Data Science on Real Estate

The first part of this assignment is to perform an analysis of a small data set. We're looking to see how you approach the analysis, the kinds of analytical techniques you use to solve the problems, and your presentation skills in delivering this analysis.

Feel free to use any analytical software you would like to perform this analysis whether it's Python, R, Tableau, SAS, Excel, etc. As this data is not proprietary you can also use cloud-based solutions in performing your analysis if you so desire.

The Data Set

This data contains a sample of real estate listings for the Twin Cities area. The fields in the data file are as follows:

Variable	Info	Description
ID	**Label**	**MLS ID Number**
Address	**Label**	**Street Address**
CITY	**Label**	**Minneapolis, St. Paul, Shoreview, Woodbury, Maplewood, West St. Paul**
STATE	**Label**	**MN (for all)**
ZIP	**Label**	**Zip Code**
ListPrice	**Response (Y)**	**Current List Price ($)**
BEDS		**# of Bedrooms**
BATHS		**# of Bathrooms (can be fractional)**
Location		**Name of neighborhood or region in the Twin Cities metro area.**
SQFT		**Square footage of home (ft.^{2})**
LotSize		**Square footage of lot (ft.^{2}) – missing for several of the homes in these data.**
YearBuilt		**Year the home was built, could be used to create a new variable called Age = 2014 - YearBuilt**
ParkingSpots		**# of Parking Spots (I assume off-street parking)**
HasGarage	**Nominal**	**Garage or No Garage**
DOM		**Days on the market, number of days the home**

		has been listed for sale.
LastSaleDate	**Date**	**MM/DD/YY of most recent previous sale of the home. Do not use!**
SoldPrev	**Nominal**	**Has the home been sold previously (Y or N), this one should be Ok to use!**
Realty		**Realty company the home is listed with**
Latitude		**Latitude (degrees)**
Longitude		**Longitude (degrees)**
ShortSale		**Is more money owed on the home than what the asking price is? (Y or N)**

The data also includes a look up table with zip code information and population estimates from the 2010 U.S. census in the 2nd sheet.

The Analysis

Just like our clients often ask us questions of their data we would like to ask you questions of this data. Please analyze the data, provide solutions to the questions we ask, and put these solutions into a thought out and easy to understand presentation.

The format of this presentation is up to you. Common methods may include a Powerpoint or Keynote presentation, a well organized Word document, a PDF produced from Adobe tools, or even an interactive website if one had time and inclination (though that’s likely a bit overkill).

Question 1:

You are home developer looking to partner with the top real estate companies to acquire and then sell a large volume of properties in the Twin Cities area (the more the better). You do not have the resources to manage too many real estate partners and a strict timeline to negotiate the deals.

What realty companies would you pick as your partners? Why would you make that choice? Demonstrate this through analysis, visual display of your results, and description of your methodology of selection.

Question 2:

All things being equal what would you predict as the listing price for a 2111 square foot house if that was the only information you had on a house in this area? How did you arrive at that estimate? Please explain.

Question 3:

Imagine you are an enterprising real estate agent who has the chance to buy up a bundle of houses for sale – but you can only pick one zip code (either 55104 or 55108).

Also, you can only buy a bundle of properties priced within the middle 50% of the values – you won't be able to buy the most expensive houses or the cheapest houses.

Assume the sample of houses in this dataset is representative for those zip codes. If you had the choice to buy 1000 homes in either 55104 or 55108 which zip code would you invest in and why? Provide your analysis and reasoning.

Question 4:
We're looking to understand what features of the home are most important to potentially predicting the list price of a house. What has the strongest relationship to listing price: square foot, lot size, or number of bedrooms? How do they compare? Please explain.

Question 5:
Short sales are sometimes good opportunities to get value for a house but there is also the risk the property will need a lot of work. How do short sales compare on the average square foot of the house, average price per square foot, and the average lot size?

Now pivot this data based on the location field. Are short sales always a better deal than regular listings? Can we say with certainty this is true for every location? Why or why not?

Question 6:
We'd like to understand how listings compare with the population in their area. Take the zip code data in the 2nd sheet and match against the house listing data. What zip codes have the highest amount of listings per the population size? Show the top 10.

Separately, what zip code has the highest listing price per person? Google that zip code and provide some hypothesis and examples as to why this might be true.

Question 7:
You've just been hired as a data scientist at the premier real estate firm in this area. They want to forecast the actual sales price for any of these listings (and future listings).

What variables from this example data do you think would be the most predictive of the actual sales price? What other kinds of data would you want have to provide the most accurate prediction of the actual sales price?

Assume you can get any data you want. Describe this data clearly and why you think it would help you build an accurate predictive model.