Course Name	**: Introduction to Data Science**
Exam Type	**:** Midterm I
Submission Deadline	**:** 05.04.2016, 13:00
Files to be submitted	**:** nameSurname.doc, nameSurname.ipynb and your data files

Student Name Surname : Harun Urhan
Student No : 1121221009
Score:

DATA MANAGEMENT & EXPLORATORY DATA ANALYSIS MIDTERM I PROJECT

This study acquaints you with the data and encourages you to develop a research question. You are free to choose the data to work with; however you are required to clearly describe your data and the association that you would like the study. You are also required to support your hypothesis by making statistical analysis on the dataset of your choice. Below are the steps to fulfil this:

Step1 Defining the dataset:

a) Choose the dataset that you would like to work with and then give the details about it.

I would like to work student alcohol consumption data which is shared in UCI Machine Learning Repository by students from University Of Camerino. It has different kinds of data of students along with their weekdays and weekend alcohol consumption. It has good that data has both male (453) and female (591) students that takes different types of courses Portuguese (social) and Math (science), good distribution in almost every variable.

b) What is it about, how many variables are there in your dataset, what are the types of each of those variables, are there any missing values, etc.

There are many variables in dataset, but only 8 of 32, [Walc, Dalc, studytime, freetime, G1, G2, G3, sex] are used in analysis. Also [Talc, Galc, GA] are derived from existing data.

- Dalc: alcohol consumption on weekdays { numeric: from 1 (very low) to 5 (very high) }
- Walc: alcohol consumption at weekends { numeric: from 1 (very low) to 5 (very high) }
- studytime: weekly study time { numeric: 1 for <2 hours, 2 for 2 to 5 hours, 3 for 5 to 10 hours, or 4 for >10 hours }
- freetime: free time after school { numeric: from 1 (very low) to 5 (very high) }
- G<n>: grade of <n>th test { numeric: from 0 to 20 }
- Talc: total alcohol consumption (Dalc + Walc) { numeric: from 2 to 10 }
- Galc: alcohol consumption group/level based on Talc { string: low (0, 3], medium (3, 7], high (7, 10] }
- GA; average grade (avg(G1, G1, G3)) { numeric: from 1.33 to 19.33 }

Step2 Managing the data: (include both detailed explanation and your python codes)

a) Does your dataset comes from separate files? If so, how did you merge them?

Yes, it does. There are two files, one for students who take Portuguese course and one for those who take Math course.

They are read from csv files into DataFrames and merged with DataFrame.append functions like df1.append(df2, ignore_index=True)

data files are student-mat.csv and student-por.csv

b) How did you handle missing values?

There wasn't any missing value.

c) Which variables (and which values) were included in your analysis, and how did you choose them?

Since hypothesis is "Students who consume more alcohol, are less successful at school", only alcohol consumption variables [Talc, Galc] and success related columns GT, studytime are included along with some supporting variables, sex and freetime.

d) Did you have to create groups of some quantitative variables?

Even though Talc is not really a quantitative but a group variable, it is grouped into more general groups (Galc) to see how it affects more clearly and make changes more visible.

e) Did you recode some variables?

No.

Step3 Providing basic descriptive statistics: (include both detailed explanation and your python codes)

ANSERWS ON IPYNB: see comments with [3a], [3b], [3c] etc.

a) Frequency distributions of counts and percentages (for numerical / categorical variables)
b) Draw univariate bar graph of each variable
c) Draw a bivariate plot that shows the relationship between two variables (for numerical variables)
d) Draw a bivariate plot that shows the relationship between two variables (for categorical variables)
e) Do the graphical results support your hypothesis?

After analysis, it can be said that the hypothesis is partially true. It is only valid for male students and counter-effect is lower than I expected (only 8%). Although it is not mentioned in the hypothesis but found during analysis, the reason is not only inebriation effect (which is the only expected) of alcohol but also time that is being spend on consuming alcohol.