EXPLORATORY DATA ANALYSIS

- We began with exploratory analysis of the training data set given to us and noticed that the data had missing values for several features. Feature variables ‘address_line_2’, ‘misc_features’, ‘average_neighborhood_price’, ‘floor_of_unit’ were removed from the data set owing to more than 70% of the data values being missing because imputation based on a marginal percentage of the data values did not make sense.
- Other variables(mentioned below) were imputed based on whether they were structurally missing or not. Those that were structurally missing were imputed by zero, others were imputed using mean of that particular column.
- Damage_code was one hot encoded

DELETED VARIABLES

1.Address_line_2
2.Misc_features
3.Average_neighborhood_price
4.floor_of_unit

IMPUTED VARIABLES	IMPUTED BY
bathrooms	0
bedrooms	0
remodel_date	transformed to categorical variables
crime_score	mean grouped by zip code
culture_score	mean grouped by zip code
floors_in_unit	0
basement	0
floors_in_building	0
schools_in_area	mean grouped by zip code
public_transit_score	mean grouped by zip code
sqft	mean grouped by subtype
overall_inspector_score	mean

MODELING

- Built a classification (logistic regression) model to identify the households that are yielding profits and the households that can be invested in, using the cleaned data set.
- Built a linear regression model to predict the final_price for the validation set
- Invested the investment amount equally amongst all potential buys.

- Calculated profits by using the predicted final price (profit = final price - investment - initial price)
- Took the records with highest profits and arranged them in ascending order of initial price such that the intial buy amount does not exceed 400,000,000.