UNIVERSITÉ DE NEUCHÂTEL

12.11.2019

Data Management

Exercise 7

Group 11:

Alexandre Ducommun
Rabeb Ben Ramdhane
Abdallah Shukor

Professeur : CIORASCU Iulian

Question 1

Description of the source data:

It is required to download all csv files for Switzerland (2 regions: VAUD and Geneva) and France (3 regions: Paris, Bordeaux, Lyon. For the following samples, we have 5 data sources. Each data sources involves a different region.
The objective behind this exercise is to be able to join multiple regions/sources in a single database. We noticed throughout the website that are several csv file for every region. So, we took the most important csv file which is "Listings". This csv file mainly contained all necessary information for every region/city. However, since each of Switzerland and France contain several regions we had to add a new column called "Region" to each of the 5 csv files in order to identify the region concerned. In addition, we realized that each country has a different currency (CHF, EURO). So, what we did is that we added for each of the 5 data sources (5 csv files: Listings.csv) a new column called Currency based on the corresponding currency of the country. Since we are also able to pay in Euro in both Countries (Switzerland and France) we could've only used one common currency which is Euro, however for this case we should create a function that converts CHF to Euro based on a specific exchange rate. Finally, we concatenated all the information in the 5 csv file in 1 main file called "combined.csv". We also created a new csv file where we stored the metadata from the downloaded csv files. This file contains the file name of the source file, the area, the city, the date of download and the url of the source file.

Questions 2

```
76  # Creating SQlite database for the following dataset
77  conn = sqlite3.connect('ex7.db')
78  c = conn.cursor()
79
80  # Create table - COMBINE
81
82  c.execute('''CREATE TABLE COMBINE
83  ([id] INTEGER PRIMARY KEY,[name] text, [host_id] INTEGER, [host_name] TEXT, [neighbourhood_group] text, [neighbourhood] text, [latitude] float,
84  [longitude] float, [room_type] TEXT, [price] INTEGER, [minimum_nights] INTEGER, [number_of_reviews] INTEGER, [last_review] DATE,
85  [reviews_per_month] float, [calculated_host_listings_count] INTEGER, [availability_365] INTEGER, [Region] text, [Currency] text)''')
86
87  # Create table - METADATA
88
89  c.execute(
90      ''' CREATE TABLE METADATA ([File_name] text Primary Key, [Date_of_download] date, [Region] text, [Country] text, [URL] text )''')
91
92  conn.commit()
93
94  read_combined = pd.read_csv(
95      r'C:\Users\USER\Desktop\ex7\combined.csv', low_memory=False)
96  read_combined.to_sql('COMBINE', conn, if_exists='append',
97                       index=False)  # Importing data to sqlite
98
99  read_metadata = pd.read_csv(
100     r'C:\Users\USER\Desktop\ex7\metadata\metadata.csv', low_memory=False)
101 read_metadata.to_sql('METADATA', conn, if_exists='append', index=False)
```

In order to create the SQlite database, we need to specify the correct field names based on what we have in the downloaded csv files in addition to any new column added to these csv files. So, after we created "combined.csv" we executed the SQlite database which contained the table

"COMBINED" having all the fields found in the "combined.csv" and then defining the data types of each field based on the corresponding csv file. We also created a table for "METADATA" which included all the information related to the downloaded files having as fields: file names, region, country, date of download and the url of the source file.

Question 3

Steps for an automated system

Step 1 – Data Extraction

The system should parse the webpage to find the sources links for download (csv files). But there are some constraints because it must only download files from the area of France and Switzerland. Moreover, each area has many sources files with different data, so the system should choose the right file (listings.csv). When the system downloads the right file, it has to rename it with the pattern "Region_listings_date" in a specified place. We also create a new csv file where we store metadata from the downloaded csv files. This file contains the file name of the source file, the area, the city, the date of download and the url of the source file.

Step 2 – Data Transformation

The next step for the system is to put the data from different sources files in a database in one or more table if needed. First the csv files are merged into one file "combined.csv" with in addition new columns that may be required like "currency" (currency differs by country) and the "region" that belongs to the csv metadata file. The problem that may be encounter is about the currency. At this point we don't know if the currency from each file is for the region or in a global currency like dollars for each region.

Step 3 – Loading data in the database

The "listings.csv" files downloaded are saved in a folder called "regions_listings" in order to combine all these csv files in one large csv file. We created another folder called "metadata"which contained the new csv file "metadata.csv" since we don't want to concatenate with the other csv files found in the "regions_listings". Once the combined csv file is created, the system will put the data in a database (Sqlite3). The system will create a database in a specified location and add a table with all the attribute that will match with the data type from the source. Then the system inserts each row from the combined csv file into the table. The data from France and Switzerland is now available in a database and ready for analysis.