STDA-project-proposal

Introduction:

The dataset that I've chosen is about the presence of foreign smartphone's sims to the OpenWifi of the Municipality of Milan. This data is open and available on the website data.gov.it. The reasons why I would like to go further with this project is that I strongly believe that are present seasonalities that can be interesting to be analysed but also can be more interesting to relate the outliers to some events that happened in the past with a certain mediatic relevance. In practice I would like to both analyse trend and seasonalities to know in which months there are more foreign people and if the trend is increasing in time and both search for outlier peaks to be related to important happenings in the Milan city. Finally I would like to forecast the possible presences in the new year in the city of Milan.

Data:

The dataset comes from the open data provided by all the municipalities of Milan. This repository is available at dati.gov.it. From this repository I selected the data going from January of 2018 to October of the 2019(approximately 2 years split into 2 different datasets).

Characteristics of the DataSet:

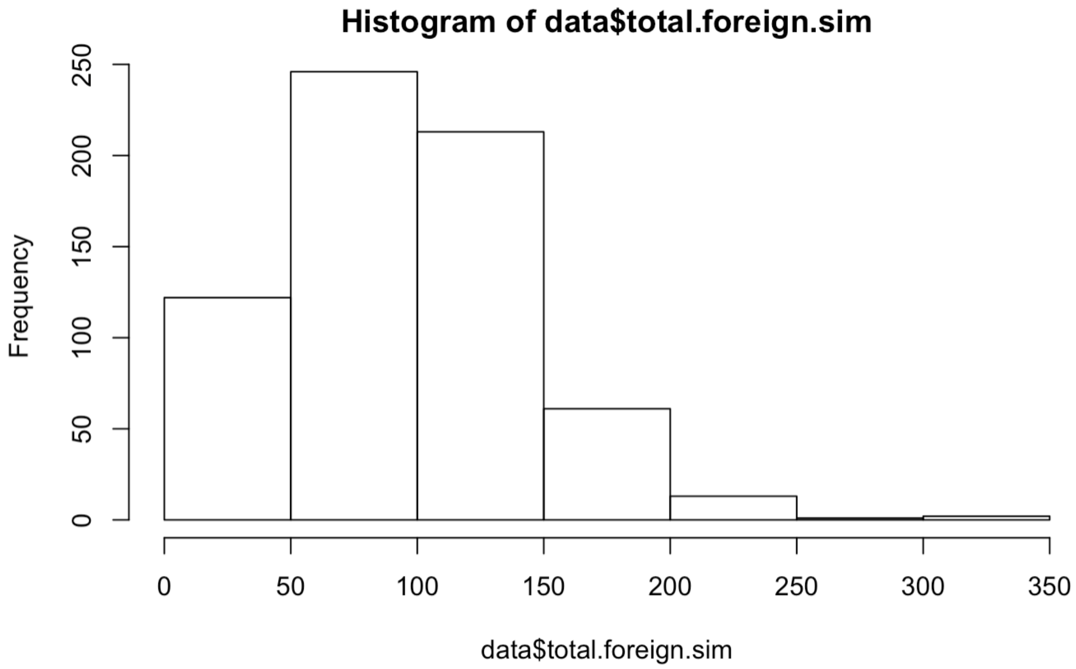
- the dataset contains 3 columns "Date, Number_of_Foreign_Sims"
- has 658 rows
- the year 2018 goes from 01/01/18 to 31/12/18
- the year 2019 goes from 01/01/19 to 30/10/19
- the datasets have no NA
- no lacking days
- the number of sims is a discrete variable about total number of foreign sims in a certain Date

Exploratory Data Analysis:

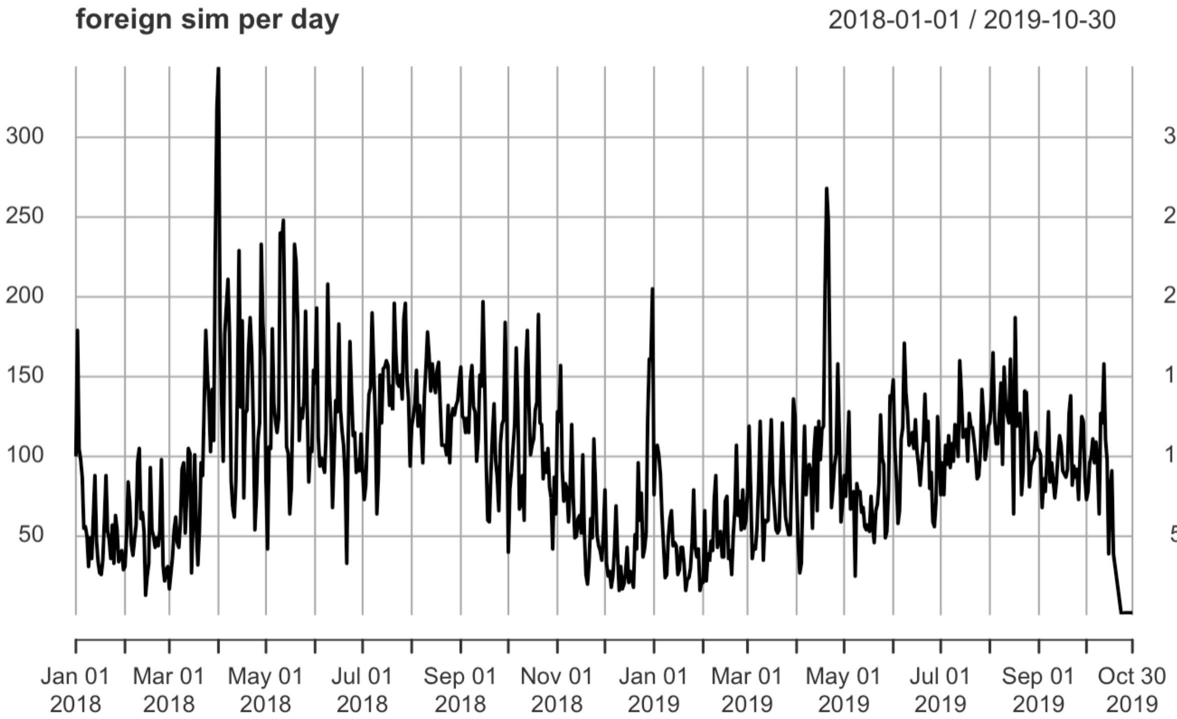
About the Number_of_Foreign_Sims column:

Min	Lower-hinge	Median	Upper-hinge	Max
1	59	95	124	344

Histogram:



The time serie:



Statistical Methods:

I would like to decompose the time serie using:

- low pass filter to find the trend
- high pass filter to detrend the TS and search for the seasonality
- I'm going to find residuals and checking them with ACF to be seasonality free
- Using an arima model to forecast