Data Wrangling
With R

The mobile phone activity dataset is composed of one month of Call Details Records (CDRs) from the city of Milan and the Province of Trentino (Italy), as provided in 1 https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/QLCABU
Only the dataset 1 is considered here.

Every time a user engages in a telecommunication interaction, a Radio Base Station (RBS) is assigned by the operator and delivers the communication through the network. Then, a new CDR is created recording the time of the interaction and the RBS which handled it. The following activities are present in the dataset:

- received SMS
- sent SMS
- incoming calls
- outgoing calls
- Internet activity

In particular, Internet activity is generated each time a user starts an Internet connection or ends an Internet connection. Moreover, during the same connection a CDR is generated if the connection lasts for more than 15 min or the user transferred more than 5 MB. The data provides CellID, CountryCode and all the aforementioned telecommunication activities aggregated every 60 minutes.

Part A

1. Load the datafile ('sms-call-internet-mi-2013-11-01.csv') into a workspace and write the R code to answer the following questions.
2. Write R code to return the first 10 rows in the dataset.
3. Write R code to determine how many unique country codes are contained in the data.
4. Plot the distribution of country codes in the data. The distribution plot is the number of rows with a given country code (y-axis) vs country code (x-axis). Note, a country code of 0 implies there is no country code either due to being unknown or privacy restrictions.
5. Add a new column to your data called "totalsms" that contains the sum of smsin and smsout. Similarly, add a new column to your data called "totalcalls" that contains the sum of callin and callout.
6. Plot the overall total of the "totalsms" column over the country code.

Part B

1. Make a heatmap. The heat map should visualize the mean of the smsin, smsout, callin, callout, and internet data columns computed over the hour of the day.

Part C

1. Run some analysis on the missing data.