Milestone report for capstone 1

This document is an interim report on my first capstone project. This document describes the:

- Problem
- Data
- Work done so far
- Work to be done

1. Problem

My capstone examines: What drives the price of residential homes?

I want to identify the key factors that drive the prices of residential homes in the United States. I want to use these features to predict the sales prices and conclude if homes are overpriced or underpriced.

My goal is to help consumers: make smarter buying/selling decisions and have better-informed conversations with real estate agents.

2. Data

I am using data from the Kaggle competition, House Prices: Advanced Regression Techniques. This data is called the 'Ames Housing' data set. The Ames Housing data is an updated and expanded version of the 'Boston Housing' data set, which we use in our Springboard exercises.

The Ames Housing data has a training and test set. The training set includes:

- About 1,500 observations of the sales of residential properties in Ames, Iowa
- About 80 features that describe different aspects of each home. For example, the number of rooms, number of bathrooms, year built, the size of the house, etc.
- The sale price of each property and the month and year sold

The test data includes another 1,500 observations, all with similar features, other than the target variable.

3. Work done so far

3-A. Data cleaning and wrangling

As the data comes from a Kaggle competition, it's already in a rather clean state. It's quite tidy for the most part: rows are observations; columns are variables/features. Columns are well-labeled. Each row has a non-null target variable. Plus, there's a well-documented data dictionary.

Still, I had to do the following to do the following cleaning:

- Create an index from existing ID
- Turn date- and time-based features into datetimes
- Clean up nulls in the non-categorical features
    - For these features, null often meant zero
- Convert ordered categorical features into integers
    - There are many categorical features with string-based values that can be represented by integers
    - Often, these features categorize quality, condition, or amount. For example, exterior quality, garage condition, or the amount of slope of a lot.
- Clean up nulls in categorical features
    - Here my initial approach was to leave most categorical nulls as NaN and clean only those where there were two or more categories that conceptually meant null. For example, if there was a column with both 'NaN' and 'none,' I made them all 'none.'
    - But later, I realized that I needed to recode NaN into a explicit category (often, a 'None' or 'No' or something similar, for example) given the specific functionality I was using in the statsmodels package.

3-B. Exploratory analysis and data story

To better understand the data, I started with my target variable--the sale price of residential homes. I looked at the following:

- Typical value: The mean sale price is about $180k. The media is about $160k. The min price is $35k, and the max is about $750k.
- Distribution: The distribution of sale price isn't normal. Rather, it's positively skewed with a longer upper tail. It's a bit bi-modal with two peaks at the human-meaningful levels of $150k and $200k. The distribution of the log of price looks significantly more normal.
- Outliers: There are outliers in sale price. The outliers are very expensive homes. There are 22 outliers three times beyond the standard deviation of sale price. There are eight outliers 1.5 IRQ above the 75th percentile.
- Sales per year: The data set covers five years. From 2006 to 2010. With 2010 as a partial year (i.e., sales for only the first six and a half months of the year). There are about 300 sales a year.
- Sales per month: The typical month has about 25 sales (mean). But there are big variations from

month to month.
- Seasonality: Sales are highly seasonal. The biggest months (i.e., May, June, and July) have 4 to 5 times the volume of the smallest months (i.e., Dec and Jan). Summer has 4 times the volume of sales as winter. Clearly, more people buy and sell homes in summer.
- Sale price by year: For this period, there isn't a clear trend of year over year sale prices increases, which isn't what I expected. Prices might be dropping, and price variability might be increasing, over these particular years. It's worth noting these years coincide with the 'Great Recession' in the United States, which included the end of the housing bubble, bank troubles caused by real estate derivatives, and other macroeconomic woes.

3-C. Statistical analysis

To start my statistical analysis, I wanted to understand which of my 80 features were most correlated with my target. To understand this, I:

- Separated non-categorical and categorical features, and for each type of feature:
- Calculated correlation statistics, f-stats, and p-values for each feature
- Visually inspected the high-correlated, statistically-significant features against sale price
- Visualized the cross-correlations between such features
- Selected top candidates i.e., high correlation with sale price, significant p-values and low correlation with other top candidates)
- Previewed a linear regression combining top candidates

For non-categorical variables, I found:

- 12 statistically-significant features with correlations (i.e., Pearson's r) higher than 50%
- Meaningful scatterplots
- High cross-correlations, unfortunately. The cross-correlation measures were sometimes 75% or more. There's real overlap across certain measures of quality (e.g., kitchen quality and exterior quality), of space (e.g., garage square footage and the number of cars), and of key properties (e.g., the number of rooms and number of bathrooms).
- Five top variables: These included overall quality, living area, garage size in cars, count of full baths, and fireplace quality.
- A disappointing preview of a regression: The adjusted r-squared of my 5-feature equation was less than the r-squared of the top individual feature's regression. While I've not yet analyzed whether the assumptions of linear regression are met, I suspect that multicollinearity is a big problem.
- Also, I found that when I repeated much of the above analysis after dropping 8 key outliers, the conclusions don't change.

For the categorical variables, I found:

- Only four statistically-significant features with 'barely any' correlation to sale price (aka, r-squared from my one-way ANOVA of more than .20%)
- Hard-to-interpret scatter plots and box plots of these categorical variables

- Surprisingly good regression preview: The adjusted r-squared of my statistically-significant model with these four categorical variables was more than 60%.

4. Work to do next

Here's what I'm looking to do next on my capstone:

- Build, score and evaluate a linear regression model to predict sale price, and submit this as a first draft of exercise 10.6.1. Here, I will start with a plain model. And if, after evaluating the plain model, I find overfitting, I'll use Lasso and Ridge to tackle such an issue.
- Build, score and evaluate a to-be-selected non-parametric technique (such as random forrest regression) to predict sale price, and submit this as second draft of exercise 10.6.1.
- Write a first draft of my final report and slides, and submit these as a first draft of exercise 10.6.3.
- Finish my final report and slides, clean up my code as needed, and submit my final drafts as the final draft of exercise 10.6.3.