Assignment: Data Prep/Cleaning

Introduction:

For this assignment and my final project, I am analyzing Zillow Housing Dataset to understand the housing market trend in the US.

Data Source Links

In this assignment and future analysis, the dataset I am referring to are from two different sources:

1. **ZILLOW:**

a) **Zillow Home Value Index (ZHVI):** Data Type = ZHVI; All Homes (SFR, Condo/Co-op); Time Series, Smoothed Seasonally Adjusted($), Geography = Metro & US.
https://files.zillowstatic.com/research/public_csvs/zhvi/Metro_zhvi_uc_sfrcondo_tier_0.33_0.67_sm_sa_month.csv?t=1676850425

b) **Inventory(For-Sale Inventory):** Data Type = For-Sale Inventory (Smoothed , All Homes, Monthly); Geography = Metro & US.
https://files.zillowstatic.com/research/public_csvs/invt_fs/Metro_invt_fs_uc_sfrcondo_sm_month.csv?t=1675705162

c) **Median Sale Price:** Data Type = Median Sale Price (Smooth & Seasonally Adjusted , All Homes, Monthly)); Geography = Metro & US.
https://files.zillowstatic.com/research/public_csvs/median_sale_price/Metro_median_sale_price_uc_sfrcondo_sm_sa_month.csv?t=1675705162

d) **Rental**: Data Type = ZORI Smoothed All Homes Plus Multifamily Time Series($); Geography = Metro & US.
https://files.zillowstatic.com/research/public_csvs/zori/Metro_zori_sm_month.csv?t=1676850425

2. **US CENSUS BUREAU:**

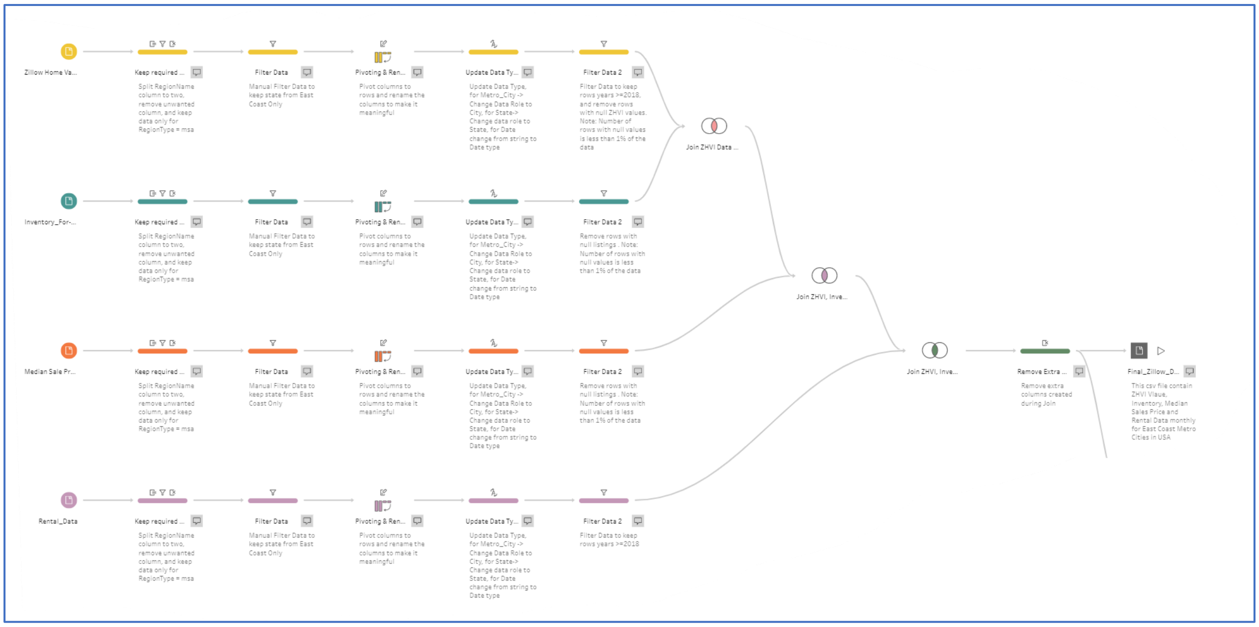
a) **Income Data:**
For the affordability analysis, I will also need the Income data. I will use the latest Income data from the US Census Bureau. The data is available for the period 2018 – 2021.
https://data.census.gov/table?t=Income+(Households,+Families,+Individuals)&g=0400000US09$1600000,10$1600000,12$1600000,13$1600000,23$1600000,24$1600000,25$1600000,33$1600000,34$1600000,36$1600000,37$1600000,44$1600000,45$1600000,51$1600000&y=2018&tid=ACSST1Y2018.S1902
In this assignment, I have considered data from Metro Cities from the East Coast States. Hence, the data I fetch from the US census bureau is only for the East Coast States. Please refer file with the naming convention as **ACSST5Y<YEAR>.S1902-Data.csv.** I have enclosed this data in the zip file.

Reading and cleaning your datasets

ZILLOW DATA:

The reading and cleansing steps for Zillow Data is as follows:

1. **Removed unwanted columns**: Removed the unwanted columns, like RegionID and SizeRank, in this step. The collection includes Country and MSA regions of two different sorts. I used a data filter and removed records where RegionType = Country. I deleted the RegionType column later. Then, I extracted the Metro City name by splitting the RegionName column, which contains the combination of the Metro City Name and State, into two. After finishing, I removed the divided column with a state and the original column.
2. **Filter Data**: I'm only focusing on Metro Cities from East Coast states for this analysis. States include Maine, New Hampshire, Massachusetts, Rhode Island, Connecticut, New York, New Jersey, Delaware, Maryland, Virginia, North Carolina, South Carolina, Georgia, and Florida. I manually selected all those states to filter the records.
3. **Pivoting and Renaming Columns:** In the data, the dates are columns. Thus, I converted the columns to rows using pivoting. I then rename every column to make it easier to understand.
4. **Update Data Type:** Next, I updated the data type of the Date column from string to Date. Then, I modified the data role to City and State for Metro City and Metro State, respectively.
5. **Filter Data:** The datasets related to ZHVI and Rental also have data before 2018. Thus, I applied a filter to gather data starting in 2018. In the tables, I also deleted rows with null values. It eliminated <1% of the rows.
6. **Join Date:** After cleaning the data, I inner-joined all the Zillow data in the order: City -> State -> Date.
7. **Remove Extra Columns**: In this step, I removed the extra columns introduced because of joining.
8. **Output:** The last phase involved producing a single file that contained all monthly Zillow Housing Statistics for Metro Cities for the East Coast States from 2018 through 2022. Final Output File Name: Final_Zillow_Data_Monthly.csv


Zillow Home Va...
Keep required ...
Split RegionName column to two, remove unwanted column, and keep data only for RegionType = msa
Filter Data
Manual Filter Data to keep state from East Coast Only
Pivoting & Ren...
Pivot columns to rows and rename the columns to make it meaningful
Update Data Ty...
Update Data Type, for Metro_City -> Change Data Role to City, for State-> Change data role to State, for Date change from string to Date type
Filter Data 2
Filter Data to keep rows years >=2018, and remove rows with null ZHVI values. Note: Number of rows with null values is less than 1% of the data
Join ZHVI Data ...
Inventory_For-...
Keep required ...
Split RegionName column to two, remove unwanted column, and keep data only for RegionType = msa
Filter Data
Manual Filter Data to keep state from East Coast Only
Pivoting & Ren...
Pivot columns to rows and rename the columns to make it meaningful
Update Data Ty...
Update Data Type, for Metro_City -> Change Data Role to City, for State-> Change data role to State, for Date change from string to Date type
Filter Data 2
Remove rows with null listings . Note: Number of rows with null values is less than 1% of the data
Join ZHVI, Inve...
Median Sale Pr...
Keep required ...
Split RegionName column to two, remove unwanted column, and keep data only for RegionType = msa
Filter Data
Manual Filter Data to keep state from East Coast Only
Pivoting & Ren...
Pivot columns to rows and rename the columns to make it meaningful
Update Data Ty...
Update Data Type, for Metro_City -> Change Data Role to City, for State-> Change data role to State, for Date change from string to Date type
Filter Data 2
Remove rows with null listings . Note: Number of rows with null values is less than 1% of the data
Join ZHVI, Inve...
Remove Extra ...
Remove extra columns created during Join
Final_Zillow_D...
This csv file contain ZHVI Vlaue, Inventory, Median Sales Price and Rental Data monthly for East Coast Metro Cities in USA
Rental_Data
Keep required ...
Split RegionName column to two, remove unwanted column, and keep data only for RegionType = msa
Filter Data
Manual Filter Data to keep state from East Coast Only
Pivoting & Ren...
Pivot columns to rows and rename the columns to make it meaningful
Update Data Ty...
Update Data Type, for Metro_City -> Change Data Role to City, for State-> Change data role to State, for Date change from string to Date type
Filter Data 2
Filter Data to keep rows years >=2018

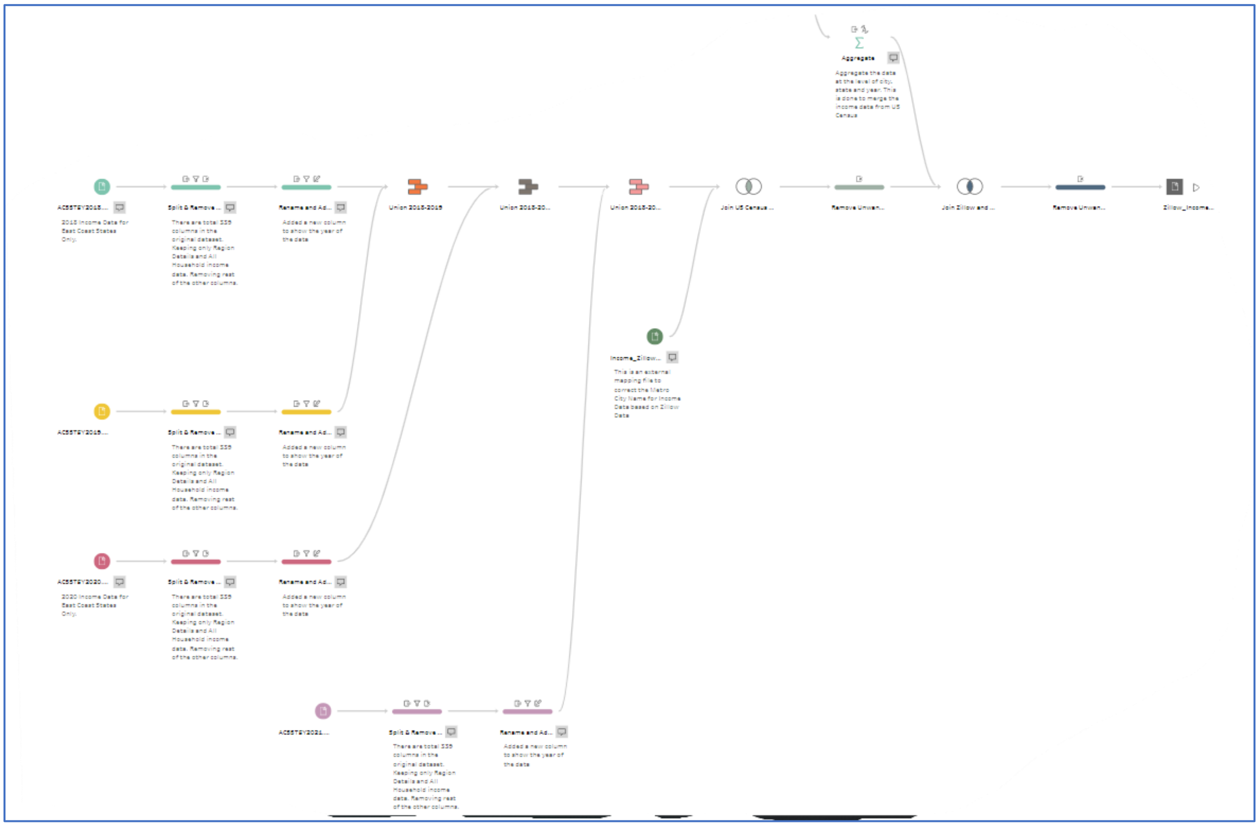
<u>US CENSUS BUREAU DATA:</u>

The reading and cleansing steps for US Census Bureau Data are as follows:

1. I have income data for the years 2018-2021. They are in four separate files with the following naming convention: **ACSST5Y<YEAR>.S1902-Data.csv**
2. Cleaning steps are same for all the files.
   Note*: The original data file for 2021 contains duplicate records for every city for some reason. I thus clean this manually.*
3. Each file contains 339 columns. So, in the first step, while reading the file, I kept only the relevant columns: Name and Mean Income in Dollars.
4. **Split & Removed Columns**: The name column is a mix of City and State in the name column. I split the column to get the city and state names. Moreover, the suffix "CDP" is present at the end of the city name. I removed it using a calculated field. Then, I removed any extra columns that were added during this process but weren't necessary.
5. **Rename and add new column:** In this step, I renamed all the columns to make them easier to understand. I also added two columns in the data, Year and State code.
6. **Union Data:** Next step is to append all the data from Year 2018-2021.
7. Zillow Data and US Census Bureau have discrepancies in their city names. To merge the data, manual intervention is needed. Hence, I created a file manually to align the city names. I have enclosed the manual file in the zip folder.
   Income_Zillow_City_ST_Mapping.csv
8. **Remove Unwanted Columns**: Removed columns that are redundant and not required.
9. **Aggregate**: Aggregated the Zillow data at the City, State, and Date levels.
10. **Join:** Joined the Zillow Data and US Census Bureau Data. Created an inner join with Condition on columns City -> State -> Date(year).

9. **Remove Unwanted Columns :** Removed all the extra columns that had appeared because of joining.
11. **Output File:** In the final step, I created an output file containing data from Zillow and US Census data. This file has yearly data from the years 2018-2021.

Final output file name: Zillow_Income_Data_State.csv



Aggregate the data at the level of city, state and year. This is done to merge the income data from US Census

2018 Income Data for East Coast States Only.

There are total 339 columns in the original dataset. Keeping only Region Details and All Household income data. Removing rest of the other columns.

Added a new column to show the year of the data

This is an external mapping file to correct the Metro City Name for Income Data based on Zillow Data

There are total 339 columns in the original dataset. Keeping only Region Details and All Household income data. Removing rest of the other columns.

Added a new column to show the year of the data

2020 Income Data for East Coast States Only.

There are total 339 columns in the original dataset. Keeping only Region Details and All Household income data. Removing rest of the other columns.

Added a new column to show the year of the data

There are total 339 columns in the original dataset. Keeping only Region Details and All Household income data. Removing rest of the other columns.

Added a new column to show the year of the data

Decide whether you are merging your datasets or keeping them separate.

I am merging the data and creating two files:

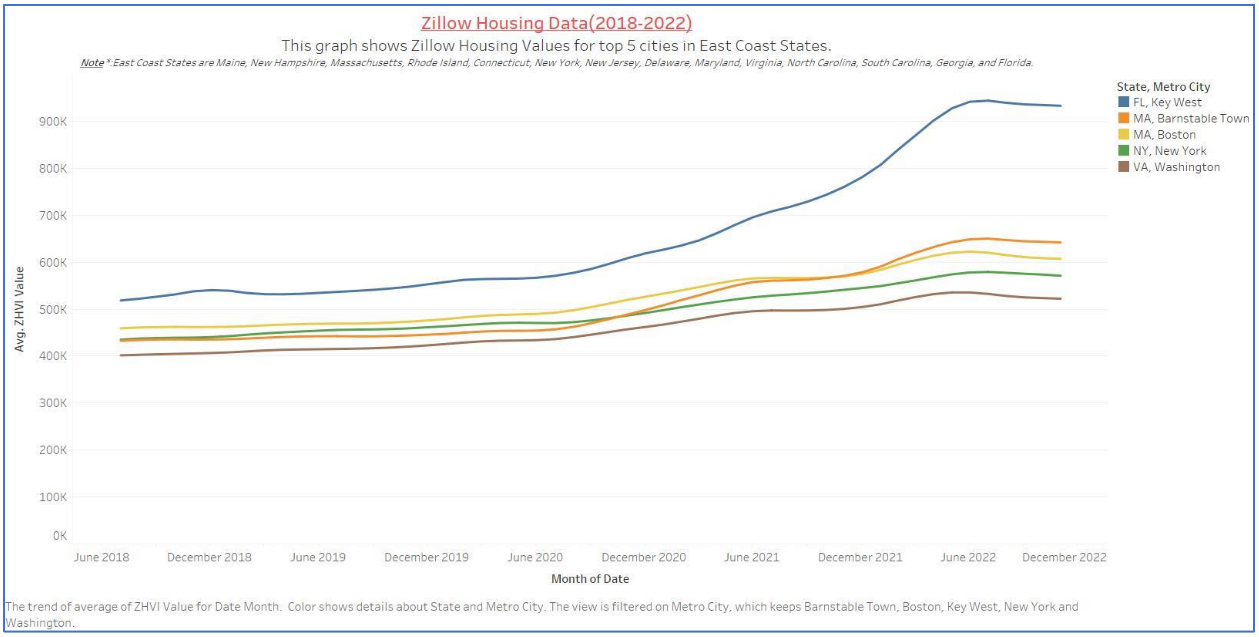
1. **Final_Zillow_Data_Monthly.csv**: This data file includes all Zillow housing metrics for the East Coast metro areas from 2018 to 2022. This data is at the monthly level.
2. **Zillow_Income_Data_State.csv**: This data file includes all Zillow housing metrics for the East Coast metro areas and Income data from US Census Bureau from 2018 to 2021. This data is at the yearly level because Income data is available only at the yearly level.

Two initial visualizations that describe your data.

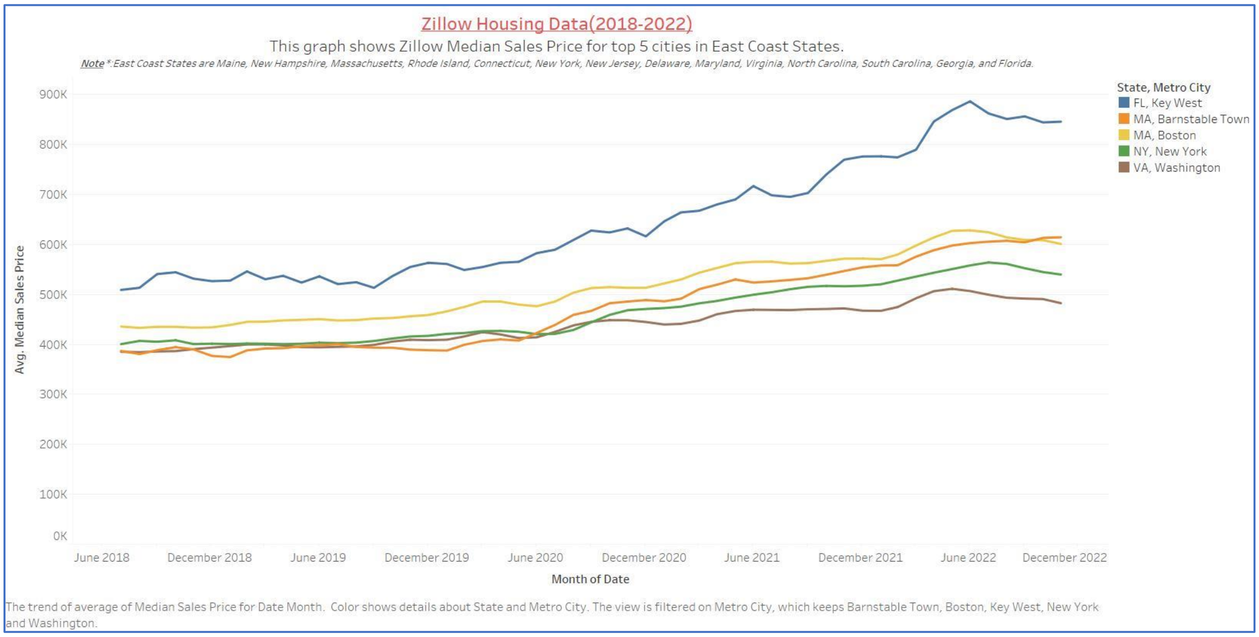
For exploratory analysis, I am considering the top 5 cities in the East Coast States based on their housing values. The top 5 cities are as follows:

- Keywest,FL
- Barnstable Town, MA
- Boston, MA
- New York,NY
- Washington, VA

1. **Zillow Housing Value Index**:
   In the graph below, we can see that Housing Value has increased over the last few years, as expected:



2. **Median Sales Price**:
   Similar to Housing Value Index, the Median Sales Price of a house has also increased in past years.



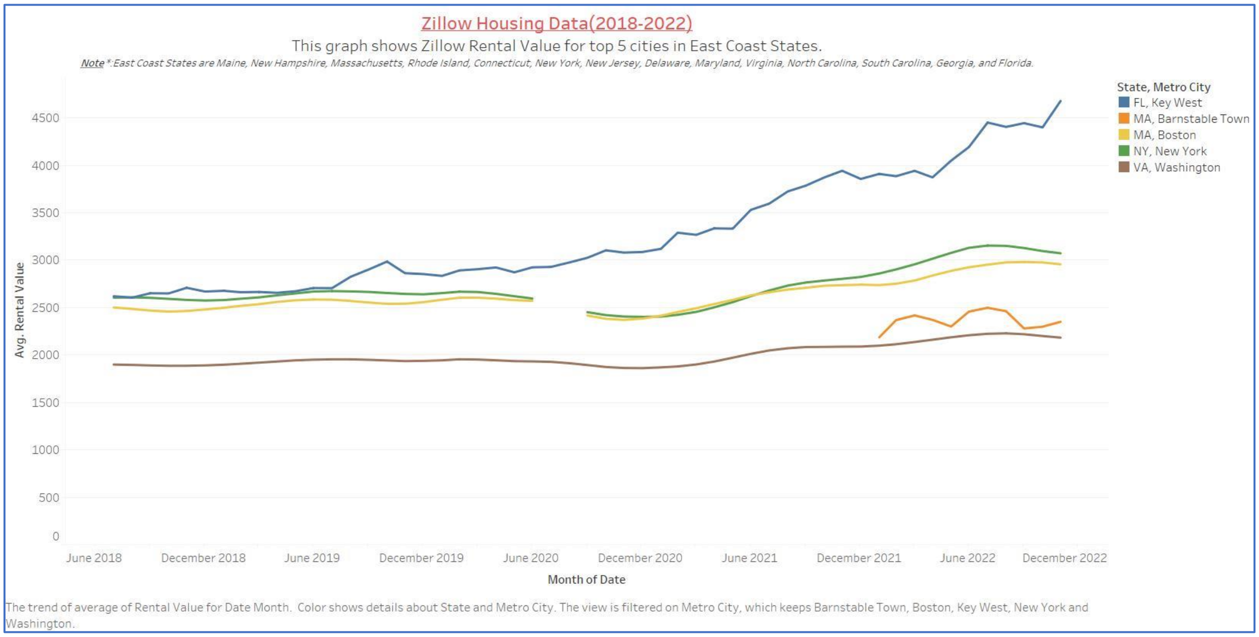
Zillow Housing Data(2018-2022)

This graph shows Zillow Median Sales Price for top 5 cities in East Coast States.

*Note *:East Coast States are Maine, New Hampshire, Massachusetts, Rhode Island, Connecticut, New York, New Jersey, Delaware, Maryland, Virginia, North Carolina, South Carolina, Georgia, and Florida.*

The trend of average of Median Sales Price for Date Month. Color shows details about State and Metro City. The view is filtered on Metro City, which keeps Barnstable Town, Boston, Key West, New York and Washington.

3. **Rental Values:**
   Rental Values have also increased in past years.



Zillow Housing Data(2018-2022)

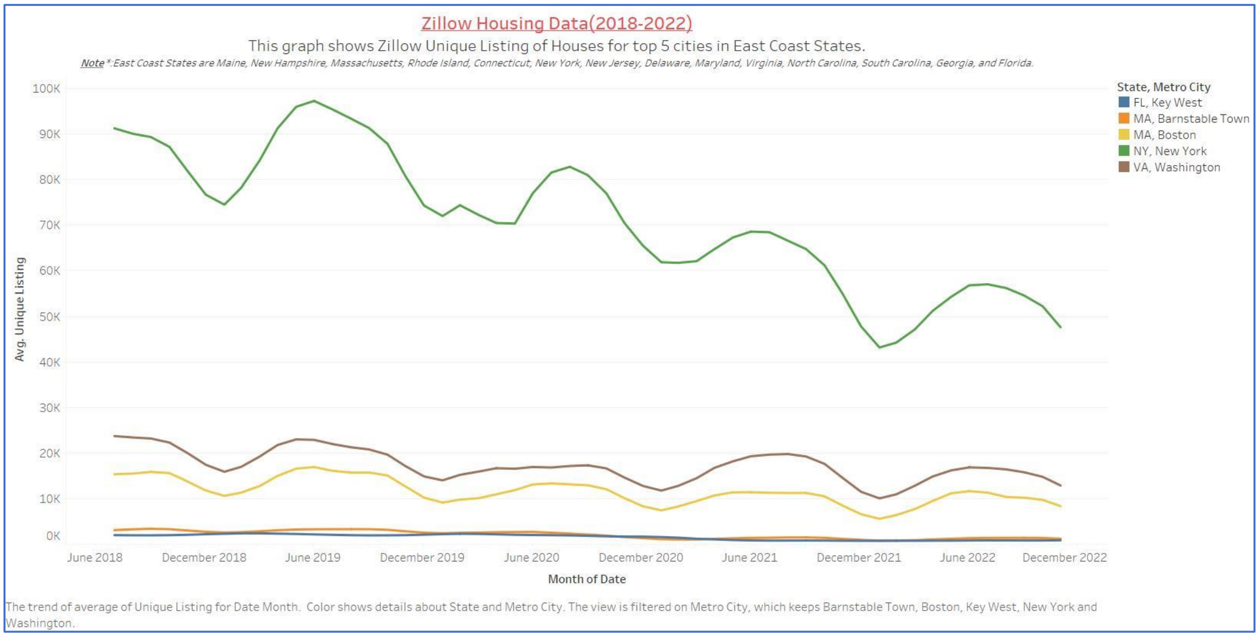
This graph shows Zillow Rental Value for top 5 cities in East Coast States.

*Note *:East Coast States are Maine, New Hampshire, Massachusetts, Rhode Island, Connecticut, New York, New Jersey, Delaware, Maryland, Virginia, North Carolina, South Carolina, Georgia, and Florida.*

The trend of average of Rental Value for Date Month. Color shows details about State and Metro City. The view is filtered on Metro City, which keeps Barnstable Town, Boston, Key West, New York and Washington.

Note: The gap is due to missing rental data for those months in the original dataset.

4. **Unique Listing**:
   In Contrast to above mentioned Housing Metrics, the listing of new houses has decreased in past years.



Zillow Housing Data(2018-2022)

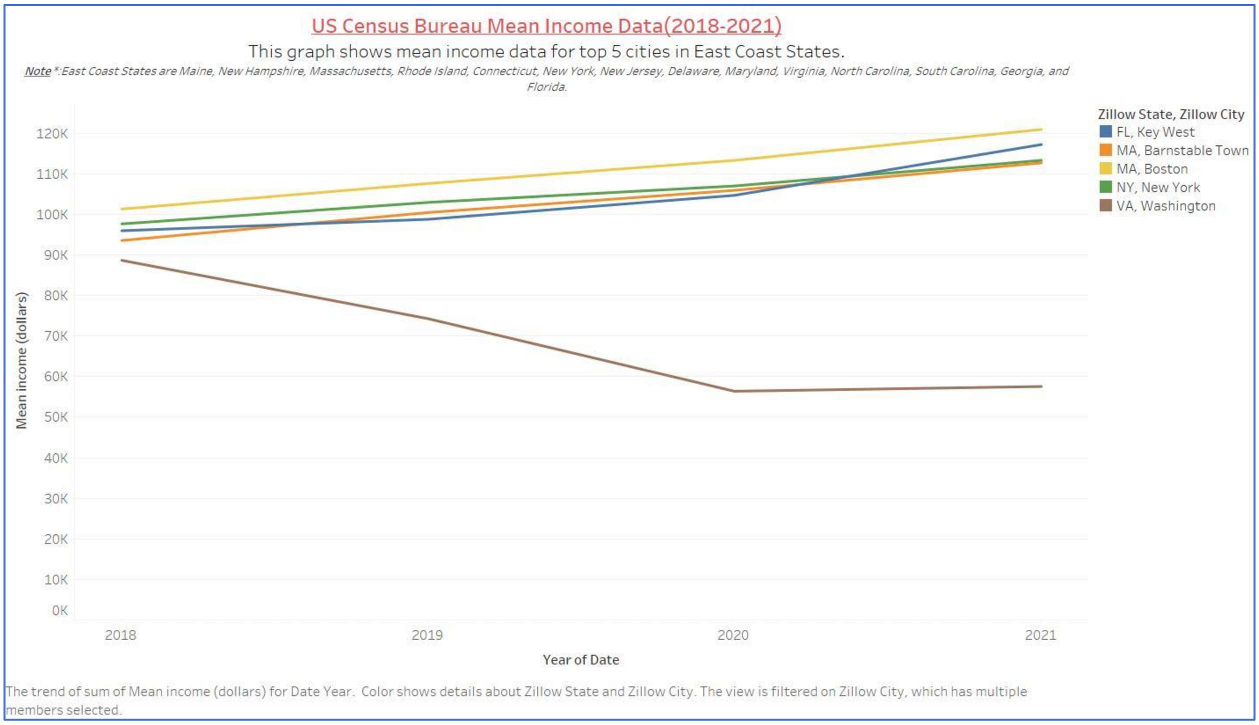
This graph shows Zillow Unique Listing of Houses for top 5 cities in East Coast States.

*Note * East Coast States are Maine, New Hampshire, Massachusetts, Rhode Island, Connecticut, New York, New Jersey, Delaware, Maryland, Virginia, North Carolina, South Carolina, Georgia, and Florida.*

The trend of average of Unique Listing for Date Month. Color shows details about State and Metro City. The view is filtered on Metro City, which keeps Barnstable Town, Boston, Key West, New York and Washington.

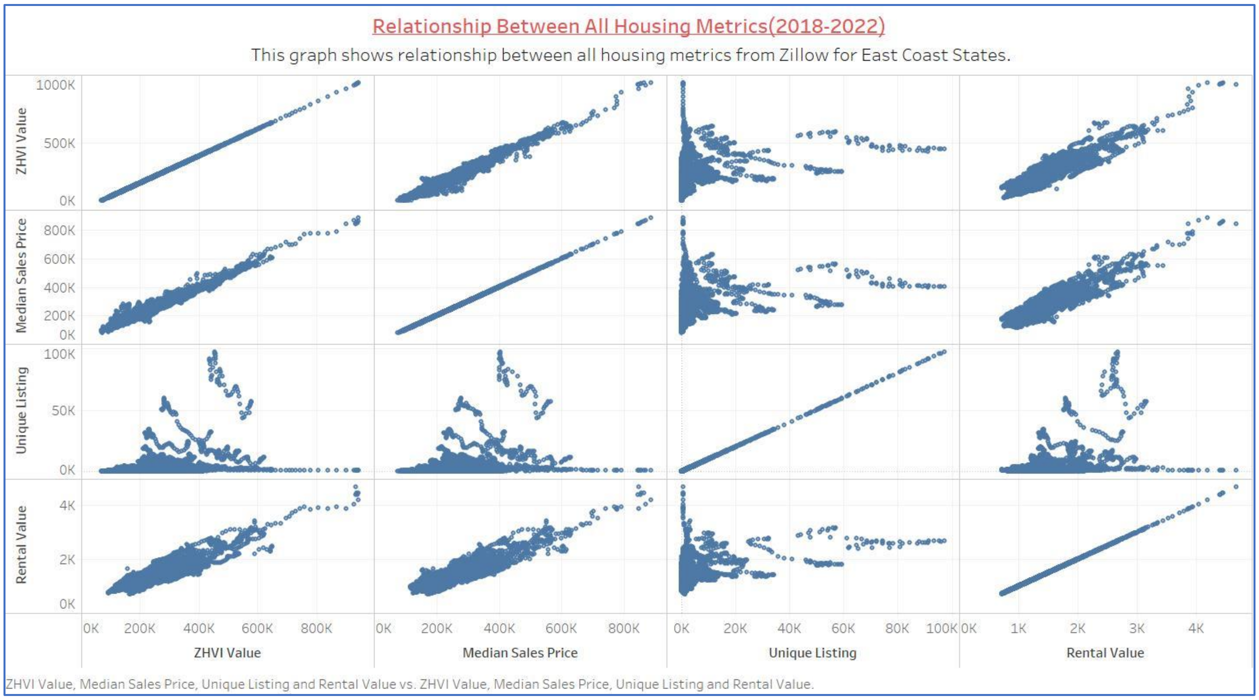
5. **Income Distribution**

This graph shows the income distribution for the last four years in the top 5 cities of East Coast State:



6. **Relationship between All Housing Metrics**:

For this graph, I considered all the data from all the cities to observe the trend between different housing metrics. As expected, there is a positive relation between Housing Value, Median Sales Price, and Rental Values. Whereas, Unique listing has a negative relationship with all other metrics.



Relationship Between All Housing Metrics(2018-2022)

This graph shows relationship between all housing metrics from Zillow for East Coast States.

ZHVI Value, Median Sales Price, Unique Listing and Rental Value vs. ZHVI Value, Median Sales Price, Unique Listing and Rental Value.

Refine the 3-4 questions.

As per the feedback received last time, this is how I am planning to refine the questions:

5.1 **What will be the home value trends for the next three years (2023 - 2025)?**

--> excellent question

5.2 **In which cities is it better to rent vs. buy?**

--> What is your criteria for "better" here?

My criteria for "better" is to identify the cheaper option between buying and renting. I will calculate the monthly mortgage payment with an assumed (and reasonable) annual Interest rate (5%) and Years of the mortgage (30 years). Then, I will then compare it with the average rental payment from the Zillow data to determine if it is cheaper to buy or rent in a particular city.

For simplicity, I am not considering the additional expenses of owning a house, like property taxes, home insurance, maintenance, etc.

5.3 **What are the emerging cities/towns for property investment?**

--> How are you planning to measure this?

--> It also looks like a lot of data, so you might want to consider downsizing and perhaps tackling a single state or a few states (eg the East coast) to narrow your scope.

Thank you for your suggestion. I have narrowed the list of cities/towns to the top cities/towns on the East coast. For the measurement, I plan to look at the Compound annual growth rate (CAGR) of the Zillow Housing Value Index (ZHVI) between 2018 and 2022. The cities/towns with the highest CAGR will be a potential list for real estate investors to invest in properties.

5.4 **What will be the inventory patterns? Are there regions with a critically low inventory?**

---> What ML will you use? Regression?

Correct. I plan to use the Regression technique to determine the inventory trend for the key cities/towns. Then, using this model, I will predict the inventory for the next month and identify which cities/towns will have an inventory increase or decrease.

5.5 **What will be the state of housing affordability, and whether policy intervention required?**

---> Maybe you can turn this last question into a clustering question.

Thank you for your suggestion. I'll convert this into a clustering problem. I plan to calculate affordability as % of income spent on the mortgage (calculated based on assumed interest rate and mortgage years) and cluster the cities based on affordability.