# Homework 1

Your Name: Srujan Shekar Shetty

Student ID: **A20529733**

**Q1. Answer the following questions by referring to the slide of the 1st class, and/or reading the syllabus [5]**

1). When is the office hours of the professor in this semester?

A-Professor Yong Zheng office hours for DataMining and Machine learning is on Tuesdays from 1PM to 3PM.

2). If you failed your middle term exam, visited the office of the professor, and requested a second chance to take the exams, what will the professor do?

A-It comes under unreasonable requests so the request shall be ignored by the professor.

3). Student A shared answers with student B, and B copies and pasted answers from student A. What are the penalties they will receive?

A-Since both the students violated the plagiarism rules for the first time both will get a zero and warning and shall be reported to the department.  
If it repeats for the second time Academic Honesty Violation report shall be filed and both the students will be expelled from IIT.

4). If you got a final grade of 74.9, which letter grade (A, B, C, D, E, F) you will get?

A-If I get 74.9 it shall be considered as C.

5). TA found that you violated the plagiarism policy for the first time in the class, and TA reported this to the professor. What will the professor do in the next?

A- Since the student violated the plagiarism rules for the first time he/she will get a zero and warning and shall be reported to the department.

**Q2. Perform data processing as requested. [35]**

Consider the data collected by a hypothetical video store for 50 regular customers. This data consists of a table which, for each customer, records the following attributes: Gender, Income, Age, Rentals (total number of video rentals in the past year), Avg. per visit (average number of video rentals per visit during the past year), Incidentals (whether the customer tends to buy incidental items such as refreshments when renting a video), and Genre (the customer's preferred movie genre). This data is available as an Excel spreadsheet. Perform each of the following data preparation tasks (each task applies to the original data):

Open the HW1.xlsx, the sheet “Data” provides you the view of the data set we are going to process; Work on Excels and Put your answers in the corresponding sheets in HW1.xlsx. Note that you can resort the excel sheets, no problems. However, for the final submissions, you should sort it back to make sure CustID is listed from 1, 2, 3, to the largest number.

1. [10 points] Use smoothing by bin means to smooth the values of the Rentals attribute. Use a bin capacity as 4, i.e., each bin has 4 elements.
2. [5 points] Use min-max normalization to transform the values of the Income attribute onto the range [1-5].
3. [5 points] Use z-score normalization to standardize the values of the age attribute.
4. [5 points] Discretize the (original) Age attribute based on the following categories: Young = 1-20; MidAge = 21-40; Old = 41+.
5. [10 points] Convert the original data into the standard spreadsheet format (note that this requires you to convert all the categorical variables in sheet “Data” to numeric variables, you should provide the final version of transformed data matrix in sheet “Answer. e).”).

Submissions: All answers should be given and submitted in the Excel sheets

**Q3. Perform data processing as requested. [60]**

Use your knowledge and Python skills to answer the following questions.

1). Repeat b), c), d), e) in Q2, by using Python on the HW1.csv data set.

2). Find the numerical variables which have strong correlations with “rentals”

3). Find variables which have dependency with “Genre”

You should submit your answers by using Q3.ipynb and Q3.html (converted from Q3.ipynb). Note that the html file is saved from Jupyter notebook, where your codes and outputs should appear in both ipynb and html files. Make sure that you clearly marked 1), 2), 3), in your ipynb and html files

Final note: you need to submit individual files, not ZIP files. Otherwise, you will be deducted for some points.