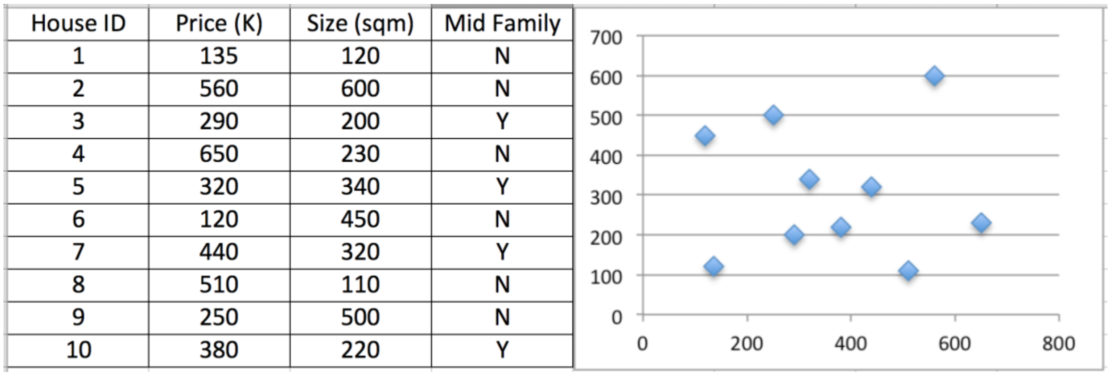
INF 550: Data Science at Scale

Homework 2, Fall 2019

Due: Sep. 23, 2019. 3:30 PM

This assignment must be done independently.

1. You have the following training dataset about houses. We want to classify if a house is a mid-class family home or not based on the price and size of a house.



House ID	Price (K)	Size (sqm)	Mid Family
1	135	120	N
2	560	600	N
3	290	200	Y
4	650	230	N
5	320	340	Y
6	120	450	N
7	440	320	Y
8	510	110	N
9	250	500	N
10	380	220	Y

   a. Define the most specific hypothesis S. Describe it as a pseudo code.
   b. Define the most general hypothesis G. Describe it as a pseudo code.
   c. Assuming the most general hypothesis, give an example of new house data for the case of
      a. A mid-class family home
      b. A false positive
      c. A false negative
   d. Now one more house is added to the training set: < 11, 450, 510, Y>. Then, will you change your hypothesis? If yes, how? If not, why not?

2. In the dataset “Grade.xlsx”, students had pass/fail grades based on their HW, midterm, and final score. Using this as a training dataset, you are supposed to make a binary classification model to decide a student with a certain score(s) would pass or fail. Especially, consider only two features in modeling. What would be your model and justify why your model is the best? What is the accuracy of your model with the training dataset? This question is not asking a specific classification algorithm but requiring a conceptual discussion to understand classification. So plain explanation with supporting numbers will be fine as the answer.

3. You have the following five transactions from a supermarket.

TID	List of Items
1	A, B, C
2	A, B, C, D, E
3	A, C, D
4	A, C, D, E
5	A, B, C, D

Suppose that the minimum support requirement is 40%. Using the Apriori algorithm, find all frequent item sets. Show your work step by step.