Capstone Milestone Report - HM

1. **Define the problem**

The problem I want to solve is how to calculate the total assessed value of a house in the Los Angeles region.

1. **Identify the client**

My clients are builders who are looking to build new houses in the area, homeowners of new houses, and real estate agents looking at new houses. All three clients could use this information to determine what the market value of a newly constructed home in a given area could be.

Builders would be able to determine if the home is worth building or not. New homeowners could determine if the price that is being asked for the house is reasonable or not. And real estate agents could see if this is a particular home that would be of interest for them to work with given the expected selling price.

1. **Describe the dataset**

This dataset has 2,985,217 records and 58 features.

As for data wrangling/cleaning, my data was relatively clean as it came from and was used in a Kaggle competition. Nonetheless, I did take a few steps:

* In 3 columns, I replaced True -> 1 and False -> 0
* In the taxvaluedollarcnt column, I removed all NaN values. The reason for this is that this will be my target variable as I get into the machine learning aspects of this project. I cannot have NaN values in my target variable. Additionally, a lot of my preliminary analysis so far has dealt with the taxvaluedollarcnt column.
* In 1 column, I replaced NaN -> 0 because it made sense. I have identified other columns to replace NaN -> 0 but will do that on a need-to-do basis only.
* I have divided my target variable taxvaluedollarcnt mentioned above by 1000. The reason being that this column has records that are in the hundreds of thousands and millions. Dividing it by 1000 makes the data easier to read and easier to use for analysis purposes.

1. **List other datasets to use**

As of now, I do not think I need to use other datasets. There is a column regionidzip. I initially thought I could find a dataset for median income as it relates to zip code or something along those lines and then join it to my dataset using this regionidzip column in my dataset. However, it turns out that the regionidzip column in my dataset is anonymized so that it basically has no real-world meaning and thus that join is not possible.

1. **Explain initial findings**

* The prices of homes with 3 or more garages are generally more expensive than houses with a hot tub/spa.
* Houses with multiple stories and houses with central air conditioning cost similar amounts
* Houses with pools and houses with decks have similar valuations
* Houses with more than the average number of bathrooms have different prices than houses with more than the average number of bedrooms
* As of now, and not mentioned in the statistical analysis report, but nothing has caught me “off-guard” as I ran through preliminary statistical analysis tests on my data. What I mean by that is that all of my basic assumptions about home valuations have been seen as true. For instance, the more bedrooms, the higher the home value; the more bathrooms, the higher the home value; the more garages, the higher the home value, etc.