Exploratory Data Analysis of U.S. Housing Trends

Thesis Project
Merve Pakcan

TABLE OF CONTENTS

INTRODUCTION

Rising housing costs and urban challenges force people to reconsider where and how they live.

PROBLEM

GOAL

Apply data analytics to uncover housing trends and guide sustainable urban choices.

1. **What drives homes to sell above their expected value in competitive urban markets?**
2. **Do cities with stronger housing demand show distinct pricing patterns or clusters?**
3. **How do housing market dynamics influence price behavior across different regions?**

RESEARCH QUESTIONS

HYPOTHESIS

Will be explained in next slide

SUSTAINABLE DEVELOPMENT GOALS

3 GOOD HEALTH AND WELL-BEING

9 INDUSTRY, INNOVATION AND INFRASTRUCTURE

11 SUSTAINABLE CITIES AND COMMUNITIES

DATA DESCRIPTION

10500 ROWS & 12 COLUMNS

The data is obtained from Zillow Research, a public platform that provides comprehensive, **regularly updated datasets** on the U.S. housing market.

Zillow

Covers the period from **March 2018 to February 2025.**

Merging Datasets(7)

- Median_sale_price.csv
- Percent_of_homes_sold_above _list.csv
- Zori_median_rent.csv
- Market_heat.csv
- Affordability_years_to_save .csv
- Mean_sale_to_list_ratio.csv
- New_con_median_sale_price.c sv

Challenges in Data Gathering

Bucharest data- Not received because of data privacy from real estate

Variables

- RegionID
- RegionName
- RegionType
- StateName
- Date
- Median Sale Price
- Pct Sold Above List
- Median Rent
- Market Heat Index
- YearsToSave
- SaleToListRatio
- NewCon Median Sale Price

HYPOTHESIS TESTING

GOAL

HYPOTHESIS 1

Cities with stronger market signals(Sale-to-List Ratio and New Construction Sale Price) tend to cluster around higher sale prices.

HYPOTHESIS 2

Cities with a higher Market Heat Index are more likely to sell homes above the list price.

Key action 1

Calculate correlations between market indicators and sale prices.

Key action 2

Apply clustering to identify high-price market groups.

Key action 1

Analyze the relationship between Market Heat Index and price performance.

Key action 2

Visualize pricing behavior across different Market Heat levels.

DATA CLEANING & PREPROCESSING

```
Shape: (10500, 12)

Column Types:
RegionID                   int64
RegionName                 object
RegionType                 object
StateName                  object
Date                datetime64[ns]
MedianSalePrice            float64
PctSoldAboveList           float64
MedianRent                 float64
MarketHeatIndex            float64
YearsToSave                float64
SaleToListRatio            float64
NewConMedianSalePrice      float64
dtype: object

Missing values per column:
NewConMedianSalePrice      1718
PctSoldAboveList            301
SaleToListRatio            291
MedianRent                 191
StateName                   84
dtype: int64
```
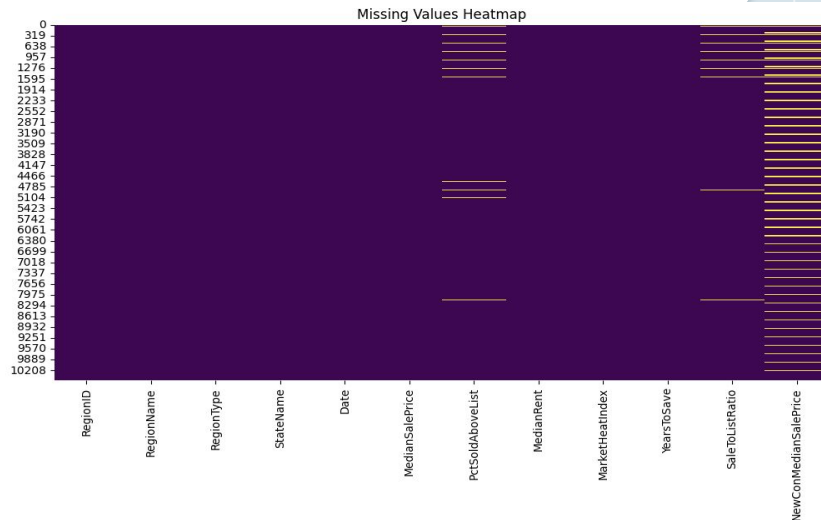
Data Types
Not dropped missing in Newcon
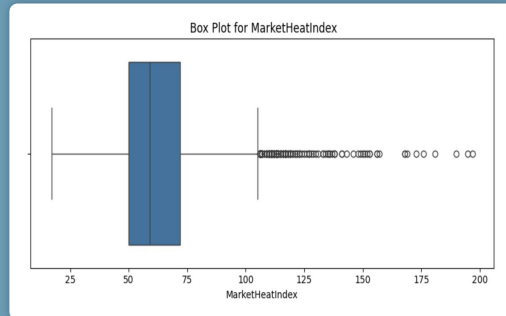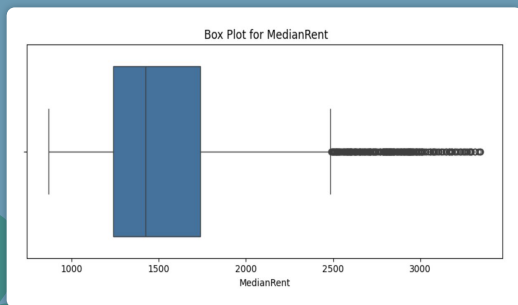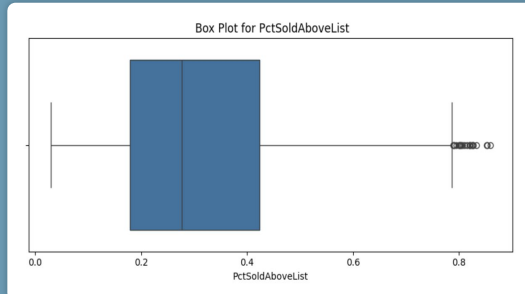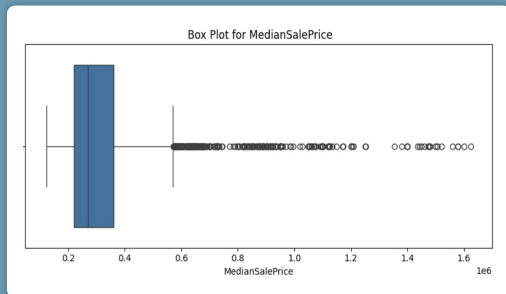
Missing Values Heatmap

```
Rows before drop: 10500
Rows after drop (excluding only key variables): 9893

Duplicate Rows: 0
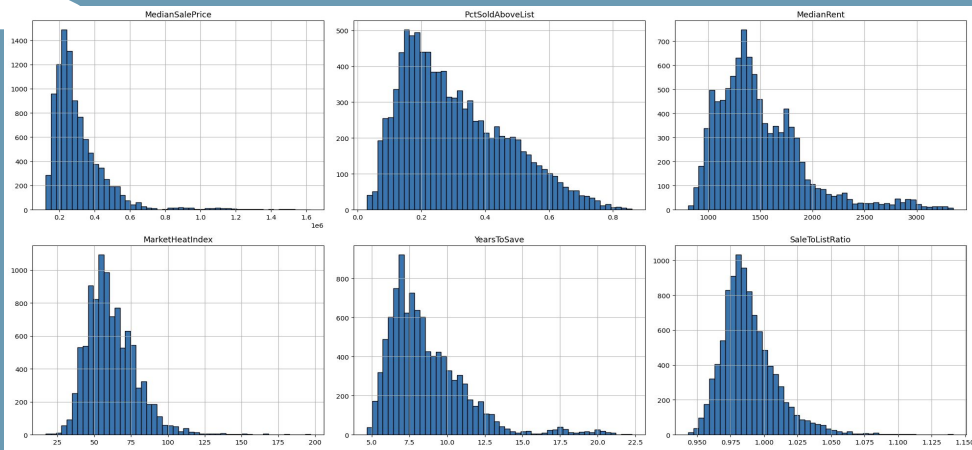```

**Missing Values**
**No duplicate**

## 10500 ROWS —> 9893 ROWS

# Outliers


Box Plot for MedianSalePrice


Box Plot for PctSoldAboveList


Box Plot for MedianRent


Box Plot for MarketHeatIndex

**Not dropped outliers, it is meaningful**

# EXPLORATORY DATA ANALYSIS

**Histograms**



**Descriptive Statistics**

```
              RegionID   Date MedianSalePrice \
count   9893.000000   9893   9.893000e+03
mean   406601.992621 2021-10-16 19:09:28.038006528   3.096715e+05
min   394312.000000   2018-03-31 00:00:00   1.219750e+05
25%   394531.000000   2020-01-31 00:00:00   2.131770e+05
50%   394792.000000   2021-10-31 00:00:00   2.650000e+05
75%   395005.000000   2023-07-31 00:00:00   3.599000e+05
max   753912.000000   2025-02-28 00:00:00   1.625000e+06
std   64112.991749   NaN   1.627796e+05

         PctSoldAboveList   MedianRent MarketHeatIndex YearsToSave \
count   9893.000000   9893.000000   9893.000000 9893.000000
mean   0.306217   1500.371539   61.627211   8.631004
min   0.029102   806.913837   17.000000   4.670567
25%   0.175565   1208.192870   50.000000   6.822396
50%   0.271981   1402.893622   59.000000   7.959976
75%   0.416761   1715.942181   71.000000   9.836859
max   0.859512   3347.407630   197.000000   22.269278
std   0.163279   429.354211   17.170322   2.625767

         SaleToListRatio NewConMedianSalePrice
count   9893.000000   8.481000e+03
mean   0.988815   4.059965e+05
min   0.943317   1.620000e+05
25%   0.975242   3.062750e+05
50%   0.985311   3.650000e+05
75%   0.998708   4.499900e+05
max   1.140964   2.510000e+06
std   0.020732   1.759764e+05
```
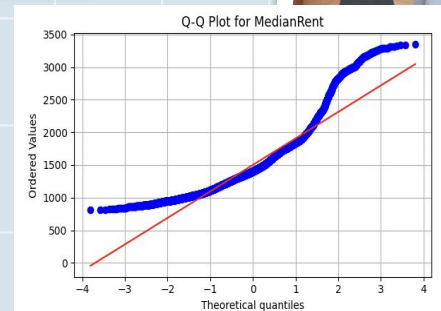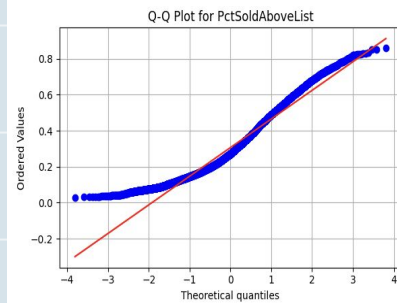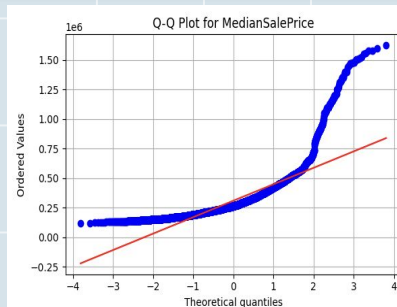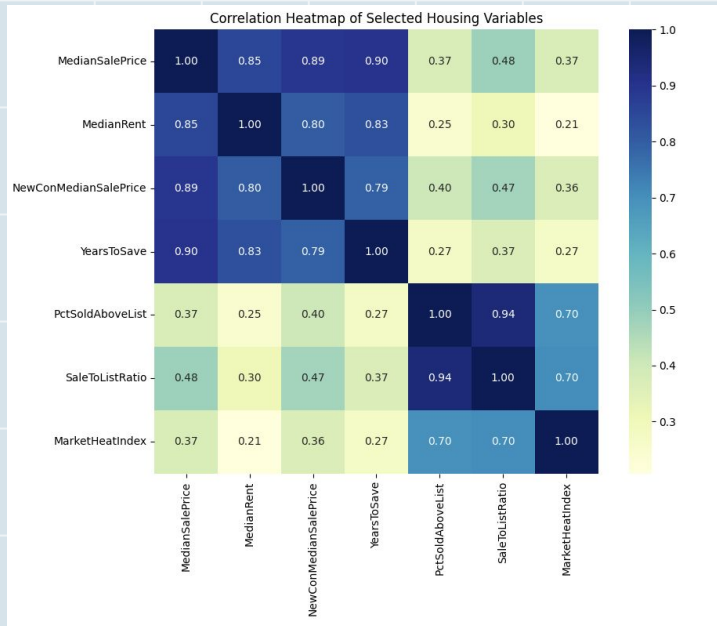
# Correlation Matrix & Q-Q Plot

Correlation Heatmap of Selected Housing Variables

Q-Q Plot for MedianSalePrice

Q-Q Plot for MedianRent

Q-Q Plot for PctSoldAboveList

Q-Q Plot for MarketHeatIndex
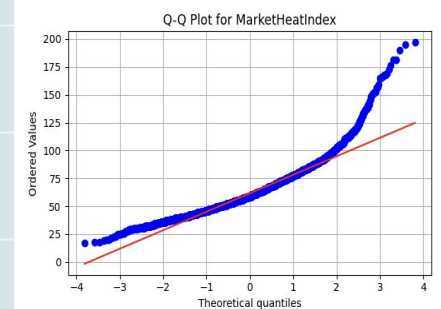
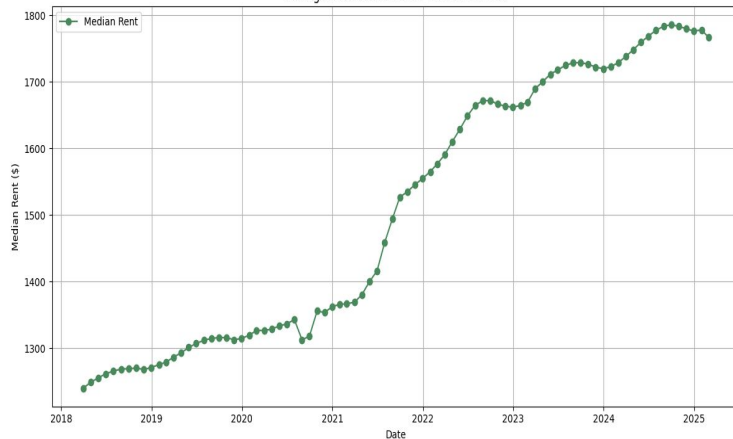**Most variables deviate from normality, as expected in housing data**

# TIME SERIES & PATTERN EXPLORATION

Average Median Sale Price Over Time in the U.S.

Average Median Rent Over Time in the U.S.
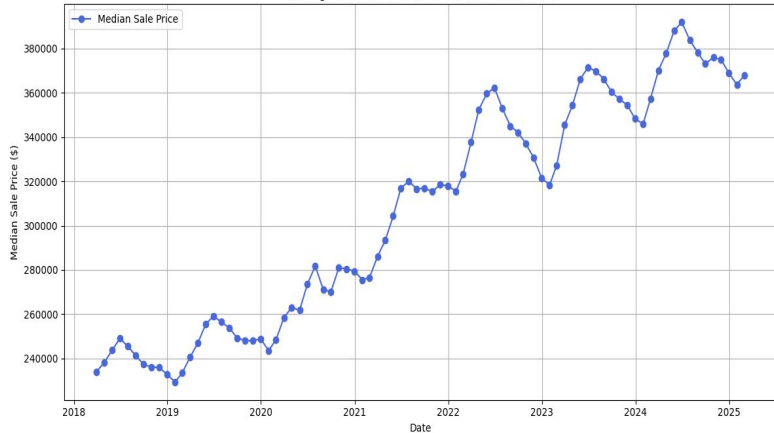
**Average median sales over time**

**Average median rent over time**

Sale prices and rents show a steady upward trend, with sharper increases during 2021–2022.
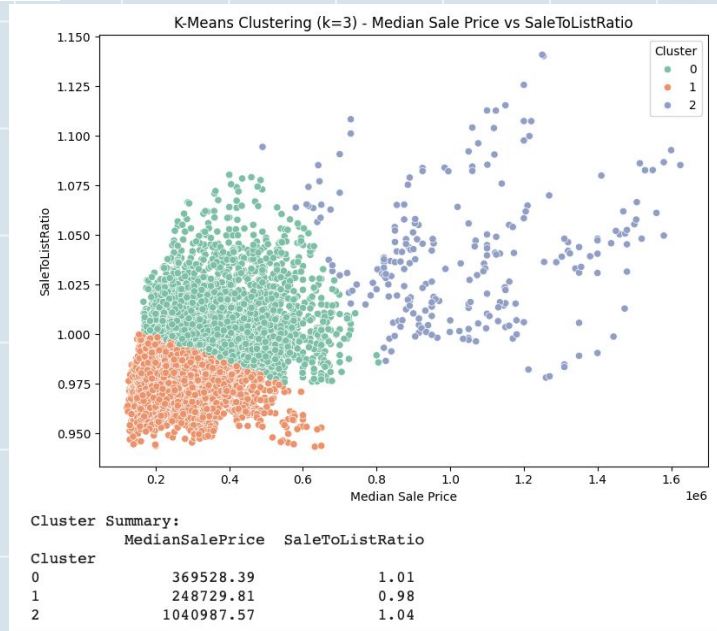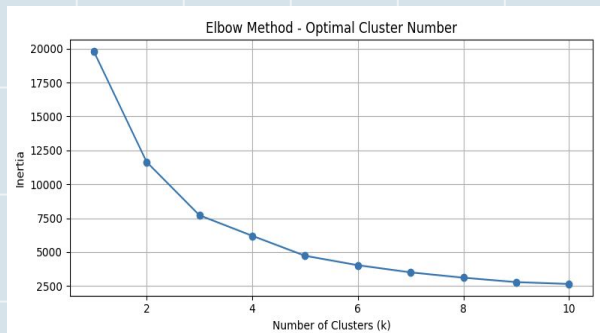
# HYPOTHESIS 1

**Cities with stronger market signals tend to cluster around higher sale prices.**

```
Correlation between Median Sale Price and SaleToListRatio:
Pearson r = 0.48 | p-value = 0.0000
```

Mean Sale to List Ratio:

The average ratio between the final sale price and the original list price.

Elbow Method - Optimal Cluster Number

K-Means Clustering (k=3) - Median Sale Price vs SaleToListRatio

```
Cluster Summary:
         MedianSalePrice  SaleToListRatio
Cluster
0              369528.39             1.01
1              248729.81             0.98
2             1040987.57             1.04
```

- **Cluster 0:** High-priced, sell near list (≈1.00)
- **Cluster 1:** Mid-priced, sell above list (>1.00)
- **Cluster 2:** Low-priced, sell below list

New Construction Median Sale Price:
The median price of newly constructed homes that were sold in a given period.

Correlation between Median Sale Price and NewConMedianSalePrice:
Pearson r = 0.89 | p-value = 0.0000

Elbow Method - Optimal Cluster Number

K-Means Clustering (k=3) - Median Sale Price vs NewConMedianSalePrice
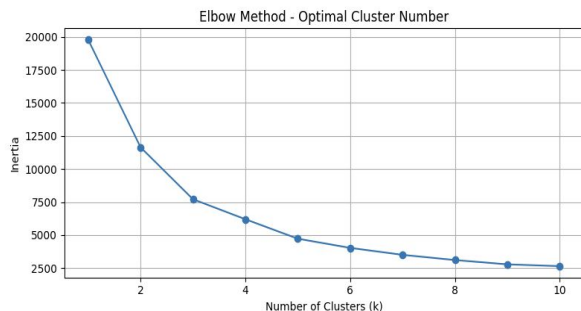
New home prices rise with overall market prices, and cities naturally group into low, mid, and high-priced clusters.
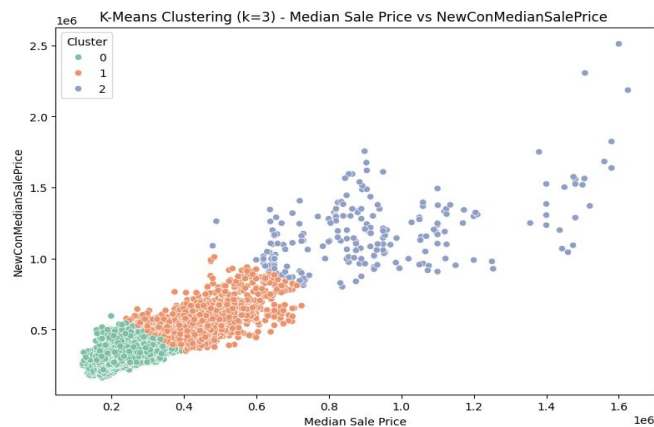
Cluster Summary:

| Cluster	MedianSalePrice	NewConMedianSalePrice
0	241471.49	330942.49
1	436789.65	528787.91
2	925075.67	1183145.89

HYPOTHESIS 2

Cities with a higher Market Heat Index are more likely to sell homes above the list price.

For Market Heat Index and Percent Sold Above List:

```
Correlation between MarketHeatIndex and PctSoldAboveList:
Pearson r = 0.70 | p-value = 0.0000
```

More competitive markets tend to push prices above the list

Elbow Method - Optimal Cluster Number

K-Means Clustering (k=3) - MarketHeatIndex vs PctSoldAboveList
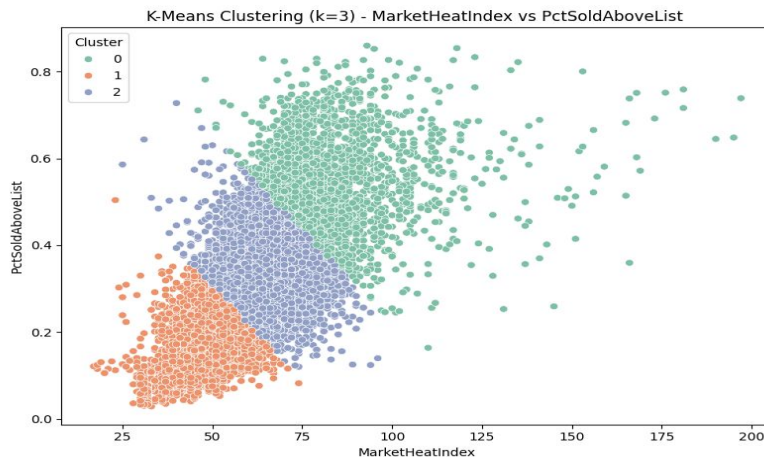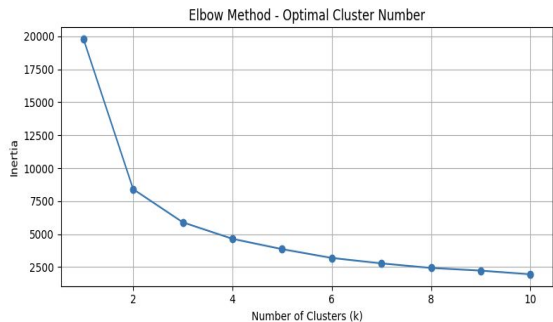
```
Cluster Summary:
           MarketHeatIndex   PctSoldAboveList
Cluster
0                84.56            0.54
1                48.11            0.17
2                64.46            0.34
```

INSIGHTS & CONCLUSION

What insights can we draw from all this?

General Comments

- Strong market signals align with higher housing prices.
- Competitive markets push prices above the list.
- Clear clustering patterns emerge across city price tiers.

EDA Challenges

- Missing data in some variables
- Non-normal distributions

Hypothesis Results

- H1 supported: Cities with stronger market signals cluster at higher prices.
- H2 supported: Market Heat Index positively correlates with homes selling above list.
- Clear clustering validated price segmentation.
- High r-values confirmed strong relationships between variables.
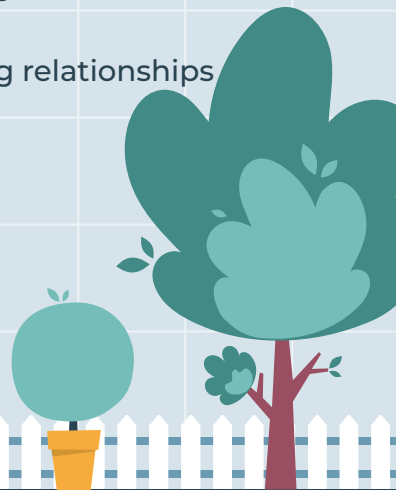
THANKS FOR LISTENING!

Do you have any questions?