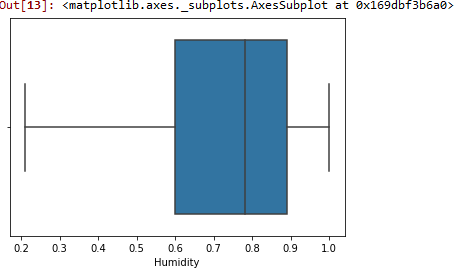
Here we take the columns which are necessary for prediction.

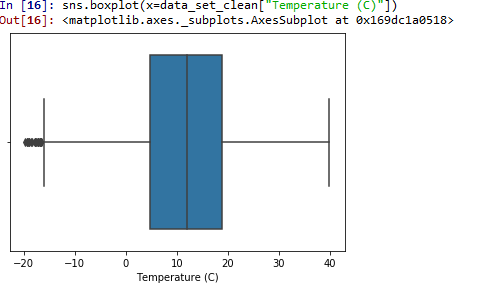


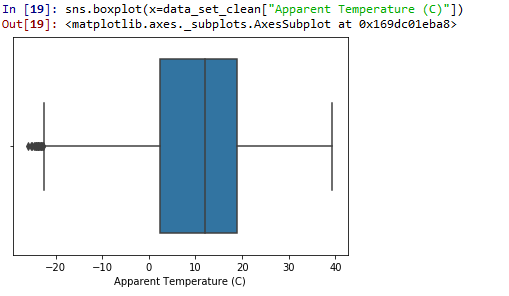


We check if there are any outliers or not with boxplot and we clean them:









Setting up the dependent and the independent variables:



Setting up dummies for the categorical variable “Precip Type”:



Splitting the data into test and train sets:

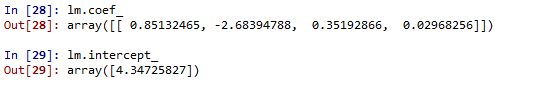


Fitting the Model:



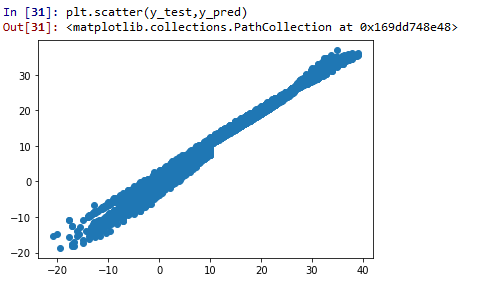


Here we see the coefficients:

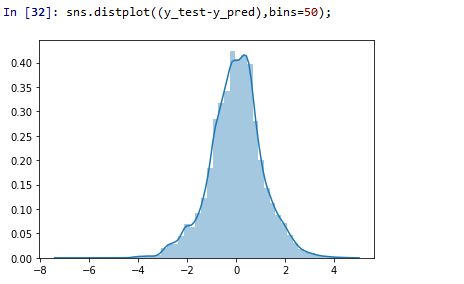


Predicting the model and viewing the residuals:





We can see the scatterplot of the predicted and the dependent variable and we can see that we have predicted quite well.



We can see the residuals are also normally distributed.

Checking the goodness of the fit :

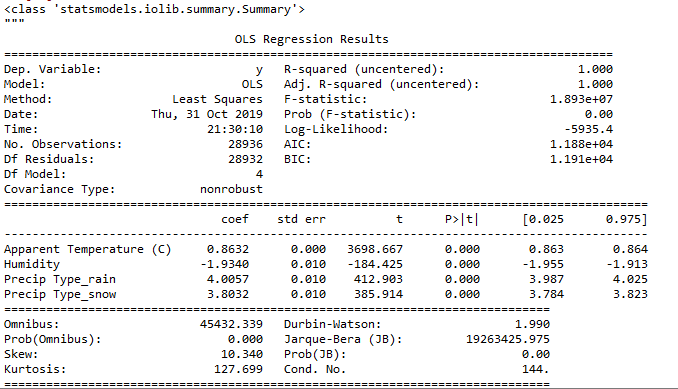




The coefficient of determination which is denoted by R2  is a key output of the regression analysis. It is interpreted as the proportion of variance in the dependent variable that can be explained by the independent variable. The value of R2  ranges from 0 to 1 and the closer the value is to 1, the better is the model. In our model, the value is 100% approximately which indicates a high amount of goodness of fit of the model.

The adjusted R2 which provides a more honest value to estimate the R-squared for the population because with the normal R-squared, the value increases with the increase in the number of predictors. The value of the adjusted R-squared in our case also comes to around 100%.In real world datasets we wont find this high R2 though.





Here we can also see the results of the model using the statsmodel.api. Here also we see the R2 and the adjusted R2 is 1.

***Multicollinearity:***





This is the most important of the assumptions of a linear model and it states that there should be no perfect linear relationship between two or more of the predictors or independent variables. This is tested with the vif function and any variable with a value of VIF significantly more than 4 will have to be removed the model. In our case multicollinearity between independent variables was present in the “Precip Type” hence we remove one of them.

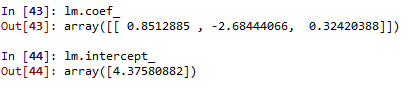




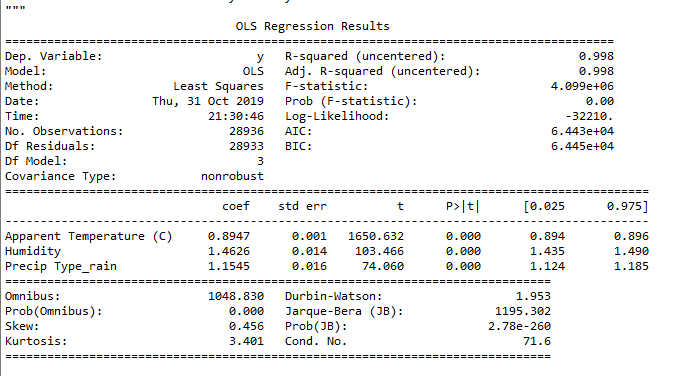
After removing the precip type snow column from the data we still see some multicollinearity present and this is due to the dataset that we have chosen to work with. We cannot remove any more independent variables as there are too few independent variables left to predict the dependent variable with.

Fitting the final model:









This is the final model and we can see the R2 is 99% which is good.

***serial correlation: -*** This assumption states that for any two observations, the residual terms should be uncorrelated or independent. This assumption is checked with the help of the Durbin–Watson test. The value of the statistic ranges from 0-4 and a value close to 2 indicates no correlation. In this case, the assumption holds with a D-W statistic value of approximately 2.

Business Interpretation:

From the model we can see that the temperature is directly affected by the other factors such as apparent temperature and humidity and precipitation type rain. From the coefficients of the independent variables we can see how much it affects the temperature.