Name: Mohammad Fahim Uddin Alvi

Student ID: 1005506668

Project Proposal

Source: (House and Mortgage data)

Available Here: [tidytuesday/data/2019/2019-02-05 at master · rfordatascience/tidytuesday · GitHub](https://github.com/rfordatascience/tidytuesday/tree/master/data/2019/2019-02-05)

Dataset

For my analysis, I will be using the state\_hpi.csv dataset, which contains information on the housing price index (HPI) for various U.S. states over time. Key variables include price\_index (the state-level house price index), us\_avg (the national average price index), year, month, and state. This dataset allows for the analysis of both time-based trends and state-specific factors influencing housing prices. The data source is a government or financial database tracking real estate trends, providing reliable and comprehensive historical data on housing prices.

Techniques

Multiple Linear Regression: I aim to use multiple linear regression to predict the price\_index (housing price index) based on variables such as year, month, and us\_avg. This analysis will help identify the relationship between time, national trends, and state-specific housing prices, providing insights into which factors most significantly impact state-level housing price fluctuations.

Binary Logistic Regression: For binary logistic regression, I plan to categorize each month’s price\_index relative to the national average. Specifically, create a binary variable indicating whether a state’s price\_index for a given month is above or below the us\_avg. Predictors for this model could include year, month, and state-specific historical data, allowing us to explore whether certain periods or states are more likely to experience above-average housing prices.

Hypotheses

Hypothesis for Multiple Linear Regression: There is a significant positive relationship between the us\_avg (national average price index) and the price\_index at the state level. I expect that states will follow national trends closely, with some variation due to local factors.

Hypothesis for Binary Logistic Regression: Certain months or years are more likely to see state-level housing prices above the national average, possibly due to economic cycles or seasonality in housing demand. For instance, I hypothesize that states with historically higher economic growth will have a price\_index above the national average.

Statistical Tests

For Multiple Linear Regression: I will use t-tests for each regression coefficient to assess the significance of predictors like year, month, and us\_avg. I will also examine the model fit using the 𝑅^2 value and an F-test for overall significance.

For Binary Logistic Regression: To evaluate the significance of predictors, I will evaluate the significance of predictors using z-tests. Additionally, I will perform a likelihood ratio test to assess the overall model fit by comparing the logistic regression model with a null model. I may also examine confidence intervals for each predictor to further confirm their significance.

Preliminary Data Analysis

Data Cleaning: Checking for missing or inconsistent values and handling them appropriately.

Descriptive Statistics: Summarizing statistics for key variables like price\_index, us\_avg, and year, and examining trends in these variables over time.

Visualization: Plotting time series for price\_index and us\_avg to identify any trends or seasonality and using box plots to detect any outliers or extreme values across states.

(Note: This proposal outlines the initial approach and techniques I plan to use for the analysis of housing price data. However, as I progress with data exploration and preliminary analysis, I may refine, add, or remove specific aspects of the analysis to better address the research objectives and enhance the accuracy and relevance of the findings. Any adjustments will be made to improve model performance or to accommodate any new patterns observed in the data.)