**Problem statement: To estimate the price of a house based on historical data of houses sold previously.**

**Refined** **Problem statement: To determine the best price of a house to be sold based on data of houses sold previously within the geography. The aim to get the highest possible price that could be deduced from the historical data so that ROI is maximized.**

**Likely output: The price of a house given the various parameters we can supply as inputs to the system being modeled.**

**Observations:**

total\_area = living\_measure  + lot\_measure  - So we could get rid of the columns ‘living\_measure, lot\_measure’ and use just 'total\_area'

dayhours - 20141107T000000 – we can shorten the values by removing the characters from ‘T’ onwards and convert this into integer value, which could give us a continuous attribute

room\_bath - 3.25 - what does the value mean? Need to strategize how to use this one

living\_measure- is this the complete area of the flat, since there are other columns as well which denote area values

lot\_measure - this should also be considered into the house are, but should it be given equal weightage since everyone might consider this to be important

yr\_renovated – Higher values could denote more value for the house

ceil - what does the value mean? Need to strategize how to use this one?

sight: How useful could this be ? Perhaps a candidate to drop off

condition - ? higher means better or worse? Need to clarify from Saandeep perhaps

quality - ? higher means better or worse? and whats the diff between this and ‘condition’

ceil\_measure: square footage of house apart from basement - ? Need to clarify from Saandeep perhaps

basement\_measure: square footage of the basement – Should we be assuming that every flat has a basement? Need to clarify from Saandeep perhaps