Analyzing Housing Price Trends and Influences:

A Statistical Exploration of the U.S. Housing Market

Mohammad Fahim Uddin Alvi

Student ID: 1005506668

STA312 - Topics in Statistics

Instructor: Nishan Mudalige

November 26, 2024

**Table of Contents** **page no.**

1. **Introduction ……………………………………………………………………………. 3**  
   Overview of the purpose, objectives, and structure of the analysis.
2. **Description of Datasets ………………………………………………………………… 3**  
   Details about the source of the data, key variables, and data cleaning steps.
3. **Methodology ……………………………………………………………………………. 4**  
   Explanation of the statistical techniques used:
4. **Multiple Linear Regression ……………………………………………………………. 4**  
   Insights from the regression analysis, including model summary and findings.

**Analysis ………………………………………………………………………… 5**

**Hypothesis ……………………………………………………………………… 6**

**Results …………………………………………………………………………...7**

1. **Binary Logistic Regression ………………………………………………………………7**  
   Classification analysis, significance of predictors, and model evaluation.

**Analysis ………………………………………………………………………….8**

**Hypothesis and Results …………………………………………………………9**

1. **Conclusion ………………………………………………………………………………. 10**  
   Final remarks on the implications of the analysis, key takeaways, and suggestions

for future research.

1. **Citations …………………………………………………………………………………. 11**
2. **Appendix ………………………………………………………………………………… 13**  
   Full R code and libraries/packages used in the analysis.
3. **Introduction**

This report analyzes state-level housing price trends in the United States using multiple and binary logistic regression models. The primary objective is to understand the relationship between the **Housing Price Index (HPI)** at the state level and factors such as **time (year and month)** and national **trends (us\_avg)**. The analysis aims to identify significant factors influencing state-specific housing price fluctuations and to determine whether certain periods or states are more likely to experience housing prices above the national average. Multiple linear regression was used to quantify the relationships between the HPI and its predictors, while Binary logistic regression was employed to classify states based on whether their HPI exceeds the national average.

The analysis will involve generating tables and graphs using RStudio and the R programming language, utilizing various libraries to support data processing, visualization, and statistical modeling. Any additional software or tools employed in the analysis will be explicitly mentioned.

1. **Description of Datasets**

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

This analysis utilizes the dataset **state\_hpi.csv** from a TidyTuesday GitHub repository. The contents of the dataset are described briefly under the Data Dictionary heading in the Source.

Additional dataset **State\_and\_US\_SA.xls** is used to clean the data of empty spaces and produce the dataframe(df). The data cleaning method is documented in the Source and the dataset is briefly described below.

**State\_and\_US\_SA.xls**:

* + A detailed dataset containing housing price data for individual states and the U.S. as a whole.
  + This dataset was cleaned to extract key variables such as:
    - price\_index: State-level HPI values.
    - us\_avg: National average HPI.
    - year and month: Temporal indicators.

The cleaned and arranged dataframe prepared for subsequent analysis is outlined below. A preview of the first few rows (head) of the dataframe is provided to demonstrate its structure and content:

**df (State-Level Housing Price Data):**

> flextable(head(df))

| year | month | state | price\_index | us\_avg |
| --- | --- | --- | --- | --- |
| 1,975 | 1 | AK | 34.63120 | 23.46149 |
| 1,975 | 2 | AK | 35.10387 | 23.61079 |
| 1,975 | 3 | AK | 35.49146 | 23.82459 |
| 1,975 | 4 | AK | 35.87223 | 24.08920 |
| 1,975 | 5 | AK | 36.28356 | 24.23425 |
| 1,975 | 6 | AK | 36.70189 | 24.21718 |

1. **Methodology**

First, we employ multiple linear regression to predict the continuous **Housing Price Index (price\_index)** at the state level based on factors such as **time (year and month)** and the **national average housing price index (us\_avg)**. This provides insights into how temporal and national trends influence state-specific housing prices. Following this, we utilize binary logistic regression to classify whether a state’s **housing price index** for a given month is above or below the **national average**. This categorical analysis allows us to identify periods and states more likely to experience above-average housing prices, offering a deeper understanding of housing market disparities.

1. **Multiple Linear Regression**

Firstly, we fit the multiple linear regression model which can be done very easily in R using the code below.

> model <- lm(price\_index ~ year + month + us\_avg, data = df)

This line fits a multiple linear regression model to predict the state-level Housing Price Index (price\_index) using year, month and us\_avg.

Calling the summary function on the model shows the following

> summary(model)

Call:

lm(formula = price\_index ~ year + month + us\_avg, data = df)

Residuals:

Min 1Q Median 3Q Max

-56.272 -8.277 -1.330 5.586 200.958

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -1.330e+03 6.214e+01 -21.396 <2e-16 \*\*\*

year 6.783e-01 3.150e-02 21.531 <2e-16 \*\*\*

month 5.246e-02 3.312e-02 1.584 0.113

us\_avg 7.418e-01 8.665e-03 85.600 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18.66 on 26873 degrees of freedom

Multiple R-squared: 0.8384, Adjusted R-squared: 0.8384

F-statistic: 4.647e+04 on 3 and 26873 DF, p-value: < 2.2e-16

From the summary above we can note the following insights.

The intercept (-1330.00) represents the expected price\_index when all predictors (year, month, us\_avg) are zero.

**year**:

* A 1-unit increase in year is associated with a 0.6783 increase in price\_index, holding other variables constant.
* The p-value (< 2e-16) indicates year is a highly significant predictor.

**month**:

* A 1-unit increase in month corresponds to a 0.0526 increase in price\_index, but the effect is not statistically significant (p = 0.113).

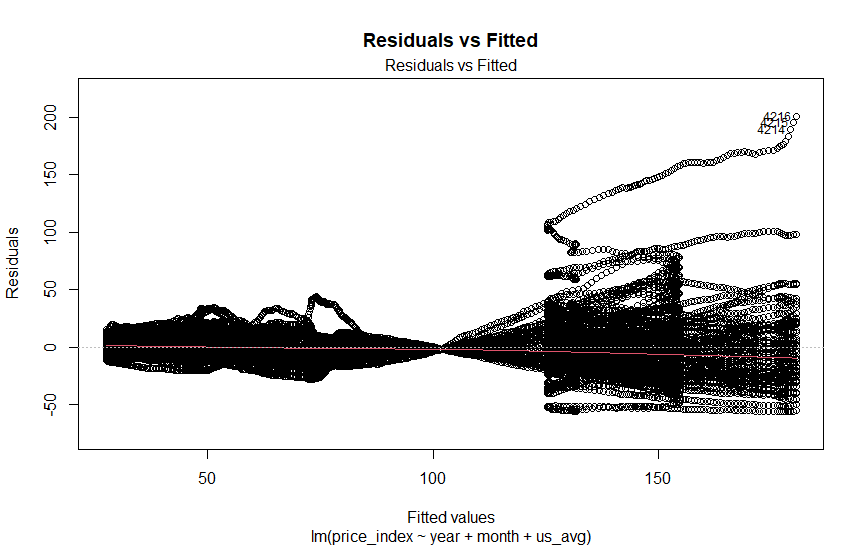
**us\_avg**:

* A 1-unit increase in us\_avg leads to a 0.7418 increase in price\_index, holding other variables constant.
* This is a strong and highly significant predictor (p < 2e-16).

R-squared is 0.8384. Meaning 83.84% of the variability in price\_index is explained by the predictors.

Checking for model fit

plot(model, which = 1, main = "Residuals vs Fitted")



As the residuals are not randomly distributed around the y = 0 line, a log transformation is applied, and a new model(model\_log) is fitted as shown below.

model\_log <- lm(log\_price\_index ~ year + month + us\_avg, data = df)

Running AIC(model, model\_log) shows that model\_log is a better fit than model as its AIC

value is less than model AIC.

**Hypothesis**

**Null Hypothesis (H₀)**: There is no significant relationship between us\_avg and log\_price\_index (β us\_avg​ = 0).

**Alternative Hypothesis (H₁)**: There is a significant positive relationship between us\_avg and log\_price\_index (β us\_avg​ > 0).

We will select between the Hypotheses in the Results section on the next page.

> summary(model\_log)

Call:

lm(formula = log\_price\_index ~ year + month + us\_avg, data = df)

Residuals:

Min 1Q Median 3Q Max

-0.88570 -0.10361 0.02508 0.10509 0.61193

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -4.349e+01 6.247e-01 -69.613 < 2e-16 \*\*\*

year 2.379e-02 3.167e-04 75.124 < 2e-16 \*\*\*

month 2.252e-03 3.330e-04 6.764 1.37e-11 \*\*\*

us\_avg 4.165e-03 8.712e-05 47.810 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1876 on 26873 degrees of freedom

Multiple R-squared: 0.8716, Adjusted R-squared: 0.8716

F-statistic: 6.08e+04 on 3 and 26873 DF, p-value: < 2.2e-16

**Results**

From above summary call, the p-value (< 2e-16) is smaller than significance level (α = 0.05)

This means we **reject** the **null hypothesis** and conclude that **us\_avg** has a significant positive relationship with **log\_price\_index**.

The multiple linear regression analysis with a log-transformed housing price index as the dependent variable confirmed a significant positive relationship between **us\_avg** (national average price index) and **log\_price\_index**. The coefficient for **us\_avg** was 0.004165 (t = 47.810, p < 2e-16), indicating that a 1-unit increase in the national average price index corresponds to a 0.4165% increase in the housing price index. The overall model explained 87.16% of the variability in housing prices, supporting the hypothesis that state-level housing prices closely follow national trends.

1. **Binary Logistic Regression**

For the binary logistic regression, the goal is to classify whether the price\_index for a given state and month is **above or below the national average (us\_avg)** and identify the factors that predict this classification.

For this regression a new binary variable **above\_avg** was created:

* **1** if price\_index > us\_avg (above the national average).
* **0** otherwise.

> df$above\_avg <- ifelse(df$price\_index > df$us\_avg, 1, 0)

**Fit the logistic regression model :**

> logit\_model <- glm(above\_avg ~ year + month + us\_avg, data = df, family = binomial)

> summary(logit\_model)

Call:

glm(formula = above\_avg ~ year + month + us\_avg, family = binomial,

data = df)

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -1.062e+02 7.156e+00 -14.838 <2e-16 \*\*\*

year 5.455e-02 3.628e-03 15.035 <2e-16 \*\*\*

month 5.221e-03 3.737e-03 1.397 0.162

us\_avg -2.686e-02 1.010e-03 -26.582 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 36834 on 26876 degrees of freedom

Residual deviance: 34554 on 26873 degrees of freedom

AIC: 34562

Number of Fisher Scoring iterations: 4

From the above summary, we find these key insights

**year**:

* A **1-unit increase in year** increases the log-odds of being above the national average by **0.05455**.
* Odds = 1.056. This means the odds of price\_index being above the national average increase by approximately **5.6% per year**.

**month**:

* The coefficient is positive but **not statistically significant** (p = 0.162), suggesting no strong seasonal effect on whether state-level prices exceed the national average.

**us\_avg**:

* The coefficient is negative, meaning a **1-unit increase in us\_avg** decreases the log-odds of price\_index being above the national average by **0.02686**.
* Odds = 0.973. This indicates that for every 1-unit increase in the national average, the odds of a state having price\_index above the national average decrease by **2.7%**.

To find a better model, we use **likelihood ratio test (LRT)** which compares:

1. The **null model** (only intercept, no predictors).
2. The **fitted model** (with predictors).

> # Fit the null model (intercept only)

> null\_model <- glm(above\_avg ~ 1, family = binomial, data = df)

>

> # Compare the null model with the full model

> anova(null\_model, logit\_model, test = "LRT")

Analysis of Deviance Table

Model 1: above\_avg ~ 1

Model 2: above\_avg ~ year + month + us\_avg

Resid. Df Resid. Dev Df Deviance Pr(>Chi)

1 26876 36834

2 26873 34554 3 2279.8 < 2.2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

The likelihood ratio test confirms that the fitted model with predictors performs significantly better than the null model (p < 0.05).

**Hypothesis and Results**

**For us\_avg:**

* **Null Hypothesis (H₀)**: β us\_avg=0 (No relationship between us\_avg and above\_avg).
* **Alternative Hypothesis (H₁)**: β us\_avg≠0 (A significant relationship exists).
* The p-value (< 2e-16) is much smaller than 0.05, so we **reject the null hypothesis**. There is a statistically significant relationship between us\_avg and the probability of a state being above the national average.

**For year:**

* The p-value (< 2e-16) indicates a statistically significant positive relationship between year and the probability of above\_avg.

**For month:**

* The p-value (0.162) is greater than 0.05, so we fail to reject the null hypothesis. There is no significant relationship between month and above\_avg.

The analysis highlights the importance of national trends (**us\_avg**) and temporal factors (**year**) in understanding state-level housing price variations. The results suggest that while time (year) has a significant positive effect, increasing national averages (**us\_avg**) reduce the likelihood of state prices exceeding the national benchmark. These findings provide valuable insights into the dynamics of housing markets and emphasize the influence of broader economic trends on state-level housing prices.

1. **Conclusion**

In this report, state-level housing price trends in the United States were analyzed using multiple linear regression and binary logistic regression to uncover the factors influencing housing price variations. The multiple linear regression analysis revealed a significant positive relationship between the national average **housing price index (us\_avg)** and **state-level prices (price\_index)**, emphasizing that state prices closely follow national trends, with a consistent upward trajectory over **time (year)**. Binary logistic regression further demonstrated that while **us\_avg** negatively impacted the odds of a state’s **housing price index** exceeding the national average, year had a positive effect, highlighting the growing likelihood of states surpassing national averages over time. The absence of significant monthly effects (month) indicated limited seasonality in housing price trends. These findings underscore the dominant role of national trends and temporal growth in shaping state-level housing prices, offering valuable insights for policymakers and stakeholders. Future studies could enhance this analysis by incorporating regional economic data or housing policy factors to gain a deeper understanding of market dynamics.

The next pages of the Report has the Citations and the appendix contains the entire R code used for this analysis including names of all the R libraries/packages used.

1. **Citations**

> # Citation for tidyverse

> citation("tidyverse")

To cite package ‘tidyverse’ in publications use:

Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J,

Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K,

Vaughan D, Wilke C, Woo K, Yutani H (2019). “Welcome to the tidyverse.” \_Journal of Open Source

Software\_, \*4\*(43), 1686. doi:10.21105/joss.01686 <https://doi.org/10.21105/joss.01686>.

A BibTeX entry for LaTeX users is

@Article{,

title = {Welcome to the {tidyverse}},

author = {Hadley Wickham and Mara Averick and Jennifer Bryan and Winston Chang and Lucy D'Agostino McGowan and Romain François and Garrett Grolemund and Alex Hayes and Lionel Henry and Jim Hester and Max Kuhn and Thomas Lin Pedersen and Evan Miller and Stephan Milton Bache and Kirill Müller and Jeroen Ooms and David Robinson and Dana Paige Seidel and Vitalie Spinu and Kohske Takahashi and Davis Vaughan and Claus Wilke and Kara Woo and Hiroaki Yutani},

year = {2019},

journal = {Journal of Open Source Software},

volume = {4},

number = {43},

pages = {1686},

doi = {10.21105/joss.01686},

}

>

> # Citation for rvest

> citation("rvest")

To cite package ‘rvest’ in publications use:

Wickham H (2024). \_rvest: Easily Harvest (Scrape) Web Pages\_. R package version 1.0.4,

<https://CRAN.R-project.org/package=rvest>.

A BibTeX entry for LaTeX users is

@Manual{,

title = {rvest: Easily Harvest (Scrape) Web Pages},

author = {Hadley Wickham},

year = {2024},

note = {R package version 1.0.4},

url = {https://CRAN.R-project.org/package=rvest},

}

>

> # Citation for readxl

> citation("readxl")

To cite package ‘readxl’ in publications use:

Wickham H, Bryan J (2023). \_readxl: Read Excel Files\_. R package version 1.4.3,

<https://CRAN.R-project.org/package=readxl>.

A BibTeX entry for LaTeX users is

@Manual{,

title = {readxl: Read Excel Files},

author = {Hadley Wickham and Jennifer Bryan},

year = {2023},

note = {R package version 1.4.3},

url = {https://CRAN.R-project.org/package=readxl},

}

>

> # Citation for janitor

> citation("janitor")

To cite package ‘janitor’ in publications use:

Firke S (2023). \_janitor: Simple Tools for Examining and Cleaning Dirty Data\_. R package version 2.2.0,

<https://CRAN.R-project.org/package=janitor>.

A BibTeX entry for LaTeX users is

@Manual{,

title = {janitor: Simple Tools for Examining and Cleaning Dirty Data},

author = {Sam Firke},

year = {2023},

note = {R package version 2.2.0},

url = {https://CRAN.R-project.org/package=janitor},

}

>

> # Citation for officer

> citation("officer")

To cite package ‘officer’ in publications use:

Gohel D, Moog S, Heckmann M (2024). \_officer: Manipulation of Microsoft Word and PowerPoint Documents\_. R

package version 0.6.7, <https://CRAN.R-project.org/package=officer>.

A BibTeX entry for LaTeX users is

@Manual{,

title = {officer: Manipulation of Microsoft Word and PowerPoint Documents},

author = {David Gohel and Stefan Moog and Mark Heckmann},

year = {2024},

note = {R package version 0.6.7},

url = {https://CRAN.R-project.org/package=officer},

}

>

> # Citation for flextable

> citation("flextable")

To cite package ‘flextable’ in publications use:

Gohel D, Skintzos P (2024). \_flextable: Functions for Tabular Reporting\_. R package version 0.9.7,

<https://CRAN.R-project.org/package=flextable>.

A BibTeX entry for LaTeX users is

@Manual{,

title = {flextable: Functions for Tabular Reporting},

author = {David Gohel and Panagiotis Skintzos},

year = {2024},

note = {R package version 0.9.7},

url = {https://CRAN.R-project.org/package=flextable},

}

>

> # Citation for dplyr

> citation("dplyr")

To cite package ‘dplyr’ in publications use:

Wickham H, François R, Henry L, Müller K, Vaughan D (2023). \_dplyr: A Grammar of Data Manipulation\_. R

package version 1.1.4, <https://CRAN.R-project.org/package=dplyr>.

A BibTeX entry for LaTeX users is

@Manual{,

title = {dplyr: A Grammar of Data Manipulation},

author = {Hadley Wickham and Romain François and Lionel Henry and Kirill Müller and Davis Vaughan},

year = {2023},

note = {R package version 1.1.4},

url = {https://CRAN.R-project.org/package=dplyr},

}

1. **Appendix**

# Libraries

library(tidyverse)

library(rvest)

library(readxl)

library(janitor)

library(officer)

library(flextable)

library(dplyr)

# Load data

state\_hpi <- readr::read\_csv("https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2019/2019-02-05/state\_hpi.csv")

mortgage\_rates <- readr::read\_csv("https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2019/2019-02-05/mortgage.csv")

recession\_dates <- readr::read\_csv("https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2019/2019-02-05/recessions.csv")

#Clean HPI

df <- read\_excel("C:/Not C/UTM/STA312/Project/State\_and\_US\_SA.xls", skip = 5) %>%

separate(Month, c("year", "month"), sep = "M") %>%

mutate\_at(vars(year, month), as.integer) %>%

filter(!is.na(year)) %>%

gather(state, price\_index, AK:WY) %>%

rename(us\_avg = `United States seasonally adjusted`) %>%

select(year, month, state, price\_index, us\_avg) %>%

mutate(price\_index = as.numeric(price\_index))

head(df)

flextable(head(df))

# Fit the regression model

model <- lm(price\_index ~ year + month + us\_avg, data = df)

# Display the model summary

summary(model)

# Residuals vs. Fitted Plot

plot(model, which = 1, main = "Residuals vs Fitted")

# Add predictions to the dataframe

df$predicted\_price <- predict(model)

# Predicted vs. Actual Plot

library(ggplot2)

ggplot(df, aes(x = predicted\_price, y = price\_index)) +

geom\_point(color = "blue", alpha = 0.7) +

geom\_abline(slope = 1, intercept = 0, color = "red", linetype = "dashed", linewidth = 1.2) +

labs(title = "Predicted vs. Actual Values",

x = "Predicted Price Index",

y = "Actual Price Index")

model\_non\_linear <- lm(price\_index ~ year + month + us\_avg + I(us\_avg^2), data = df)

summary(model\_non\_linear)

df$log\_price\_index <- log(df$price\_index)

model\_log <- lm(log\_price\_index ~ year + month + us\_avg, data = df)

summary(model\_log)

model\_interaction <- lm(price\_index ~ year \* us\_avg + month, data = df)

summary(model\_interaction)

AIC(model, model\_log,model\_non\_linear,model\_interaction)

plot(model\_log, which = 1) # Residuals vs. Fitted

plot(model\_log, which = 2) # Q-Q Plot

df$above\_avg <- ifelse(df$price\_index > df$us\_avg, 1, 0)

table(df$above\_avg)

# Fit the binary logistic regression model

logit\_model <- glm(above\_avg ~ year + month + us\_avg, data = df, family = binomial)

summary(logit\_model)

# Fit the null model (intercept only)

null\_model <- glm(above\_avg ~ 1, family = binomial, data = df)

# Compare the null model with the full model

anova(null\_model, logit\_model, test = "LRT")