Capstone Project 1: Data Wrangling

Data Collection and Cleaning

The Telecom Italia dataset chosen for this project consists of 59 TSV files (one file for each data, data was collected for 2 months from Nov to Dec 2013) consisting of spatially aggregated Telecommunication activity (Internet, sms and calls) at an interval of 10 mins, each file is about 300 MB. The total size of the dataset is 18.6 GB, due the huge size of the chosen dataset all of the data cannot be loaded at once into python (RAM constraints of a single system). To overcome this, each file was loaded into a pandas DataFrame in a loop.

Once one TSV is loaded to a temporary dataframe, following steps are performed on this dataframe:

1. Time interval column is converted to CET timezone so that further resampling can be done with ease
2. The NaN values are updated with 0 assuming that there were no Telecommunication activity during that time interval, since all of the values (SMSin, SMSout, callIn, callOut, Internet) are integers.
3. The Internet, SMS and Call activity is aggregated to 10 mins time interval ignoring the GridId and CountryCode. GridId and CountryCode is ignored for this analysis since the size of the file is too big

The processed dataframe is appended to a consolidated dataframe, which will contain aggregated information of all the other dataframe processed previously. A copy of the original dataframe is created (cdf2). Based on this copied data frame a new dataframe is created, this dataframe is created using the resample function on the datetime indexed column of the dataframe (here it is timeInterval). The resampling function should be used along with an aggregation function, here I have used sum() aggregation function to sum up the remaining columns per day (cdfDaily)

Following variables were derived on the daily aggregated dataframe (cdfDaily):

1. SMS: sum of SMSIn and SMSout. This will help in identifying the total SMS per day
2. Call: sum of callIn and callout. This will help in identifying the total calls per day
3. Day: day of the week (0-Monday, 6-Sunday). I have derived this variable to later identify if there is a correlation between the telecommunication activity and the day of the week
4. Weekend: whether a particular was recorded on a weekday or a weekend (derived based on the Day column, above, 5 / 6 belong to weekend)

Identifying Outliers

A box plot was generated for each of the variables in the consolidated dataframe to identify outliers (Fig1 below). It was observed that there were some outliers in the SMSout category. On further analysis it was identified that outliers were only on one day 25-12-2013 between 10:00 AM and 12:20 PM. This peak SMSout is on **Christmas day** which is an expected behaviour during this time and these outliers does not seem like an data collection error and hence no further action is taken on these outliers.

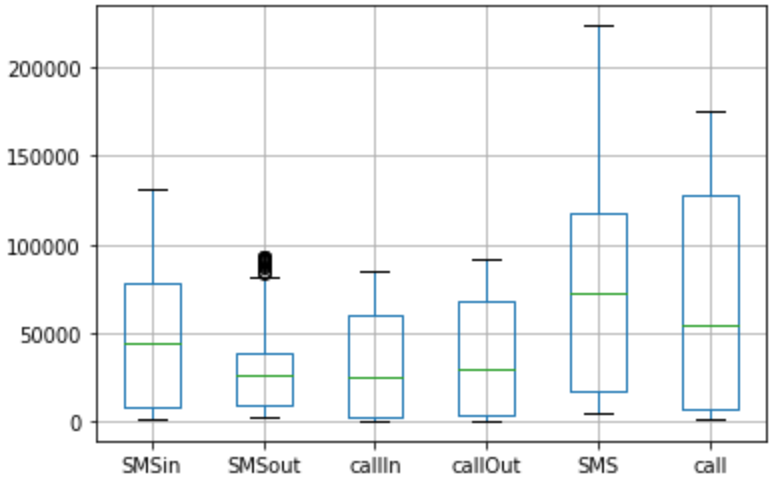

Fig1: Box plot on consolidated dataframe