Lab 1: Data Import, Wrangling and Exploration.

During this lab you will gain practical knowledge of the data import, wrangling and exploration techniques that have been presented during Lectures 1-4. To refresh the material, please consult the notes available for the module on Moodle.

You are provided with two supplementary files: "BSCY4.csv" and "BSCY4.sql". Each of the files contains a partition from the "Avocado Prices" dataset available on Kaggle[1]. You can get the detailed description of the fields of the two partitions following the link provided as footnotes.

In this lab, you will need to use **pandas** module of Python to import contents of both of the files, cleanse the data and merge it into a single data frame. Note that some of the formatting and otherwise errors have been introduced into the data. Use the techniques presented to you in class to correct those errors. Upon completion of this lab you will need to submit python code that help identify problems in data from both files as well as the code that corrects errors.

Please, ***comment your code sufficiently*** to avoid possible misunderstandings during the marking process. Additionally in your comments ***you will need to answer a number of questions*** regarding to the errors you find. The questions, description of the task can be found together with the marking scheme below.

Import and wrangling "BSCY4.csv" (4 points)

Step 1. Import data from "BSCY4.csv". **(1 point)**

Step 2. Cleanse information in the "date". Do all rows follow the same format when it comes to "date"? What formats are there and how many entries per each format? **(1 point)**

Step 3. Cleanse the data in the field "type". How many genuine categories are present? Do you see problems with how the categories represented? How many entries have errors? **(1 point)**

[1] https://www.kaggle.com/neuromusic/avocado-prices

Step 4. Cleanse the content of the field "average price". How many genuine missing values are there? How many entries have erroneous string-based representation. **(1 point)**

Import and wrangling "BSCY4.sql" (4 points)

Step 5. "BSCY4.sql" contains a dump from a MySQL database. Install MySQL on your computer and import the dump file.

Use Command: mysql -u <user> -p < BSCY4.sql

Step 6. Now your mysql installation contains BSCY4 database that contains 1 table, AVOCADO. Use **pymysql** module to import contents of the table via **pandas. (1 point)**

Step 7. Cleanse the content of the field "region". What can you say about the regions represented? How many different regions there are? Are there problems with this variable, if yes, what are the problems and how many? **(1 point)**

Step 8. Cleanse the content of the field "year". What years are represented? Describe any errors that you see in data. How many rows are affected? **(1 point)**

Step 9. Cleanse the content of the field "type". What avocado type are represented? Describe any errors that you see. How many rows are affected? **(1 point)**

Data Consolidation (2 points)

Step 10. Perform Visual Inspection of the results of the two previous imports. Are the two data frames suitable for consolidation? What problems do you see? Correct the problems. **(1 point)**

Step 11. What method should you use to consolidate the two frames correctly? Perform the consolidation. **(1 point)**