We chose project 3, and its main requirements are:

3. **STUBHUB**

   - Dataset: Andrew Sweeting's Stubhub database available at: https://www.journals.uchicago.edu/doi/suppl/10.1086/669254
   - **Do not contact Andrew Sweeting**
   - Predict game attendance and ticket prices based on characteristics
     - Do the prices vary by team (home, visiting) record, time of year, day of week, time of day, stadium, section, etc.?
   - For your favorite team, provide a recommendation on seller strategy on Stubhub, including which tickets are best to resell and when you should post your tickets.

For the dataset, we have following question:
a. there are 3 datasets in the paper:
2 Stubhub datasets, which consist of daily listing (not transaction) information on the "buy" page for each game from January 6th, 2007 to September 30th, 2007, collected using an automated script. One is cross-sectional data, another is panel data. Do we need to merge these 2 datasets?

Another dataset is an attendance dataset, which is used by researchers to predict attendance. But this dataset is not matched with these 2 Stubhub datasets. Can we ignore this dataset because it is not collected from Stubhub and it cannot be merged?

b. There is no stadium data in the dataset. Can we create it by ourselves?

We can get the home field stadium for each team by ourselves.

c. The datasets have no time of day for a game variable required by professors.

The numerical date in the datasets, such as 17267, can be converted into date "2017-04-11". No time of day information.

d. In order to analyse the impact of characteristics such as section and stadium (string variables), we need to create many dummy variables. Is too many dummy variables an issue? Will it cause the code computational intensive?

e. The question “Do the prices vary by team (home, visiting) record, time of year, day of week, time of day, stadium, section, etc.? “

Is this a classification problem?