**DATASET/PLAN FOR DATA (4 points)**

**Data Sources (links, attachments, etc.):**

Our real estate sales data is aggregated data on single-family home sales in the Houston metro area for the past five years. It was downloaded from the agent section of the Houston Association of Realtors’ website (har.com) on February 10, 2024 and spans the period December 08, 2018 to February 09, 2024.

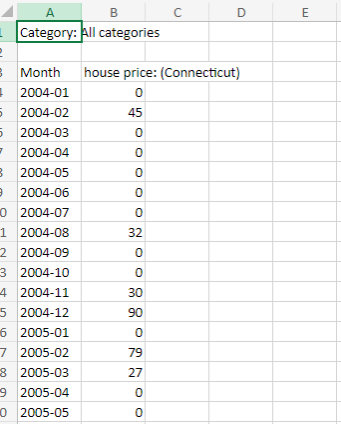
<https://www.har.com/>

Our internet search data is aggregated statistics on real estate-related searches on Google for the same period. We consider a variety of search terms typically used by retail house-buyers, and consider search statistics at the country-wide, Texas, and Houston levels geographically.

<https://trends.google.com/trends/explore?date=all&geo=US-CT&q=housing%20prices&hl=en-US>

**Data Description (describe each of your data sources, include screenshots of a few rows of data):**

The real estate data from har.com has around 363 variables and more than 120,000 rows. The raw data is more than 400 MB in size. This data would need to undergo the preparation steps to bring it to a stage where it can be utilized to generate good models, do research and for further steps in the data modeling process. A snapshot of the raw data given below:

                   **Fig. 1 Snapshot of the housing market data in Houston                           Fig. 2 Google trends data** 

For the internet search data, we will use Google Trends data for a variety of search terms relating to home sales (This may involve additional research and analysis regarding the time frame since the home buying process takes more time than a typical consumer product). The search terms will include topics a typical home buyer may use in home buying from the start of the home buying process to its completion.

**Key Variables: (which ones will be considered independent and dependent? Are you going to create new variables?** **What variables do you hypothesize beforehand to be most important?)**

Dependent variable: “sale amount” in real estate data

Variable selection: The predictors that we think would be important for this data, at this initial stage would be Total sq feet, number of bathrooms, bedrooms, garage, appliances, heating, cooling, neighborhood, school district, swimming pool, and year built. This is our initial hypothesis of the predictors to be important to predict selling prices. We may have to create new variables that we foresee in two distinct kinds of situations. First, there are some predictors, for example, that have comma separated values. We may have to separate these into their own predictors. We will also have to create dummy variables for some predictors which have categorical data.

The real estate data from har.com has around 363 variables and more than 120,000 rows. The raw data is more than 400 MB in size.

First step we did was to determine the type of information the variables have, are there missing values and if so what is their percentage, are there information which we don’t think are pertinent at this initial stage. (We may always add it later after building our base models)

The first step of the cleaning process was when we realized that there were a lot of blank “” values. The way we decided to handle these were to convert these blank values to NA’s. After performing this step, we have the data dimensions as around 128 K rows and 363 variables.

The next step in this data cleaning process was to determine the missing percentage of data in the columns. At this point, we decided that if the threshold was 50% or more data were missing then those columns will be deleted. We ran the code in R, named this data as **data\_cleaned\_stage1**, and checked the dimensions which came to around 128 K rows and 147 variables.

Now, we have 147 variables, we decided to manually examine these variables and select which we deemed, were the best candidates for predictors. (A disclaimer here is that, we my revisit this decision later on, but use this to build our first base models)

We decided to use the following variables for now, that is, Acres, "BathsTotal", "BedsTotal", "SqFtTotal", "CloseDate", "ClosePrice", "Connections", "Siding", "CoolSystem", "CountyOrParish", "Dishwasher", "Disposal", "Energy", "Exterior", "Floors", "ForLease", "GarageDesc", "HeatSystem", "HOAMandatory", "Interior", "GeoMarketArea", "Microwave", "NewConstruction", "OvenType", "PoolPrivate", "PropertyType", "DwellingType", "Restrictions", "Roof", "SchoolDistrict", "RangeType", "WaterSewer", "YearBuilt", "PostalCode" and named this subset as **data\_cleaned\_stage2. T**he dimensions for this came to around 128 K rows and 34 variables.

After this, there are columns in this dataset - data\_cleaned\_stage2 – which has comma separated values. For example, the “Connections” variable has values in it like “"Electric Dryer Connections, Gas Dryer Connections, Washer Connections". Here , we want to separate the unique values into its own columns and then have 1 or 0 based on the presence of the individual values. There are a few variables which we will have to do cleanup like this. We have created functions and then loop it through them to get the desired results. The outcome for this step is added to the dataset called **data\_cleaned\_stage3. T**he dimensions for this came to around 127 K rows and 184 variables.

The intent to doing this is also to allow for our team members to select the data from the cleaning step they would need. This allows for greater flexibility for our team to try out different models with different predictors for research and analysis.

The next step would be to remove some additional columns which had missing percent greater than 15%. This was done because if we remove the NA’s keeping these variables in we get around 52K rows where as if we take them out and remove NA’s we get around 126k rows. The columns removed in this step were "Acres", "Dishwasher", "Disposal", "Interior", "Microwave", and "OvenType". We named this dataset as **data\_cleaned\_stage4** with dimensions of 126k rows and 178 variables.

Now, from **data\_cleaned\_stage4,** we have added a new column called HouseAge which is a calculation between the current year and year the house was built. We have also just taken the rows where ForLease = False since we are not interested in leased home but only home ownership. We have also removed PropertyType (since it has only Single family home values), PostalCode (redundant), and CloseDate. Converted the categorical columns of CloseDate. Converted the categorical columns of Geomarket area, county, schooldistrict into dummy variables. We have now named this dataset as **data\_cleaned\_stage5** and we are ready to run any models on this one as needed as a base step.

This will serve as the base dataset for now. Team members have the option to take any of the datasets in these steps or create their own to start building models.