**Introduction:**

For our project 1 analysis, we will be taking a closer look at identifying trends within the United States Real Estate industry utilizing a well-known Real Estate listings database. We chose to segment the data to represent certain criteria including:

1. Three distinct states (Pennsylvania, Wisconsin & Washington) to represent three distinct regions of the United States (East, Midwest & West)
2. 5 bedrooms or less
3. Listings from 2015 onwards (~10 years)

We subsequently performed regression and correlative analysis to determine and visualize a number of relationships:

1. Linear regression to determine any correlation between listing price vs. number of bedrooms, number of bathrooms, land size (Acres) and living space (Sq Ft.)
2. OpenWeather and Geoapify API utilization to retrieve and run associated linear regressions to determine any correlation between listing price vs. city population and city maximum temperature

Further analysis consisted of graphically representing average home price and total number of homes over time.

Files incorporated for submission with this analysis include two executed Jupyter notebooks containing the exploration, cleaning, examination and visualization process (“Regression\_Analysis\_Project\_1.ipynb & “Time\_Analysis\_Project\_1.ipynb”), two folders containing image files (“image\_files” & “image\_files\_2), this word document summarizing the project’s scope (Introduction, Questions of Interest, Limitations & Acknowledgements), a word document providing detailed written analysis and a PDF file of presentation slides.

**Questions of Interest:**

1. Given the structure of the database with an “outcome” variable of price and several associated “input” variables, what correlations exist for each input and how much impact do they have on price? Which variable has the largest impact? How do these correlations compare amongst the geographical areas we chose to analyze? What are the measures of central tendency for pricing in these areas?
2. Number of Bedrooms vs. Price
3. Number of Bathrooms vs. Price
4. Land Size (Acres) vs. Price
5. House Living Space (Sq\_Ft) vs. Price
6. City Population vs. Price
7. City Maximum Temperature vs. Price
8. Can we create a predictive model to estimate the price of listings based on the above variables?
9. Housing prices and the real estate industry can be extremely volatile and highly dependent on worldwide events and resulting socioeconomic impacts. Given the Covid-19 pandemic, we hypothesize a notable impact taking place around 2019/2020. Can we visualize its impact on the U.S. real estate market within our selection criteria? What are the effects of this event as well as the time variable as a whole in terms of the real estate market’s supply and demand metrics and specifically in regards to listings categorized as “sold” vs. “for sale”?
10. What are some implications we can put forth to an investor, real estate developer and the like given our analysis? Given our selection criteria and specific states studied, where would we suggest a market participant invest and why? Are there any industries outside the real estate market specifically that can glean actionable insight from our analysis?

**Limitations:**

**Dataset integrity** – It should be noted that the U.S. Real Estate listings dataset used within this analysis was directly sourced from the popular real estate listing website, “Realtor.com” (<https://www.realtor.com>). As such, we’d like to acknowledge there exists a rare degree of data entry discrepancy that results in some inconsistencies. For example, a home’s address digits that were incorrectly logged as the number of bedrooms for the listing causing a datapoint that represents a 444 bedroom home snugly situated on a 0.34 acre lot. While these inconsistencies proved to be rare, standard outlier detection and exclusion was performed to assist removing such incorrect listings. We’d also like to note this dataset and the resulting analysis are at the sacrifice of typical considerations that assess the data’s architecture such as format and speed of access/updating.

**Distribution of Housing Price Data** – Due to the highly complex and multivariate nature of the real estate industry and specifically housing prices, the data isn’t normally distributed. As such, it’s important to acknowledge that when employing linear regression modeling and other statistical tests that assume normal distribution, it’s critical to consider if the data itself is roughly linear. Because we are sampling from state specific data that contains data produced under roughly similar conditions (housing prices), we can assume linearity.

**Sample Size Discrepancy** – Given each state naturally has a different number of associated listings, we’d like to highlight there exists difference in the sample sizes used to determine the correlation for each state’s input variables respectively.

**API Reliability** – As no API is perfect, our analysis includes instances where each of the two APIs used were unable to retrieve accurate responses populating the desired results. Although consideration was taken to subsequently remove null values, acknowledgement of this instance is important. Also, please note there exists slight changes in associated regression analysis and correlations each time the regression notebook code is executed, due to the live aspect of pulling data from a central repository. As such, the temperature regression specifically, may slightly differ from written analysis metrics.

**Acknowledgements:**

**Dataset:**

<https://www.kaggle.com/datasets/ahmedshahriarsakib/usa-real-estate-dataset/data>

**Publisher:**

Ahmed Shahriar Sakib

<https://www.kaggle.com/ahmedshahriarsakib>

**Outside Collaboration and Assistance:**

-Academic Tutors

-Copilot AI

-Xpert Learning Assistant AI

-Chat GPT AI