TSA Data Science and Analytics

Gulf Coast, TX

January 2025

Team Member ID: 2015075

Team #2015-2

Table of Contents

Introduction

The trajectory of human civilization has always been marked by inflection points that profoundly reshape societal structures. Technological innovations, global challenges, and unexpected disruptions have consistently rewritten the norms of our existence, work, and relations. The early 21st century is no exception to this ongoing evolution, with global events accelerating changes that would otherwise take decades to unfold.

The COVID-19 pandemic emerged as one such pivotal moment, dramatically reshaping traditional norms in virtually every aspect of human life. What began as a public health crisis quickly became a catalyst for widespread societal transformation, particularly regarding how we view work and living spaces. Remote work, once considered a niche arrangement, rapidly transitioned from an emergency band-aid during the pandemic era to a viable, long-term strategy for many organizations.

This fundamental shift in workplace dynamics inevitably streamed into other domains, with one of the most glaring effects being felt in the residential real estate sector. Public health mandates that rendered traditional office work inviable spurred reevaluation of geographical priorities and in turn, housing preferences. As technological infrastructure continued to develop and facilitate virtual collaboration, conventional geographic constraints of employment began to truly dissolve for many.

In the context of this paper, the research examines how demographic factors - incomes and population density - affected trends in the housing market prior to and after the pandemic. By analyzing real estate data from 2000 to 2024, we seek to measure and gain understanding of the size of these transformative changes, as well as explore how such lifestyle changes brought about by COVID-19 have potentially reconfigured long-standing patterns in housing markets.

Data Dictionary

Data Dictionary for Housing Price Increase Analysis

Field Name	Data Type	Description	Field Example
Density	Integer	The population density of a zip code area, measured in people per square mile.	1500 (people/sq mi)
Median family income (2018)	Integer	The median family income for a zip code area	100000 (USD)
Percentage_Price_Increase_PrePandemic	Float	The percentage change in housing prices from 1/31/2000 to 2/28/2020	129.7 (%)
Percentage_Price_Increase_PostPandemic	Float	The percentage change in housing prices from 2/28/2020 to 11/30/2024	-8.4 (%)

Purpose

The primary objective of this analysis is to investigate how population density and income levels serve as predictive factors for housing market trends, comparing their predictive effectiveness in the pre-pandemic and post-pandemic periods. This study specifically examines whether COVID-19 fundamentally changed the traditional correlations between these demographic indicators and housing market behavior, and if so, to what extent these relationships have shifted.

By conducting a comparative analysis of these periods, this study seeks to quantify the impacts of the pandemic on the dynamics of the residential real estate market, focusing especially on the effects the widespread recent adoption of remote work policies, due to them being imperative for workplace safety during the peak of the pandemic. The assimilation of such policies decreased the need for on-site work, lessening the need for physical proximity to one's place of work, which could potentially lead to shifts in residential housing preferences.

The findings of this research project will be useful to several stakeholders within the real estate ecosystem. To investors, the analysis will provide critical insight on emerging market opportunities and related risk factors. Housing market analysts will benefit from a better understanding of how societal disruptions can reshape established market patterns. Additionally, prospective home buyers will gain valuable information on how demographic factors may influence property values in different areas under different societal conditions in the coming years.

Furthermore, this research contributes to the holistic understanding of how major societal shifts can restructure the relationship between demographic factors and housing market results. These findings are particularly relevant as the workplace continues to evolve toward hybrid and remote models, potentially resulting in long-term changes in residential preferences and hence property valuations.

Methods

Data Sources and Collection

Three open-source files were used in this study to analyze the relationship between real estate price fluctuations, household income levels, and population density:

1. Zillow Home Value Index (ZHVI): This index offered monthly measurements of typical home values between the 35th and 65th percentile range from January 2000 to November 2024.
2. Missouri Census Data Center (MCDC): Median household income data from 2018, allowing the categorization of areas into higher (> $100,000) or lower income (≤ $100,000).
3. Forefront.us: Population density data, enabling the grouping of areas as higher density (>1,000 persons/sq mi) or lower density (≤1,000 persons/sq mi) following the NCRCRD urban definition.

Data Cleansing

Processing of data was conducted using the Python Pandas library. The workflow included:

1. Initial data integration achieved by joining datasets on a common identifier zip code
2. Data cleaning procedures included:
   a. Removal of all rows that included null or 0 values in columns crucial to the analysis
   b. Removal of irrelevant columns (e.g., longitude, latitude, water area, etc.)

Data Integration

Data preparation steps completed before any analysis was conducted included:

1. Percent_Price_Increase columns were created for both pre and post pandemic
2. Segmenting the data into 4 different groups based on their Income Level and Population Density
3. Calculating and capping each group's max values at their 95^{th} percentile to help account for extreme outliers that were distorting the color visualization in the scatter plots

Statistical Analysis

Analysis compared two periods: 1/31/2000-2/29/2020 ("pre-pandemic") and 2/29/2020-11/30/2024 ("post pandemic") relying on Python's SciPy and Excels charting feature. The analysis consisted of:

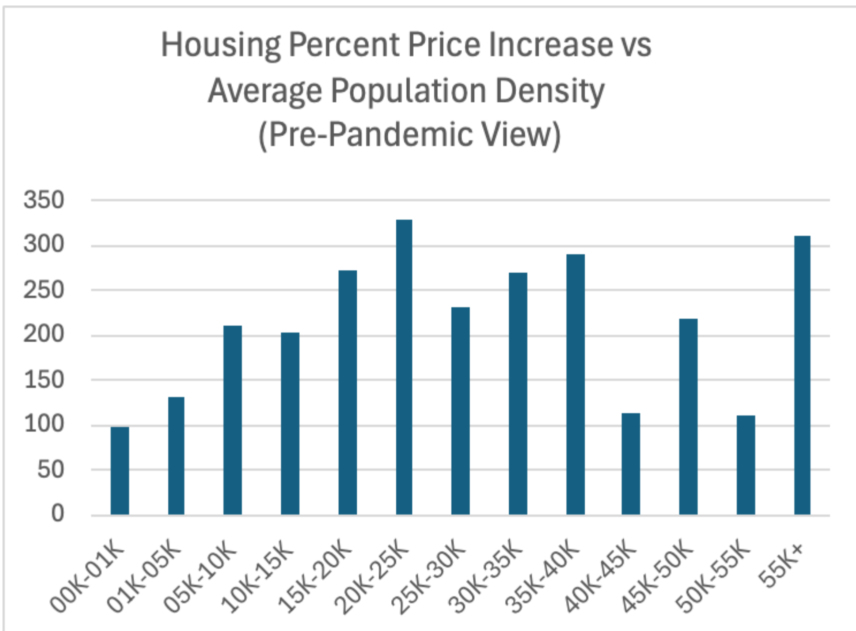
1. Individual Influence Analysis: The analysis involved examining the relationship between each variable (population density and median household income) and percentage price increase independently, using bar graphs. In these graphs, the x-axis represented one variable, while the y-axis showed the percentage price increase. The goal was to identify trends in pre-pandemic data and then determine whether those trends intensified, weakened, or reversed after the pandemic.
2. Combined Influence Analysis: The combined analysis involved computing the mean percentage price increases for each categorical group and using Welch's T-tests to determine whether differences between regions with varying characteristics were statistically significant. Data visualization was conducted using Python's Matplotlib and Microsoft Excel, which generated multi-panel scatter plots for each categorical group. These plots displayed population density on the x-axis, median family income on the y-axis, and the percentage price increase represented through color gradients using the 'coolwarm' colormap, with an alpha transparency of 0.75 to enhance visibility of overlapping points. Additionally, comparative visualizations were created to highlight temporal trends before and after 2020, as well as shifting relationships between population density, household income, and price appreciation.

Results

This study examined the relationship between an area's population density and income level on the changing pricing for housing ever since the pandemic took full effect (operationally defined as 2020 in accordance with the year the World Health Organization (WHO) declared the COVID-19 virus a pandemic. The study interpreted how the following variables affected the housing market pre COVID-19 and post COVID-19 and then analyzed if their relationship was significantly different.

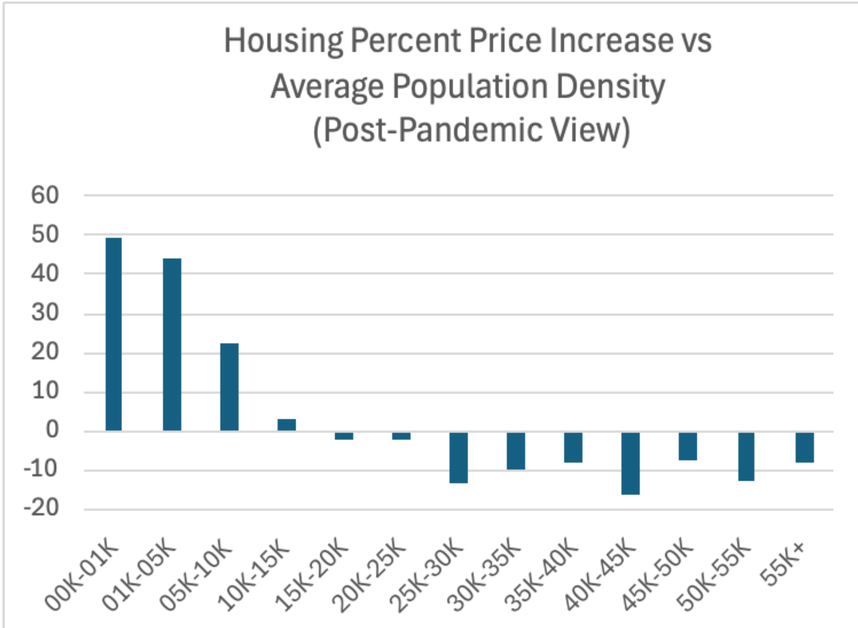
Population Density (defined as the number of people per square mile)

Pre-Pandemic Analysis



The housing price data from 2000-2020 reveals a positive correlation between population density and real estate price appreciation across zip codes. Starting with the lowest density areas at around 100% price increase in this time period, the graph illustrates an upward trajectory as density increases, with price appreciation reaching more than 300% in the highest density areas. This major difference in price growth between low and high-density areas may reflect the increasing trend of urbanization during these two decades. With more people and most importantly businesses gravitating toward city centers, urban housing became more in demand, thereby driving up property values in areas of high concentration. In general, higher density corresponds with greater price increases, although this relationship is somewhat variable and has local peaks across the density spectrum. This variation, especially near the highest levels of concentration, is particularly important to note in the case of this study, as it alludes to certain drawbacks of extreme urban density even before the pandemic.

Post-Pandemic Analysis



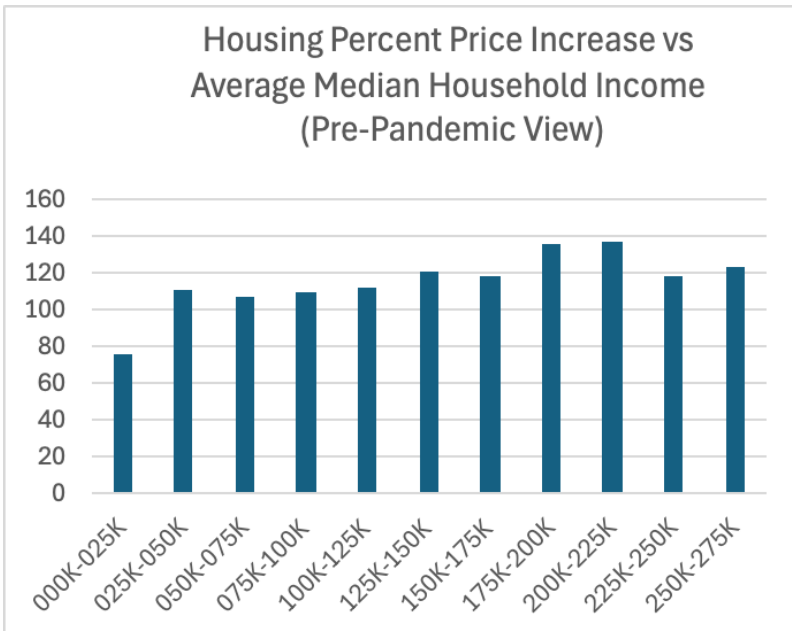
The housing price data from 2020-2024 reveals a strong negative correlation between population density and real estate price appreciation across zip codes. Starting at the lowest density areas with approximately 50% price increases, the graph shows a steep downward trajectory as density increases, with price valuation declining to negative values in the highest density areas. Such a strong divergence in price growth between low and high-density areas might reflect a significant shift in housing preferences during this period. With the rise of remote work and correlative changing lifestyle preferences, housing demand shifted notably toward less densely populated areas, leading to an inflation of property values in less dense regions. Though this relationship between property value and area concentration levels experiences some more variation within the moderate to extreme density range, in general, lower density corresponds with greater price increases, showing a clear inverse relationship compared to historical trends. The latter point in particular carries special significance in the case of the current study, as it suggests that areas of higher levels of concentration became substantially less attractive to buyers, leading to valuation drops of up to 20% since the pandemic took the world by storm.

Pre Versus Post Pandemic Analysis

The stark contrast between these two graphs reveals a fundamental shift in housing market dynamics before and after 2020. Between 2000-2020, housing prices demonstrated a strong positive correlation with population density, reflecting traditional preferences for urban living and the advantages of proximity to business centers. However, the data from 2020 through 2024 appears to be the complete inverse of this trend, with lower-density areas gaining significant price appreciation and high-density urban areas experiencing significant price decline. This dramatic shift can be largely attributed to the COVID-19 pandemic's catalytic effect on remote work adoption. While the pandemic initially made companies and organizations embrace remote work as an emergency measure, it actually catalyzed the development of strong digital infrastructure, improved remote collaboration tools, and new organizational strategies that have outlived the immediate public health emergency. As a result, many workers are no longer bound by the traditional need to live near urban centers of employment. Instead, the data shows they are actively choosing less densely populated areas with more space and offer different potentially preferable lifestyle options. This transformation appears to be more than a mere temporary pandemic response as the 2020-2024 trend continues past the acute phase of the pandemic, indicating that the widespread adoption of remote work and the tools that enable it, have fundamentally changed the relationship between population density and housing demand.

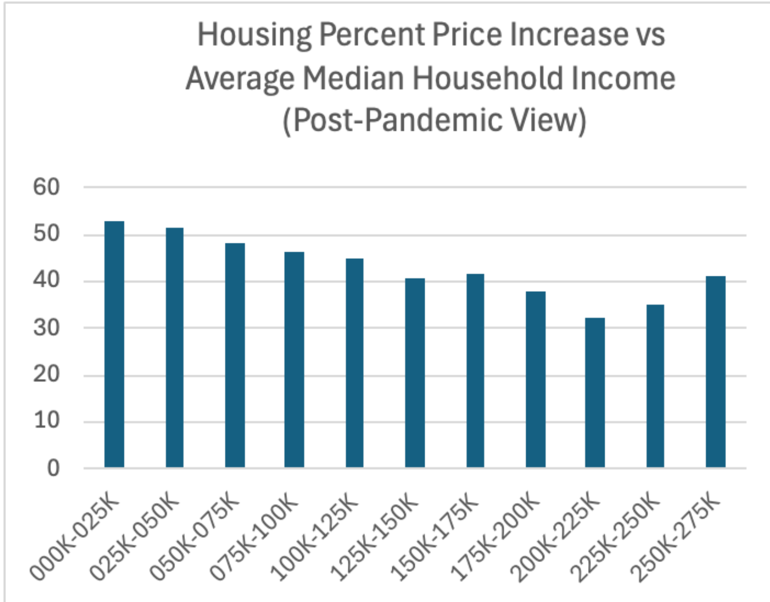
Income Level (defined as the Median Household Income of the area)

Pre-Pandemic Analysis



The housing price data from the years 2000-2020 reflects a modest positive correlation between median household income and the appreciation of real estate prices across various income brackets. Starting with the lowest income areas, with just about a 75% increase in price, the graph illustrates a gradual upward trajectory as median income increases, reaching increases near 135% in the higher income brackets. This moderate difference in price growth between low and high-income areas suggests that while wealth does play some role in housing price variation, the relationship is not as strong as one might otherwise expect. With price increases ranging from 75% to 135% across all income levels, the data illustrates relative stability in appreciation rates, with only a slight edge towards higher-income areas.

Post-Pandemic Analysis



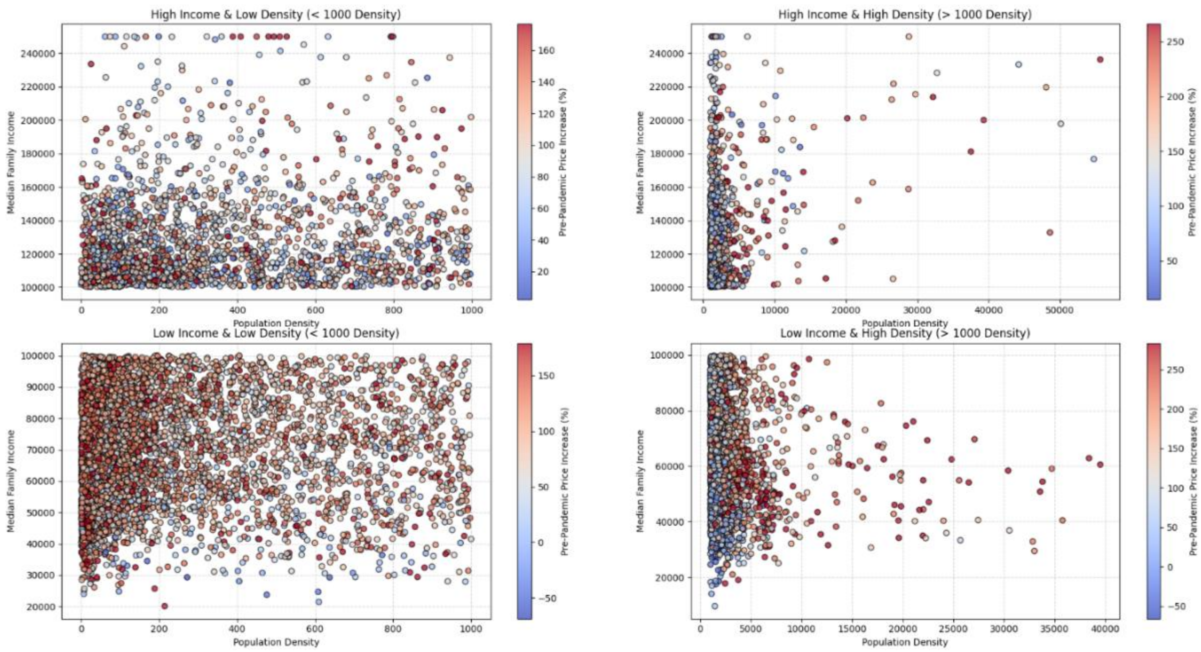
The housing price data from 2020-2024 reflects a notable negative correlation between median household income and real estate price appreciation across income brackets. Starting with the lowest income areas at around 50% price increases, the graph trends downward as median income rises, with price appreciation falling to approximately 30-35% in the uppermost income ranges. This inverse relationship between income levels and price growth presents an interesting shift from historical patterns, suggesting a possible democratization of housing market gains during this period. Since lower-income areas experienced the highest appreciation rates, the data indicates a possible correction or rebalancing in the housing market. In general, lower median household income corresponds with greater price increases, with the relationship showing a consistent negative trend across the income spectrum. Of relevance to the present study, this finding would suggest that traditionally less affluent areas may have become more attractive to buyers, likely due to factors of affordability combined with changing patterns of employment in the wake of the pandemic.

Pre Versus Post Pandemic Analysis

The comparison between housing price trends in 2000-2020 and 2020-2024 reveals an interesting reversal in the relationship between median household income and real estate appreciation during the inflection point, year 2020. While data up until 2020 demonstrates a moderate positive correlation, and the post-2020 period shows a modest negative correlation, this inversion suggests that the traditional relationship between wealth and residential property value growth were challenged during the pandemic era. Part of this shift can be attributed to the pandemic-driven rise in remote work, which freed many employees from the necessity of living in city centers next to their place of employment, allowing them to explore other, potentially lower cost options. With workers able to maintain their income regardless of their location, housing demand shifted to areas atypical of major employment hubs, contributing to significant price growth in previously more affordable markets. That said, both modest trends allude to other variables also play a key role in determining the housing market's appreciation trends.

Population Density & Income Level Combined Impact

Price Increase Trends Across Different Income and Population Density Categories (2000-2020)



Basic Descriptive Statistics

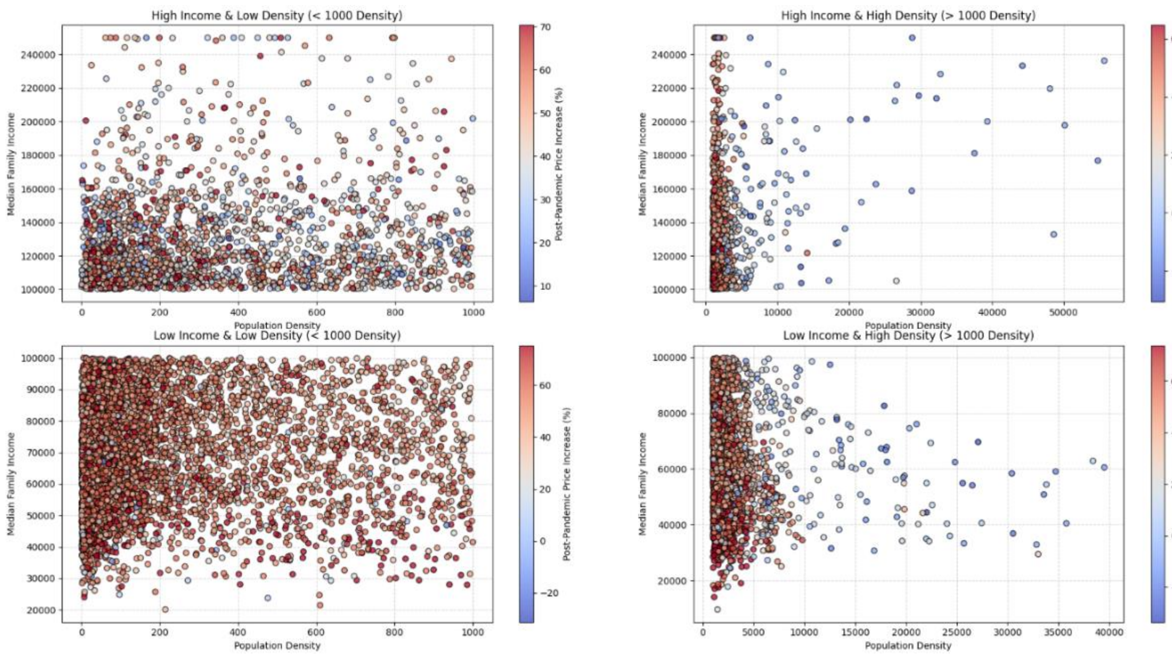
1. Mean Results Analysis (from 1/31/2000-2/28/2020)
   a. High Income & High Density: 150%
   b. High Income & Low Density: 102%
   c. Low Income & Low Density: 99%
   d. Low Income & High Density: 140%

   The category with the highest average percentage increase is High Income & High Density.
2. t-tests: To prove the difference between this category and the others were statistically significant multiple Welch's T-tests were run comparing High Income & High Density with the other categories with the following results:
   a. t-test with High Income & Low Density: t-stat = 21.31, p-value = 0.00000
   b. t-test with Low Income & Low Density: t-stat = 24.01, p-value = 0.00000
   c. t-test with Low Income & High Density: t-stat = 3.87, p-value = 0.00011

   T-Test Analysis: The very low p-values suggest that the differences in the data are highly unlikely to be due to chance. The high absolute t-stats highlight a clear and statistically significant gap in percentage price increases between High Income & High Density and the other categories, confirming that the disparities aren’t random but rather reflect meaningful shifts in market behavior or economic dynamics.

Additionally, the noticeably higher absolute t-stats in both low-density categories suggest that Population Density may have played a stronger role in driving the observed changes.

<u>Price Increase Trends Across Different Income and Population Density Categories (2020-2024)</u>



1. Mean Results Analysis (from 2020-2024)
   a. High Income & High Density: 34%
   b. High- Income & Low Density: 47%
   c. Low Income & Low Density: 50%
   d. Low Income & High Density: 44%

   The category with the lowest observed average percentage price increase is High Income & High Density. This is starkly contrasted with the 2000-2020 study where it came out as the highest increasing category over all the others by over 10 percent points.

2. T-Tests: T-Test Results comparing High Income & High Density with other groups:
   a. T-Test with High Income & Low Density: t-stat = 17.65, p-value = 0.00000
   b. T-Test with Low Income & Low Density: t-stat = 23.37, p-value = 0.00000
   c. T-Test with Low Income & High Density: t-stat = 12.55, p-value = 0.00000

   T-Test Analysis: The extremely low p-values indicate that the observed differences in the data are highly unlikely to be due to chance. The high absolute t-stat values further confirm a significant and substantial difference in percentage price increases between High Income & High Density and other categories, supporting the conclusion that the disparities reflect meaningful variations in market behavior or underlying economic factors, rather than random variation.

Conclusion

The findings of the research reveal a dramatic shift in the relationship between demographic factors and housing price appreciation before and after the COVID-19 pandemic. The data demonstrates that once stable, traditional relationships among population density, income levels, and housing price appreciation have not only weakened but, in many cases, have fundamentally inverted.

Prior to 2020, high-density, high-income areas generally outperformed other market segments, showing an average price appreciation of 149% compared to appreciation rates ranging from 97-138% in other categories. This trend aligned with the general notion about urban centers driving economic growth and property values. However, post-2020 data suggested a radically different story is told, with high-density, high-income areas showing the lowest appreciation rate (35%) compared to other categories (45-51%).

The most jarring finding is the inversion of population density correlation with housing prices. From 2000 to 2020, density showed a strong, positive correlation with price appreciation, and from 2020 to 2024, it exhibited an equally strong negative correlation. This reversal would suggest that the pandemic-induced shift to remote work has fundamentally altered the traditional premium placed on urban proximity. The persistence of this trend well beyond the acute phase of the pandemic indicates that this dynamic is not merely a temporary response to health concerns but rather represents a deeper structural change in housing market dynamics.

Income level correlations, while also showing a reversal, reflected a relatively modest shift. This result suggests that while wealth remains a factor in housing market behavior, its influence may be weakening as remote work opportunities enable a spread of access to high-paying jobs to more geographic areas previously seen as less desirable. The stronger performance of lower-income neighborhoods in the post-2020 period hints at a possible rebalancing of the housing market, though further research would be needed to determine whether this trend reflects a mere temporary rebound or a permanent structural shift.

These findings could have significant implications for real estate investors, urban planners, and policymakers. The diminishing premium for high-density areas may indicate that future urban development should focus on quality-of-life considerations rather than mere proximity to city centers with business districts. Moreover, the stronger performance of traditionally less expensive markets points to potential opportunities for economic development in previously overlooked areas while also raising concerns about displacement and affordability in these emerging markets.

The research ultimately supports the hypothesis that high-density, high-income areas experienced dramatically lower price appreciation in the post-pandemic period compared with their pre-pandemic trajectory. While COVID-19 initially catalyzed shifts in work and lifestyle patterns, these changes have evolved into durable preferences extending well beyond the temporary pandemic-related concerns. The persistence and stability of these new housing market dynamics, even as pandemic restrictions have lifted, strongly suggests that this shift represents a lasting structural change rather than a temporary market disruption.

Next Steps

The findings of this study regarding the shifting relationships between population density, income levels, and housing values during the pandemic era present several possible avenues for further research studies and practical applications.

First, conducting a longitudinal study examining these correlations across a longer post-pandemic time frame would be beneficial in determining if the noted changes represent a permanent shift or merely a temporary market adjustment. It is essential for such research to monitor these metrics across several subsequent years to identify any possible signs back to pre-pandemic norms, or of a stable "new" normal. Moreover, particular attention must be paid to emerging corporate policies mandating a return to the office, since the decisions regarding workplace flexibility made by major employers could greatly influence housing preferences, as evidenced by this study, and may modify some of the observed trends.

Additional demographic factors should be looked at to further understand the shifting dynamics of the housing market. Specifically, future research should account for:

1. Age distribution within zip codes to understand how generational differences influence housing choices
2. Educational attainment levels to determine if there is correlation with remote work adoption and housing preferences
3. Industry concentration in different areas to determine how certain industries facilitate or diminish the effects of remote work on demand within the residential market

A comparative analysis across countries could also offer valuable insights into whether the observed trends are unique to the United States or indicative of a global shift in housing demand. This global outlook would be particularly helpful in understanding how various policy responses to the pandemic and differing levels of remote work adoption have influenced housing markets around the world.

Future studies could also investigate the influence of technological infrastructure on residential preferences. Specifically, studies should examine how access to high-speed internet and other digital resources influenced property values in less dense regions, as these factors become increasingly important for remote workers.

Finally, the development of predictive models incorporating the newly discovered correlations between density, income, and housing prices may help stakeholders to more accurately forecast future market trends. These models must consider both traditional market drivers and emerging factors such as rates of remote work adoption, technological infrastructure development, and the potential impacts of corporate return-to-office policies.

These next steps will expand on the current findings to enable a more complete understanding of how the pandemic has reshaped the relationship between demographic variables and the appreciation of the housing market, offering useful insight to policymakers, investors, and potential home buyers.

TECHNOLOGY STUDENT ASSOCIATION

COVID-19s Impact On Traditional Relationships Between Demographic Factors and Housing Valuation

ID: 2015075, Team #2015-2
Katy, Texas

Introduction

The COVID-19 pandemic served as a catalyst for an unprecedented shift in American society, transforming emergency remote work policies into a permanent restructuring of where and how people live. While housing preferences had historically been tied to job locations, the widespread adoption of remote work dissolved these geographic constraints for many. This report explores how population density and income level interacted to influence the dynamics of the housing market from 2000 through 2024, centering on how changes brought by the pandemic shifted such relations and to what extent.

Purpose

The purpose of this research is to examine whether the onset of remote work due to COVID-19 fundamentally altered the traditional relationships between population density, income levels, and housing market trends. Through comparative analysis of housing data prior to and after the pandemic, particularly among metropolitan areas, this study quantifies changes in residential preferences and property valuations. The findings contribute to an improved understanding of how large-scale workplace transformations reshape housing markets, informing real estate investors, homebuyers, and policymakers.

Methods

Data was compiled from sources including Zilllow, Fourfront US, and the Missouri Census Data Center.

Results

Density – Average price increases showed positive correlation pre-COVID and reversed post-COVID (see figures 1-2).

Figure 1 - Housing Percent Price Increase vs Average Population Density (Pre-Pandemic View)

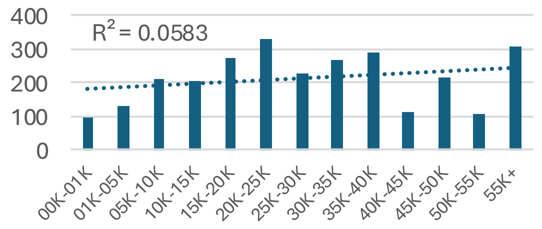
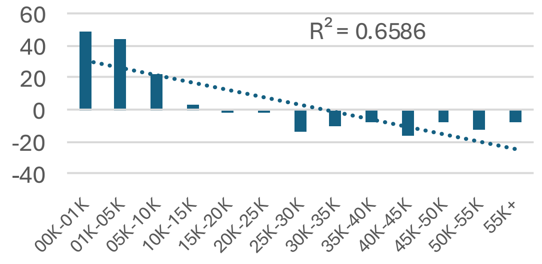


Figure 2 - Housing Percent Price Increase vs Average Population Density (Post-Pandemic View)



Results

Income – Average price increases also showed positive correlation pre-COVID and reversed post-COVID (see figures 3-4).

Figure 3 - Housing Percent Price Increase vs Average Median Household Income (Pre-Pandemic View)

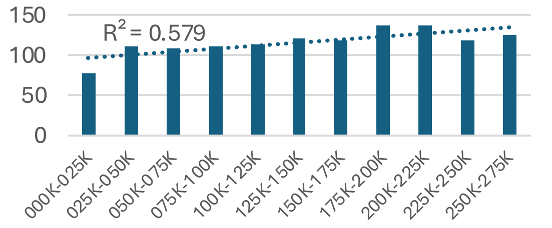
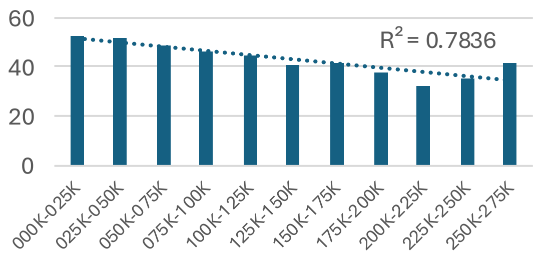


Figure 4 - Housing Percent Price Increase vs Average Median Household Income (Post-Pandemic View)



Combined Income and Density – High income with high density had the greatest increases pre-pandemic whereas this category had the lowest increases post-pandemic.

Conclusion

Analysis reveals a dramatic reversal in housing market dynamics after 2020, with high-density, high-income areas shifting from highest to lowest price appreciation rates. This persistent trend suggests a structural market change driven by remote work, rather than temporary pandemic effects.

Next Steps

- Conduct a longitudinal study to determine if the observed changes in housing market dynamics are permanent or temporary.
- Research demographic trends such as age distribution, educational attainment, and industry concentration.
- Develop models that relate the new correlations between density, income, and housing prices to predict future market trends.

References

FourFront. *(2022, December 9). US Population Density Dataset by Zip, City/State, Coordinates.* https://www.fourfront.us/data/datasets/us-population-density/

Zillow. (2024). *Housing Data - Zillow Research.* Zillow Research. https://www.zillow.com/research/data/

Rice, G. (2018). *ZIP Code Lookup - MCDC.* University of Missouri. https://mcdc.missouri.edu/applications/zipcodes/

Bednarik, Z. (2022, January 20). *Rural-Urban Classification used by NCRCRD.* North Central Regional Center for Rural Development. https://ncrcrd.ag.purdue.edu/2022/01/20/rural-urban-classification-used-by-ncrcrd/

World Health Organization. (2025). *Coronavirus Disease (COVID-19) Pandemic.* World Health Organization. https://www.who.int/europe/emergencies/situations/covid-19

Pandas. (2018). *Python Data Analysis Library.* Pydata.org. https://pandas.pydata.org/

Matplotlib. (2012). *Matplotlib: Python plotting — Matplotlib 3.1.1 documentation.* Matplotlib.org. https://matplotlib.org/

SciPy. (2020). *SciPy.org — SciPy.org.* Scipy.org. https://scipy.org/

Bibliography

FourFront. *(2022, December 9). US Population Density Dataset by Zip, City/State, Coordinates. https://www.fourfront.us/data/datasets/us-population-density/*

Zillow. (2024). *Housing Data - Zillow Research*. Zillow Research. https://www.zillow.com/research/data/

Rice, G. (2018). *ZIP Code Lookup - MCDC*. University of Missouri. https://mcdc.missouri.edu/applications/zipcodes/

Bednarik, Z. (2022, January 20). *Rural-Urban Classification used by NCRCRD*. North Central Regional Center for Rural Development. https://ncrcrd.ag.purdue.edu/2022/01/20/rural-urban-classification-used-by-ncrcrd/

World Health Organization. (2025). *Coronavirus Disease (COVID-19) Pandemic*. World Health Organization. https://www.who.int/europe/emergencies/situations/covid-19

Pandas. (2018). *Python Data Analysis Library*. Pydata.org. https://pandas.pydata.org/

Matplotlib. (2012). *Matplotlib: Python plotting — Matplotlib 3.1.1 documentation*. Matplotlib.org. https://matplotlib.org/

SciPy. (2020). *SciPy.org — SciPy.org*. Scipy.org. https://scipy.org/

Appendix

A. Dataset Sourcing

In this study, three datasets were merged using Python. Below are the descriptions of each dataset, and links provided in the Bibliography section of this paper.

1. **US Population Density Dataset**
   a. Source: FourFront (2022)
   b. Data on population density by ZIP code, city/state, and coordinates
2. **Housing Market Dataset**
   a. Source: Zillow (2024)
   b. Data related to housing prices across various locations in the United States
3. **Median Household Income Dataset**
   a. Source: Rice, G. (2018)
   b. Geographical and demographic information by zipcode

B. Data Set Merging

Below is the Python code that was used to combine the three datasets:

zillow_data.py ● | Merged_Data_Filtered_Removed_Cleaned_Added.csv ● | Merged_Data_TSA_Filtered_Removed_Cleaned.csv | Merged_D

C: > Users > lucas > zillow_data.py > ...

```python
import pandas as pd
import matplotlib.pyplot as plt
from scipy import stats

# Load datasets
prices_df = pd.read_csv('C:\\Users\\lucas\\Desktop\\TSA Project\\2000-2024-Prices-Per-Month.csv')
population_df = pd.read_csv('C:\\Users\\lucas\\Desktop\\TSA Project\\Population-Density-Final-Edit.csv')
income_df = pd.read_csv('C:\\Users\\lucas\\Desktop\\TSA Project\\Median_Household_Income_2018.csv')

# make column names same to merge
prices_df.rename(columns={'Zip': 'ZIP'}, inplace=True)
population_df.rename(columns={'Zip': 'ZIP'}, inplace=True)
income_df.rename(columns={'ZIP Code': 'ZIP'}, inplace=True)

# mwrge datasets
merged_data = pd.merge(prices_df, population_df, on='ZIP', how='inner')
merged_data = pd.merge(merged_data, income_df, on='ZIP', how='inner')

```

C. Data Cleansing

Below is the Python code that was used to cleanse the data to make it analyzable:

```python
# remove rows with blank or 0 values in important columns
columns_to_check = ['density', 'Median family income (2018)', '11/30/2024', '2/29/2020', '1/31/2000', 'ZIP']
merged_data = merged_data[(merged_data[columns_to_check].notnull().all(axis=1)) & (merged_data[columns_to_check] != 0).all(axis=1)]

# convert 'Median family income (2018)' from string to integer
merged_data['Median family income (2018)'] = merged_data['Median family income (2018)'] \
    .replace(r'[^\d]', '', regex=True) \
    .astype(int)

# create percentage increase columns
merged_data['Percentage_Price_Increase_PrePandemic'] = (
    (merged_data['2/29/2020'] - merged_data['1/31/2000']) / merged_data['1/31/2000']
) * 100

merged_data['Percentage_Price_Increase_PostPandemic'] = (
    (merged_data['11/30/2024'] - merged_data['2/29/2020']) / merged_data['2/29/2020']
) * 100

# keep only needed columns
merged_data = merged_data[columns_to_check + ['Percentage_Price_Increase_PrePandemic', 'Percentage_Price_Increase_PostPandemic']]

```

D. Data Preparation Below is the code that created the Percentage_Price_Increase columns for each respective time frame, classified each zip code into one of 4 different categories based on Population Density and Income level, and capped the data at each individual category's 95th percentile to account for exceptions that would later throw off the color visualization of the scatterplots.

```
# define group
high_income_low_density = merged_data[(merged_data['Median family income (2018)'] > 100000) & (merged_data['density'] <= 1000)]
high_income_high_density = merged_data[(merged_data['Median family income (2018)'] > 100000) & (merged_data['density'] > 1000)]
low_income_low_density = merged_data[(merged_data['Median family income (2018)'] <= 100000) & (merged_data['density'] <= 1000)]
low_income_high_density = merged_data[(merged_data['Median family income (2018)'] <= 100000) & (merged_data['density'] > 1000)]

# find 95th percentiles for each group pre
high_income_low_density_95th_pre = high_income_low_density['Percentage_Price_Increase_PrePandemic'].quantile(0.95)
high_income_high_density_95th_pre = high_income_high_density['Percentage_Price_Increase_PrePandemic'].quantile(0.95)
low_income_low_density_95th_pre = low_income_low_density['Percentage_Price_Increase_PrePandemic'].quantile(0.95)
low_income_high_density_95th_pre = low_income_high_density['Percentage_Price_Increase_PrePandemic'].quantile(0.95)

# cap values at their 95th percentiles for better color visualization pre
high_income_low_density['capped_pre'] = high_income_low_density['Percentage_Price_Increase_PrePandemic'].clip(upper=high_income_low_density_95th_pre)
high_income_high_density['capped_pre'] = high_income_high_density['Percentage_Price_Increase_PrePandemic'].clip(upper=high_income_high_density_95th_pre)
low_income_low_density['capped_pre'] = low_income_low_density['Percentage_Price_Increase_PrePandemic'].clip(upper=low_income_low_density_95th_pre)
low_income_high_density['capped_pre'] = low_income_high_density['Percentage_Price_Increase_PrePandemic'].clip(upper=low_income_high_density_95th_pre)
```

E. Data Visualization

This paper reflects two approaches for data visualization:

1. Scatterplots: The following code created a 2x2 grid that allowed for a total of 4 graphs to be created (one for each individual category), facilitating the creation of all the scatter plots for the pre pandemic period.

```
fig, axes = plt.subplots(2, 2, figsize=(14, 12))
```

   a. Below is the code that created the High Income & Low-Density scatterplot for the pre and post pandemic eras:

```
# High Income & Low Density
scatter1 = axes[0, 0].scatter(
    high_income_low_density['density'],
    high_income_low_density['Median family income (2018)'],
    c=high_income_low_density['capped_pre'],
    cmap='coolwarm',
    edgecolor='k',
    alpha=0.75
)
axes[0, 0].set_title('High Income & Low Density (< 1000 Density)')
axes[0, 0].set_xlabel('Population Density')
axes[0, 0].set_ylabel('Median Family Income')
axes[0, 0].grid(True, linestyle='--', alpha=0.5)
fig.colorbar(scatter1, ax=axes[0, 0], label='Pre-Pandemic Price Increase (%)')
```

```
# High Income & Low Density
scatter1 = axes[0, 0].scatter(
    high_income_low_density['density'],
    high_income_low_density['Median family income (2018)'],
    c=high_income_low_density['capped_post'],
    cmap='coolwarm',
    edgecolor='k',
    alpha=0.75
)
axes[0, 0].set_title('High Income & Low Density (< 1000 Density)')
axes[0, 0].set_xlabel('Population Density')
axes[0, 0].set_ylabel('Median Family Income')
axes[0, 0].grid(True, linestyle='--', alpha=0.5)
fig.colorbar(scatter1, ax=axes[0, 0], label='Post-Pandemic Price Increase (%)')
```

   b. Below is the code that created the High Income & High-Density scatterplot for the pre and post pandemic eras:

```python
# High Income & High Density
scatter2 = axes[0, 1].scatter(
    high_income_high_density['density'],
    high_income_high_density['Median family income (2018)'],
    c=high_income_high_density['capped_pre'],
    cmap='coolwarm',
    edgecolor='k',
    alpha=0.75
)
axes[0, 1].set_title('High Income & High Density (> 1000 Density)')
axes[0, 1].set_xlabel('Population Density')
axes[0, 1].set_ylabel('Median Family Income')
axes[0, 1].grid(True, linestyle='--', alpha=0.5)
fig.colorbar(scatter2, ax=axes[0, 1], label='Pre-Pandemic Price Increase (%)')
```

```python
# High Income & High Density
scatter2 = axes[0, 1].scatter(
    high_income_high_density['density'],
    high_income_high_density['Median family income (2018)'],
    c=high_income_high_density['capped_post'],
    cmap='coolwarm',
    edgecolor='k',
    alpha=0.75
)
axes[0, 1].set_title('High Income & High Density (> 1000 Density)')
axes[0, 1].set_xlabel('Population Density')
axes[0, 1].set_ylabel('Median Family Income')
axes[0, 1].grid(True, linestyle='--', alpha=0.5)
fig.colorbar(scatter2, ax=axes[0, 1], label='Post-Pandemic Price Increase (%)')
```

c. Below is the code that created the Low Income & Low-Density scatterplot for the pre and post pandemic eras:

```python
# Low Income & Low Density
scatter3 = axes[1, 0].scatter(
    low_income_low_density['density'],
    low_income_low_density['Median family income (2018)'],
    c=low_income_low_density['capped_pre'],
    cmap='coolwarm',
    edgecolor='k',
    alpha=0.75
)
axes[1, 0].set_title('Low Income & Low Density (< 1000 Density)')
axes[1, 0].set_xlabel('Population Density')
axes[1, 0].set_ylabel('Median Family Income')
axes[1, 0].grid(True, linestyle='--', alpha=0.5)
fig.colorbar(scatter3, ax=axes[1, 0], label='Pre-Pandemic Price Increase (%)')
```

```python
# Low Income & Low Density
scatter3 = axes[1, 0].scatter(
    low_income_low_density['density'],
    low_income_low_density['Median family income (2018)'],
    c=low_income_low_density['capped_post'],
    cmap='coolwarm',
    edgecolor='k',
    alpha=0.75
)
axes[1, 0].set_title('Low Income & Low Density (< 1000 Density)')
axes[1, 0].set_xlabel('Population Density')
axes[1, 0].set_ylabel('Median Family Income')
axes[1, 0].grid(True, linestyle='--', alpha=0.5)
fig.colorbar(scatter3, ax=axes[1, 0], label='Post-Pandemic Price Increase (%)')
```

d. Below is the code that created the Low Income & High-Density scatterplot for the pre and post pandemic eras:

```python
# Low Income & High Density
scatter4 = axes[1, 1].scatter(
    low_income_high_density['density'],
    low_income_high_density['Median family income (2018)'],
    c=low_income_high_density['capped_pre'],
    cmap='coolwarm',
    edgecolor='k',
    alpha=0.75
)
axes[1, 1].set_title('Low Income & High Density (> 1000 Density)')
axes[1, 1].set_xlabel('Population Density')
axes[1, 1].set_ylabel('Median Family Income')
axes[1, 1].grid(True, linestyle='--', alpha=0.5)
fig.colorbar(scatter4, ax=axes[1, 1], label='Pre-Pandemic Price Increase (%)')
```

```python
# Low Income & High Density
scatter4 = axes[1, 1].scatter(
    low_income_high_density['density'],
    low_income_high_density['Median family income (2018)'],
    c=low_income_high_density['capped_post'],
    cmap='coolwarm',
    edgecolor='k',
    alpha=0.75
)
axes[1, 1].set_title('Low Income & High Density (> 1000 Density)')
axes[1, 1].set_xlabel('Population Density')
axes[1, 1].set_ylabel('Median Family Income')
axes[1, 1].grid(True, linestyle='--', alpha=0.5)
fig.colorbar(scatter4, ax=axes[1, 1], label='Post-Pandemic Price Increase (%)')
```

e. Below is code that aided the visual representation of the graphs and caused the output of the graph:

```python
plt.tight_layout()
plt.show()
```

f. The following code created a 2x2 grid that allowed for a total of 4 graphs to be created (one for each individual category), facilitating the creation of all the scatter plots for the post pandemic period:

```python
fig, axes = plt.subplots(2, 2, figsize=(14, 12))
```

g. Below is code that aided the visual representation of the graphs and caused the output of the graph:

```python
plt.tight_layout()
plt.show()
```

2. Bar Graphs: Microsoft Excel was used to:
   a. Assign each zipcode's Population Density into groups increasing in increments of 5,000 people per square foot
   b. Assign each zipcode's Median Household Income into groups increasing in increments of $25,000
   c. Create pre and post pandemic pivot tables to determine the average percent increases in housing prices across the various Population Density and Median Household Income groupings

d. Create bar graphs showing the average percent increase in housing pricing by Population Density and Median Household Income, for pre and post pandemic views.

e. Embedded Excel file:

Merged_Data_Filtered_Removed_Cleaned.›

F. Statistical Analysis

1. The following code calculates the mean percentage price increase for each respective category in the pre and post pandemic eras:

```
# Analysis pre
column_name = 'Percentage_Price_Increase_PrePandemic'
avg_high_income_high_density_pre = high_income_high_density[column_name].mean()
avg_high_income_low_density_pre = high_income_low_density[column_name].mean()
avg_low_income_low_density_pre = low_income_low_density[column_name].mean()
avg_low_income_high_density_pre = low_income_high_density[column_name].mean()

```

```
# Analysis post
column_name = 'Percentage_Price_Increase_PostPandemic'
avg_high_income_high_density_post = high_income_high_density[column_name].mean()
avg_high_income_low_density_post = high_income_low_density[column_name].mean()
avg_low_income_low_density_post = low_income_low_density[column_name].mean()
avg_low_income_high_density_post = low_income_high_density[column_name].mean()
```

2. t-tests:
   a. After observing that High-Income, High-Density areas had a significantly greater average percentage increase than the other categories, the following codes t-tests were run to determine if this was by mere chance or rather reflected a significant dynamic in the market.

```
t_stat_high_low_density_pre, p_val_high_low_density_pre = stats.ttest_ind(
    high_income_high_density[column_name], high_income_low_density[column_name], equal_var=False
)
t_stat_high_low_income_pre, p_val_high_low_income_pre = stats.ttest_ind(
    high_income_high_density[column_name], low_income_low_density[column_name], equal_var=False
)
t_stat_high_high_density_pre, p_val_high_high_density_pre = stats.ttest_ind(
    high_income_high_density[column_name], low_income_high_density[column_name], equal_var=False
)

print("Pre-Pandemic Analysis")
print("1. Average Percentage Price Increase From 2000-2020")
print(f"a. High Income & High Density: {avg_high_income_high_density_pre:.2f}%")
print(f"b. High Income & Low Density: {avg_high_income_low_density_pre:.2f}%")
print(f"c. Low Income & Low Density: {avg_low_income_low_density_pre:.2f}%")
print(f"d. Low Income & High Density: {avg_low_income_high_density_pre:.2f}%")
print("2. T-Test Results comparing High Income & High Density with other groups:")
print(f"a. T-Test with High Income & Low Density: t-stat = {t_stat_high_low_density_pre:.2f}, p-value = {p_val_high_low_density_pre:.5f}")
print(f"b. T-Test with Low Income & Low Density: t-stat = {t_stat_high_low_income_pre:.2f}, p-value = {p_val_high_low_income_pre:.5f}")
print(f"c. T-Test with Low Income & High Density: t-stat = {t_stat_high_high_density_pre:.2f}, p-value = {p_val_high_high_density_pre:.5f}")

```

   b. After observing that High-Income, High-Density areas had a significantly low average percentage increase than the other categories, the following codes t-tests were run to determine if this was by mere chance or rather reflected a significant dynamic in the market

```python
t_stat_high_low_density_post, p_val_high_low_density_post = stats.ttest_ind(
    high_income_high_density[column_name], high_income_low_density[column_name], equal_var=False
)
t_stat_high_low_income_post, p_val_high_low_income_post = stats.ttest_ind(
    high_income_high_density[column_name], low_income_low_density[column_name], equal_var=False
)
t_stat_high_high_density_post, p_val_high_high_density_post = stats.ttest_ind(
    high_income_high_density[column_name], low_income_high_density[column_name], equal_var=False
)

print("Post-Pandemic Analysis")
print("1. Average Percentage Price Increase From 2020-2024")
print(f"a. High Income & High Density: {avg_high_income_high_density_post:.2f}%")
print(f"b. High Income & Low Density: {avg_high_income_low_density_post:.2f}%")
print(f"c. Low Income & Low Density: {avg_low_income_low_density_post:.2f}%")
print(f"d. Low Income & High Density: {avg_low_income_high_density_post:.2f}%")
print("2. T-Test Results comparing High Income & High Density with other groups:")
print(f"a. T-Test with High Income & Low Density: t-stat = {t_stat_high_low_density_post:.2f}, p-value = {p_val_high_low_density_post:.5f}")
print(f"b. T-Test with Low Income & Low Density: t-stat = {t_stat_high_low_income_post:.2f}, p-value = {p_val_high_low_income_post:.5f}")
print(f"c. T-Test with Low Income & High Density: t-stat = {t_stat_high_high_density_post:.2f}, p-value = {p_val_high_high_density_post:.5f}")
```