Swimming Pool Prediction

code link: Github

Introduction:

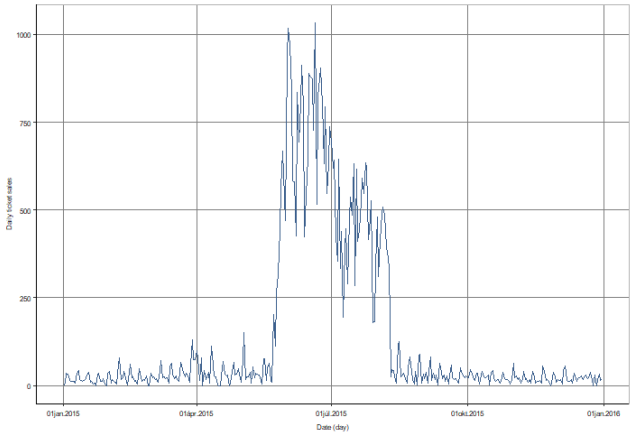
The aim of this analysis to predict a year of usage of outside swimming pool in quantity based on the previous years admissions. The analysis uses a time series data on type of admission, location, quantity, admission cost and date. This report will detail the data preparation process including the cleaning and filtering, the model creation and choosing, and finally, the evaluation of the prediction. The prediction will be made for 2016 based on the chosen model and the data from the previous years.

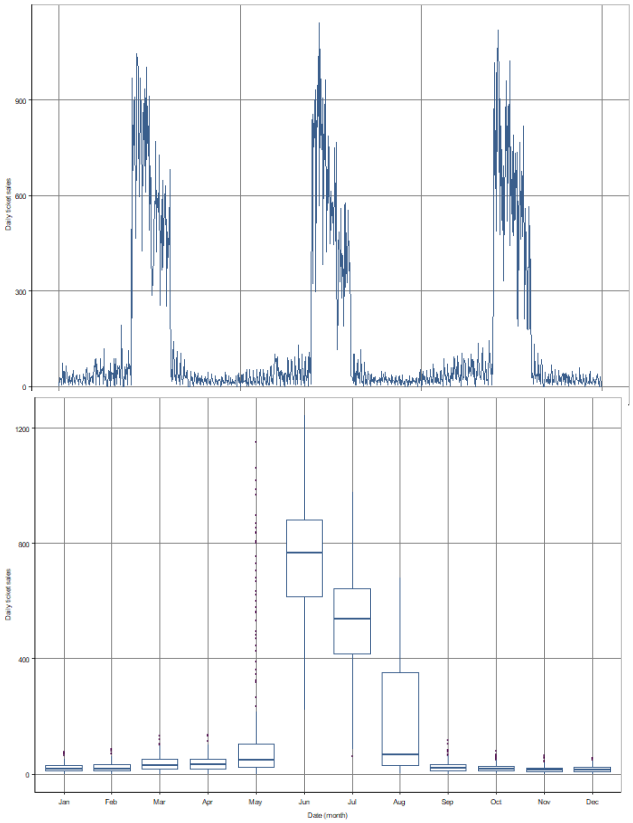
Data Preparation:

To begin with, because the prediction focuses on the open-air swimming pools, therefore those were filtered for. Also, the date column was set for date to be able to aggregate the quantity and day. After that, different columns were made for year, month, day, the day of the week. Also, columns were made with Boolean or binary variable to measure whether the date was on summer- or simple holiday, or on weekend. On the other hand, to avoid having infinite values at the log transformation, where the quantity was lower than 1 it was made 1 otherwise left as it was before.

Descriptive Charts:

Before starting the model creation and prediction, a basic inspect of the time series pattern through the years may help for understanding the data. First of all, for a year, there is a visible increase in the sales of tickets for pools in 2015 that can be observed through in the previous years as you can see in the below line charts. While the analysis uses the sales data from 2010 until 2015, the chart only visualizes from 2012 to 2014. A clear seasonality pattern can be observed through the 3 years.





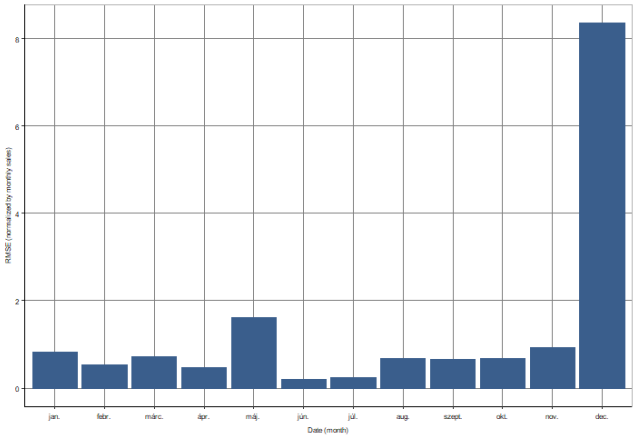
Similar seasonality can be noticed in monthly sales in all of the mention years with boxplot. However, the extreme values exceed the 50% of the sales distribution in each month that may be considered in the long run.

Model Evaluation:

For the prediction, three different simple multiple regression models were created to see which one works better for the prediction for whole 2016 sales. The first one only includes the trend, the month and day of the week, the second model adds to the first model the holiday, the interactions between the summer holiday and day of week and between weekend and month. And finally, the third model compares the log transformed quantity to the trend, the month, the day of the week and the interaction between the school holiday and day of the week. All of the models were trained on the training set that resulted in the following RMSEs which helps to see how well the model fits the data:

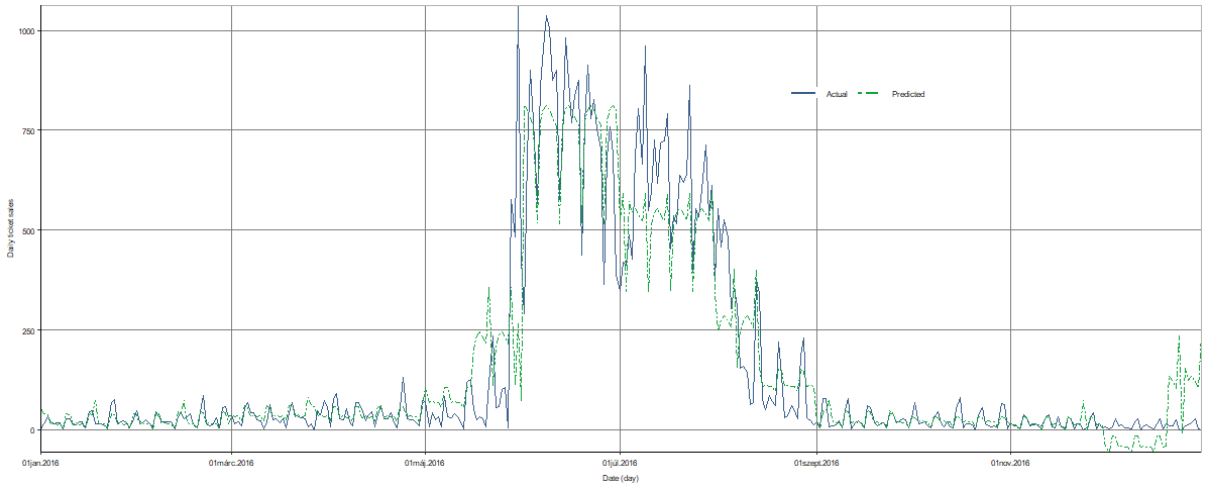
	model 1	**model 2**	**model 3**
RMSE	121.7983	109.6712	164.8688

As it can be seen that the model 2 has the lowest RMSE that indicates a better fit considering that RMSE measure the distance of the actual values from the regression line. This validation method in which the holdout set is separated from the training set by a date is called as supervised learning. However, it gave and 99.36 for the RMSE that illustrates that the model fits better for reality. It must be mentioned that the REMSE tends to increase significantly by end of the year:

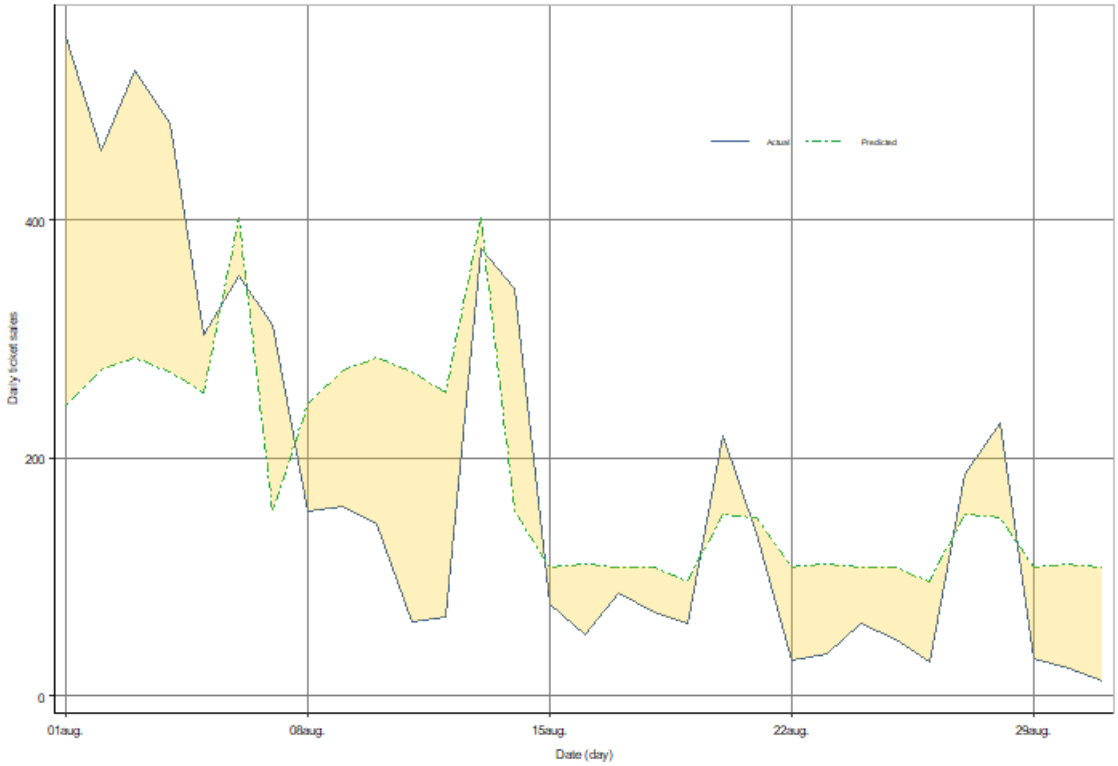


Prediction Evaluation:

As it was mentioned before, the prediction was made for 2016 which resulted in the following:



It can be seen that the prediction captured the pattern considerably well however it is unsuccessful to predict the extreme values through time to time. Also, there is a significant difference between the actual and predicted values at the end of the year that can be explained by the long-run prediction but by the high increase in RMSE in December that was visualized before. In the following line chart, the difference between the predicted and actual values can be seen better for a month sale in August.



Conclusion:

To sum up, the aim of this analysis was to predict a daily sale of open-air swimming pool admission sales for a year. Three different multiple regression models were created from which the best one was picked based on the RMSE values. In order to see the external validity of the model, a supervised learning was implemented that gave an insight of a good fit for other datasets as well. The prediction itself captured the seasonality well however it lacks of covering the extreme values through time to time. Also, it presented a high variety by the end of the year.