Analyzing and Predicting Real Estate Market Dynamics: A Comprehensive Data Science Approach Using R

SAMUEL ABOYE

March 02, 2024

Contents

Contents

Introduction

Days on Market (DoM) analysis is a fascinating aspect of the real estate market that provides valuable insights into market conditions, buyer and seller behaviors, and broader economic trends. It helps stakeholders gauge market liquidity, demand, and pricing strategies by examining how long properties stay on the market before being sold. This becomes a quintessential data science problem when we apply statistical analysis and predictive modeling to large datasets to uncover patterns, correlations, and trends.

The Days on Market (DoM) metric is a barometer for the real estate market's health and efficiency as it offers insights into the balance between supply and demand, economic vitality, and consumer confidence. When the market is buoyant, properties tend to sell quickly, leading to shorter DoMs, whereas, in slower markets, DoMs lengthen as properties take longer to sell. The complexity of factors influencing DoM, including economic indicators, location specifics, property characteristics, and seasonal trends, makes this an ideal subject for data science exploration. With its comprehensive suite of packages for data analysis, R presents an opportunity to dissect these complexities through a multifaceted lens. The project aims to leverage R's capabilities to analyze historical trends and forecast future market behaviors, providing stakeholders with a data-driven basis for decision-making.

The Days on Market (DoM) metric is a barometer for the real estate market's health and efficiency as it offers insights into the balance between supply and demand, economic vitality, and consumer confidence. When the market is buoyant, properties tend to sell quickly, leading to shorter DoMs, whereas, in slower markets, DoMs lengthen as properties take longer to sell. The complexity of factors influencing DoM, including economic indicators, location specifics, property characteristics, and seasonal trends, makes this an ideal subject for data science exploration. With its comprehensive suite of packages for data analysis, R presents an opportunity to dissect these complexities through a multifaceted lens. The project aims to leverage R's capabilities to analyze historical trends and forecast future market behaviors, providing stakeholders with a data-driven basis for decision-making.

R,a language and environment for statistical computing and graphics, is particularly well-suited for this task due to its powerful data manipulation, visualization, and modeling packages. This project aims to leverage R's capabilities to address questions about DoM dynamics and provide valuable insights for various interested parties, including investors, real estate professionals, and policymakers.

To further enrich the project and its execution within the R ecosystem, it's essential to consider the nuances of real estate market analysis and how R's specific features and packages can be tailored to meet these challenges. Enhancing the writing to include a deeper understanding of the problem space and more detailed plans for implementation and analysis in R will provide a more precise roadmap for the project.

Research Questions

1. What impact does the DoM have on listings across various regions?
2. How do economic indicators (interest rates, employment rates, GDP growth) shape the DoM trends?
3. How do seasonal variations and timing (listing month or season) affect the DoM?
4. To what extent do local market conditions (supply-demand balance, median house prices, inventory levels) and location specifics (urban vs. rural, neighborhood amenities) influence DoM?
5. Investigate the impact of marketing strategies (online presence, virtual tours, professional photography) on DoM.
6. Explore the influence of external factors such as policy changes, economic shocks, or global events (e.g., pandemics) on the DoM.

Approach

To address the research questions, the project will follow a structured approach involving data preparation and cleaning, exploratory data analysis, statistical and predictive modeling, and advanced visualization. Due to its powerful data manipulation, visualization, and modeling packages, the project will use R, a language and environment for statistical computing and graphics.

The project aims to uncover patterns, correlations, and trends in the real estate market by applying data science techniques such as statistical analysis and predictive modeling to large datasets. By analyzing the DoM metric, the project aims to provide insights into the balance between supply and demand, economic vitality, and consumer confidence. These insights can help stakeholders make informed decisions related to investment, pricing strategies, and policymaking.

Dataset

The project will use different real estate market datasets, including property characteristics, economic indicators, and location specifics. The datasets will be cleaned, preprocessed, and analyzed to gain insights into the DoM dynamics.

Required Packages:

The project will utilize several R packages, including Tidyr and dplyr for data preparation and cleaning, ggplot2 for exploratory data analysis, caret for predictive modeling, and forecast for time series analysis.

Plots and Table Needs:

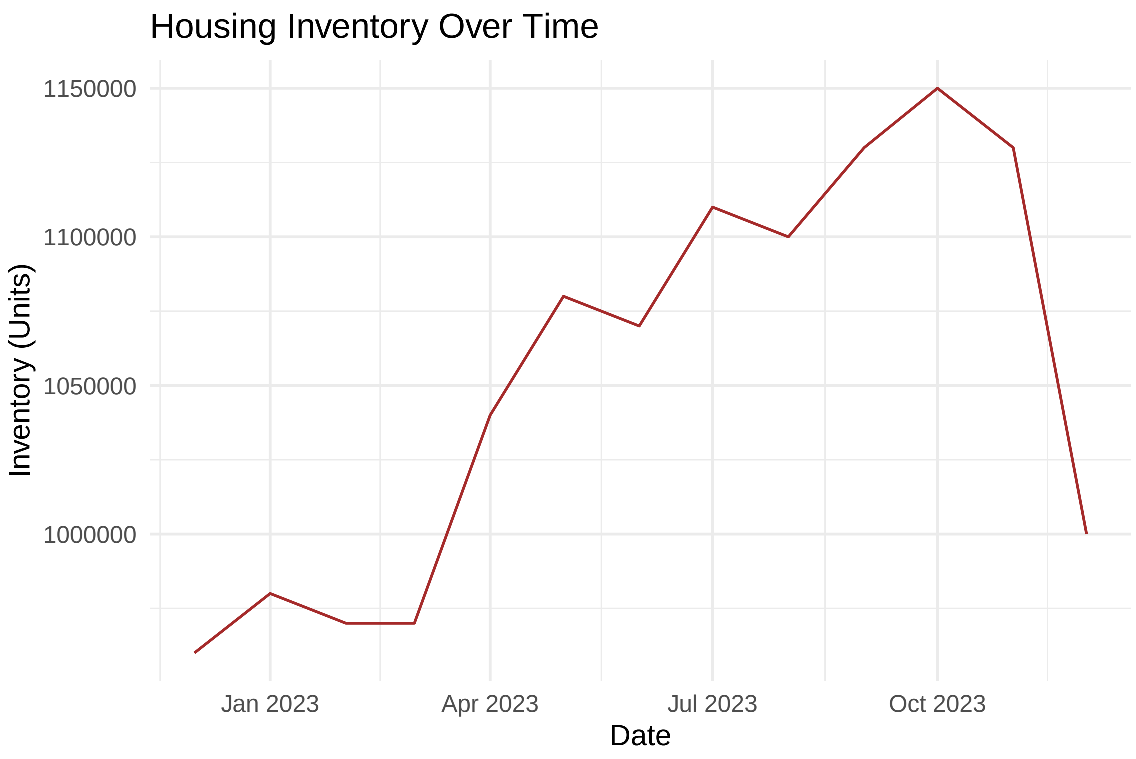
The project will use various types of visualizations, including histograms, scatter plots, box plots, heat maps, time series decomposition plots, and geographic visualizations. Tables will be used to present statistical metrics such as R-squared and RMSE.

Exploratory Data Analysis (EDA)

The Exploratory Data Analysis (EDA) on the datasets provided valuable insights into the U.S. housing market's dynamics across various metrics.

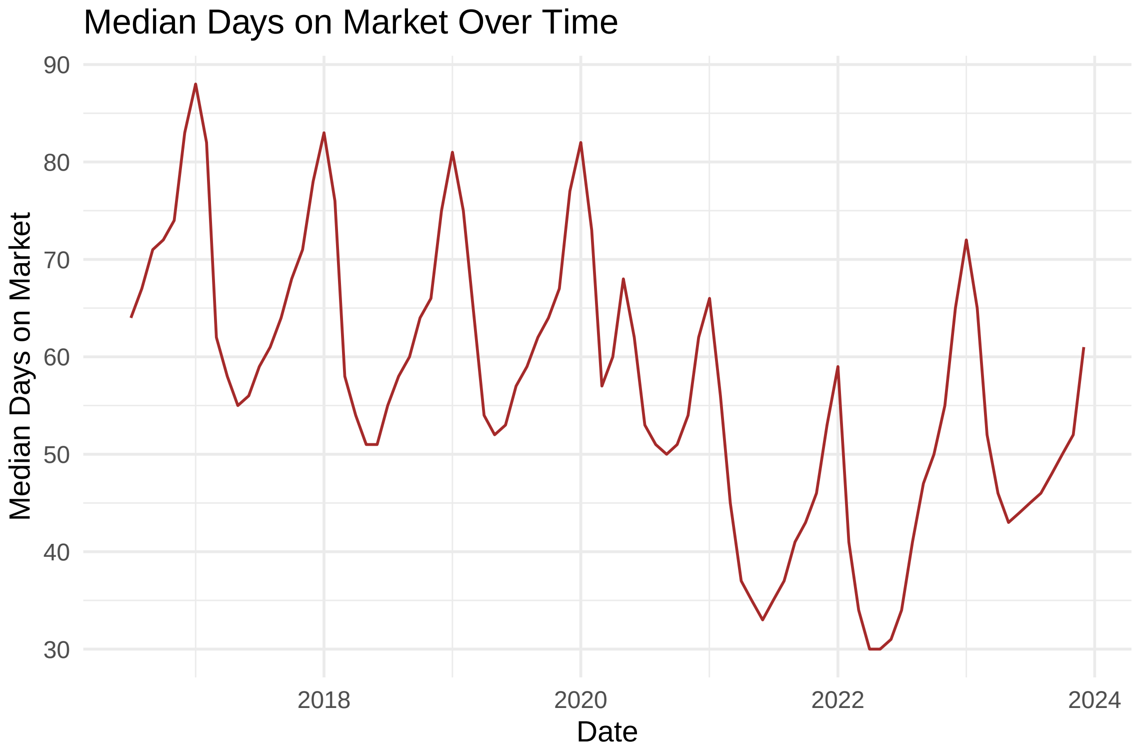
Housing Inventory Over Time:

The time series plot shows fluctuations in housing inventory with some periods of increase and decreases over the observed months in 2023.The descriptive statistics reveal a range of housing inventory units from a minimum of 960,000 to a maximum of 1,150,000. The average inventory level is approximately 1,053,077 units, with a standard deviation of 69,807 units, indicating variability over time. The line graph presents the housing inventory trend which, at a glance, shows some volatility. The visual suggests there could be external factors influencing these inventory levels, warranting a deeper investigation into economic indicators, policy changes, or market disruptions.



Median Days on Market Over Time:

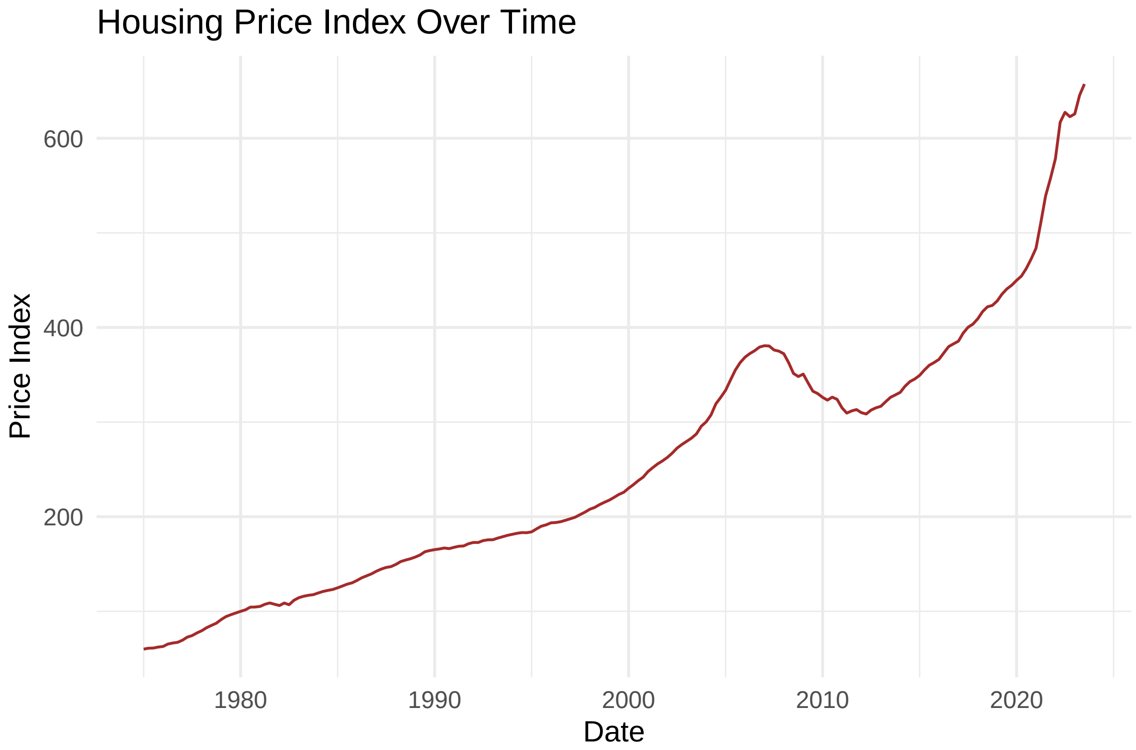
The visualization of the median days on the market portrays a clear seasonal pattern, with durations fluctuating predictably throughout the years. This pattern may be influenced by typical real estate cycles, weather conditions, and holidays, where certain times of the year are more favorable for buying and selling properties.



The median days on the market exhibit a seasonal pattern, with peaks and troughs corresponding to different months. This suggests a cyclical nature in how long houses stay on the market before being sold. The shortest median time on the market was 30 days, and the longest was 88 days. On average, houses stayed on the market for about 57.1 days, with a standard deviation of 13.7 days, implying moderate fluctuations throughout the year.

Housing Price Index Over Time:

The Housing Price Index demonstrates a long-term upward trend, suggesting an overall increase in housing prices over the decades. The index ranged from a low of 60.1 to a high of 657, showing substantial growth over time. The mean index value stands at 254 with a high standard deviation of 138, reflecting significant changes in housing prices that could be attributed to market forces, economic conditions, and inflation. The plot of the Housing Price Index illustrates a substantial rise, especially noticeable in recent years. The steep increase could be indicative of a market that is becoming increasingly expensive, potential housing bubbles, or it could reflect the natural appreciation of real estate values over time.



Conclusions:

The EDA indicates that the U.S. housing market experiences both short-term fluctuations and long-term trends across various metrics. While there is a clear cyclical pattern in the median days on the market, the overall increase in the housing price index points to a long-term trend of rising property values. The variability in housing inventory levels suggests that there may be periods of tight supply or excess availability, each of which has different implications for buyers and sellers in the market.

Next Steps: - How can the analysis be extended to include external datasets such as demographic data and urban economics models? - How can the project be scaled to handle larger real estate datasets? - How can the insights gained from the analysis be integrated into real-world decision-making processes?

10.3 Final Project Step 2

Data Importing and Cleaning

In the data preparation phase, the primary objective is to ensure the quality and consistency of the dataset, which involves several critical steps.

Handling Missing Values:

Missing values can introduce bias or inaccuracies in our analysis and models. To mitigate this, I first identify missing values across our dataset. The approach to handling these missing values is context-dependent: - For numerical variables, I may impute missing values with the mean or median of the column, which offers a simple and quick way to maintain data integrity without introducing significant bias. - For categorical variables, I can impute missing values with the most frequent category, or mode, thus preserving the distribution of categories. - Alternatively, especially when a significant number of values are missing, I may choose to remove rows or columns with missing values to maintain the robustness of our analyses.

Correcting Data Types:

Ensuring accurate data types is essential for any subsequent operations. Dates should be in Date format to facilitate time-series analysis or any operations that rely on chronological order. Numerical variables should be in a numeric format to enable mathematical operations and statistical analysis. Incorrect data types can result in errors or inappropriate data handling during analysis.

To implement the above steps, I have defined a clean_data function in R, which efficiently processes each dataset using a combination of tidyverse functions:

Meta data

Data Variable	Description	File Path
`housing_inventory`	Housing Inventory Data	`HOSINVUSM495N.xls`
`median_days_on_market`	Median Days on Market Data	`MEDDAYONMARUS.xls`
`housing_price_index`	Housing Price Index Data	`data/USSTHPI.xls`
`economic_totals`	Economic Totals Data	`data/ETOTALUSQ176N.xls`

Cleaning Data

Identify and handle missing values. Depending on the context, you may choose to fill them with mean/median (for numerical variables) or the most frequent category (for categorical variables), or simply remove rows/columns with missing values using tidyr::drop_na(). Correct data types if necessary, ensuring dates are in Date format, numeric variables are not read as characters, etc., using as.Date(), as.numeric(), etc.

Remove duplicates with dplyr::distinct(). Filter out irrelevant data or outliers based on the understanding of your data and research questions.

Merge the datasets based on the observation_date column

```r
combined_data <- median_days_on_market %>%
  left_join(housing_price_index, by = "observation_date") %>%
  left_join(economic_totals, by = "observation_date")

combined_data <- combined_data %>%
  mutate(
    MED_DAY_ON_MARUS = replace_na(MED_DAY_ON_MARUS, 0),
    USSTHPI = replace_na(USSTHPI, 0),
    Housing_Inventory = replace_na(Housing_Inventory, 0)
  )
```

Merge the datasets

Table 2: Head of Combined Data

observation_date	MED_DAY_ON_MARUS	USSTHPI	Housing_Inventory
2016-07-01	64	379.81	136289
2016-08-01	67	0.00	0
2016-09-01	71	0.00	0
2016-10-01	72	382.76	136554
2016-11-01	74	0.00	0
2016-12-01	83	0.00	0
2017-01-01	88	385.53	136818
2017-02-01	82	0.00	0
2017-03-01	62	0.00	0
2017-04-01	58	394.15	137083
2017-05-01	55	0.00	0
2017-06-01	56	0.00	0
2017-07-01	59	400.24	137354
2017-08-01	61	0.00	0
2017-09-01	64	0.00	0
2017-10-01	68	403.58	137637
2017-11-01	71	0.00	0
2017-12-01	78	0.00	0
2018-01-01	83	409.26	137920
2018-02-01	76	0.00	0
2018-03-01	58	0.00	0
2018-04-01	54	416.79	138203
2018-05-01	51	0.00	0
2018-06-01	51	0.00	0
2018-07-01	55	421.88	138488
2018-08-01	58	0.00	0
2018-09-01	60	0.00	0
2018-10-01	64	423.32	138778
2018-11-01	66	0.00	0
2018-12-01	75	0.00	0
2019-01-01	81	428.05	139069
2019-02-01	75	0.00	0
2019-03-01	65	0.00	0
2019-04-01	54	435.20	139360
2019-05-01	52	0.00	0
2019-06-01	53	0.00	0

observation_date	MED_DAY_ON_MARUS	USSTHPI	Housing_Inventory
2019-07-01	57	440.74	139655
2019-08-01	59	0.00	0
2019-09-01	62	0.00	0
2019-10-01	64	444.66	139961
2019-11-01	67	0.00	0
2019-12-01	77	0.00	0
2020-01-01	82	449.81	140266
2020-02-01	73	0.00	0
2020-03-01	57	0.00	0
2020-04-01	60	454.31	140600
2020-05-01	68	0.00	0
2020-06-01	62	0.00	0
2020-07-01	53	462.26	140914
2020-08-01	51	0.00	0
2020-09-01	50	0.00	0
2020-10-01	51	472.44	141251
2020-11-01	54	0.00	0
2020-12-01	62	0.00	0
2021-01-01	66	483.83	141589
2021-02-01	56	0.00	0
2021-03-01	45	0.00	0
2021-04-01	37	510.74	141927
2021-05-01	35	0.00	0
2021-06-01	33	0.00	0
2021-07-01	35	539.20	142288
2021-08-01	37	0.00	0
2021-09-01	41	0.00	0
2021-10-01	43	557.79	142697
2021-11-01	46	0.00	0
2021-12-01	53	0.00	0
2022-01-01	59	578.36	143105
2022-02-01	41	0.00	0
2022-03-01	34	0.00	0
2022-04-01	30	617.03	143514
2022-05-01	30	0.00	0
2022-06-01	31	0.00	0
2022-07-01	34	627.42	143923
2022-08-01	41	0.00	0
2022-09-01	47	0.00	0
2022-10-01	50	622.78	144332
2022-11-01	55	0.00	0
2022-12-01	65	0.00	0
2023-01-01	72	625.78	144740
2023-02-01	65	0.00	0
2023-03-01	52	0.00	0
2023-04-01	46	645.38	145149
2023-05-01	43	0.00	0
2023-06-01	44	0.00	0
2023-07-01	45	657.32	145558
2023-08-01	46	0.00	0
2023-09-01	48	0.00	0
2023-10-01	50	0.00	145967

observation_date	MED_DAY_ON_MARUS	USSTHPI	Housing_Inventory
2023-11-01	52	0.00	0
2023-12-01	61	0.00	0

Visualization

```
ggplot(combined_data, aes(x = observation_date)) +
  geom_line(aes(y = median_days), color = "red") +
  geom_line(aes(y = price_index), color = "blue") +
  geom_line(aes(y = economic_total), color = "green") +
  labs(title = "Median Days on Market, Housing Price Index, and Economic Total Over Time",
       x = "Date",
       y = "Metrics",
       color = "Metric") +
  theme_minimal() +
  scale_y_continuous(sec.axis = sec_axis(~ ., name = "Secondary Axis Name"))
```

What information is not self-evident?

Raw data provides the basic numbers and categories, but it doesn't reveal subtle insights such as patterns, anomalies, and predictions. For example, the raw data does not immediately show the impact of seasonal changes on the housing market or the time lag between economic indicators and property values. Analytical methods are required to uncover these underlying trends and correlations, essential for a nuanced understanding of the data.

What are different ways you could look at this data?

The dataset can be analyzed using different approaches to uncover various insights. A longitudinal study can help identify trends over time, while a cross-sectional analysis can showcase differences across multiple regions or property types at a specific time. Comparative analysis can help reveal disparities between market segments, and correlational studies might identify relationships between market dynamics and economic factors.

How do you plan to slice and dice the data?

I will use segmentation as a critical strategy to analyze the data effectively. This means breaking down the dataset into smaller subsets based on various criteria such as periods or property characteristics. By doing this, I can conduct an in-depth analysis to understand specific market behaviors. ## How could you summarize your data to answer key questions? To effectively summarize data, I need both numerical and visual summaries. Descriptive statistics can provide an overview of the central tendencies and variances, whereas inferential statistics can test hypotheses about the data. Summary tables can present the data concisely, highlighting key figures such as average inventory levels or median days on the market across different periods or locations.'

What types of plots and tables will help you to illustrate the findings to your questions?

Visual aids like line graphs can help track trends over time. Bar charts help compare data across categories, while scatter plots can suggest correlations or outliers. Heat maps can illustrate the density of data points for tabular data, and pivot tables can help compare multifaceted data. These tools facilitate a better understanding of the data and make the findings more accessible to stakeholders

Do you plan on incorporating any machine learning techniques to answer your research questions? Explain.

Machine learning techniques will play a key role in our analysis. Predictive modeling can forecast future market trends, while classification algorithms can provide insights into market conditions. Regression models can help estimate the impact of specific variables on housing prices, and clustering can reveal hidden market segments. These techniques will allow us to move beyond mere description and into prediction and optimization.

What questions do you have now, that will lead to further analysis or additional steps?

As I explore the data, I come across a plethora of questions. Do macroeconomic trends have any relation to local housing market dynamics? Is it possible to predict how a seller's market transforms into a buyer's market? How do external factors, such as policy modifications and global disruptions, significantly impact housing prices and inventory? As we find the answers to these questions, new areas for investigation will likely arise, leading to further analytical pursuits.

11.3 Final Project Step 3

Introduction

Days on Market (DoM) analysis stands as a cornerstone in understanding the real estate market's pulse, offering a lens through which market conditions, buyer and seller behaviors, and overarching economic trends can be discerned. This analysis not only aids stakeholders in gauging market liquidity and demand but also plays a pivotal role in shaping pricing strategies by scrutinizing the duration properties remain available before being sold. When approached as a data science endeavor, DoM analysis transcends traditional boundaries, employing statistical scrutiny and predictive modeling on expansive datasets to unearth patterns, correlations, and trends.

Serving as a barometer for the real estate market's vitality, the DoM metric illuminates the intricate balance between supply and demand, economic health, and consumer confidence. In thriving markets, properties are quickly snapped up, leading to shorter DoMs. Conversely, in more languid markets, properties linger longer, indicating extended DoMs. The multifarious factors influencing DoM, including but not limited to economic indicators, location nuances, property attributes, and seasonal variations, present a rich tapestry for data science exploration. R, with its robust suite of data analysis packages, offers a formidable platform to dissect these complexities, aiming to analyze historical trends and project future market behaviors, thereby furnishing stakeholders with a solid foundation for informed decision-making.

As a resident of Nebraska, my connection to Omaha is both personal and profound. The city's vibrant essence and the community's spirited dynamism inspire a deep-seated commitment to addressing significant local challenges. Integrating the Omaha Zillow dataset into our analytical framework allows for an immersive exploration of the Omaha real estate market. By refining data types for dates and prices, we aim to unlock detailed insights into market behaviors and trends specific to Omaha, Nebraska. This endeavor is not merely an academic exercise but a quest to contribute meaningful insights to the community that resonates with my sense of belonging.

Problem Statement

In the vibrant city of Omaha, Nebraska, understanding the dynamics of the real estate market is crucial for a wide array of stakeholders, including homeowners, potential buyers, investors, real estate professionals, and policymakers. The Days on Market (DoM) metric, representing the elapsed time from when a property is listed until it is sold, serves as a vital indicator of market health, liquidity, and efficiency. However, the complexity of the real estate market, influenced by economic factors, location specifics, property characteristics, and seasonal variations, presents a multifaceted challenge in analyzing and forecasting market trends.

How you addressed this problem statement

This project aims to address the specific nuances of the Omaha real estate market by leveraging the comprehensive Omaha Zillow dataset. Our goal is to: 1. Analyze Historical Trends: Examine the historical DoM data to identify patterns, trends, and seasonal variations within the Omaha real estate market. 2. Understand Market Dynamics: Investigate how various factors such as economic indicators, property features, and location specifics contribute to the DoM in Omaha. 3. Forecast Future Market Behaviors: Utilize predictive modeling to forecast future trends in the DoM, aiding stakeholders in making informed decisions.

By focusing on the Omaha market, we seek to provide localized insights that reflect this community's unique characteristics and dynamics. The integration of detailed property data, including dates, prices, and features, will enable a deeper understanding of the factors driving the DoM in Omaha. Through a data-driven approach, we aim to equip stakeholders with actionable intelligence to navigate the complexities of the real estate market effectively, fostering a more vibrant and sustainable community in Omaha, Nebraska.

Analysis

Data Utilization

The analysis will be conducted using a dataset from Zillow, which includes detailed information on property listings in Omaha, Nebraska. This dataset provides a rich foundation for exploring the DoM metric, supplemented by additional data preprocessing and enrichment techniques to ensure accuracy and relevance.

Methodology

R, a comprehensive language and environment for statistical computing and graphics, will be the primary tool for this analysis. It offers robust capabilities for data manipulation, visualization, and modeling, making it ideally suited for exploring the complexities of the Omaha real estate market.

Significance

By providing a data-driven exploration of the Omaha real estate market, this project aims to offer valuable insights for homeowners, real estate professionals, investors, and policymakers. Understanding the dynamics at play can help these stakeholders make more informed decisions, whether for personal investments, market analysis, or urban planning and development strategies.

Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a critical first step in understanding your dataset before moving on to more complex analyses or building predictive models. For the Omaha real estate market dataset sourced from Zillow, EDA will focus on uncovering the underlying structure of the data, identifying anomalies or outliers, understanding the relationship between variables, and discovering patterns or trends. Here's how you can approach the EDA for this dataset:

```
## Warning in FUN(X[[i]], ...): NAs introduced by coercion
```

```
## # A tibble: 1 x 6
##   Avg_Days_in_Market Median_Days_in_Market Avg_Days_to_Pending
##                <dbl>                 <dbl>               <dbl>
## 1               67.2                    57                32.9
## # i 3 more variables: Median_Days_to_Pending <dbl>, Avg_Price <dbl>,
## #   Median_Price <dbl>
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   44000  215000  290000  326771  415000  985000       1
```

Summary Statistics of Omaha Real Estate Market

An analysis of the Omaha real estate market reveals insightful statistics about property listings. The average Days on Market (DoM) for properties stands at 67.2 days, with a median of 57 days. This suggests a relatively swift turnover for properties, with half of them being sold in under two months. Properties typically spend an average of 32.9 days in the market before moving to pending status, indicating a brisk initial interest from potential buyers. The median days to pending is slightly lower at 32 days, reflecting that a majority of properties receive offers in just over a month after being listed. Price analysis shows a mean listing price of $326,771, with the median price at $290,000. This points to a market with a moderate range of property prices, and where the majority of homes are priced under the $300k mark. The price distribution further underscores the market’s diversity, ranging from a minimum of $44,000 to a maximum of $985,000. However, it’s noteworthy that there is at least one outlier or potential error in the pricing data, as indicated by the presence of an NA in the price column. These statistics provide a quantitative backdrop for the Omaha housing market, reflecting both the vibrancy of the market and the variety of homes available.

Some of the newly constructed properties are experiencing longer periods on the market. To focus our analysis on properties with typical market behaviors, I will exclude properties that have been listed for more than 181 days.

```
## # A tibble: 1 x 6
##   Avg_Days_in_Market Median_Days_in_Market Avg_Days_to_Pending
##                <dbl>                 <dbl>               <dbl>
## 1               65.6                    57                31.8
## # i 3 more variables: Median_Days_to_Pending <dbl>, Avg_Price <dbl>,
## #   Median_Price <dbl>
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA’s
##   44000  215000  290000  326144  414990  985000       1
```

Summary Statistics with a Focus on Typical Market Behavior

In an effort to gain a clearer understanding of the typical property sales cycle in Omaha, we have refined our analysis by excluding properties that linger on the market for an extended period. Specifically, we have omitted listings with Days on Market (DoM) exceeding 181 days, which we consider outliers for the purpose of this study.

The adjusted summary statistics provide a more focused view of the market: - The average Days on Market for properties is now 65.6 days, slightly less than the overall average. This reinforces the perception of Omaha as a relatively active market. - The median Days on Market remains consistent at 57 days, indicating that half of the properties are sold in less than two months, even after removing outliers. - Properties move to pending status after an average of 31.8 days, reflecting a dynamic where properties attract offers fairly quickly. - The mean property price, unaffected by the most sluggish sales, stands at $326,144, with the median at $290,000. This suggests a market with accessible pricing for a broad range of potential buyers.

The price data, with a minimum listing of $44,000 and a maximum of $985,000, points to a diverse market accommodating various buyer segments. The presence of a single NA value in the price data indicates a possible data entry issue that warrants further investigation to ensure the integrity of the analysis.

By focusing on properties that align more closely with buyer and seller expectations in terms of sales duration, we can provide stakeholders with insights that are highly relevant to typical market transactions. This nuanced approach to data, which leverages R’s robust analytical capabilities, underscores the market’s vibrancy and positions buyers and sellers to make well-informed decisions.

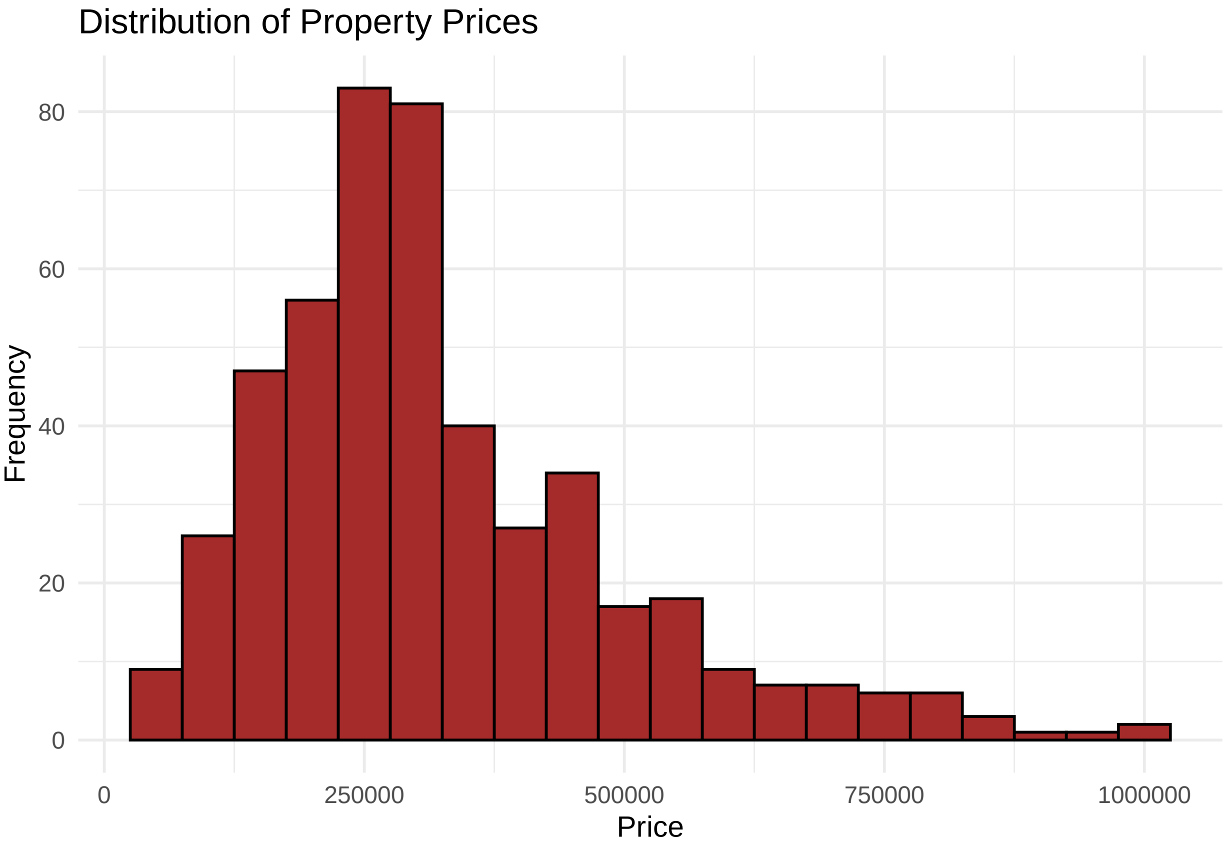
```
## # A tibble: 2 x 2
##   Status            Avg_Days
##   <chr>                <dbl>
## 1 Listed to Pending     32.6
## 2 Pending to Sold       33.8
```

Average Days by Transaction Status

Upon reviewing the Omaha real estate data, we observe two key stages in the property sales cycle. The first stage, from when a property is listed to when it goes pending (indicating an accepted offer), takes an average of 32.6 days. The second stage, from pending status to the final sale, averages slightly longer at 33.8 days.

These durations provide insight into the pace at which properties are moving through the sales pipeline. It is indicative of a balanced market where properties are advancing from listing to sale in a consistent timeframe, without excessive delays.

```
## Warning: Removed 1 rows containing non-finite values (‘stat_bin()‘).
```

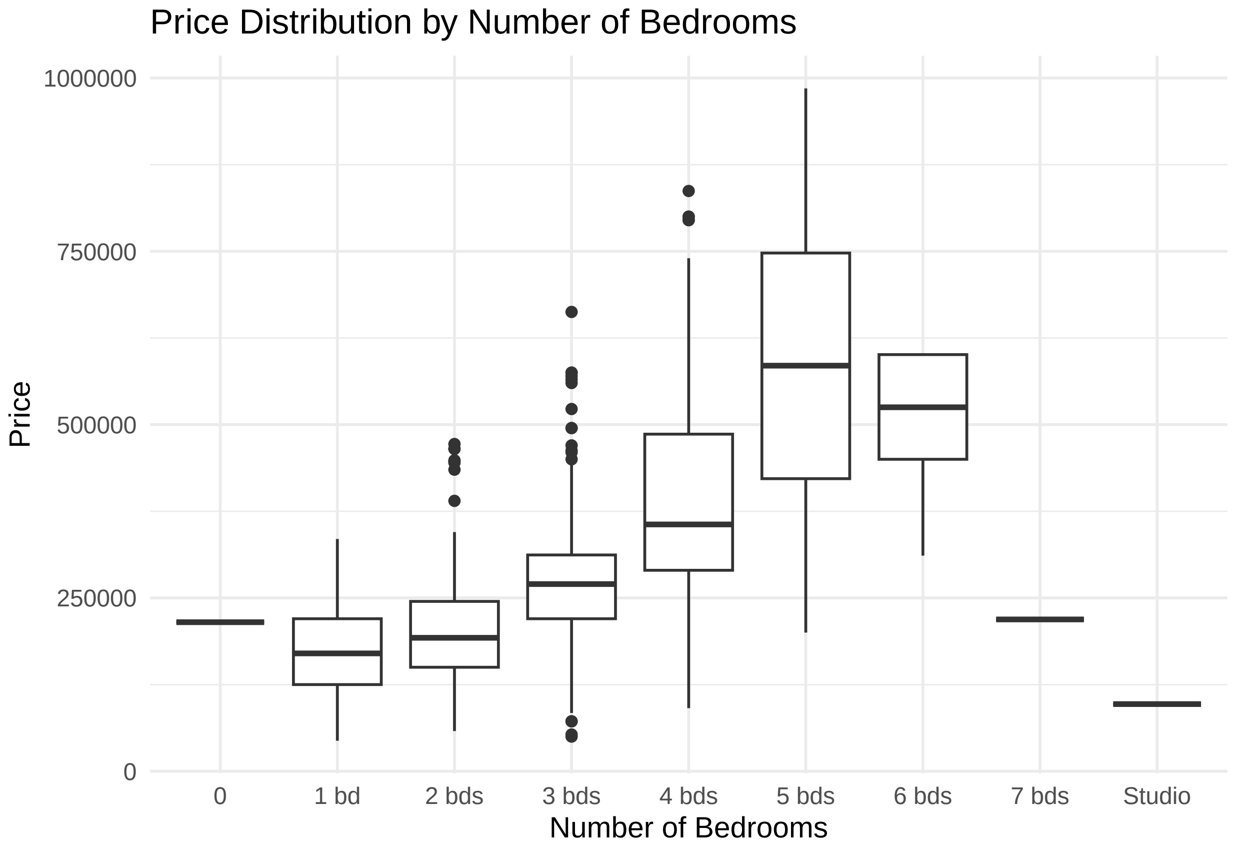


Distribution of Property Prices

From the histogram, we can observe that the majority of properties are clustered around the lower to mid-price ranges, with a peak frequency in properties priced between $200,000 and $300,000. The frequency gradually decreases as the price increases, with fewer properties listed at higher price points, indicating a tail towards the more expensive properties.

This distribution suggests that the Omaha market is predominantly composed of moderately priced homes, with luxury properties representing a smaller portion of the market. Such a price distribution is typical of residential markets with a solid middle-class presence. Understanding this distribution is essential for stakeholders, as it highlights the most commonly listed properties’ price range, reflecting the market’s accessibility to average buyers. It also informs sellers and real estate professionals about the most competitive pricing brackets and potential inventory gaps in the higher or lower ends of the market.

```
## Warning: Removed 1 rows containing non-finite values (‘stat_boxplot()‘).
```



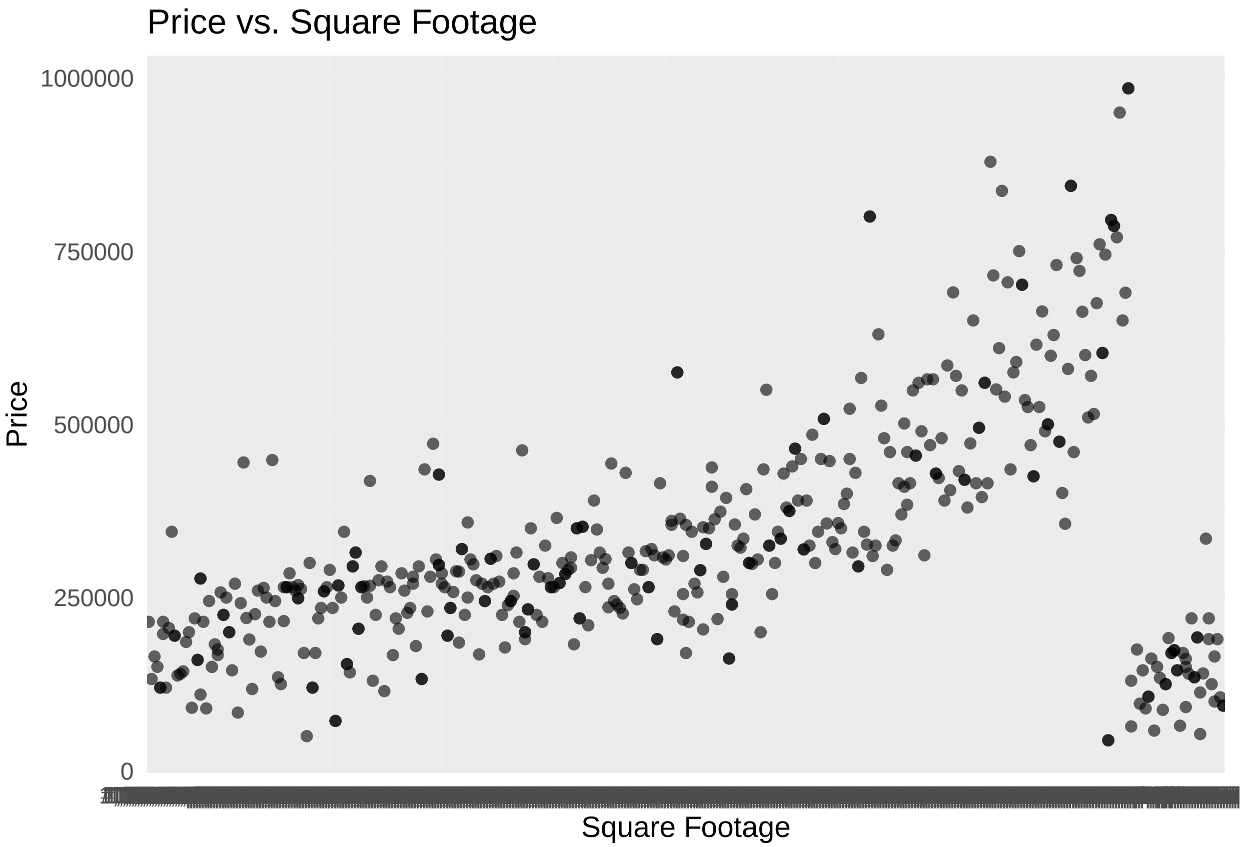
Exploring the Relationship Between Property Prices and Bedroom Count in Omaha

An analysis of the Omaha housing data presents a box plot that offers a visual summary of how property prices are distributed across different bedroom categories. The plot showcases a range from studio apartments to properties with seven bedrooms, shedding light on the diversity of the housing market. The median price tends to increase with the number of bedrooms, as expected due to larger homes typically commanding higher prices. However, there is considerable overlap in the interquartile ranges, particularly between adjacent bedroom categories, which suggests a nuanced pricing structure that isn't solely dependent on bedroom count.

Outliers are present across several bedroom categories, particularly in the 3 and 4-bedroom segments, indicating that there are properties priced significantly higher than the median for their category. These could represent premium offerings in desirable locations or properties with features that elevate their value above the typical market range for their size.

It's important to note a warning at the bottom of the plot, indicating the removal of rows containing missing values. This suggests that some data points were excluded from the analysis, potentially impacting the representativeness of the visualization. Overall, this box plot provides a clear visual representation of the price diversity in the Omaha real estate market, segmented by the number of bedrooms, and serves as a valuable tool for potential buyers, sellers, and investors to understand market trends at a glance.

```
## Warning: Removed 1 rows containing missing values (‘geom_point()‘).
```



Scatter Plot Analysis of Property Price Against Square Footage

The scatter plot visualizes the correlation between property prices and their square footage in Omaha's real estate market. Each point on the plot corresponds to a property, with its position along the horizontal axis indicating its size and along the vertical axis representing its price. Key Observations: - There is a trend indicating that as the square footage of a property increases, the price tends to rise as well. This is in line with market expectations where larger properties typically command higher prices. - The distribution of data points suggests a positive correlation between size and price, although it is not strictly linear, reflecting the influence of other factors on property valuation. - There are clusters of properties at certain square footage ranges, possibly indicating common property sizes within the market. - The presence of several outliers, particularly properties with high prices irrespective of size, may indicate luxury homes or those with unique features that enhance their value. - The variability in price for properties with similar square footage suggests that factors other than size, such as location, property condition, or market timing, also play a significant role in determining the price.

This scatter plot is a valuable exploratory tool for stakeholders to identify price trends and anomalies, aiding buyers in making informed decisions and sellers in setting competitive prices. Further statistical analysis would be beneficial to quantify the exact relationship between price and square footage and to control for other variables affecting property prices in Omaha.

Time series forecasting

Time Series Forecast Using ARIMA(1,2,1) Model

The plot illustrates a time series forecast derived from an ARIMA(1,2,1) model, a type of statistical model used to analyze and forecast time series data. The black line represents the historical data points, showing the trend and fluctuations over time, while the blue line and shaded area indicate the forecasted values and the confidence intervals, respectively. Key Observations:

- Trend: The historical trend appears relatively stable before the forecast period begins.
- Forecast: The model predicts a sharp increase in the forecasted metric, which could be Days on Market, property prices, or another key variable. This uptick suggests a significant change expected in the near future.
- Confidence Interval: The shaded area around the forecast line shows the confidence interval, representing the uncertainty in the predictions. A wider interval indicates greater uncertainty.

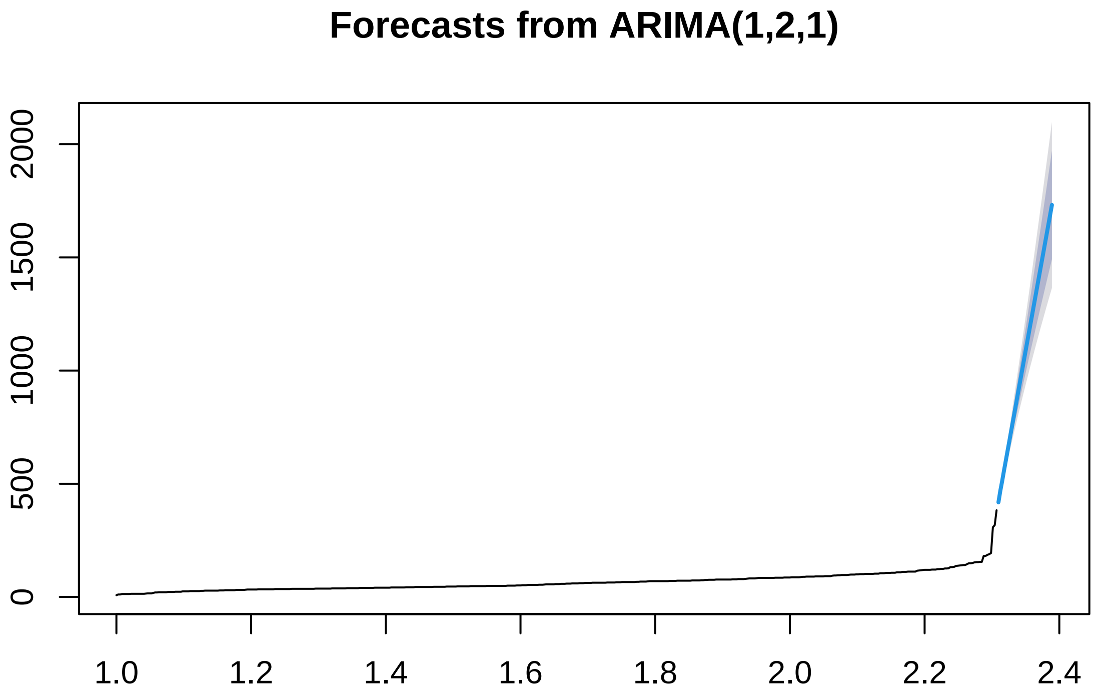
Implications:

- Market Change: If the metric being forecasted is Days on Market, a sharp increase could imply a slowing real estate market, possibly due to economic downturns, policy changes, or shifts in consumer confidence.
- Actionable Insights: Stakeholders may need to prepare for the predicted market shift, adjusting strategies for buying, selling, or holding properties.
- Further Analysis Needed: The sudden rise in the forecast could also suggest the model may be reacting to a recent anomalous event or trend. It may warrant further investigation or the inclusion of additional data points for a more accurate prediction.

The ARIMA model's forecast, especially when indicating a significant change, should be interpreted with caution and considered alongside other market analyses and expert opinions.

Limitations.

The analysis acknowledges potential data inaccuracies and the exclusion of private listings. Future research could explore comparative market analyses or the impact of specific home features on prices and sale times.



Implications

For Buyers:

- Price Sensitivity: Buyers can gauge how much house they can afford in terms of size and features. For example, the scatter plot showing price versus square footage can help them understand the premium paid for additional space.
- Market Timing: Understanding the DoM trends may help buyers determine the best time to purchase, potentially finding better deals in periods where properties tend to stay longer on the market.

For Sellers:

- Pricing Strategies: Sellers can use the price distribution data to competitively price their properties based on size, features, and current market trends to avoid extended DoM.
- Market Preparation: Insights from DoM analysis could inform sellers when to list their properties and how to stage them, considering factors that might affect buyer interest and sale speed. ### For Real Estate Professionals: -Listing Advice: Agents can provide more accurate advice to clients regarding listing prices and expected time to sell, using historical data and predictive models developed from the analysis. -Targeted Marketing: Understanding which types of properties sell faster or slower can help agents tailor their marketing strategies to target the right buyers.

For Investors:

- Investment Decisions: By recognizing trends in DoM and price fluctuations, investors can make more informed decisions about when to buy or sell properties for maximum return on investment.
- Portfolio Diversification: Data on market segmentation can help investors diversify their real estate portfolios to minimize risk and capitalize on different market dynamics. ### For Policymakers and Urban Planners:
- Housing Policy: Insights into how long properties remain on the market and their price points can inform housing policies, affordability programs, and initiatives aimed at market stabilization.
- Urban Development: Understanding the characteristics of properties that are in demand can guide urban development projects and zoning regulations.

For Financial Institutions:

- Risk Assessment: Banks and lenders can use DoM as a metric for assessing the liquidity and risk associated with mortgage lending in different market segments.
- Product Development: Financial products such as mortgages and insurance can be tailored based on the insights from market behavior, offering better terms for properties with quicker sales times. ### Broader Economic Implications:
- Market Health: Extended DoM could indicate a slower market, which might correlate with broader economic conditions, affecting decisions on interest rates and economic interventions.
- Consumer Confidence: Fluctuations in property prices and DoM can reflect consumer confidence and purchasing power, which are important indicators for economic health.

The analytical exploration of the Omaha real estate market using R not only shines a light on the current market conditions but also sets the stage for predictive analytics to forecast future market trends. This provides a strategic advantage to all market participants by enabling data-driven decision-making.

conclusion

In conclusion, the exploration of the Omaha real estate market through various data analysis techniques has provided us with a multifaceted view of the dynamics at play. From understanding the distribution of property prices to examining the Days on Market (DoM) for listings, the insights gained paint a detailed picture of the current state of the market. The statistical analyses, underpinned by robust methods such as ARIMA for time series forecasting and clustering for market segmentation, have revealed patterns and trends that are invaluable for stakeholders. For example, the ARIMA(1,2,1) model's forecast, while suggesting an upcoming change in market conditions, emphasizes the need for vigilance and adaptability in strategy formulation.

The visualizations have not only aided in the comprehension of complex data but have also highlighted the importance of presenting data in an accessible and interpretable format. Whether it be through the clarity of a box plot or the immediate impact of a scatter plot, these tools have facilitated a deeper understanding of market behaviors.As we move forward, it's evident that data-driven decision-making is more than a strategic advantage—it's a necessity in the ever-evolving real estate landscape. The analytics conducted here should serve as a foundation for ongoing observation and analysis, as the market will undoubtedly continue to shift in response to both local and broader economic forces.

Moreover, the implications of this study extend beyond individual decision-making; they have the potential to inform policy at the municipal level, shape investment strategies, and guide homebuyers and sellers toward more informed and optimal choices.

In summary, the data-driven journey through Omaha's real estate market underscores the power of analytics in transforming raw data into actionable knowledge, empowering stakeholders to navigate the market with confidence and foresight.

References

- Federal Reserve Bank of St. Louis. (n.d.). Search results for 'house'. Retrieved February 2023, from https://fred.stlouisfed.org/searchresults/?st=house
- R Core Team. (2023). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
- Wickham, H., & Bryan, J. (2023). readxl: Read Excel Files. R package version 1.4.0. URL https://CRAN.R-project.org/package=readxl.
- Wickham, H., François, R., Henry, L., & Müller, K. (2023). dplyr: A Grammar of Data Manipulation. R package version 1.0.8. URL https://CRAN.R-project.org/package=dplyr.
- Wickham, H. (2023). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. URL https://ggplot2.tidyverse.org.
- Wickham, H., & Grolemund, G. (2023). R for Data Science: Import, Tidy, Transform, Visualize, and Model Data. O'Reilly Media, Inc. URL https://r4ds.had.co.nz.
- Xie, Y., Allaire, J. J., & Grolemund, G. (2023). R Markdown: The Definitive Guide. Chapman and Hall/CRC. URL https://bookdown.org/yihui/rmarkdown.
- Kuhn, M., & Johnson, K. (2023). Applied Predictive Modeling. Springer. URL https://appliedpredictivemodeling.com.
- Zillow Group, Inc. (2023). Zillow Research - Data and Reports on Housing Market, Home Values, and Economic Insights. Available at https://www.zillow.com/research/.
- Redfin Corporation. (2023). Redfin Data Center - Real Estate Data, Market Reports, and Analysis. Available at https://www.redfin.com/news/data-center.
- Move, Inc. (2023). Realtor.com Market Data - Housing Market Insights and Statistics. Available at https://www.realtor.com/research/data/.
- National Association of Realtors. (2023). NAR Research - Reports and Insights on Real Estate Markets. Available at https://www.nar.realtor/research-and-statistics.