# Advances in Data Science Midterm Exam Assignment

**Remarks:**

Write the code yourself. ***Cheating is strictly forbidden.***

For each problem write your code in the function format and give the names of the functions as problem numbers, for example for the solution of problem1:

def problem1(input): return something

Put the codes for all problems into one python file and name that file in the following format: NameSurname\_AdvDataScience\_midterm.py

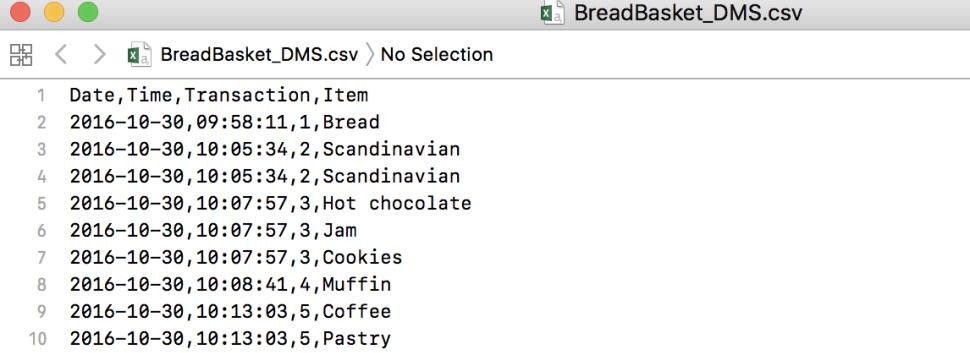
Give as much as documentation for your script using comments.

Please upload your solutions to ninova system. If you are having problems with the ninova system you can send your solutions to [badays@itu.edu.tr](mailto:badays@itu.edu.tr) e-mail address.

Note: You can use PANDAS or NumPy libraries for the following problems.

# Problem 1 (30 Points).

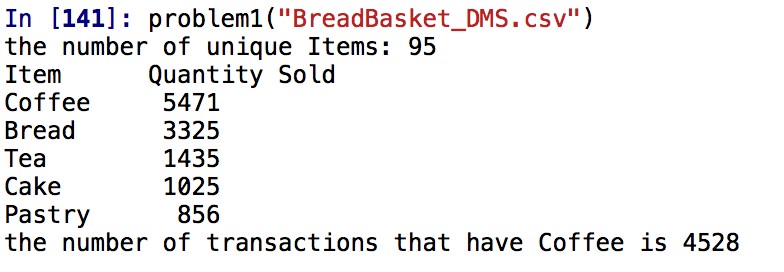
You are given “BreadBasket\_DMS.csv” file that contains sales informations of a bakery. The colums are: Date,Time,Transaction and Item. Transaction can be thought as the order id.



Write a function that takes this file as an input and to the following:

1. Print the number of unique items (10Points)
2. Print how many times each item was sold. (10Points)
3. How many transactions contain Coffee. Note that you may have more than one coffee in one transaction. (10Points)

Your output should be like the following:



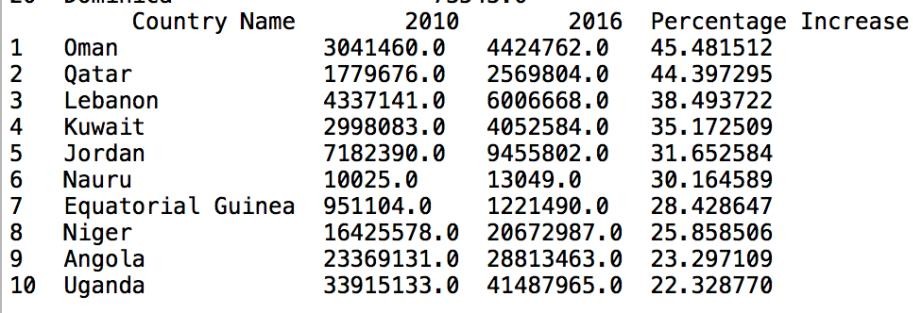
# Problem 2 (40 Points).

You are given “country\_population.csv” file that contains population information of countries between the years 1960 and 2016. Write a function that takes this file as an input and do the following analysis:

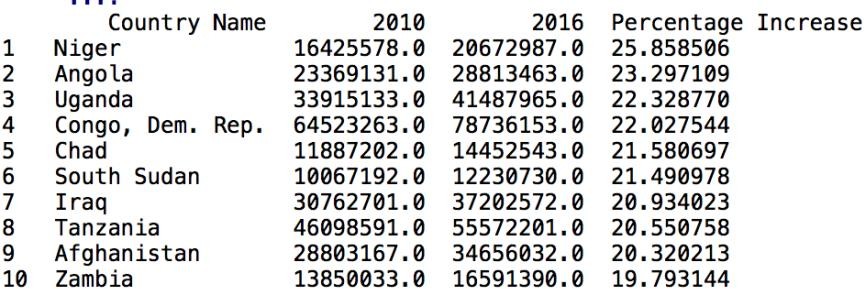
1. Print the bottom-10 countries having the least number of population in 2016. The output should be like the following figure: (10Points)



1. Calculate the percentage increase of the 2016 populations compared to 2010 populations. Print the top-ten countries having highest increase. (10Points)



1. Do the same task asked in problem2.b but this time only for the countries that have more than 10 million population. (10Points)

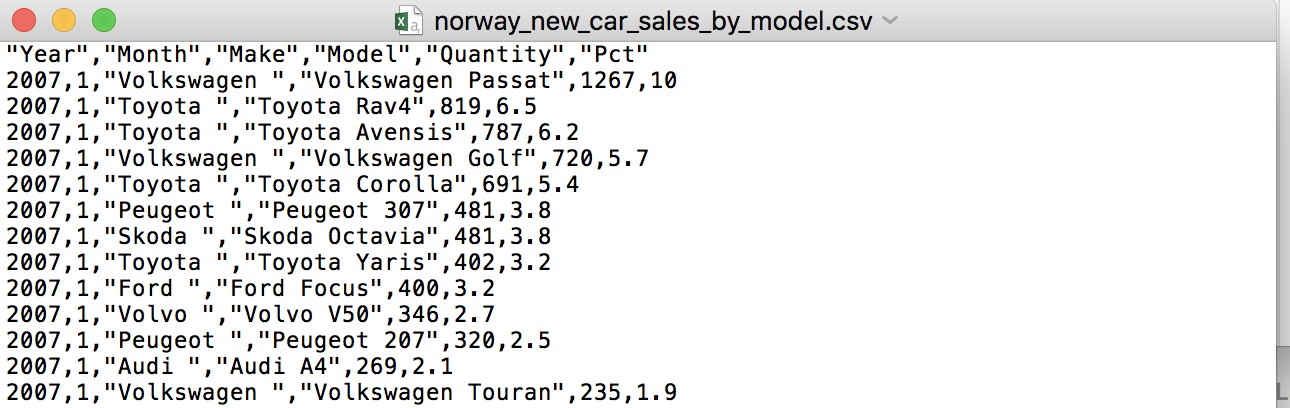


1. Find the year that world population has the highest increase compared to previous year. Note that there is a line just for the whole world which has a country name “World”. (10Points)



# Problem 3 (30 Points).

Norway\_new\_car\_sales\_by\_model.csv file contains information of the new car sales in Norway between the years 2007-2017. The dataset was obtained from [www.kaggle.com](http://www.kaggle.com/) web site. The dataset comprises of monthly car sale quantity for various manufacturers and models. Make columns shows the manufacturer and Pct column shows the percent share in monlty total sales.



Before doing analysis replace “\xa0Mercedes-Benz” with “Mercedes-Benz” in the make column. Also change all make and model column to lowercase.

For each manufacturer find the model that has highest sales fluctuation over the years. You can calculate fluctuation as standart deviation of the yearly total sales for each model. Your output should look like the following.

