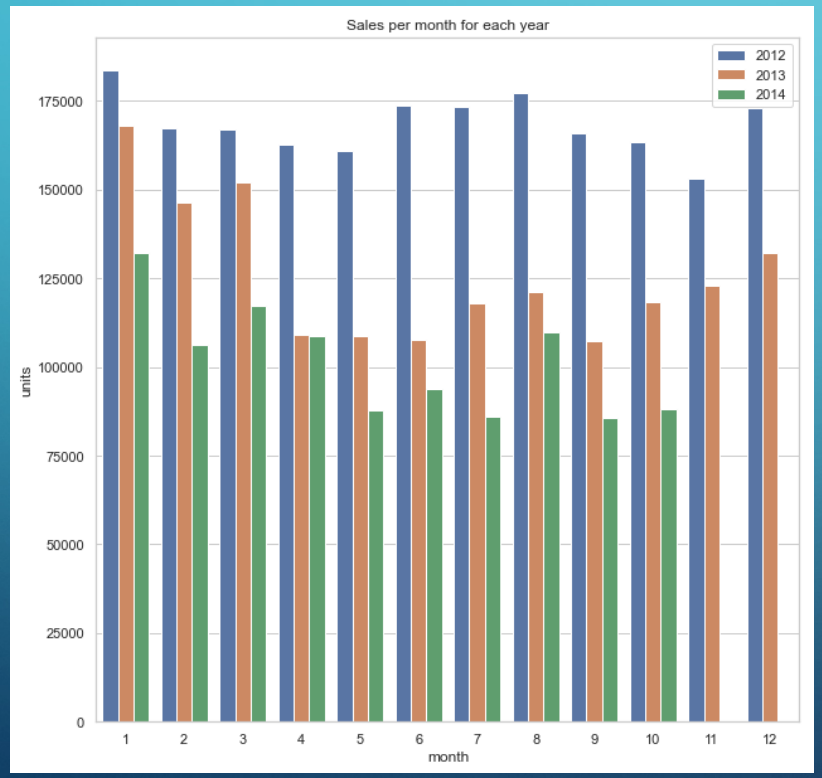
Machine learning EX2

Guy Yehezkel
Aviv Lugasi

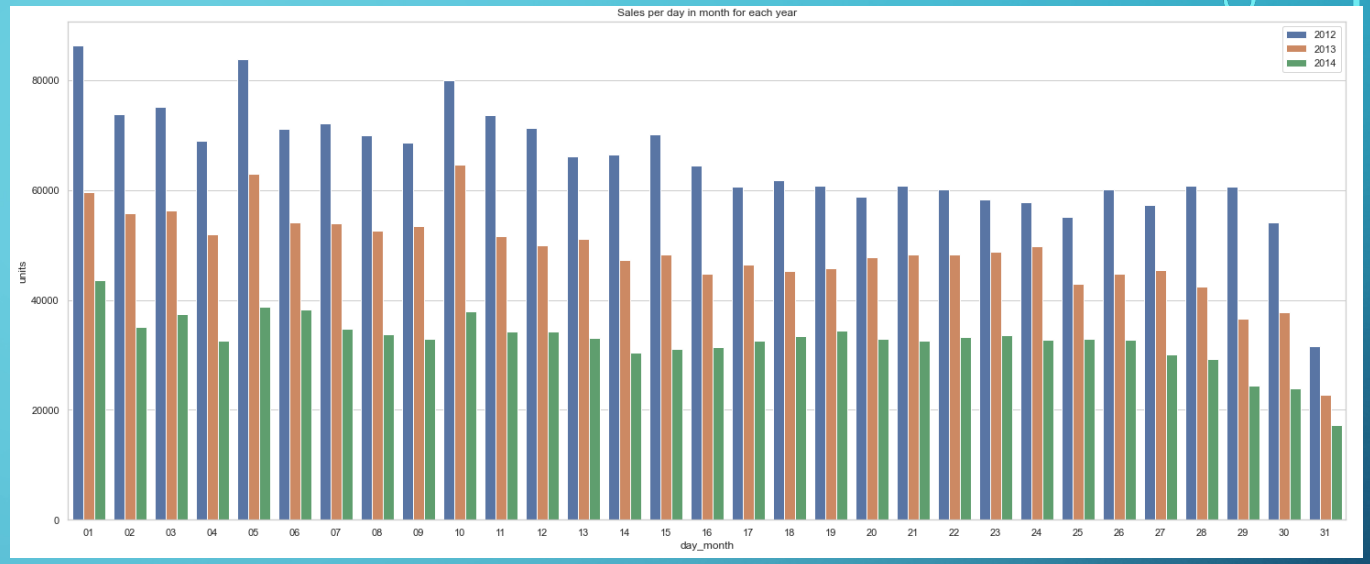
Section A - Exploration

First, we will present a few interesting findings that came up during the exploration process

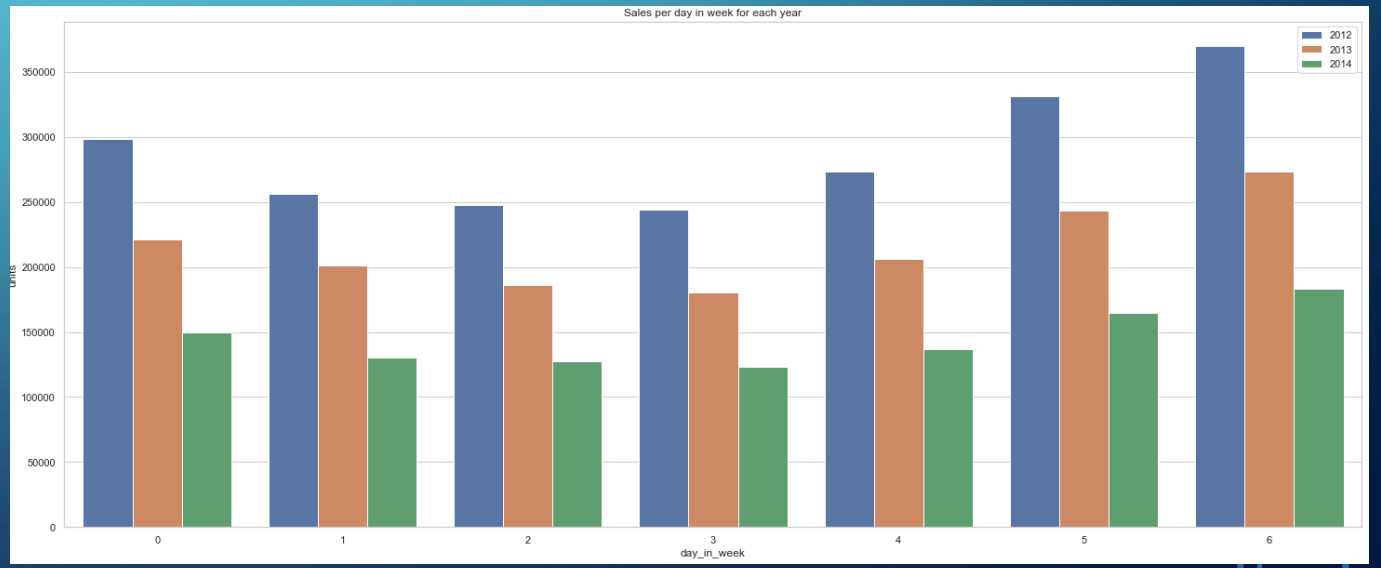
1. We have found that there has been a consistent decline in product sales over the years. Although the data describe sales of only three years, it can be predicted that the downward trend will continue in the coming years as well.



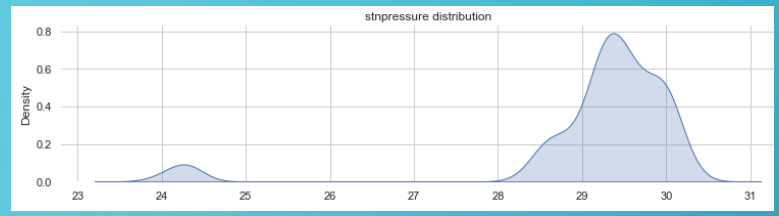
2. We examine the sales along the days in a month in each year. As we can see the highest sales are at the beginning of month, and that the sales at the end are drastically lower for each year. The decline in sales between the first five days of the month and the last five days of the month is estimated at about 51% on average. Then there are some peeks at the 1th, 5th and the 10th days of the month, also for each year. We assume that the reason for that peeks is that these days are the days which the salary come in to all employees accounts.



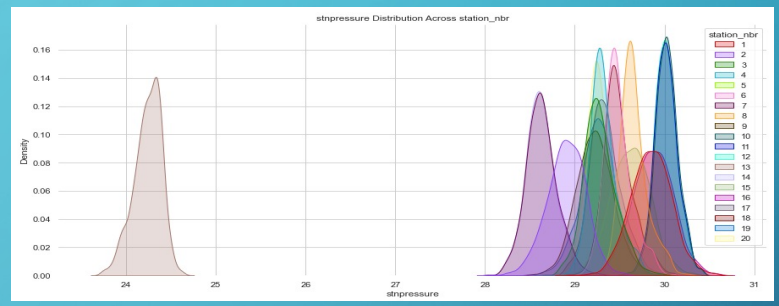
3. We also examine the sales along the week. As we can see, sales are higher on Fridays, Saturdays and Sundays which is the weekend, compared to midweek.



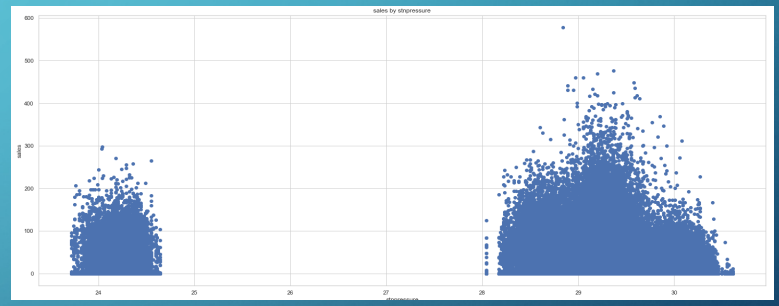
4. Examining the numeric features of weather data, we found that the pressure are divided to two clusters.



Explore the distributions across the stations, we found that the pressure value at station 13 was extremely low compare to the other stations.



We therefore chose to examine the sales data of station 13, which in the rest of the features is more or less the same as the other stations, and to investigate the effect of the pressure on actual sales.



<u>Findings</u>

The pressure feature divides our stores into 2 clusters:

1. Station number 13.
2. All other stations.

We found that the average sales of stores associated with Station 13 (1.167) is 20% higher than the average sales of stores (0.976) associated with other stations.

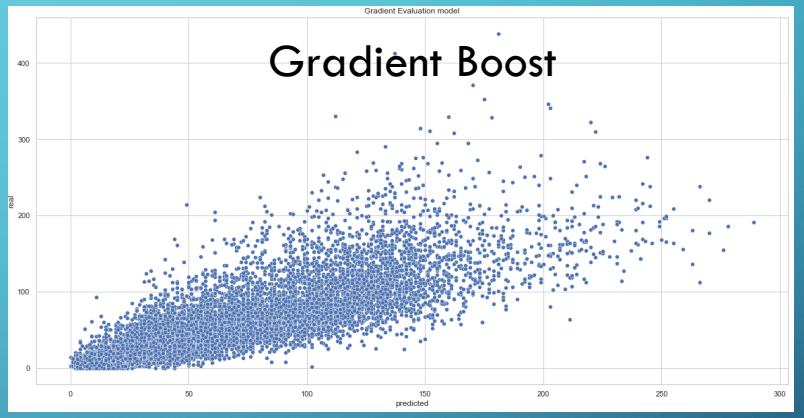
It can be concluded that low pressure causes an increase in sales.

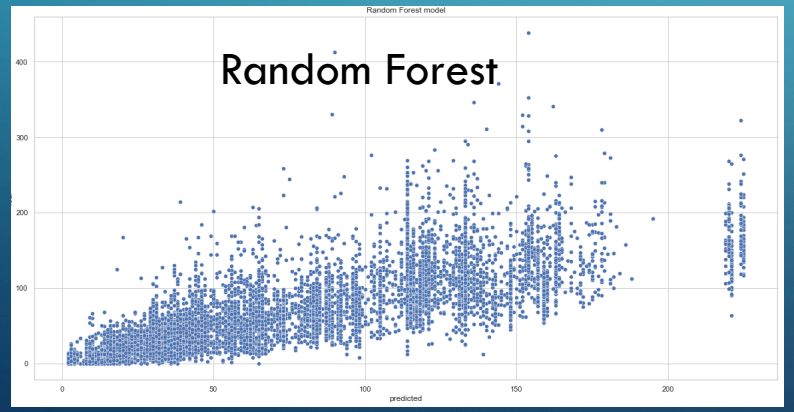
Also, we found that the importance of the pressure feature is the highest of all the features in predicting keysum in section C

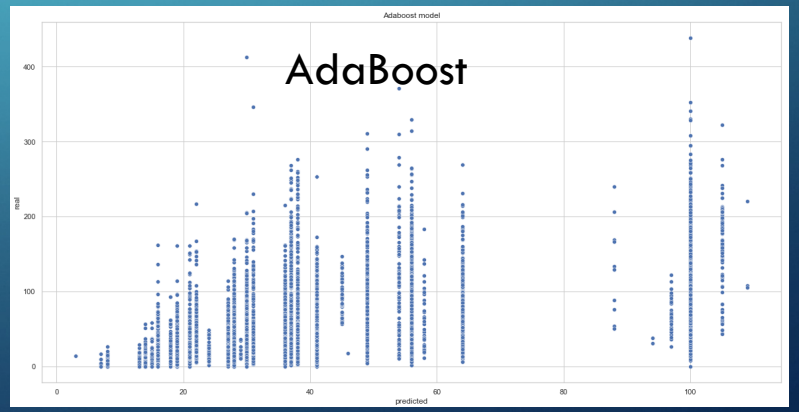
Section C

Here we have chosen to use all the models learned in class in order to maximize the effectiveness of the prediction. The models we chose to use are: Regression Tree, Gradient Boost, Random Forest, AdaBoost.

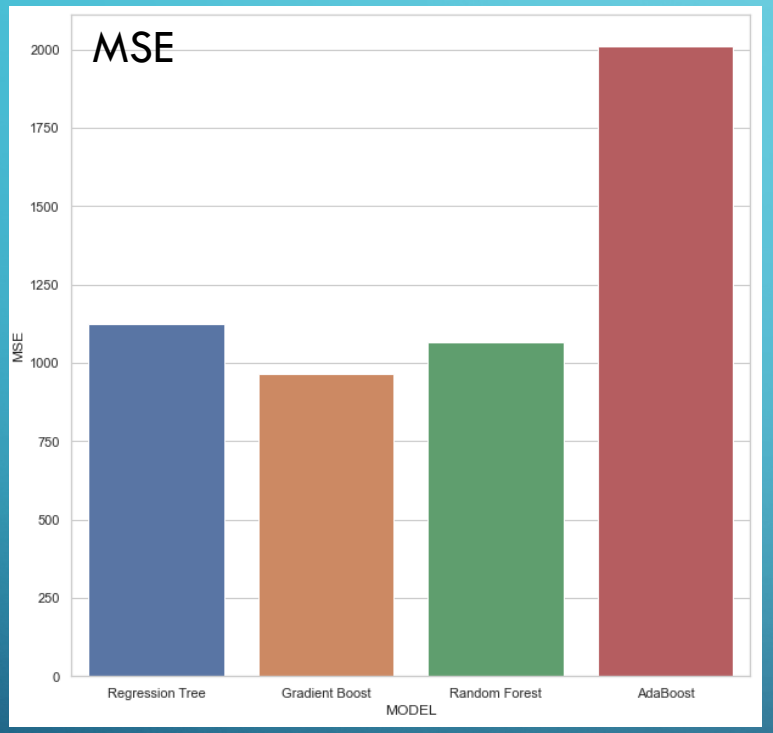
The following graphs describe the success of the prediction. The Y axis is the true values, and the X axis is the predicted values. The more the X, Y values of the points close, the more accurate the model prediction is.
Here are the results :



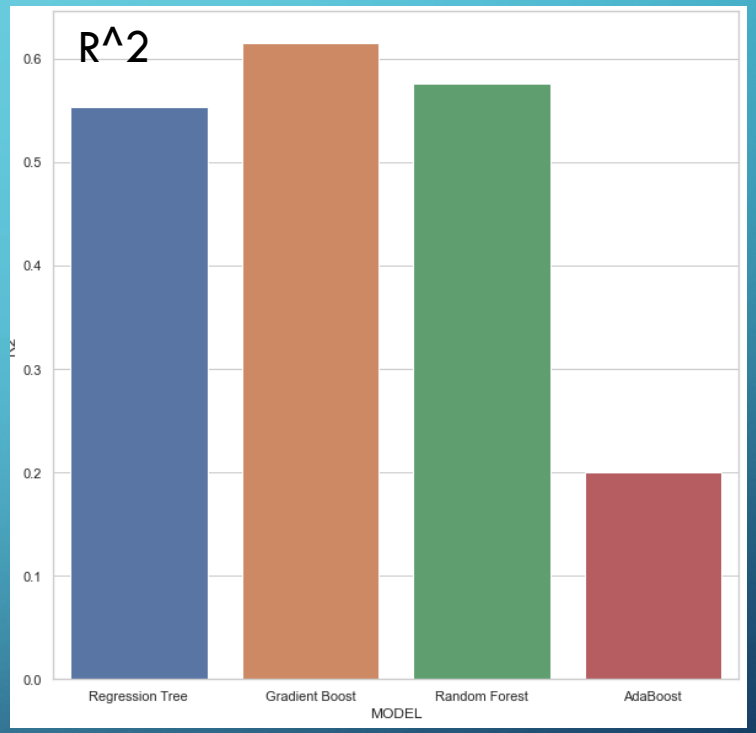




As can be seen in the graphs, the model that gave the best results is a Gradient Boost, followed by a Random Forest. In addition, we chose to compare the models by the metrices : R squared and MSE.
Even in these indices, it can be clearly seen that the results of the Gradient Boost model are the best.
Here are the results:



Reggresion tree	Gradient Boost	Random Forest	AdaBoost
1116.955	932.651	954.43	1871.19



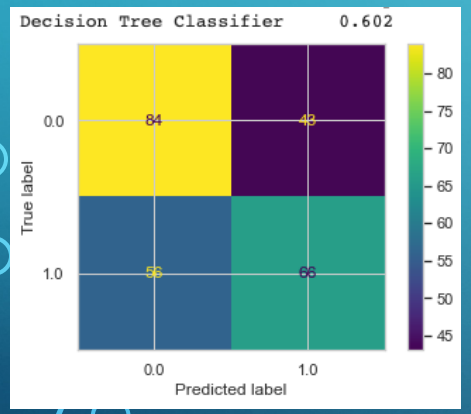
Reggresio n tree	Gradient Boost	Random Forest	AdaBoost
0.556	0.629	0.620	0.256

Section D

An examination of the sales data of Station 11, we found that during the three years only the products 9,61,68,86,87,110 were sold.
We wanted to check if there are products whose increased sales indicate precipitation. We found that products number 110 (7.1%), 61 (21%), 86 (158%), 87 (8.9%) were sold more on days with precipitation.

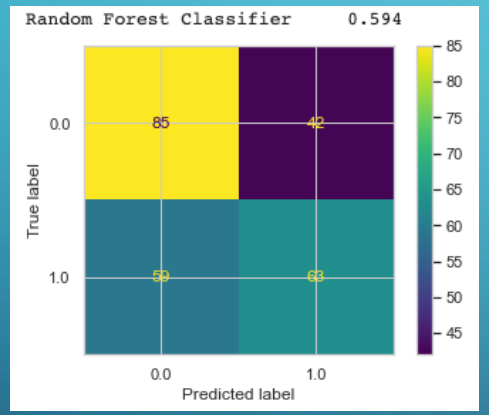
After pre-processing we chose to compare the following models to predict whether there was precipitation or not: Classification tree, Random Forest, Gradient Boost, AdaBoost, AdaBoost
The models that gave the best results are KNN and Classification Tree.

Classification tree



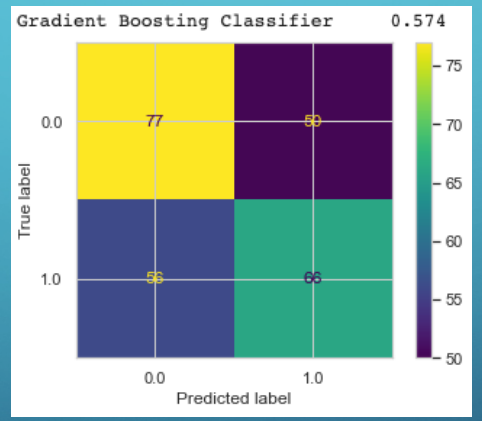
Accuracy	sensitivity	specificity
0.602	0.661	0.541

Random Forest



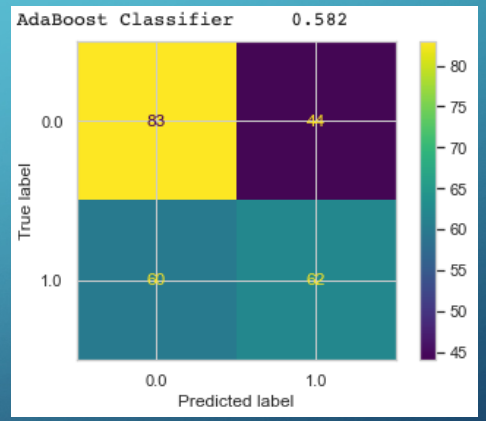
Accuracy	sensitivity	specificity
0.594	0.669	0.516

Gradient Boost



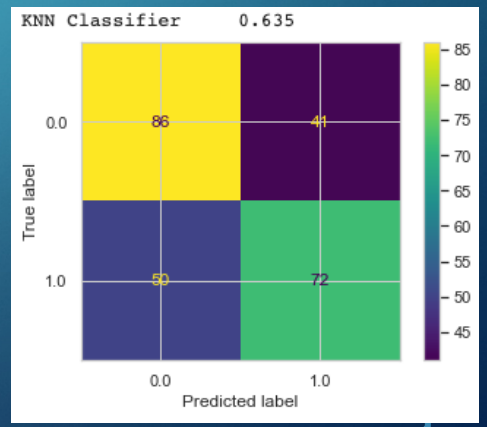
Accuracy	sensitivity	specificity
0.574	0.606	0.541

AdaBoost



Accuracy	sensitivity	specificity
0.582	0.654	0.508

KNN

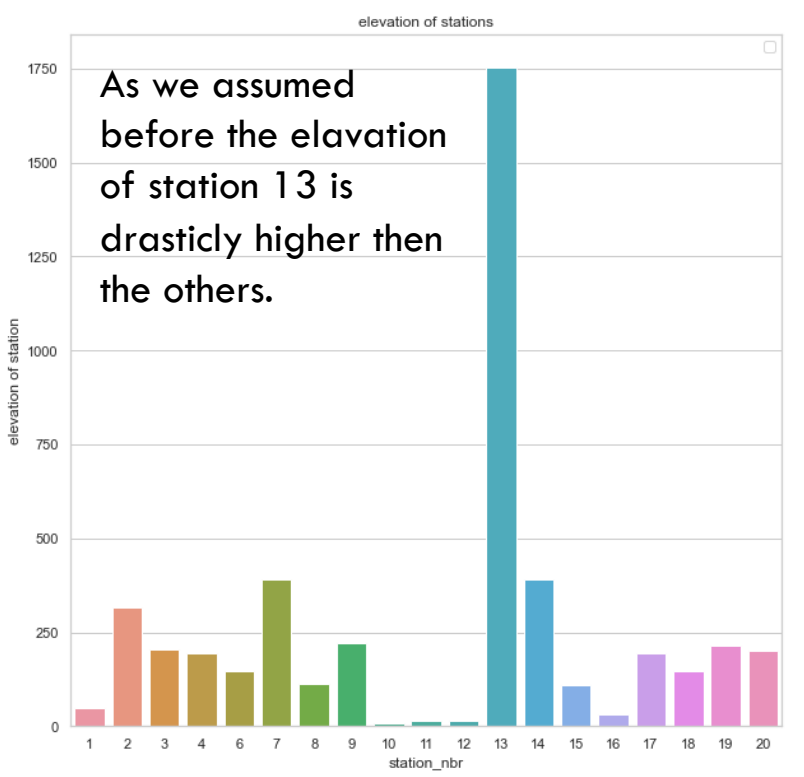


Accuracy	sensitivity	specificity
0.635	0.677	0.590

Section E

Barometric formula

The barometric formula, sometimes called the exponential atmosphere or isothermal atmosphere, is a formula used
to model how the pressure (or density) of the air changes with altitude. The pressure drops
approximately by 11.3 pascals per meter in first 1000 meters above sea level. formula source



$$P = P_b \left[\frac{T_b + (h - h_b)L_b}{T_b} \right]^{\frac{-g_0 * M}{R * L_b}}$$

where:

P_b = static pressure (pressure at sea level) [Pa]

T_b = standart temperture at sea level[K] = 288.15

L_b = standart temperture lapse rate [K/m] = -0.0065

h = height

h_b = height at bottom of atmospheric layer(0 at sea level)

R = universal gas constant [N*m/mol*K] = 8.31432

g_0 = gravitational acceleration rate constant [m/s^2] = 0.0289644

M = molar mass of earth's air [Kg/mol] = 0.0289644

if we isolate the height that we want to calculate we get:

$$h = h_b + \frac{T_b}{L_b} * \left[\left(\frac{P}{P_b} \right)^{\frac{-RL_b}{g_0 * M}} - 1 \right]$$

Section F

The data we want to display consists of 17 features, ie it is represented in 17 dimensions.
In order to make data virtualization we must reduce the dimensions to a maximum of 3.
We chose to perform a PCA in order to reduce the dimensions.
We will select the 2 PCs that take up most of the overall variance of the data.
Our goal is to represent the data in 2 dimensions and see how it is divided into clusters.
We performed a PCA algorithm and saw that PC1 captures the most variance compare to all other PCs. So, we examine the 4 most important features in the creation of PC1, are they : 'tavg', 'wetbulb', 'tmin', 'tmax'. We will now perform a clustering algorithm on the data and present the results in a two-dimensional graph where the features describing the axis will be each pair out of 4 in the above features.

Let's have a look on the GMM results: