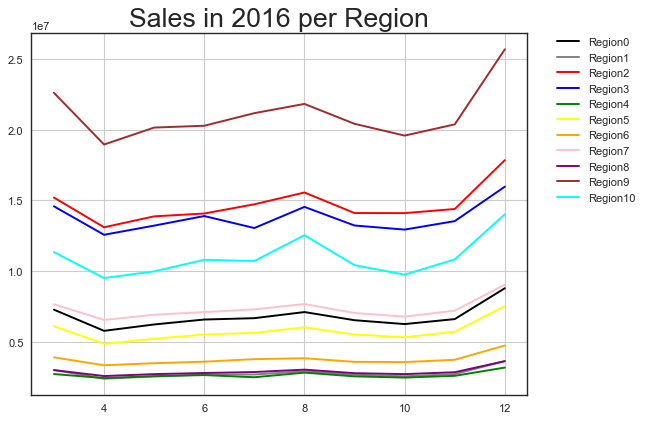
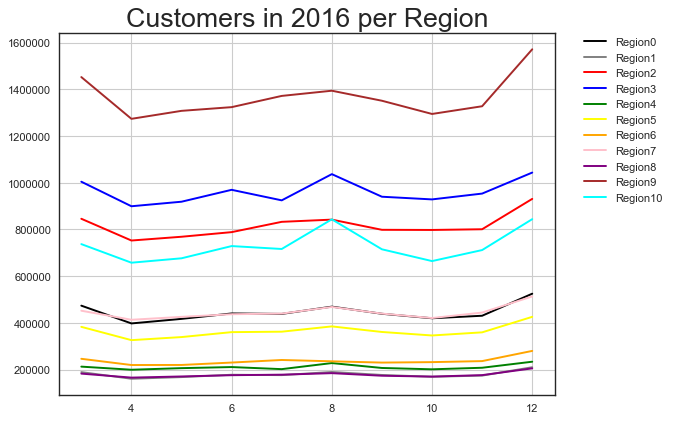
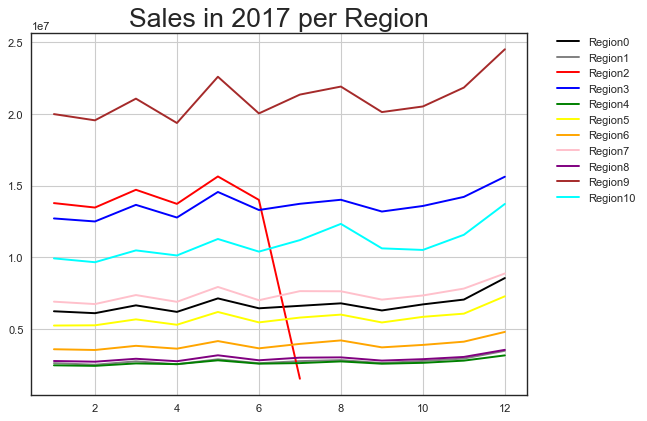
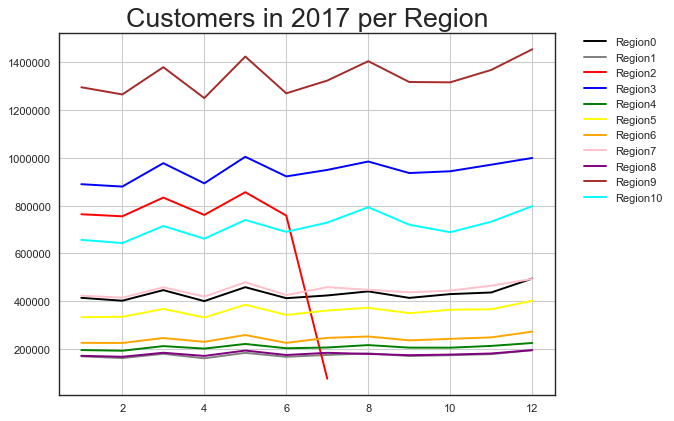
**Data Mining and Text Mining project**

Sales forecast

1. **Data Exploration**

In order to have a clear idea of our dataset we started to explore it, looking for some relationships among different regions. Therefore, we plotted the trend of **total customers** and **total sales** per month for every region in the dataset. Here we attach our results.



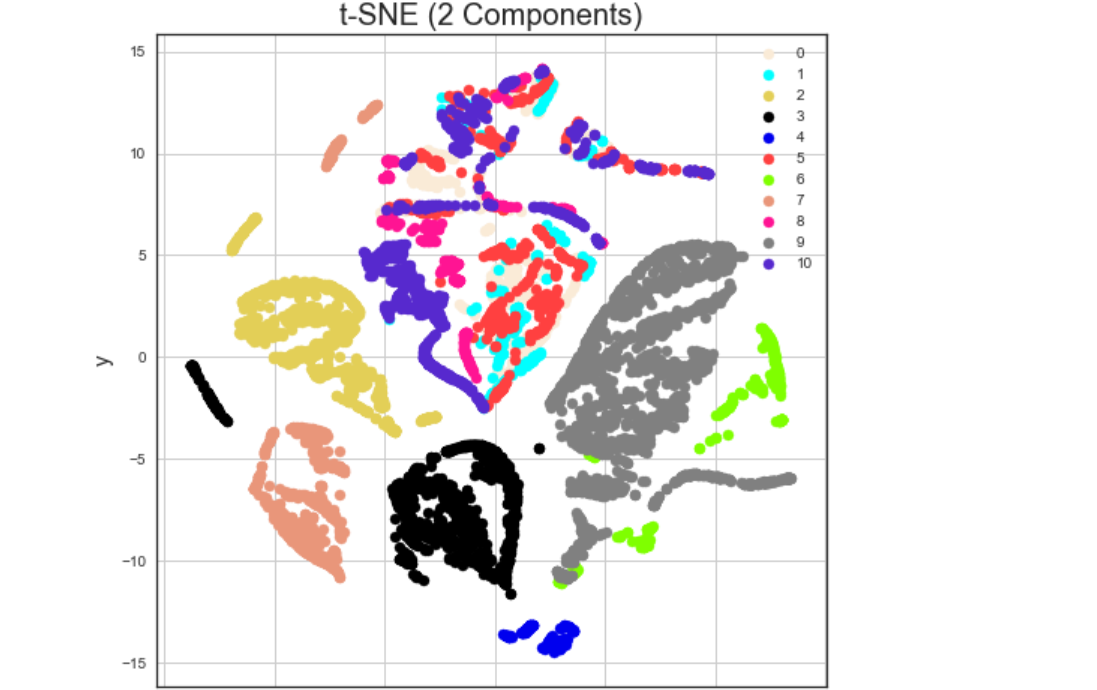


The very first we notice is a huge fall of Region 2 from July 2017: by exploring the dataset we discovered we were missing those tuples from that month up to January 2018.

As the graph shows and as expected, **the number of customers and the number of sales in one month are strictly correlated.**

Furthermore, there is an increase of sales in March, May, August and in December. Region 2 asides, no significant differences of trends is found between 2016 and 2017 (please notice that data from 2016 starts from March).

As a second step, we run a t-SNE plot, pointing out the different regions:

As we notice, there is a “cloud” of regions which looks pretty similar:

**Region 0, Region 1, Region 5, Region 8, Region 10.**

Then, **Region 2 and Region 3** are quite distinct from the others.

Finally **Region 9 and Region 6** look similar.

**Region 7 and Region 4** are distinct as well, but the dataset does not contain many tuples from these regions: we will explain later how we treated them.

Then, we plotted a clustermap to analyze the feature’s correlations.

By analyzing the dendograms, we saw that the following variables are quite incorrelated from the target variable and very correlated to others. That’s why we dropped them: *Max\_Dew\_PointC, Min\_Dew\_PointC, Max\_Sea\_Level\_PressurehPa, Mean\_Sea\_Level\_PressurehPa, Max\_Gust\_SpeedKm\_h.*

1. **Preprocessing**

**Missing Values**

Listing all the attributes with missing values, this was the result:

CloudCover 41181

Events 124098

Max\_Gust\_SpeedKm\_h 409947 //already dropped

Max\_VisibilityKm 11338

Mean\_VisibilityKm 11338

Min\_VisibilitykM 11338

By asking to the domain expert, we were told that when “Events” is missing, that means no event occurred, so we imputed those values as “Not Specified”.

Then, we imputed "CloudCover" with the mean, making a distinction when it misses along with Events and when it misses on its own.

Finally, we noticed that the number of missing values of Max/Mean/Min\_VisibilityKm is the same and when one of them misses, also the others do: we imputed them with the respective mean.

**Dealing with outliers**

By plotting the trends of the numerical attributes, we applied **winsorization** in order to eliminate outliers.

The following is the function we used, where:

* col: the attribute to apply winsorization on;
* quant: the quantile to consider for modify the outlier;
* lambda function, to distinguish between case ‘>’ and case ‘<’

def q(col, quant, f):

t = sales[col].quantile(quant)

print(f'col {col} at {quant}-th quantile => {t}')

sales.loc[f(sales[col], t), col] = t

**Normalization**

We applied a np.log1p to all the numerical variables with a skewness greater than 0.75 (target variable and NumberOfCustomers excluded).

**IsOpen drop**

By exploring the rows with the attribute “IsOpen” set to 0, it turned out that the sales were 0 also.

Therefore we dropped all the rows with the store closed, since they only caused noise in the model. We will simply set NumberOfSales to 0 in the submission dataset.

1. **Model building**

**Feature manipulation**

In order to deal with the **date**, we splitted the feature we were given in “Day\_Of\_Week” and “Month”, and considered them as categorical features. We decided not to considered “Year” as an attribute since the sales trend is the same between 2016 and 2017.

Then, in order to better fit the general trend of the data, we added the following features to the dataset:

* AvgSalesForMonth: the mean of the total sales for the StoreID of the considered tuple in that month.
* VarSalesForMonth: the variance of the total customers for the StoreID of the considered tuple in that month.
* AvgCustomersForMonth: the mean of the total customers for the StoreID of the considered tuple in that month.
* VarCustomersForMonth: the variance of the total customers for the StoreID of the considered tuple in that month.

**Feature Selection**

//qui ci va roba di santa

**How do we predict?**

We decided to split the prediction in two steps:

1. Prediction of NumberOfCustomers
2. Prediction of NumberOfSales

*First step*

We decided to split the dataset in different groups, putting together similar regions. The groups we created were the ones computed by t-SNE:

1. Region 0, Region 1, Region 5, Region 8, Region 10 + Region 4 (its sales trend is very similar to the one of Region 8) and Region 7 (its sales trend is very similar to the one of Region 0)
2. Region 2
3. Region 3
4. Region 6, Region 9

Therefore, we created 4 different models, each of them trained on the data coming from these different groups.

In this step, we trained the model using all the features selected by the stepwise selection.

*Second step*

Now we have a prediction of NumberOfCustomers, we attach them to our data and try to predict the NumberOfSales, building a single model trained on the whole dataset, using only the feature which are mostly correlated with NumberOfSales:

NumberOfCustomers, Region\_AreaKM2,HasPromotions,IsHoliday,Region\_GDP, StoreID.