HOMEWORK ASSIGNMENT I

- Work individually.
- Submit a code and a short explanation along the code. The names of the files should be YOURNAME.py (or YOURNAME.ipynb) and YOURNAME.docx (e.g. OrsanOzener.py, OrsanOzener.docx)
- Any type of plagiarism will not be tolerated and will lead to disciplinary actions.
- **Due Date: TBD. Please submit your report through LMS. Only ONE submission per person.**

1 Introduction

In this assignment, you are expected to use data from the Health Care Survey assembled by WHO as part of the Year 2000 World Health Report (WHOData.csv).

The variables in the file are:

COMP = composite measure of health care attainment; LCOMP = logCOMP

DALE = Disability adjusted life expectancy (other measure); LDALE = logDALE

YEAR = 1993,...,1997; TIME = 1,2,3,4,5; T93, T94, T95, T96, T97 = year dummy variables

HEXP = per capita health expenditure; LHEXP = logHEXP; LHEXP2 = log-squaredHEXP

HC3 = educational attainment; LHC = logHC3; LHC2 = log-squaredHC3; LHEXPHC = logHEXP * logHC3

SMALL = indicator for states, provinces, etc. SMALL ¿ 0 implies internal political unit, = 0 implies country observation

COUNTRY = number assigned to country

STRATUM = another country indicator

GROUPTI = number of observations when SMALL = 0. Usually 5, some = 1, one country = 4.

OECD = dummy variable for OECD country (30 countries)

GINI = gini coefficient for income inequality

GEFF = world bank measure of government effectiveness*

VOICE = world bank measure of democratization of the political process*

TROPICS = dummy variable for tropical location

POPDEN = population density*

PUBTHE = proportion of health expenditure paid by bublic authorities

GDPC = normalized per capita GDP; LGDPC = logGDPC; LGDPC2 = log-squaredGDPC

You are expected to predict the target variable "COMP" (so please remove COMP and LCOMP from the predictors) using necessary tools (regression, eliminating correlated features, clustering, so basically whatever you think is relevant). The criterion will the MSE over the test set. Please use the following lines for train and test split.

```
np.random.seed(123)
from sklearn.model_selection import train_test_split
train_x, test_x, train_y, test_y = train_test_split(df, labels, test_size = 0.25, random_state=42)
```

If you have any questions, please send an email to: orsan.ozener@ozyegin.edu.tr