MAT 345 - PROJECT #1
due Wednesday, October 3, 2018 at 10:00PM.

OBJECTIVE: In this project you will work with an existing data set: load and clean data; conduct exploratory data analysis to find missing values, outliers etc.; conduct exploratory data analysis to visualize the data; draw meaningful observations and note patterns; communicate your results.

GRADING: The assignment is worth 5% of your course grade.

INSTRUCTIONS: Students will work individually on this project, but they may ask questions and clarification from classmates and the instructor. Students must submit their projects on Moodle.

SUBMIT THE FOLLOWING: A copy of your code and a presentation/report of your findings. Make sure your name is on all files submitted.

PROJECT: You will analyze housing data from `http://realdirect.com`. Download the file housing-data.zip from Moodle, which contains one year worth of housing data for various New York boroughs. In this project, you will analyze housing data for **sales in Manhattan.** Your project must contain, but is not limited to:

1. Load in and clean up the data:
   (a) make sure your data is formatted correctly
   (b) check for wrong or missing data (check square footage, price, date of sale etc.)
   (c) consider only actual sales
   (d) remove outliers that do not look like actual sales
2. Perform some simple exploratory data analysis to visualize data:
   (a) across neighborhoods,
   (b) across time.
3. Estimate parameters for sale price, overall and per neighborhood:
   (a) range
   (b) median
   (c) average
   (d) standard deviation
4. Summarize your findings in a report, including the plots you produced for part (2).

Note: This exercise is based on a case study presented in the book *Doing data science,* by R. Schutt, C. O'Neil. Sample R code analyzing the Brooklyn housing data is available on Moodle.