Residential Market Analysis NY Tri-State

by

BIA 610B – Team 3

Suleesh Sugathan

Nayan Sarkar

Ramneek Sharma

CONTENTS

INTRODUCTION

Business Understanding

The trajectory of New York Tri-State region real estate market since the recession defies national trends. The real estate market has surged ahead, even as markets languished in the rest of the country. In this report we have performed a statistical data analysis of trends and insight into the ever-changing market to track the fluctuations, conditions, trends, and policies. We have focused on key parameters such as inventory, sales, pending sales and months of supply among many other measures. Visualization of the data will help investors and home buyers make informed decisions when it comes to making purchases in the New York Tri-State area.

Modeling and Evaluation

Below are the models that haven been used to study and evaluate the data:

- Scatter plot
- Tree Map
- Stacked Bar chart
- Line graph
- Packed Bubble Chart
- Box-and-Whisker Plot
- Dual-Axis Combination Bar Charts with Trendlines
- Point Map
- Dendrogram
- Sankey Diagram
- Line Chart combination with Forecasting

Data Source

The data has been sourced from **Redfin**, a renowned real estate brokerage firm in the US for the years 2015 to 2020.

MODEL 1: NEW LISTING VS TOTAL INVENTORY
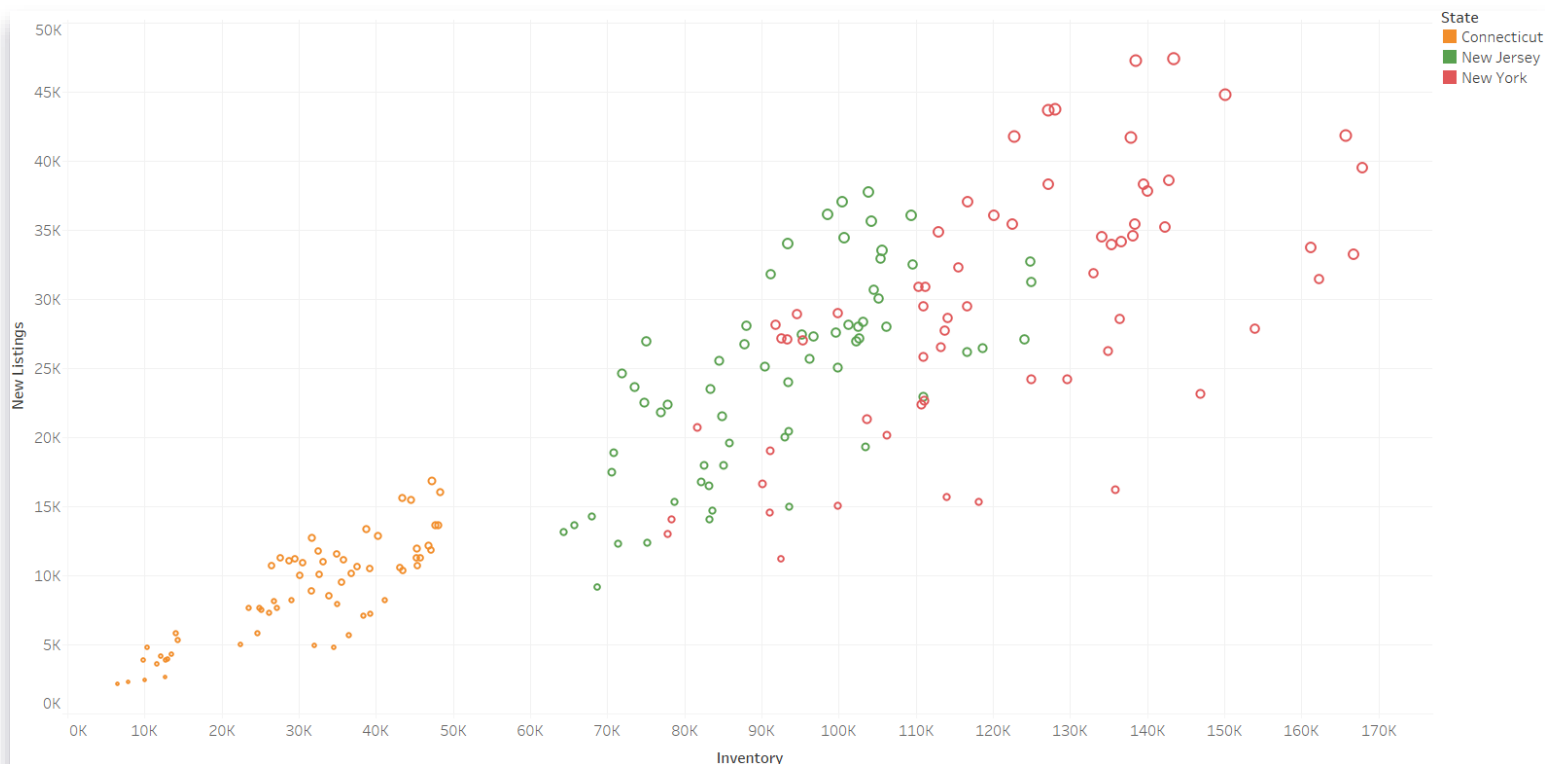
First model is a **Scatter Plot** visualization that is comparing the **New Listing** against the **Total Inventory** available. Scatter plots offer a good way to do ad hoc analysis. Another benefit is it is one of the few visualizations that allow you to view many marks in a small space and make it easy to identify outliers and the correlations.

Here the *New Listings* are the total number of new houses listed for sale during a given period and the *Inventory* is the total number of active house listings on the last day of a given period. As we can see amongst all the states, New York has the highest inventory but relatively low new listings, but Connecticut has the lowest inventory, but has relatively high new listings compared to others.

MODEL 2: PRICE PER SQUARE FOOT
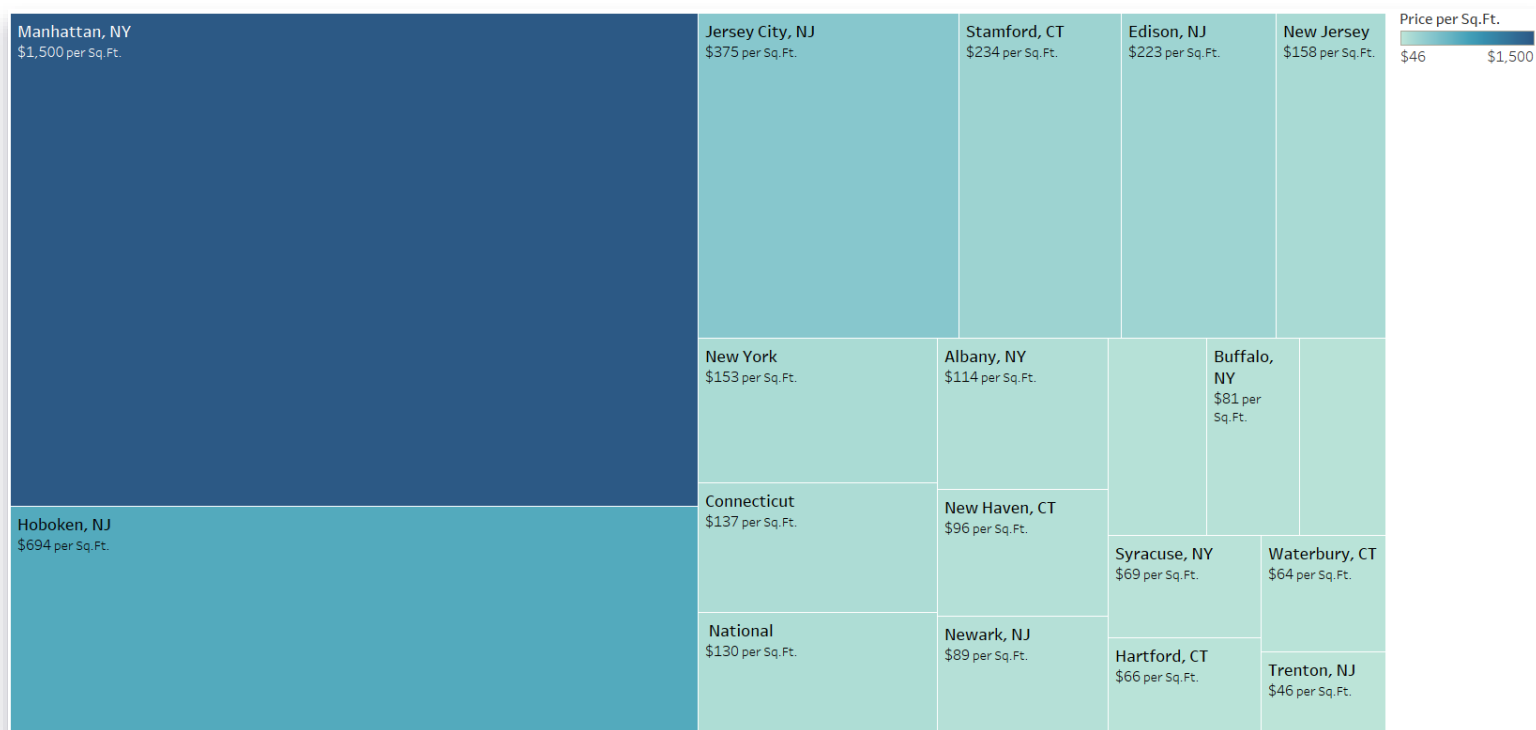
Second model is a **Tree Map** visualization that uses rectangles in hierarchies to display the average list **Price per Square Foot** of all active listings as across the Tri-State region along with the national average. The rectangles are easy to visualize as both the size and shade of the color of the rectangle reflect the value of the measure. Tree maps also allow you to reasonably communicate and consume hundreds of marks on a single view. This makes it easier to spot patterns and relationships that you would not otherwise be able to see.

As we can see, Manhattan has the highest price per square foot across the Tri-State region and it more than 10x greater than the National average of $130 per sq. ft. The visualization also lets us easily decipher that New York has the most expensive rates while New Jersey and Connecticut have some overlap with Trenton been the lowest at $46 per sq. ft.

MODEL 3: MONTHS OF MARKET SUPPLY
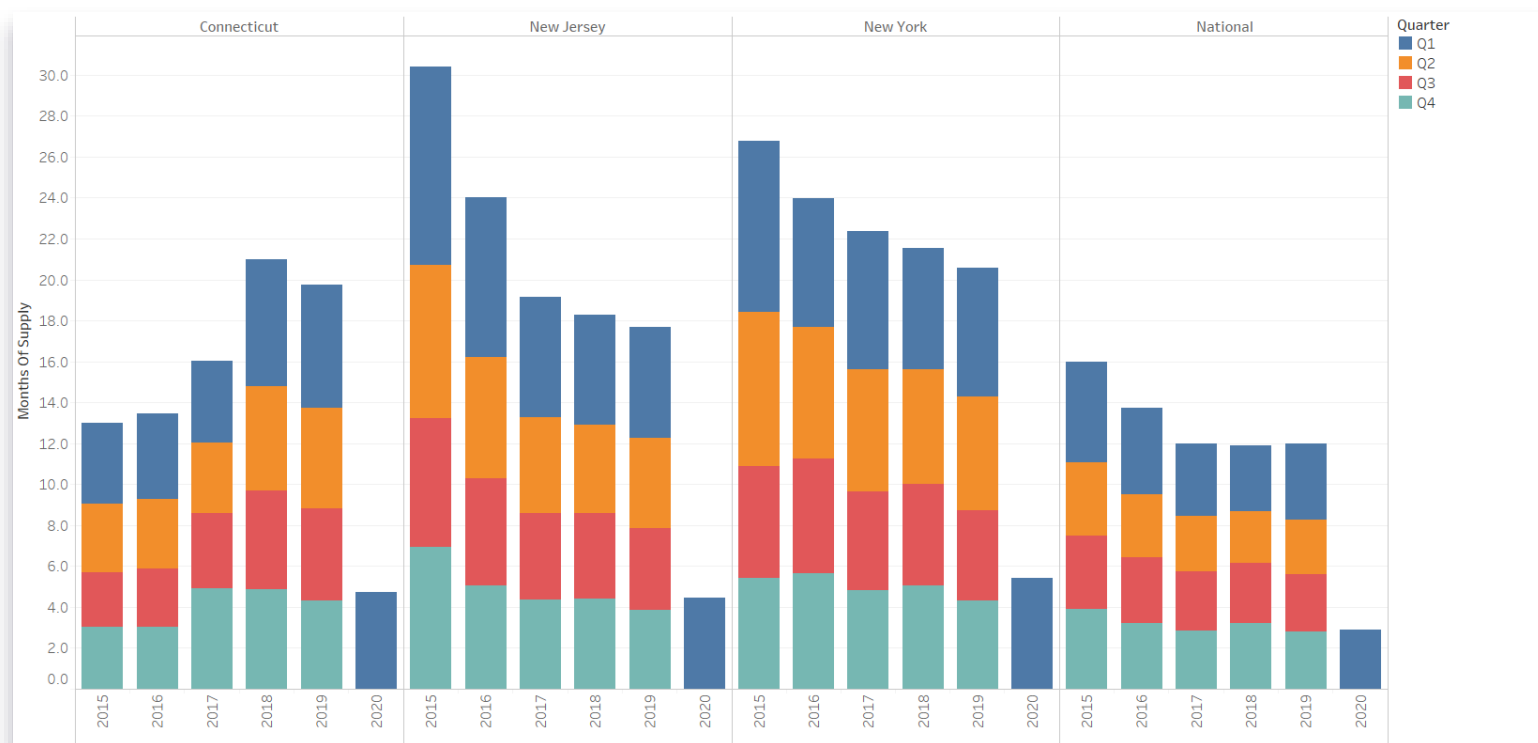
Third model is a **Stacked Bar Chart** that displays the average **Months of Market Supply** across the Tri-State regions along with national average that is distributed across all four quarters of a year. A stacked bar chart, also known as a stacked bar graph, allows us to break down and compare parts of a whole. Each bar displays a total amount that is then further broken down into sub-amounts.

Here the *Months of Market Supply* refers to how long it would take for the supply to be bought up if no new homes came on the market. It is the inventory divided by home sales for monthly data. We can see that the months of market supply has been on a decline since 2015 in New York, New Jersey and at a National level but Connecticut has seen an inverse trend with an increasing supply.

MODEL 4: DAYS ON MARKET
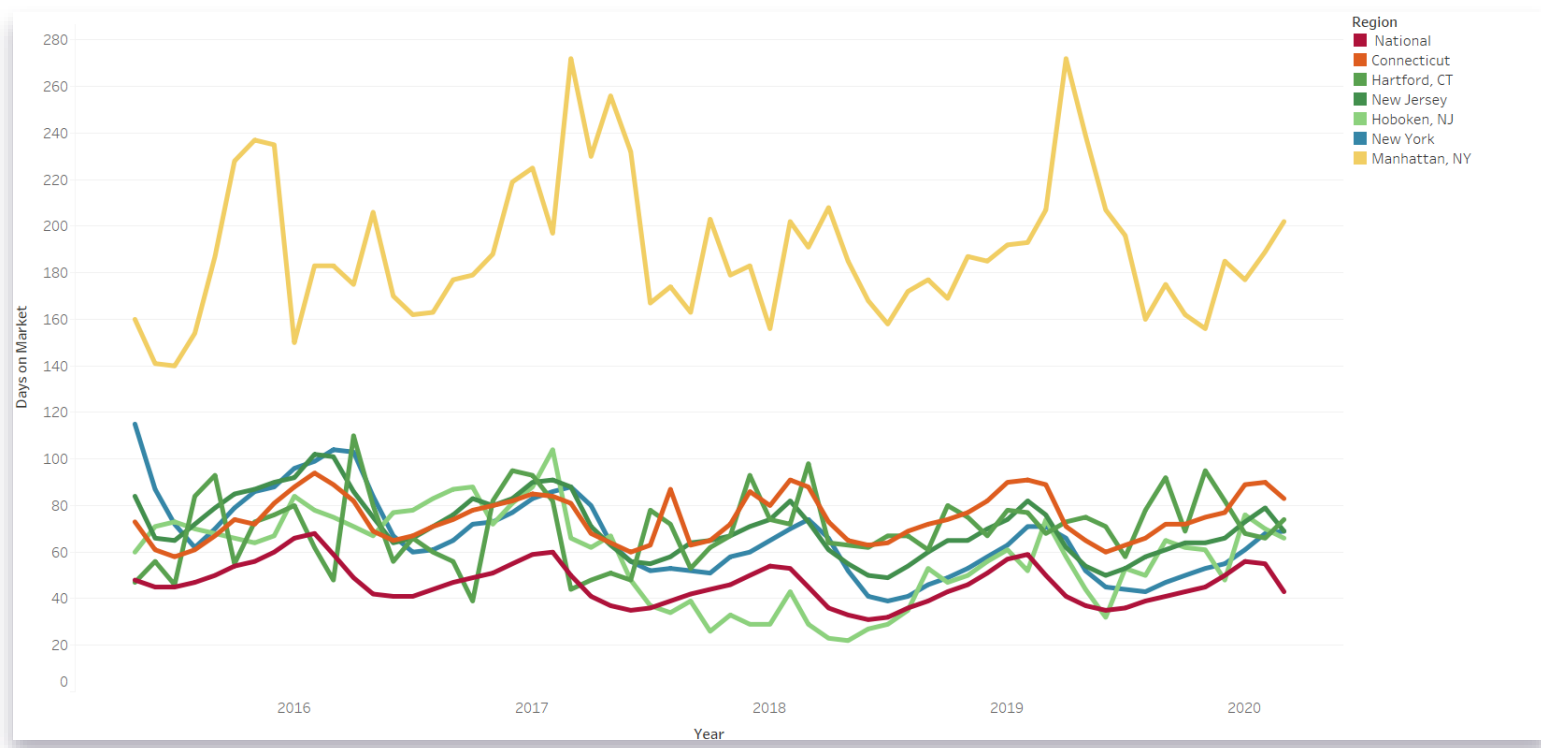
Fourth model is a **Line graph** that is used to compare the average **Days on Market** across few major cities in the Tri-State along with the national average. A line graph, also known as a line chart, is used to visualize the value of something over time. They can be used when you are plotting data that has peaks (ups) and valleys (downs), or that was collected in a short period.

Here the *Days on Market* is the number of days between the date the home was listed for sale & when the home went off-market/pending sale. Excludes homes that sat on the market for more than 1 year. As evident, Manhattan has listings that sit on the market way longer before been sold than other cities and even the national average that is far less than almost all the major Tri-State cities. One reason for this would be the high price for Manhattan listings and hence they sit on the market for longer.

MODEL 5: LISTING OFF MARKET IN 2 WEEKS
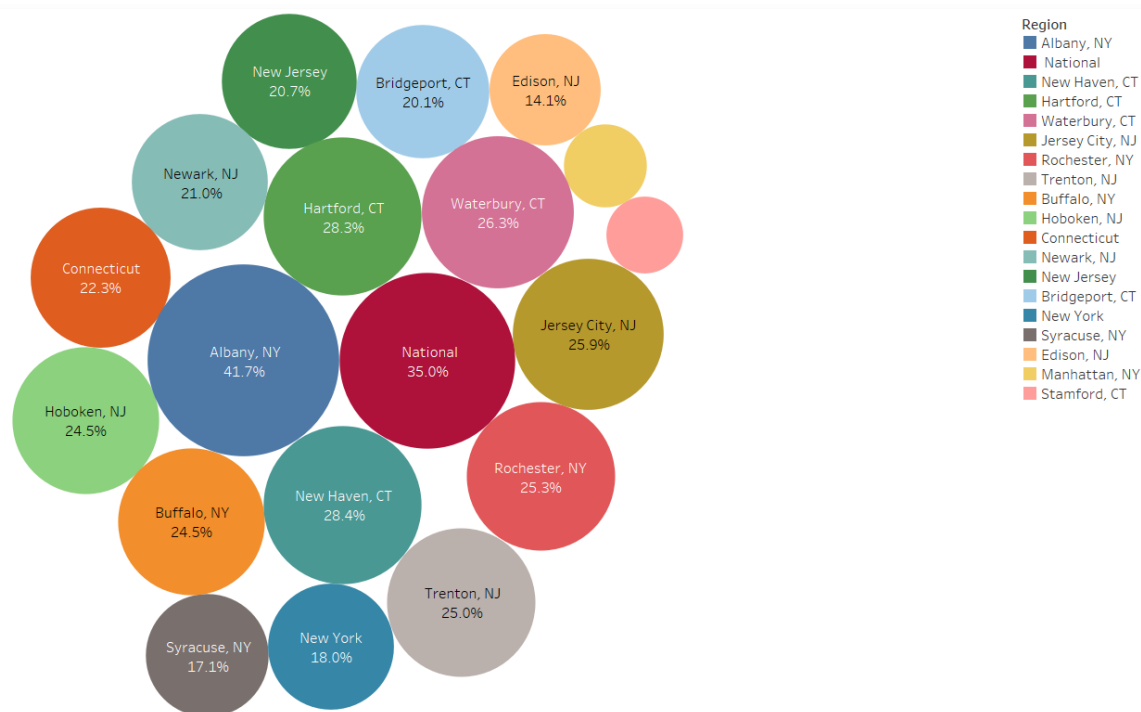
Fifth model is a **Packed Bubble Chart** that allows us to quickly visualize the places were house listings go off the market in under 2 weeks or less. A bubble chart is used to show data in bubbles or circles with varying sizes and according to a color scheme. It is a generalization of the scatter plot, replacing the dots with bubbles. The overall picture of Bubble Charts can be used to analyze for patterns/correlations. Too many bubbles can make the chart hard to read, so Bubble Charts have a limited data size capacity.

Here having a *House Listing off Market in 2 weeks* refers to the ratio of total number of listings that went from active to pending/off-market against the total number of new listings during a given period. As represented, Albany has a higher percentage of having the house off market quickly while Manhattan and Stamford have a lower chance, and this mainly comes down the pricing of the houses in these regions.

MODEL 6: PRICE DROP FOR ACTIVE LISTINGS
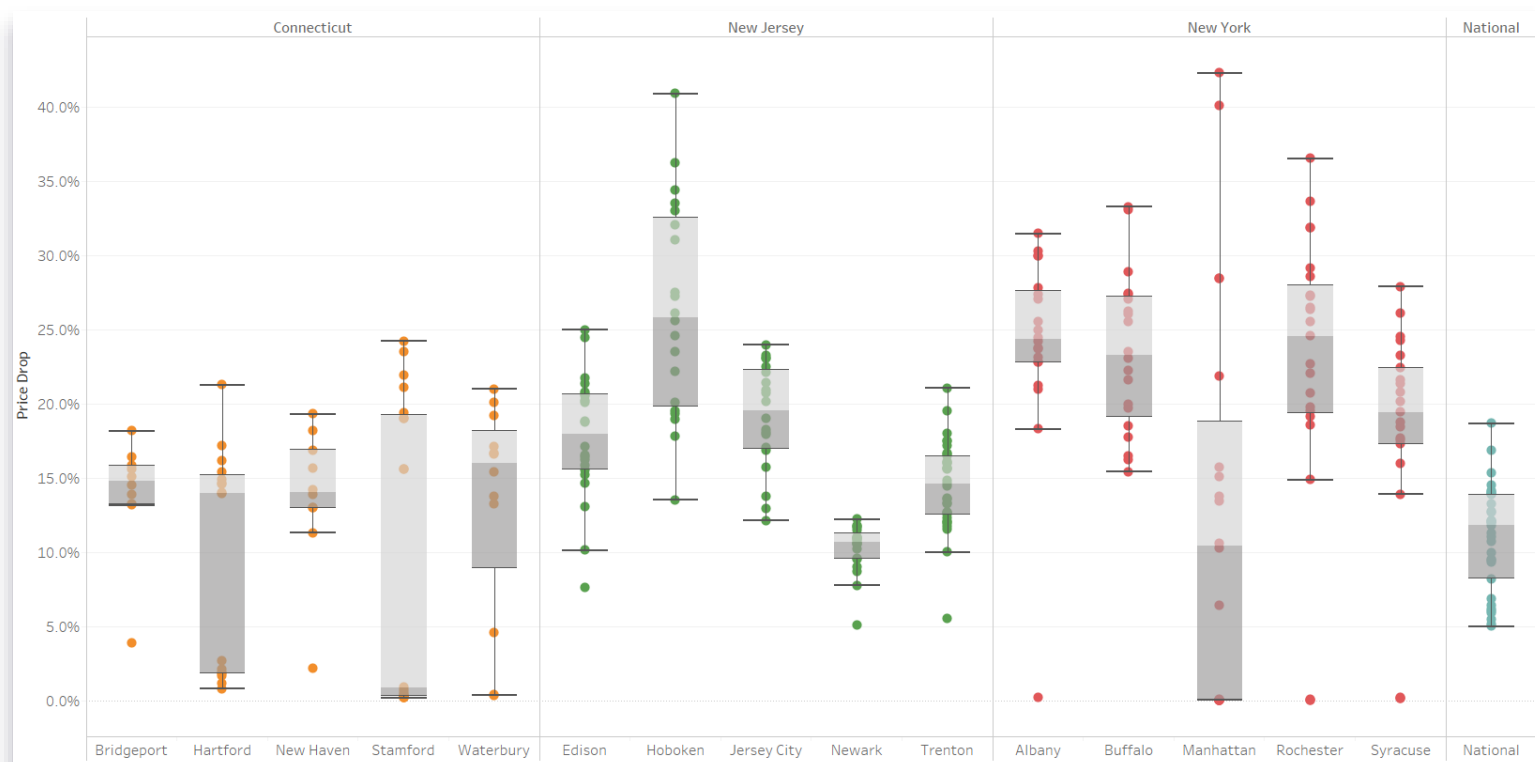
Sixth model is a Box-and-Whisker Plot, or **Box Plot**, that we have used to illustrate the average **Price Drop** percentage for active listings across the Tri-State and the national level over a period. Box Plot is an effective visualization choice for illustrating distributions. They work particularly well when you want to compare the distributions across two different dimension members side-by-side. Boxes indicate the middle 50 percent of the data and we can configure lines, called whiskers, to display all points within 1.5 times the interquartile range, or all points at the maximum extent of the data.

Here the average *Price Drop percentage for Active Listings* tells us how many active listings dropped their price during a period. In the Tri-State region, New York cities have the highest price drop percentages and it has seen some big price drops over the years while Connecticut cities have resisted big price drops, especially Hartford.

MODEL 7: LIST PRICE VS SALE PRICE
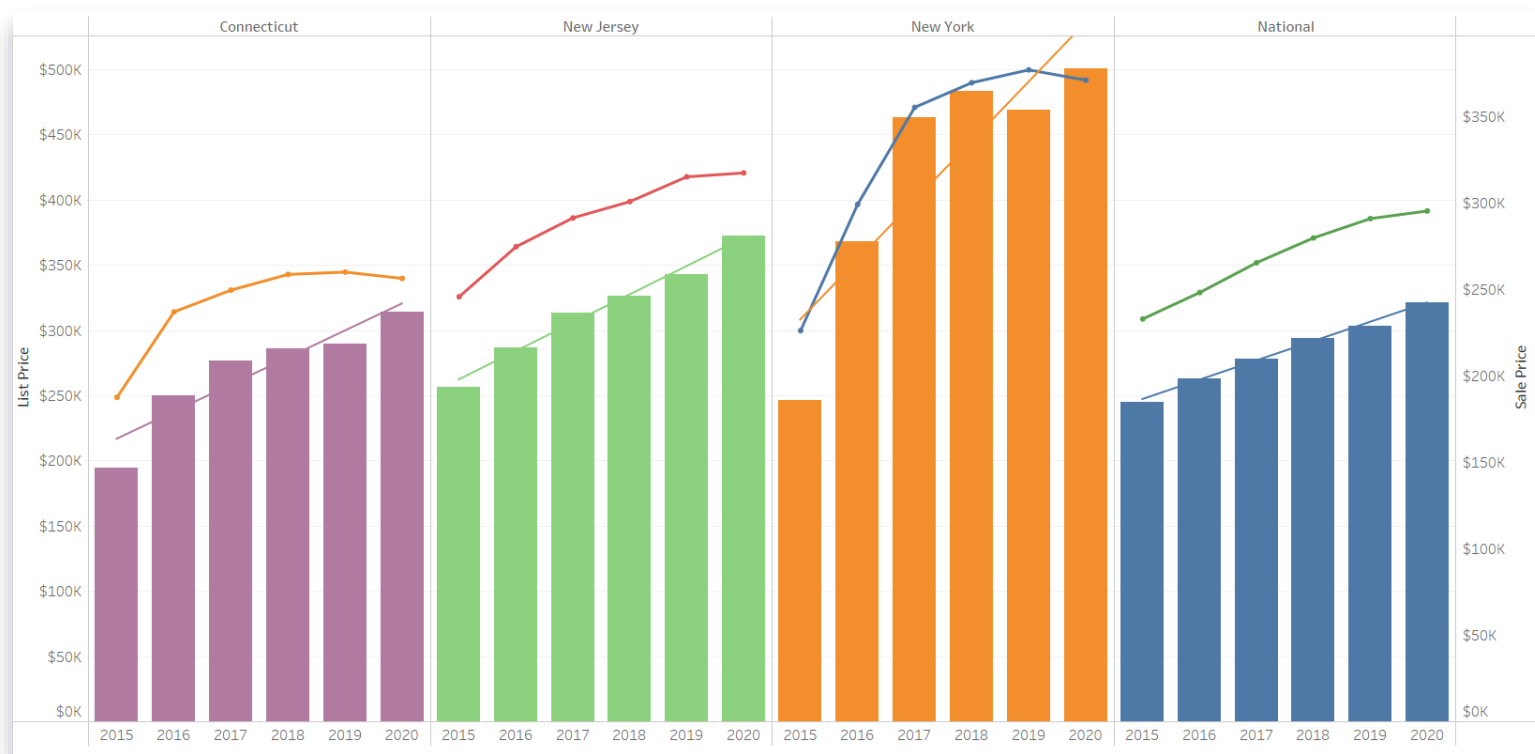
Seventh model is a **Dual-Axis Combination Bar Charts with Trendlines** that compares the average **List Price** for a listing and the average **Sale Price** for listings in the Tri-State and National level over a period. Dual-axis combination charts have two axes and they display a combination of different mark types. Along with it, Trend lines allows to highlight trends in the data.

Here the average *List Price* is the most recent listing price covering all homes with a listing date during a given time period and it is depicted using a bar chart, while the average *Sale price* is the final home sale price covering all homes with a sale date during a given time period and is depicted using a line graph. As illustrated, the final sale price of listing tends to remain higher than the list price but New York cities did see a dip in that during the first quarter of 2020 and the trendline suggests that it list price is only going to go higher and hence we might see higher sale prices for all of Tri-State region.
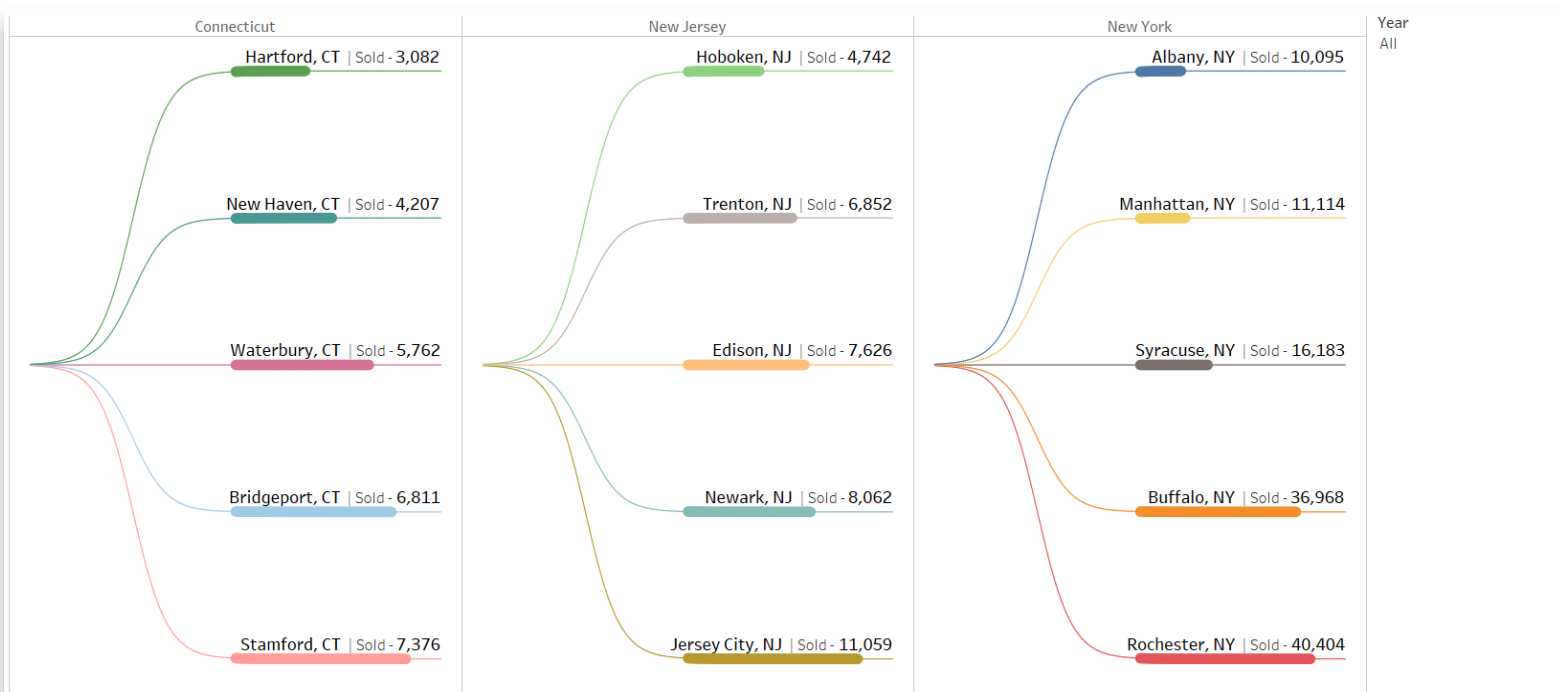
MODEL 8: PENDING HOUSE SALES
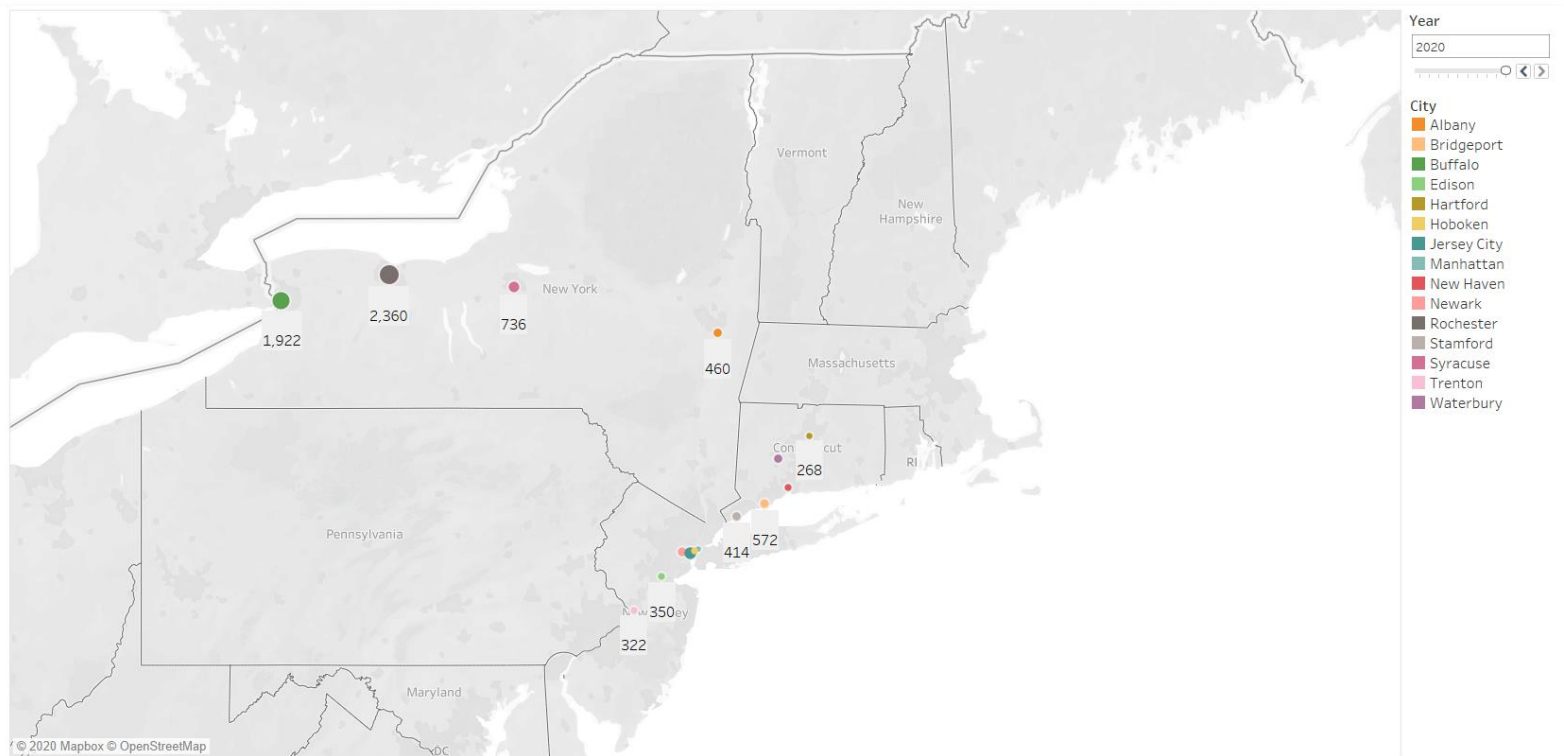
Eighth model is a simple **Point Map** to highlight the cities across the Tri-State region that have the highest number of **Pending House Sales.** Maps are one of the most effective chart types and are also among the easiest chart types to create. They are technically scatter plots with points at the combination of each latitude-longitude pair and an image of a map in the background. They help us to see patterns between geographic locations that may otherwise be challenging to discover.

Here the *Pending House Sales* tells you the total number of homes that went under contract but has not closed on during a given period and were on the market for less than 90 days. As we can see, for the year 2020, Buffalo, NY has the highest number of listings that are pending sales while the areas near the Metropolitan region of New York City has way lesser number of listings. Reasoning for that can be the demand for houses mean that closing on the sale is faster than other regions.

MODEL 9: TOTAL HOUSES SOLD

Connecticut	New Jersey	New York	Year All
Hartford, CT | Sold - 3,082	Hoboken, NJ | Sold - 4,742	Albany, NY | Sold - 10,095	
New Haven, CT | Sold - 4,207	Trenton, NJ | Sold - 6,852	Manhattan, NY | Sold - 11,114	
Waterbury, CT | Sold - 5,762	Edison, NJ | Sold - 7,626	Syracuse, NY | Sold - 16,183	
Bridgeport, CT | Sold - 6,811	Newark, NJ | Sold - 8,062	Buffalo, NY | Sold - 36,968	
Stamford, CT | Sold - 7,376	Jersey City, NJ | Sold - 11,059	Rochester, NY | Sold - 40,404	

Ninth model is a **Dendrogram** that shows the hierarchy based on the **Total number of Houses Sold** in all three states of the NY Tri-State. Dendrograms are essentially tree diagrams that show the hierarchical relationship of a data set. It is constituted of a root node that gives birth to several nodes connected by edges or branches. The last nodes of the hierarchy are called leaves. The main use of a dendrogram is to work out the best way to allocate objects to clusters.

As we can see, the model tells us the total number of homes with a sale date during a given period for the major cities in the Tri-States. Rochester, NY has the highest number of houses sold not only in New York state but also across the Tri-State. We can also see that the residential housing market in New York state is much bigger than other two states as the number sold across different cities is substantially higher.

MODEL 10: TOTAL HOUSES SOLD PERCENTAGE
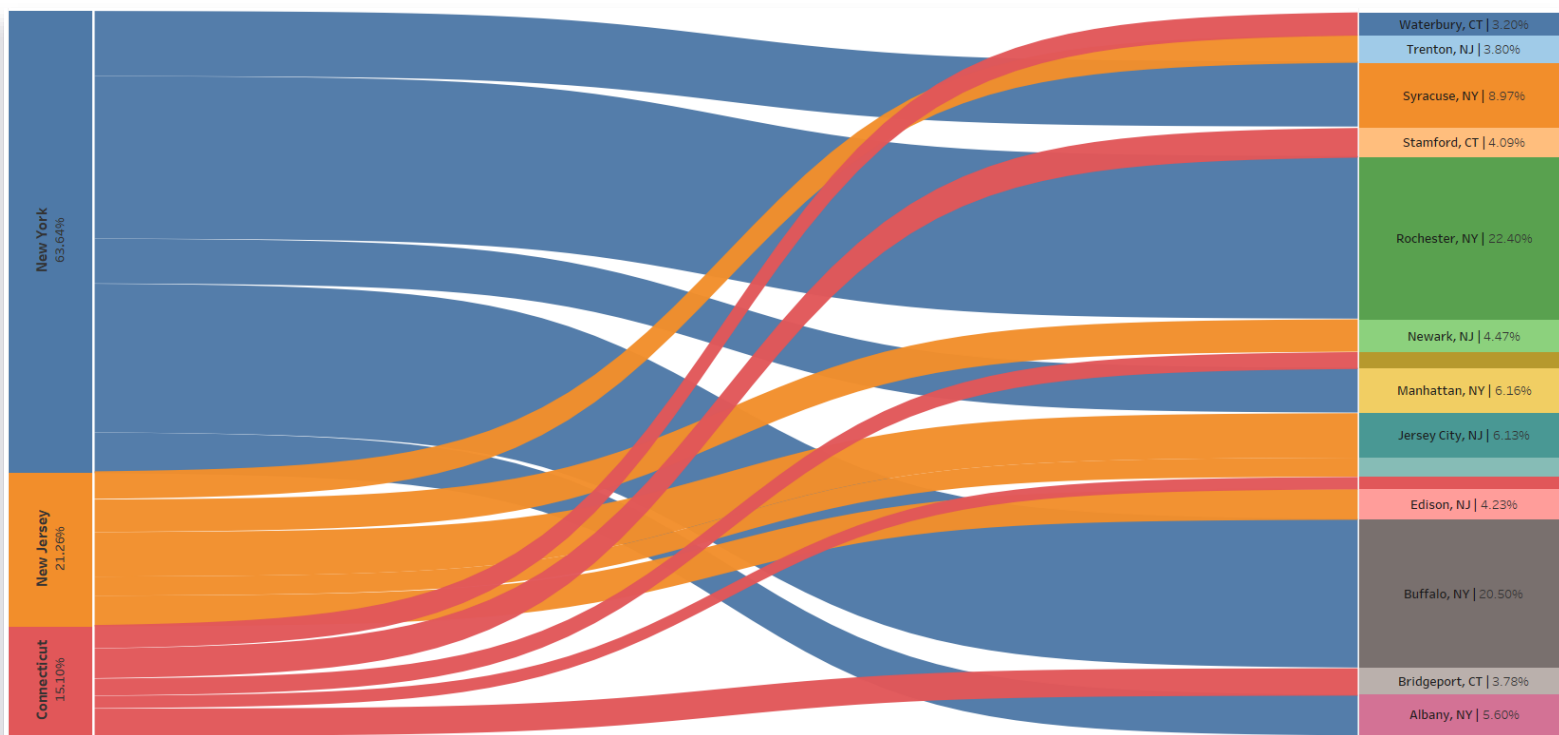
Tenth model is a **Sankey diagram** that we use to depict the distribution of **Percentage of Houses Sold**. Sankey chart diagram is a visualization used to depict a flow from one set of values to another. The things being connected are called nodes and the connections are called links. In Sankey diagrams, the width of the arrows is proportional to the flow quantity it represents. So, attention is drawn immediately to the most important flows in the processing system. It is best used when you want to show a many-to-many mapping between two categorical dimensions.

Here we have used the Sankey diagram to simply show how we can visually represent the percentage of houses sold can be divided at a state level across the major cities. For a future analysis, we could use this model to depict house the various real estate properties that could be residential or commercial units are spread across the Tri-State.

MODEL 11: SOLD ABOVE AND TO LIST PRICE
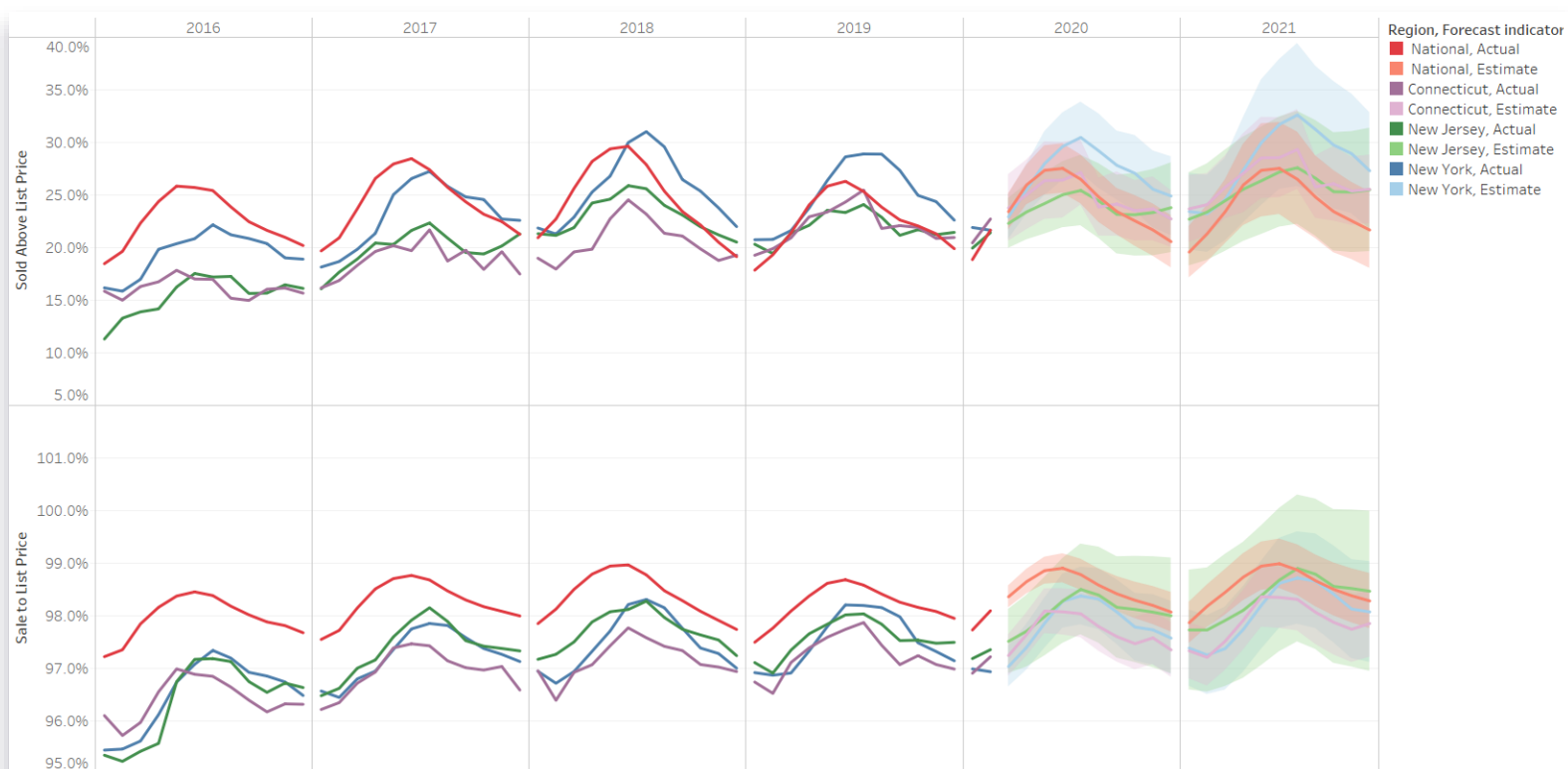
At last, the eleventh model is a **Line Chart combination with Forecasting** to compare the percentage of Houses that **Sold Above List Price** and the percentage of Houses that **Sold to List Price**. The line chart allows to depict the value of percentages over time while forecasting is a means of predicting what will happen in the future based on historical data. Though forecasting can be, to some extent, guesswork, most businesses use forecasting to help predict where the market will head.

Here the *Sold Above List Price* refers to the percentage of homes sales with a sale price greater than their latest list price covering all homes with a sale date during a given period and the *Sold to List Price* refers to the mean ratio of each home's sale price divided by their list price covering all homes with a sale date during a given period. In the model, it is estimated that New York cities will have a higher chance of selling above list price up to almost 40% higher while cities in New Jersey will have listing near to the list price and not very high.

EVALUATION AND DEPLOYMENT

As we could see, though each of the Tableau models varied from the other but the results of each model was very easy to understand and infer upon to gain excellent insights into the residential real estate market of NY Tri-State and how it differs when compared to the rest of the country. Though the models are based on the past data, but they can be used to evaluate the market and devise good strategies for a buyer or seller on how to handle and account for the fluctuations or trends in the market.

One thing to note is that these analytical models do not account for any situation that could be extraordinary say like the current pandemic situation or a recession that could affect the real estate markets.

In the year 2020, New York Tri-State region has been hit hardest by the pandemic and there is bound to be some correction with respect to the real estate market as well and we could see that towards the first quarter of 2020 as well that showed in the current market conditions, it seems a pricing plateau has arrived in the region. There is increased real sales activity, but no one can predict when this increased activity will begin to push prices upward. Hence these analytical models can be improved upon with further data sourcing with regards to the COVID-19 situation and be used to make broad predictions on how the market is going to move ahead in the coming years and buyers/sellers can be recommended when is the best time to enter or exit the real estate market.

Models Published

All the models shown in the report have been published on **Tableau Public** and are available for interaction at: **Real Estate Market across NY Tristate [2015-2020]**