First Draft

Project Title: California Housing Price Trends and Predictive Modeling
Course: MATH 120
Author: Elliot Wiley

This project investigates housing price trends and predictive modeling using California Zillow Home Value Index data. The refined research questions guiding this analysis are:

1. How have median home values changed over time in major California cities over the last 10–15 years?
2. Which California cities have experienced the fastest housing price growth, and which have grown the slowest?
3. Did the COVID-19 pandemic cause noticeable changes in California housing price trends?
4. Can a simple linear regression model reasonably predict future home values based on historical data?

The dataset used in this project comes from the Zillow Home Value Index (ZHVI), which provides monthly median home values by city and region across the United States. For this project, the dataset is filtered to include only California cities. The data was downloaded from the Zillow Research Data website and uploaded to Google Drive for use in a Google Colab notebook. Key variables in the dataset include:

- Region Name (city or region name)
- State
- County
- Metro area
- Monthly median home values across many years (time-series format)

The dataset is structured in a wide time-series format, where each row represents a city and each column represents a month of housing values. The size of the raw dataset is very large, containing thousands of regions and hundreds of monthly values. Limitations include missing values for some cities, varying lengths of available data across regions, and the large file size, which requires filtering and cleaning before analysis.

To prepare the data for analysis, I will first filter the dataset to include only California cities. Next, I will remove rows with excessive missing values and use forward-filling techniques to handle smaller gaps in the data. The clean dataset will then be saved for reproducibility. For visualization, I plan to create:

- Time-series line plots to show how housing prices change over time for individual cities
- Multi-city comparison plots to compare price trends across major California regions
- Growth rate calculations to quantify long-term price increases

For modeling, I will use linear regression to build a predictive model for future home values based on historical time-series data. Model evaluation will include visual comparisons of predicted vs. actual values and error metrics such as mean absolute error. These methods will help determine how effective the model is at capturing long-term housing trends.

So far, a full GitHub repository has been created with a proper folder structure, a formatted README file, and a working base version og the notebook. The Zillow dataset has been successfully loaded into Google Colab using Google Drive. I have confirmed that the dataset loads correctly and that California filtering works as intended. A time-series plot of housing prices for Los Angeles has already been generated, showing clear long-term growth and I will also add an analysis into the notebook for it. Also, a basic linear regression model has been implemented and tested for one city.

The main challenge so far has been the large size of the Zillow dataset, which can slow processing and visualization. Another challenge has been missing or inconsistent data for some cities, which limits the number of usable regions at least from what I have seen so far. Additionally, selecting appropriate time ranges and cities for fair comparison has required careful filtering. These issues are being addressed by focusing on a smaller group of major California cities, removing rows with excessive missing values, and using simple baseline models before attempting more complex approaches.

Before the final draft is submitted, I will:

- Expand the analysis to include multiple California cities
- Compare regional growth rates
- Refine and test the regression model
  Add additional visualizations
  Finalize documentation and code cleanup in the GitHub repository
- Strengthen the interpretation of the model results