# Homework 1

Your Name: \_\_\_\_\_\_TsuCheng Lu\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Student ID:\_\_\_\_\_\_\_\_\_\_\_A20525040\_\_\_\_\_\_\_\_\_\_\_\_\_

**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?

Tuesdays 12:40 - 13:40, 15:00 – 15:30

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

You will not have a second chance

3). Recently, you submitted a late assignment, and want to waive the late submission penalty due to that you have a medical issue, what should you do next?

Send an email to the professor and talk about your situation and provide your medical document. If not, you might get a 15% penalty on your assignment.

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

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?

student will be graded as a zero score.

**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

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 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”

In conclusion, age and income have a strong correlation with rentals.

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

If p-value is less than 0.05, then we can conclude that they have a dependency with "Genre". In conclusion, Income, Age, Rental, and Incidentals have a 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