U.S. Rental Price Analysis Report

Author: Yingjun Zhong

May 9, 2023

1 Introduction

The housing market in the United States is constantly evolving, with changes in rent prices being one of the most significant and closely watched trends. Accurately forecasting rent prices is vital for a range of stakeholders, including tenants, landlords, investors, and policymakers. For tenants, it can help them plan their budgets and determine whether they can afford to stay in their current residence or need to look for more affordable housing options. For landlords and investors, understanding future trends in rent prices can inform decisions on property investments and rental strategies. Policymakers also need accurate information on rent prices to develop effective policies to ensure affordable housing options for all.

This report aims to provide insights into the forecasting of U.S. rent prices and the relationship between different socioeconomic indicators and rent prices in the country. By analyzing historical data and identifying key factors influencing rent prices, this report offers stakeholders a better understanding of how the market operates and what drives fluctuations in prices. This knowledge can help investors and policymakers develop more informed strategies and policies to address challenges related to affordable housing and housing affordability.

1.1 Objective

The analysis aims to cover the following objectives:

- **National Rental Price Forecasting**

  The first objective is to create precise time series models that can forecast U.S. rental prices. By analyzing historical data and identifying trends, these models can provide valuable insights to stakeholders, including investors, landlords, and policymakers, who can use this information to make informed decisions about the housing market.

- **Alternative Data Predictor Identification**

  The second objective is to supplement traditional data sources with alternative data predictors. Stakeholders can identify unique features that may impact rental prices by exploring external data sources, such as social media, satellite imagery, and public records. These features can help stakeholders to understand better the relationship between different variables and rental prices, which can, in turn, help them make more informed decisions.

2 Data Collection, Descriptive Analysis, & Cleaning

In this section, we will discuss the data collection, cleaning, and feature engineering processes for the three main datasets used in this analysis: the Zillow Rent Index, the American Community Survey (ACS), and the Internal Revenue Service (IRS) tax data.

2.1 Data Sources

The following are the datasets and their brief description used in this analysis:

- **Zillow Rent Index:** The Zillow Rent Index dataset includes attributes like Rental Price, Number of bedrooms, Area of property, location, and zip code.
- **American Community Survey, U.S. Census Bureau:** The survey dataset holds attributes like Population, Gini Index, and GDP per capita.
- **Internal Revenue Services:** IRS dataset consists of features like Wages and Salaries, number of returns, number of personal exemptions, adjusted gross income, and interest received from individual income tax returns filed every year.

2.2 Data Cleaning Operations

We performed several data cleaning operations to tidy the data, which involved the following steps:

- **Discarding Irrelevant Features:** We eliminated several features that were deemed irrelevant from an analytical standpoint, such as RegionID, RegionName, Metro, CountryName, and SizeRank.
- **Imputing Missing Values:** We encountered several instances of missing data in the rental pricing attribute, which could lead to incomplete analysis. To address this issue, we applied linear interpolation as well as forward-fill and backward-fill functions. Interpolation is particularly effective for time series attributes. Attached below shows missing data instances in the data visualized using missingno package in python.

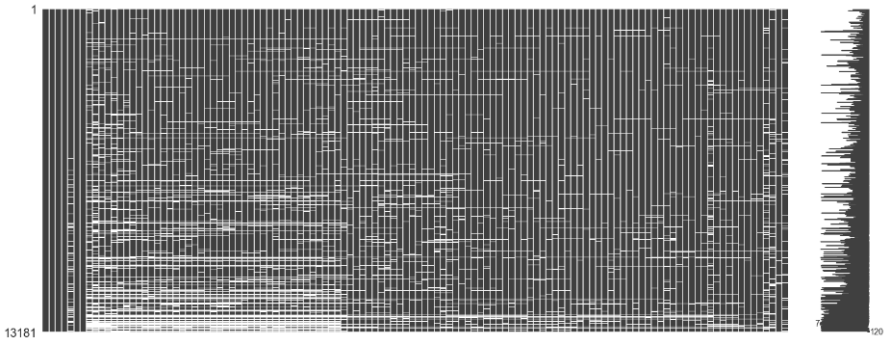

Figure 1: Missing Value Visualization

- **Data Wrangling:** We consolidated date columns and price columns into single attributes, effectively reformatting the data into a time series structure with date and price as key features. Fig 2 and Fig 3 shows the dataset before and after transformation applied.

	RegionID	RegionName	City	State	Metro	CountyName	SizeRank	2010-09	2010-10	2010-11	...	2019-04	2019-05	2019-06	2019-07	2019-08	2019-09	201 1
0	61639	10025	New York	NY	New York-Newark-Jersey City	New York County	1	3031.0	3058.0	3031.0	...	3785.0	3788.0	3786.0	3784.0	3766.0	3779.0	3843.
1	84654	60657	Chicago	IL	Chicago-Naperville-Elgin	Cook County	2	1790.0	1787.0	1784.0	...	2039.0	2070.0	2105.0	2140.0	2168.0	2185.0	2125.
2	61637	10023	New York	NY	New York-Newark-Jersey City	New York County	3	3269.0	3304.0	3320.0	...	3874.0	3898.0	3917.0	3929.0	3931.0	3963.0	Na
3	91982	77494	Katy	TX	Houston-The Woodlands-Sugar Land	Harris County	4	1547.0	1549.0	1560.0	...	1765.0	1755.0	1751.0	1752.0	1754.0	1759.0	1764.
4	84616	60614	Chicago	IL	Chicago-Naperville-Elgin	Cook County	5	1922.0	1925.0	1921.0	...	2245.0	2289.0	2332.0	2372.0	2398.0	2412.0	2348.

Figure 2: Before Transformation

	City	State	year	price
0	Austin	TX	2010-09	1219.0
1	Boston	MA	2010-09	2048.0
2	Honolulu	HI	2010-09	2038.0
3	Houston	TX	2010-09	1052.0
4	Las Vegas	NV	2010-09	1082.0

Figure 3: After Transformation

2.3 Descriptive Analysis

This section provides an overview of the dataset used in the analysis. A table with details of data type of each attribute and the corresponding data statistics summary of each numeric field is included below. For a more detailed description of the data, please refer to the code.

```
zip                          int64
City                        object
State                       object
Metro                       object
CountyName                  object
datetime            datetime64[ns]
zri                        float64
year                         int64
total_firms                  int64
job_creation_rate          float64
job_destruction_rate       float64
startup_firms                int64
dtype: object
```

(a) Data Type of Indicator Dataset

```
City      object
State     object
year      object
price    float64
dtype: object
```

(b) Data Type of Pricing Dataset

Figure 4: Data Type of Different Data-sets

	zip	zri	year	total_firms	job_creation_rate	job_destruction_rate	startup_firms
count	93744.000000	93744.000000	93744.000000	93744.000000	93744.000000	93744.000000	93744.000000
mean	43726.884025	1575.035645	2016.500000	39210.235151	13.626151	11.322587	3481.655658
std	33913.714419	660.802069	1.707834	47418.416684	2.094381	1.840758	4482.759391
min	1013.000000	504.000000	2014.000000	574.000000	6.466000	5.695000	19.000000
25%	11368.000000	1069.000000	2015.000000	12689.000000	12.003000	10.193000	865.000000
50%	33160.500000	1430.000000	2016.500000	20715.000000	13.894000	11.174000	1672.000000
75%	78703.000000	2009.000000	2018.000000	44272.000000	14.938000	12.255000	4348.000000
max	99654.000000	5136.000000	2019.000000	196469.000000	30.168000	36.560000	18899.000000

Figure 5: Descriptive Statistics

We utilized a box plot to visualize the rental ranges for various prominent cities, and discovered that San Francisco, Boston, and Los Angles had the highest apartment rental prices. Additionally, Boston had a significant number of outlier prices. Cities like Jacksonville, Houston, New Orleans has the lowest apartment rent.

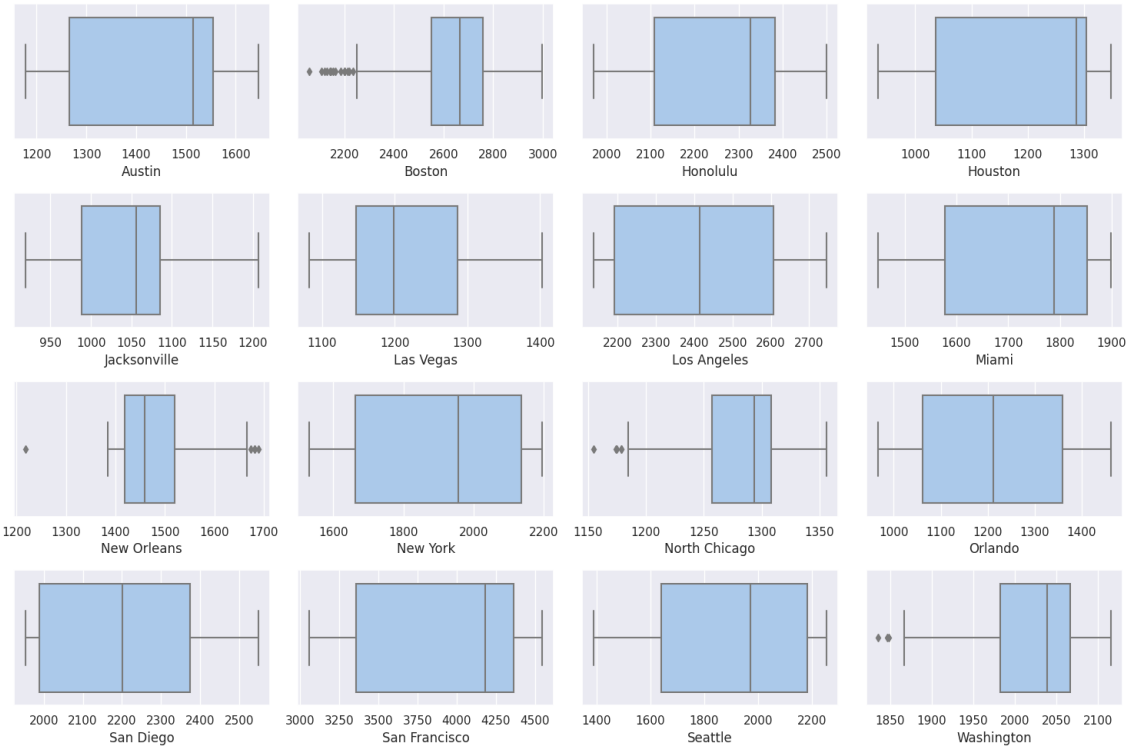

Figure 6: Descriptive Statistics

3 Data Visualization & Insights

This section presents the insights derived from various visualizations created during the analysis. As previously stated, the main objective of this analysis is to provide valuable insights to aid in informed decision-making when it comes to buying or renting property, from a variety of perspectives. By examining the data in different ways and presenting it in visual form, we can identify patterns, trends, and relationships that might not be apparent from the raw data alone. Through these visualizations, we aim to provide actionable insights that can be used by a range of stakeholders, including renters, buyers, investors, and policymakers, to make informed decisions about the housing market.

3.1 Visualizing Rental Price Over Time & Forecast

The purpose of this visualization is to gain insights into the trends in apartment rental prices over time in different cities and analyze the variations among them. This can provide valuable information to stakeholders such as tenants, landlords, and investors, who can make informed decisions based on the observed patterns.

To forecast the rental prices for the next two years, we utilized an ARIMA model in Python. This model is widely used for time-series forecasting and is particularly useful when there is a

trend or seasonality in the data. The ARIMA model considers the historical data and applies mathematical techniques to predict future values.

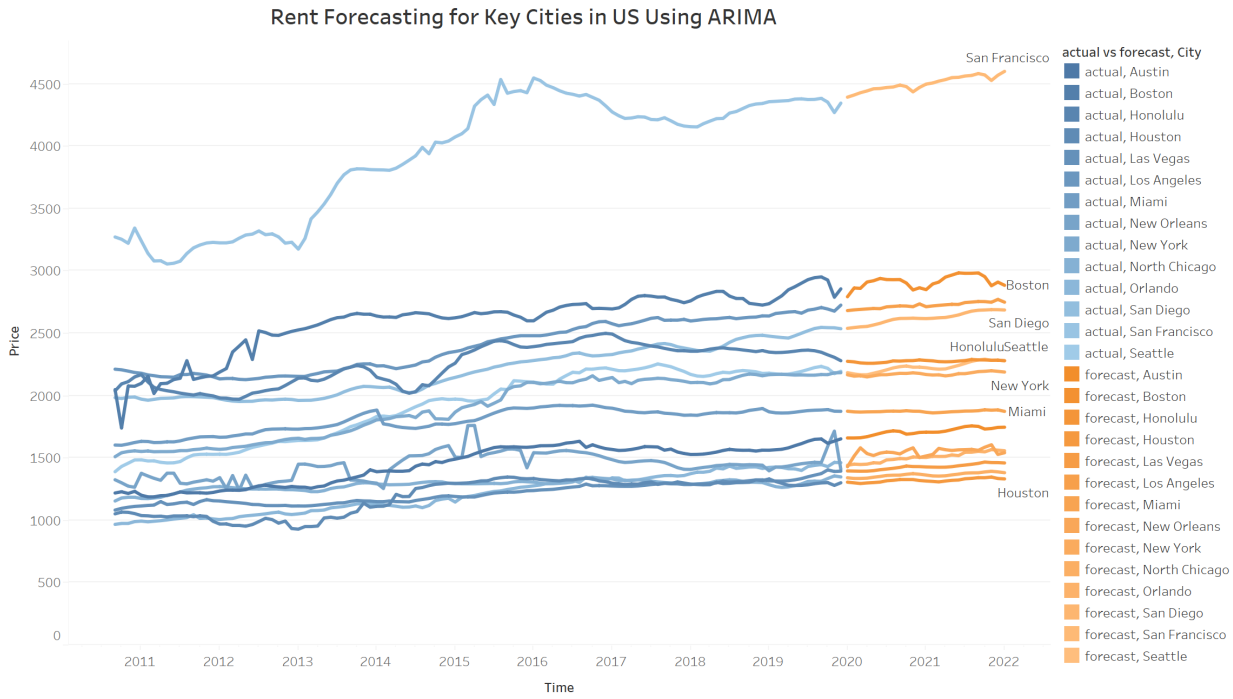

Figure 7: Rent Trend & Forecast

Following are some insights that we have obtained using the above rental forecast:

- San Francisco has the highest rent of all times as compared to other cities and the forecast shows even higher prices in the coming years.
- Most of the cities has rent varying between 900 and 2700 USD.
- Forecast indicates a slow increase in the rent prices over the next two years.
- Boston, Las Vegas, and San Diego has the highest rent after San Francisco over the time
- Cities like Houston, New Orleans, and JacksonVille has the lower rent over the time as compared to other cities.

3.2 Visualizing Variations using Parallel Coordinate Plot

Our main objective now is to investigate the interdependence of various attributes and how they influence each other. We decided to utilize the Parallel Coordinate Plot from the plotly library to achieve this goal. This interactive plot serves as an excellent tool to create a visual representation that demonstrates the relationships between different attributes. By using this plot, we can observe how changes in one attribute affect the values of other attributes, and how they are correlated with each other. This visualization can provide us with valuable insights into the data and help us identify patterns and trends that might not be apparent through simple numerical analysis. This is similar to visualizing the correlation between the features but in an interactive manner.

The plot below is a parallel coordinate plot between different features influencing the rent of the property.

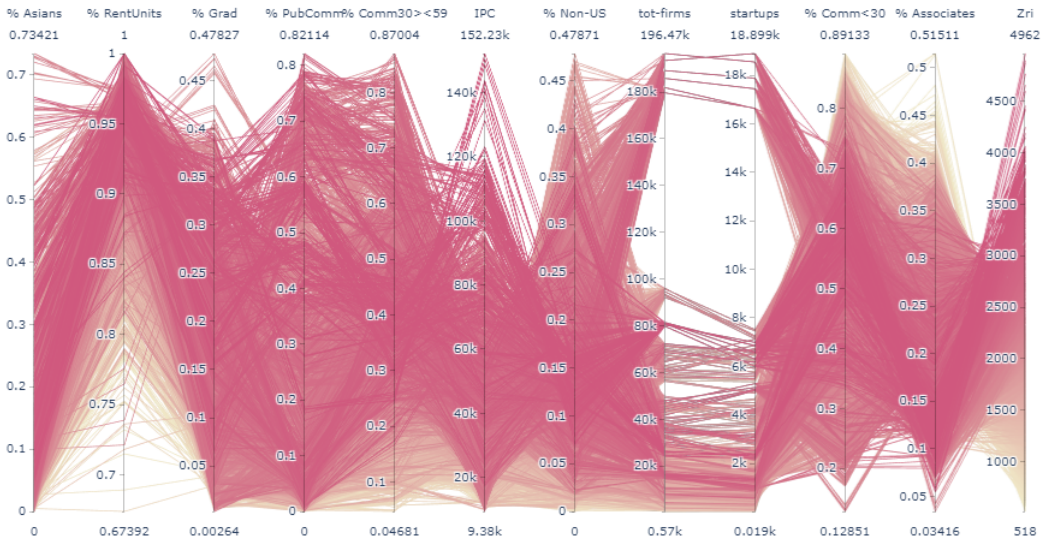

Figure 8: Parallel Coordinate Plot

Following are some insights that we have obtained using the parallel coordinate plot:

- Percent Asians, Percent Units Occupied, Public Commuting, Income Per Capita, Percent Non-US, total firms, and startups features shows a positive correlation with the ZRI (Rental Price) feature.
- Percent commute less than 30 minutes and associate degree features shows a negative correlation with the ZRI feature.

3.3 TSNE Visualization of Cities

The goal of this section is to use visualization techniques to identify cities that are similar based on a high number of dimensions. To accomplish this, we will utilize a TSNE-based visualization method to compress the data into two dimensions and create a visualization that displays the cities in a way that makes it easy to compare and contrast them.

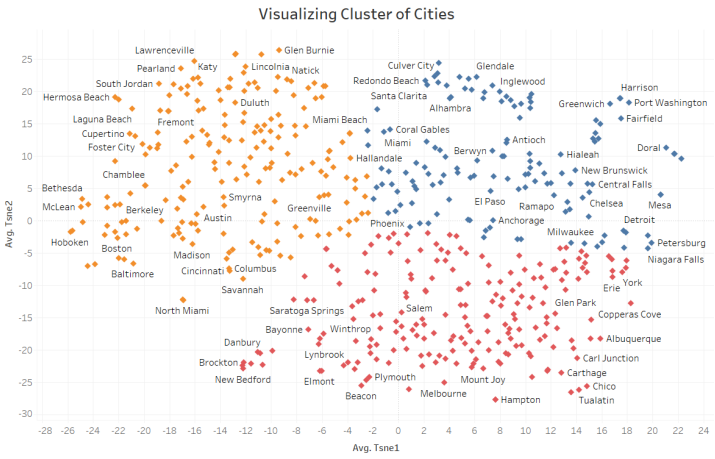

Figure 9: TSNE Cluster Visualization

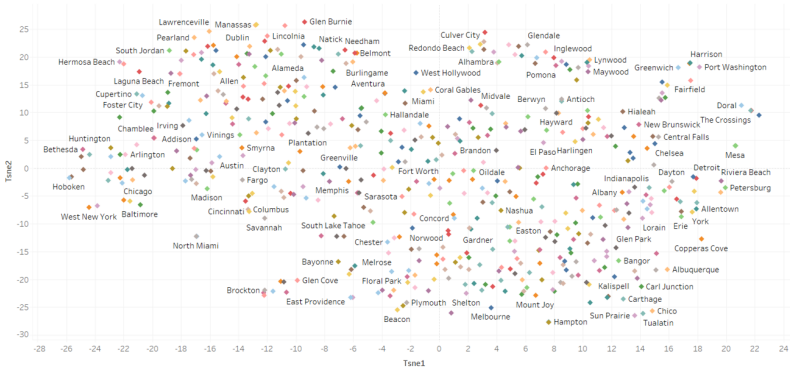

Figure 10: TSNE Cities Visualization

Following are some insights that we have obtained using the attached TSNE cluster and TSNE normal plot:

- Big Cities like New York, Boston, Chicago, San Francisco are clustered together in figure 10.
- Most of the big tech cities are clustered in the Orange cluster of figure 9.
- Cities like Dallas, Austin, Houston, and Denver are clustered together indicating group of non-expensive cities presumably in figure 10.
- All non-famous cities are clustered in Blue and Pink color in figure 9.

3.4 Correlation Heatmap

In this section, we aim to delve deeper into the relationship between all the numeric features in our analysis by creating a correlation heatmap. This visualization technique will provide us with a clear and concise representation of the correlation values between each attribute, allowing us to gain a better understanding of their relationships.

A correlation heatmap is a graphical representation of the correlation matrix, where each cell of the matrix is represented by a color that corresponds to its correlation value. The correlation matrix provides a way to measure the linear relationship between pairs of variables. In our case, each variable represents a specific numeric feature in our dataset, and the correlation matrix will show us how each feature is related to one another.

The correlation values range from -1 to 1, with negative values indicating a negative relationship and positive values indicating a positive relationship between the attributes. The closer the correlation value is to -1 or 1, the stronger the relationship between the two attributes. A correlation value of 0 indicates that there is no linear relationship between the two attributes.

Following are some insights that we have obtained using the attached Correlation Heatmap:

- Percent Asians, Percent Units Occupied, Public Commuting, Income Per Capita, Percent Non-US, total firms, and startups features shows a positive correlation with the ZRI (Rental Price) feature.
- Percent commute less than 30 minutes and associate degree features shows a negative correlation with the ZRI feature.

- Poverty Rate, Workforce unemployed, and age group between 0 and 17 has near zero correlation indicating no strong relationship.

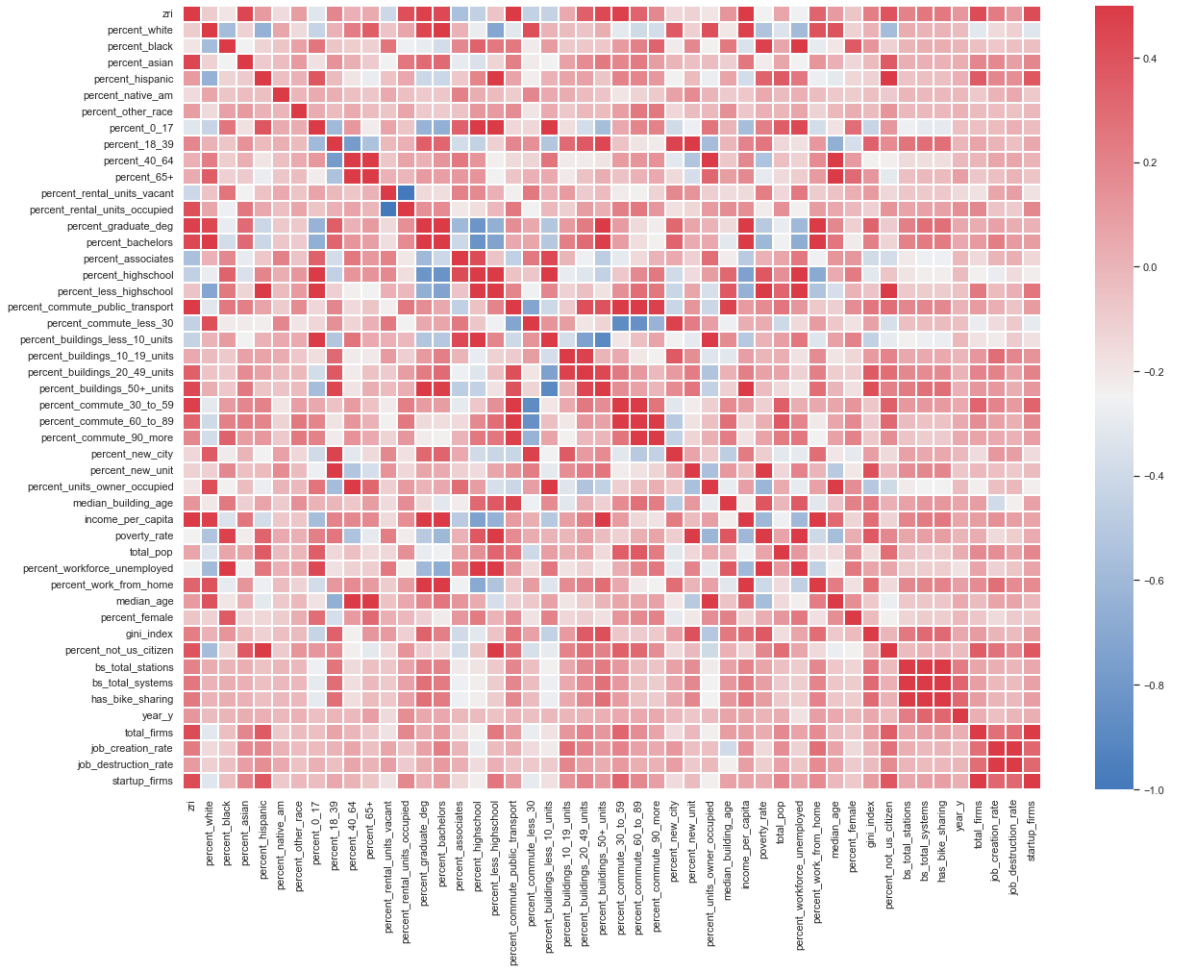

Figure 11: Correlation Heatmap

3.5 World Map & Treemap

In order to gain a comprehensive understanding of the rental market across different states in the US, we have dedicated this section to visualizing the average rent, along with other relevant statistics. By creating two different visualization charts, we hope to provide a clear and informative representation of the data.

The first visualization chart is a US map that partitions the states by their respective average rent values. This will allow us to quickly identify the states with the highest and lowest average rent. Additionally, we can use this map to compare the average rent values across different states and look for patterns or trends.

The second visualization chart is a treemap that highlights the most expensive cities within specific states. This will give us a better understanding of which cities are driving the high average rent values in each state. We can also use this chart to compare the rent values across different cities within a state and identify any outliers.

By combining these two visualization charts, we can gain valuable insights into the rental market in different states, including which states have the highest rent values, which cities

within those states are the most expensive, and how these values compare to each other. Furthermore, we can use these insights to make informed decisions regarding rental properties and investments in different states and cities.

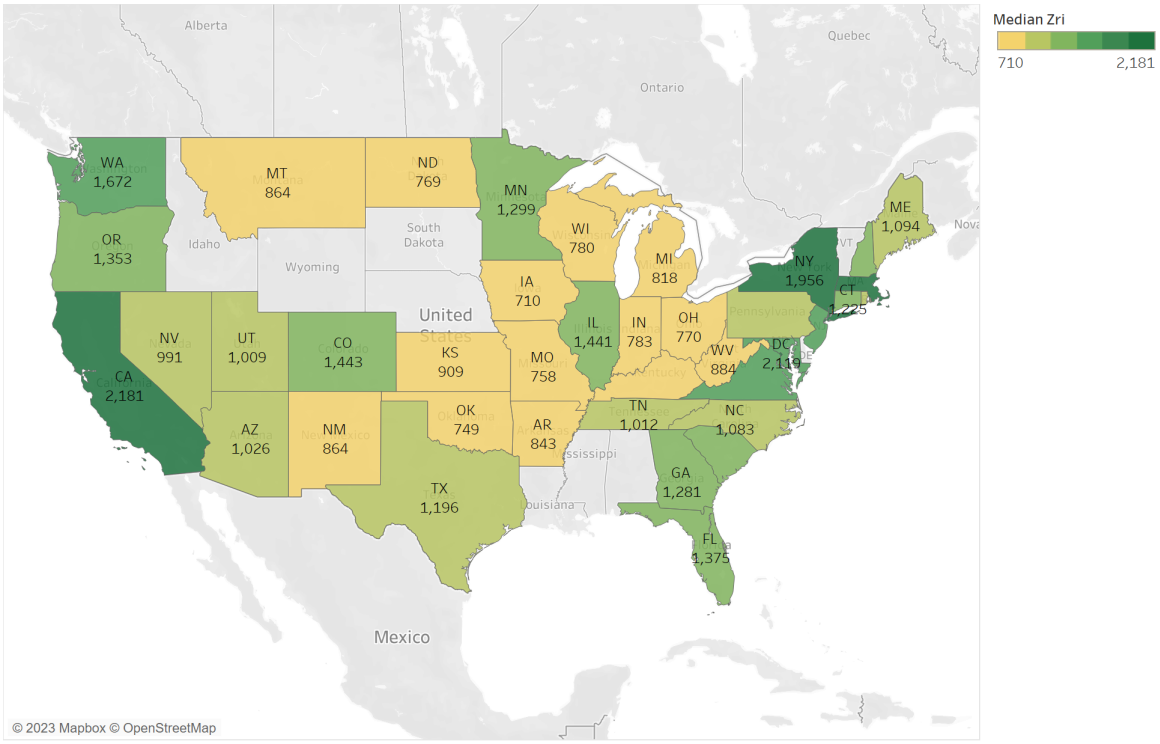

Figure 12: Rental Price Heat Map

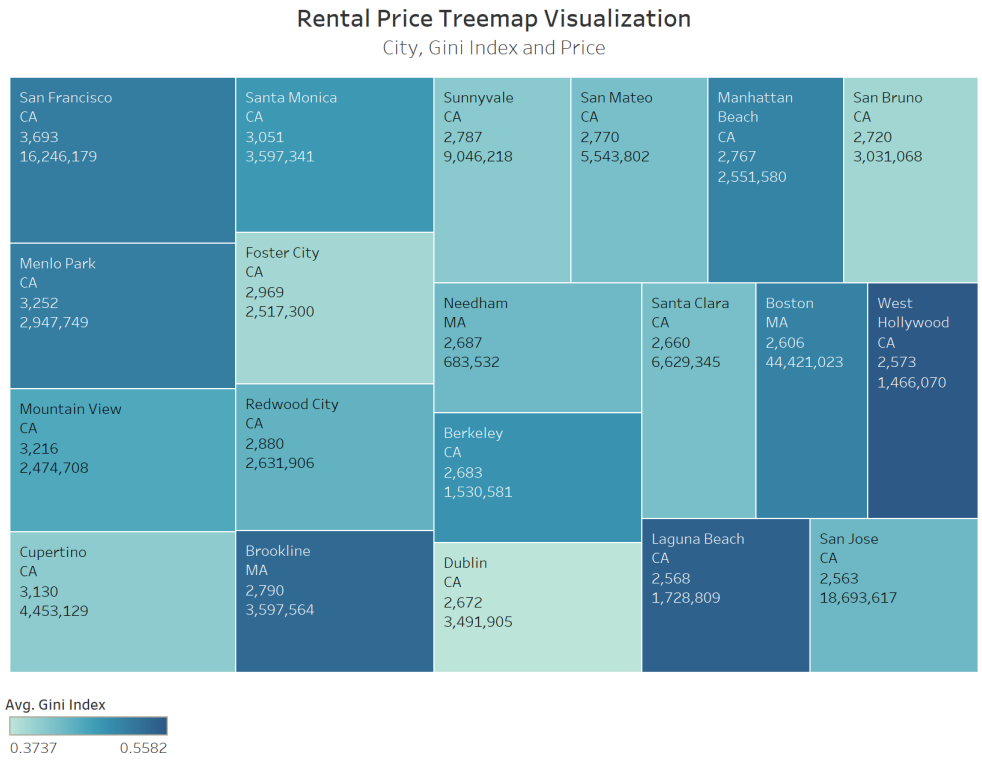

Figure 13: Tree Map

- Clearly California, Massachusetts, and New York are among the top most expensive states in US in terms of housing rent.
- Montana, North Dakota, Idaho, Ohio, Wisconsin, and Minnesota are among the states having low rent.
- Cities like San Francisco, Menlo Park, and Mountain View has the highest rent and belongs to California state and some are from MA.