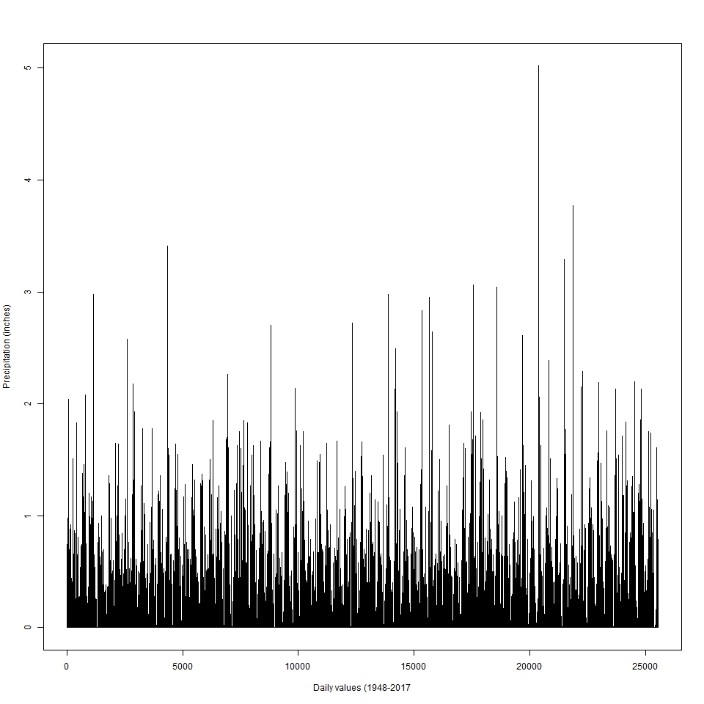
**Summary of the analyses:**

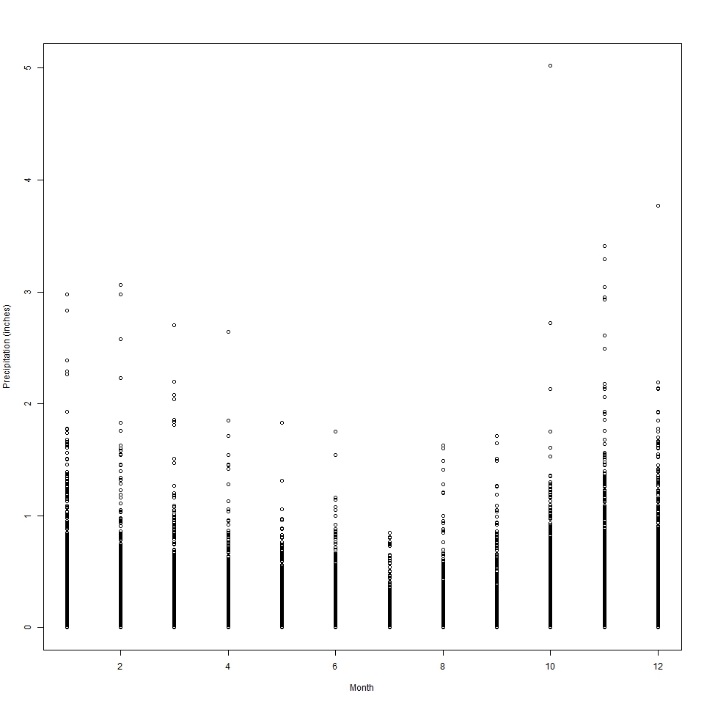
**Objective:** Can we predict if today will rain based on yesterday data?

**1. Exploratory data analysis**

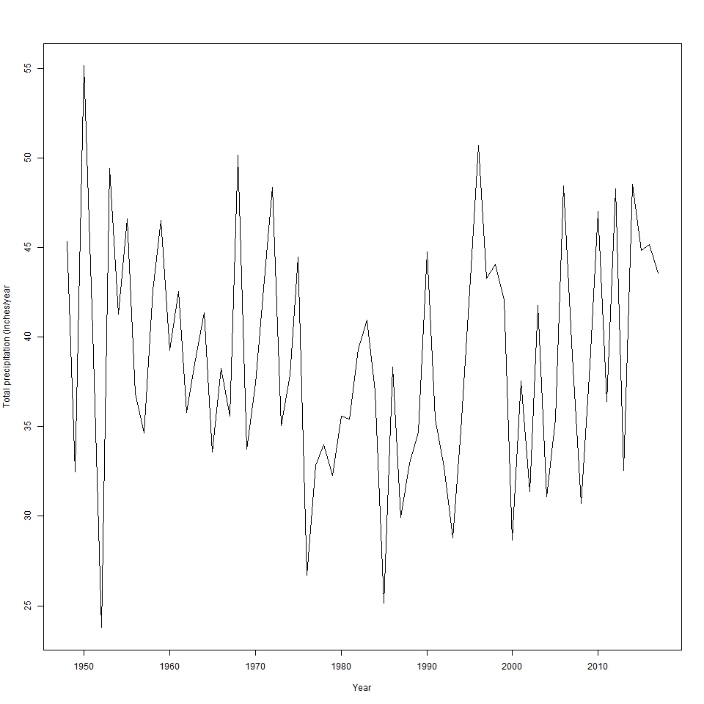
There is a cycle in the amount of daily rain:



The amount of precipitation is month dependent:



There is wide variation in the total amount of precipitation from year to year, which implies that it will somewhat difficult to construct a highly predictive model



**2. Modeling approach**

**Model:**

Rain\_today(T/F) = function(Rain\_yesterday(T/F) + Precipitation\_yesterday(inches) + Temp\_max\_yesterday (F) + Temp\_min\_yesterday (F) + Month (1-12) )

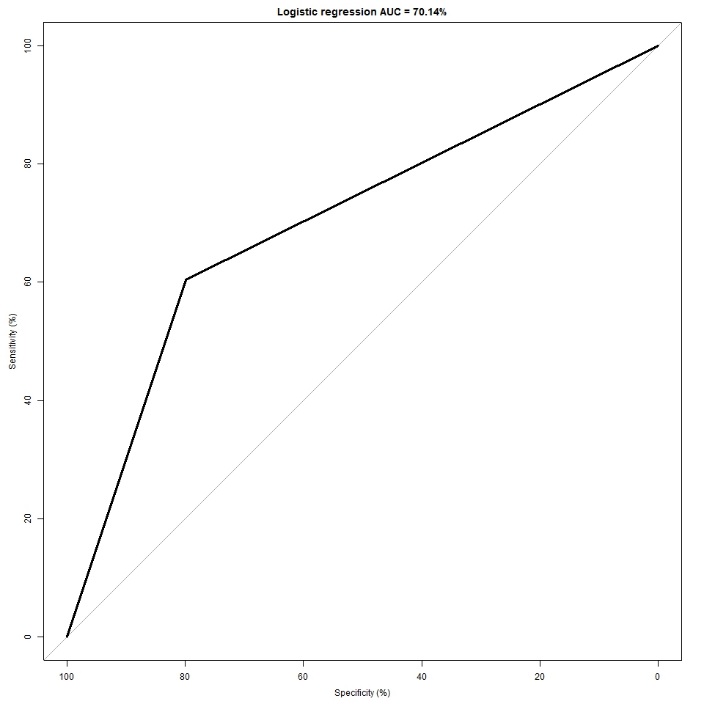
First we fitted the model using several machine learning algorithms in R. The dataset was randomly split in two subsets, the training and the testing datasets. In this case a 60:40 training to testing ratio was used, but other ratios can be employed. After training using the selected ML methods (logistic regression, Naïve Bayes, decision trees, bagging, Random Forests, Support Vector Machine (SVM), and Artificial Neural Networks (ANN) algorithms), the models performance was evaluated using AUC and ROC curves. While there is a wide variety of metrics that can be used to evaluate performance, these are the most common used in ML classification. Briefly, the receiver operating characteristic (ROC) curve plots the true positive rate vs. the false positive rate. AUC (area under the curve) actually refers to the AUROC (area under the ROC curve) and ranges between 0.5 - classifier with no predictive power - and 1.0 - perfect classifier ([Lantz, 2015](#_ENREF_1)). As a word of caution, probably a wider selection of metrics should be used in model selection, since AUC values may fail to entirely capture the model quality ([Lobo et al., 2008](#_ENREF_2)).

The analysis showed that logistic regression, decision trees, bagging and ANN resulted in fairly good classifiers (above 0.7) considering the law number of predictor variables, while NB, RF, SCM resulted in poor classifiers (AUC values between 0.6 and 0.7).

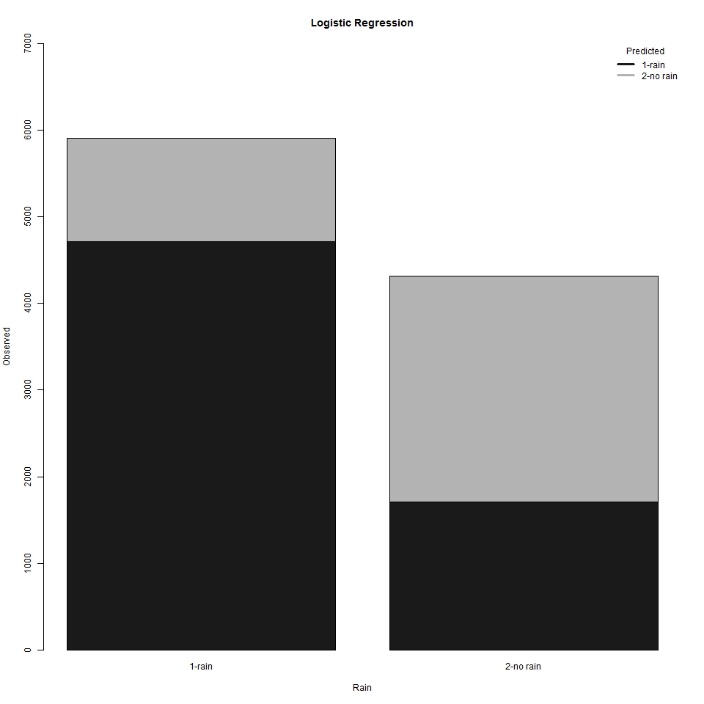
Considering its simplicity, we chose the model proposed by the logistic regression (AUC value of 0.7014, marginally smaller than the best model 0.7036 built using ANN).

The summary plots for each ML method considered are loaded in the folder.

ROC plot:



Observed vs. predicted: how to read it: there were 5905 rainy days and about 4315 days without rain in the test dataset. From the rainy days, the logistic regression model predicted 4715 and misclassified about 1190. For the clear days, the model predicted correctly 2608 and misclassified 1707.



Lantz, B., 2015. Machine Learning with R, 2nd ed. Packt Publishing Ltd., Birmingham, UK.

Lobo, J.M., Jiménez-Valverde, A., Real, R., 2008. AUC: a misleading measure of the performance of predictive distribution models. Global Ecol Biogeogr 17, 145-151.