**Manhattan Sales Observations**

The focus of this study was to examine any predictable attributes that could help determine the sale price of a family home in Manhattan, New York. Variables in the dataset describe the neighborhood, burrough, square footage, year built, sale date and more for real estate sold for over the past year.

Data cleaning steps include removing leading digits for sale.price, gross square feet and land square feet. The analysis was conducted on the log of the sale.price variable as well as the log of gross.square.feet.

The manhattan\_sales data suggests that as the square footage increases, so too does the price. While the study focuses on 1, 2, and 3 family homes, plots for the overall dataset suggest the same pattern exists for all real estate types. A box plot on the log sale price of family homes shows that the min is around 5.6, the max is around 7.6, the first quartile is around 6.4, the 3rd quartile is around 6.9 with the median around 6.7.

Analysis was conducted to detect whether there is a time effect on the data. The results show that the sale price is evenly spread across the different months for the given time period and therefore no significance was attributed to time.

Further analysis is needed on the rest of the categorical variables to see what role they have in predicting sale price. A box plot on neighborhood suggests that different neighborhoods will have a significant effect on sale price. Grouping addresses together per unit such as block might provide a more finite analysis result for predicting sale price over small geographic distributions. The results from the preliminary analysis indicate that a regression analysis would be appropriate to run on this data set. The resulting regression equation would be a good tool for predicting the sale price of homes in Manhattan, New York.