Problem Identification

Problem Statement:

- Where are the best cities/towns in the United States to invest in real estate based on projected ROI over 5 years?

Context:

- I am analyzing real estate data so investment firms and homebuyers can use a data-driven approach to identify the best cities for them to invest in real estate or buy a home. They will be able to see the expected ROI for each town and determine where the best places will be to realize the most ROI.

Criteria for success:

- This will be successful if we are able to rank the top cities/towns in the US by ROI and show projected ROI over a 5 year period

Scope of Solution:

- The focus of this will be to determine and rank average real estate ROI over the next 5 years for US cities. I will use historical data around home sales, market conditions (buyers/sellers market), sales over/under asking price, and home values to determine ROI.

Constraints:

- Outside factors (economy, regulations, natural disasters/climate Covid) can skew data for certain periods that could affect ROI
- Data is not available about improvements made to houses that lead to greater value like adding rooms and updating features

Stakeholders:

- Real estate investment firms – Using a data-driven approach to target specific locations to invest in real estate
- Personal Homebuyers – being able to compare towns to purchase a home in and see how the value of their investment will increase in the next 5 years

Data sources:

- Zillow Public Housing Data:
  - Home Values
  - Home Value Forecast
  - Sales
  - Market Heat Index

Outline

Data Wrangling:

- First, I will import and clean the data from Zillow. This will include filling in missing values and organizing data in correct formats
- I noticed missing values in average home sales price per month in some towns and plan to fill those values with a mean of the month prior and after.

EDA:

- Next, I plan to use PCA to find relationships between variables that lead to better ROI. I will also use correlation mapping for these relationships
- In handling the data, I will fill in or exclude missing values and determine if there are any outliers that should be excluded as well

Preprocessing & training:

- I plan to use train/test splits to train and test the data with cross validation tests. I plan to use a regression model and random forest model
- I will also measure the performance of each model to decide which to move forward with for the actual modeling of the data

Modeling:

- Finally, after choosing the model I will do 5 fold cross validation to retrain the chosen model and evaluate if we can accept the model as good data.