Student Name: METU Dept. Of Business Administration
Student Number: BA4318 FAll 2019

TEST 01 / 6.11.2019

Part 2 (70 Points, Take-Home, Due 13.11.2019 08:30, Submission through ODTUCLASS)

You must take Part 1 of the exam in order to submit this part.

Question 3 (Extracting the dataset, 20 points)

You are to acquire two weather related datasets from Kaggle.

- Hourly Weather Surface - Brazil (Southeast region), from -- https://bit.ly/2JS727D -- note that this is a very large file (2 GB).
    - We are interested in the "Instant Air Temperature (celsius degrees)" column
- Weather Madrid 1997 – 2015, from -- https://bit.ly/36yHJ4m – this is a smaller file (571 KB).
    - We are interested in the "Mean TemperatureC" column.

You shall open the files for these datasets and transform them into two separate dataframe objects, i.e. df-brazil and df-madrid.

Hint: The Brazil dataset is very large. Maybe you should **modify how you load the datasets, so that you load only the columns which you would need**.

Question 4 (Transforming the data set, 40 points)

Here are some problems with the data sets we have.

- Date ranges for these datasets are not exactly the same. You should drop dates where we do not have data for both locations.
- The Brazil dataset has multiple measurements in the same date because the dataset originates from multiple sources and has hourly measurements instead of daily measurements. We need daily averages for both.

What we need at the end is a dataframe, i.e. df-final with the following columns

- A "date" column, which is filled with dates on which both datasets had measurements
- A "temp-brazil" column, which is **the average of all measurements** on that particular date.
- A "temp-madrid" column, which is **the single measurement** for Madrid.

You must **define at least one function** to manage the complexity of this task. Most probably you will define two functions.

Hint: You should **first process the Brazil dataset, grouping measurements for each day, and getting their mean**. The result could be assigned into a new dataframe.

Question 5 (Simple Query, 10 points)

We'd like to see if there is a statistical correlation between daily average temperatures of Brazil and daily average temperatures of Madrid.

So, calculate the correlation between these columns.

Hint: Dataframe objects can do simple statistical calculations.

Bonus (Interpretation of Results, 10 points)

Look at the result for correlation, and explain it. Your explanation should be at the end of your program as comment lines.

General Notes:

- Your source file (i.e. .py file) should include your student number in the file name such as ba4318-test1-part2-123456.py.
- Add your student name and number as comments in the source file as well.
- You should keep your solution a secret until the end of the submission. So **wait and do not commit on your Github** until then.