Assignment #3: Machine Learning: supervised and unsupervised learning
Spring 2016
CS3943/9223
Prof. Rumi Chunara
Total: 30 points

All questions must be completed in R. Implement and comment your code so that anyone reading the file can reproduce the code easily (e.g. set the file path once at the beginning of the script where it can be easily changed). Save the code as an R markdown file, and upload it to NYU classes.

In all three parts of this question we will use property sales data from New York City. You can read about and download the data here: http://www1.nyc.gov/site/finance/taxes/property-rolling-sales-data.page. Let's use the Manhattan data for analyses.

1. Data Exploration (10 points total)
a) Load in and clean up the data. The data will be messy. Once you get it loaded in, conduct exploratory data analysis in order to find out where there are outliers or missing values, decide how you will treat them, make sure the dates are formatted correctly, making sure values you think are numerical are being treated as such, etc. (4 points)
b) Once the data is in good shape. Conduct exploratory data analysis to make comparisons (i) across neighborhoods, and (ii) across time. You can use descriptive statistics and/or visualizations (6 points)

2. Supervised Learning (10 points total)
a) Analyze the sales data using regression with predictors you feel are relevant. Justify why regression was appropriate to use and which predictors you include in the model. (2 points)
b) Visualize the coefficients and fitted regression model. (2 points)
c) Predict the neighborhood using a k-nearest neighbors classifier. Use a 10-fold cross validation, and report the error. (3 points)
d) Report and visualize your prediction findings. (2 points)
e) Describe any decisions that could be made or actions that could be taken from this analysis. (1 point)

3. Unsupervised Learning (10 points total).
a) Perform a PCA on the data after scaling the variables to have standard deviation equal to one.
b) Plot the first three principal component score vectors in order to visualize the data. One way to do this is to generate one plot with the first two vectors as the x and y axes, and a second plot with the first and the third. How do you choose how to color each point?