**Stat 602 Class Project Weekly Updates**

**Competition: NYC Real Estate Price Competition**

**Team Kaggle Name (if Relevant): k fold k times**

**Team Members:** Reid-Vincent Paris (Vinny),Subrata Pal, Amin Shirazi

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Week 12 (April 12-16) Summary of Activity and Progress:**

**Amin:**

We spent most of our time on understanding the data, cleaning the data, and dealing with NA’s in the data to come up with tidy data. One major problem was the abundance of zeros in the data for ‘residential units’, ‘commercial units’, ‘land square feet’, and ‘gross square feet’. Instead of removing them from the data, I tried some naïve imputation approaches. I tried to replace the zeros with the median of the variables in the same area. i.e., I first filtered on the properties in the neighborhood in which there are non-zero values for the four variables I mentioned, and then tried to take their medians to come up with some values for the zeros. I am not sure if that is the best approach, but this might be helpful. I also worked on the NA’s for factors in the data such as tax class and building class. We ran a multiple linear regression on the raw data first (where we put aside all rows with NA) and then on the data after replacing the zeros and NA’s.

**Vinny:**

**Subrata:**

**Week 13 (April 19-April 23) Summary of Activity and Progress**

**Week 14 (April 26-30) Summary of Activity and Progress**

**Week 15 (May 3-May 7) (Finals Week) Summary of Activity and Progress**