US Housing Market Trends

Department of Applied Data Science, San Jose State University

DATA 228: Big Data Technologies & Applications

Professor: Andrew H. Bond

December 13, 2021

By,

Team Dive In Data

Aswini Pusuluri - 015310269

Heer Parekh - 015270320

Sowmya Ravichandran - 015337400

Vamshi Krushna Lakavath - 015351310

Abstract

The housing market in the United States is one of the quickest developing areas and most secure spots to contribute for many people.. The real estate market has encountered critical repetitive instability in the course of the last a quarter century because of major underlying changes and financial vacillations. These days, we can see a great deal of instability in the real estate market. It may very well be because of the COVID-19 pandemic, telecommuting, rising pay in few industries, lack of inventory, high timber costs, and record-low home loan rates. Housing prices have touched an all-time high in most US cities while the affordability issue is worsening and increasing job losses. The sudden growth of house prices looks like a bubble because in the past we have seen some housing market crashes like in the 1980s and 2008. The latest housing bubble was observed in 2008. Housing prices touched the peak in 2016 and started falling from early 2007 and reached a record low in 2012. Our main objective in this project is to analyze the changes in trends during the period of five years from 2016 to 2021. Through our analysis, we will find the market's steadiness and observe the overall average growth of the US housing market from 2016 to 2021.

Keywords: US Housing Market, COVID-19, Average Growth, Bubble

Table of Contents

Abstract 3

1. Introduction 5
- 1.1 Project Background 5
- 1.2 Project Goals and Motivation 5
- 1.3 Project Application and Impact 5
- 1.4 Expected Deliverables 6

2. Background and Related Work 7
- 2.1 Literature Review 7

3. Data Preparation 8
- 3.1 Data Description 8
- 3.2 Data Cleaning and Exploration 13

4. Architecture 18
- 4.1 AWS S3 19
- 4.2 AWS Glue 19
- 4.3 AWS Redshift 20
- 4.4 AWS Sagemaker 20
- 4.5 Tableau 20

5. Security Management Services 21
- 5.1 Identity And Access Management 21
- 5.2 Virtual Private Cloud 21
- 5.3 AWS Secret Manager 22
- 5.4 Endpoint 22

6. Implementation 23

7. Data Visualization 30
- 7.1 Charts 31

8. Conclusion 42

9. Future work 43

10. References 44

1.Introduction

1.1 Project Background

The United States Housing Market is one of the fastest-growing sectors and safest places to invest.Nowadays, we can see a lot of volatility in the housing market. Housing prices have touched an all-time high in most US cities.The sudden growth of house prices looks like a bubble because in the past we have seen some housing market crashes like in the 1980s and 2008.So we want to analyze the housing market data.For that, We have used latest US housing market data from 2016 to 2021.For analyzing the prices with demand we have collected US cities' population from 2016 to 2021.

1.2 Project Goals and Motivation

The main purpose of this project is to answer the following analytical questions in order to give insights into US market trends and help investors make better decisions. The price variations from 2016 to 2021 are examined first. The percentage change in the market trend over the last five years might help determine if property prices have risen or fallen. It's also crucial to know which cities are the cheapest and most expensive to buy a property in terms of price and region. There were also major swings during the COVID-19 pandemic, which had a huge impact on the property market because of a big demand-supply discrepancy. It is necessary to investigate the link between population and price per square foot and also a recommendation for the ideal months to buy real estate.

1.3 Project Application and Impact

This project is targeted to any person who is looking forward to making his investments in the housing market. Because the market is so unpredictable, looking at previous patterns might help you predict how the market will respond to the current scenario. House prices are soaring as

a result of high demand and limited supply. Although there are just a few residences on the market, the rivalry to invest has become fierce. Investors are debating whether or not to invest at this high price, and whether or not they will be able to benefit in the future. This initiative will assist investors in making more informed judgments and prudent investments.

1.4 Expected Deliverables

The deliverables of this project is to create an AWS pipeline for processing massive numbers of housing data. Because the data is expected to be available in any format, it's critical that we preprocess it so that we can better see market movements. Since this market is always changing, the data collected is loaded into the cloud services, the pipeline is executed, and the visualizations provide immediate insights. A report detailing the complete pipeline with screenshots.

2. Background and Related Work

2.1 Literature Review

According to Kouwenberg et al tendencies.'as in the US housing market (2011), fundamentalists and chartists are two types of investors. Fundamentalists believe that based on rentals, the housing price will revert to its basic value, whereas chartists extrapolate historical price patterns. An Empirical Heterogeneous Agent Model for the Housing Market was used by the author to investigate the dynamics of housing prices and to produce better housing market projections for policymakers, lenders, and consumers. They used the Freddie Mac repeat-sales house price index dataset to conduct their research.The assessed model can create blast and busts cycles endogenously, initiated by the boundedly levelheaded conduct of the financial backers. In spite of the fact that the model is amazingly straightforward and adapted in nature, it can conjecture the decay of the public U.S. house value list from 2006 onwards. The heterogeneous specialist model beats a few notable benchmark models in an appraisal of viewing for out-of-test figures. Heterogeneous specialist models may in this manner not simply
be of hypothetical premium, yet additionally a helpful anticipating apparatus for real estate market members furthermore.

3. Data Preparation

3.1 Data Description

For this project, we have used two separate data sources for our analysis. It consists of historical data of fifteen years. The first data source is "US Real Estate Market Trends from 2016 to 2021" on the Kaggle website. The Dataset name is RDC_Inventory_Core_Metrics_Zip_History.csv. It has data of the US cities housing market from July 2006 to July 2021, this dataset has a total of 40 columns and 900 thousand records. The fields of the dataset are provided below.

Month_date_yyyymm: Represents the date on which the data is updated

Postal_code: Postal code is the postal code of a city

Zip_name: Represents the name of the city or area

Flag: Data values are outside of their typical range.

Median_listing_price: The median listing price of the houses listed within the specified geographic area in that month.

Median_listing_price_mm: The percentage change in the median listing price from the previous month in that area.

Median_listing_price_yy: The percentage change in the median listing price from the previous year in that area.

Active_listing_count: The count of active listings within the specified geography

during the specified month. The active listing count tracks the number of for-sale properties on the market.

Active_listing_count_mm: The percentage change in the active listing count from the previous month in that area.

Active_listing_count_yy: The percentage change in the active listing count from the previous year in that area.

Median_days_on_market: The median number of days a property listing stays on the market within the specified geography during the specified month.

Median_days_on_market_mm: The percentage change in the median number of days from the previous month in that area.

Median_days_on_market_yy: The percentage change in the median number of days from the previous year in that area.

New_listing_count: The number of new listings added to the market within the specified geography.

New_listing_count_mm: The percentage change in the new listing count from the previous month.

New_listing_count_yy: The percentage change in the new listing count from the same month in the previous year.

Price_increased_count: The number of listed houses that have had their price

increased within the specified geography in that month.

Price_increased_count_mm: The percentage change in the price increase count from the previous month**.**

Price_increased_count_yy: The percentage change in the price increase count from the same month in the previous year.

Price_reduced_count: The number of listed houses that have had their price reduced within the specified geography in that month.

Price_reduced_count_mm: The percentage change in the price reduced count from the previous month.

Price_reduced_count_yy: The percentage change in the price reduced count from the previous year.

Pending_listing_count: The number of pending listings within the specified geography during the specified month.

Pending_listing_count_mm: The percentage change in the pending listing count from the previous month.

Pending_listing_count_yy: The percentage change in the pending listing count from the same month in the previous year.

Median_listing_price_per_square_foot: The median listing price per square foot within the specified geography during that month.

Median_listing_price_per_square_foot_mm: The percentage change in the median listing price per square foot from the previous month.

Median_listing_price_per_square_foot_yy: The percentage change in the median listing price per square foot from the same month in the previous year.

Median_square_feet: The median listing price per square feet within the specified geography during that month.

Median_square_feet_mm: The percentage change in the median listing square feet from the previous month.

Median_square_feet_yy: The percentage change in the median listing square feet from the same month in the previous year.

Average_listing_price: The average listing price within the specified geography during that month.

Average_listing_price_mm: The percentage change in the average listing price from the previous month.

Average_listing_price_yy: The percentage change in the average listing price from the same month in the previous year.

Total_listing_count: The total of both active listings and pending listings within the specified geography during that month.

Total_listing_count_mm: The percentage change in the total listing count from the previous

month.

Total_listing_count_yy: The percentage change in the total listing count from the same month in the previous year.

Pending_ratio: The ratio of the pending listing count to the active listing count within the specified geography during that month.

Pending_ratio_mm: The change in the pending ratio from the previous month.

Pending_ratio_yy: The change in the pending ratio from the same month in the previous year.

The second dataset that we used for this project is Population. That has been collected from the UScensus.com website. The dataset consists of the population of cities of the US in a specified month of the specified year. Primarily this dataset has more than 60 columns. Each month of the year is a column that has been transposed to rows for each city of the state. So we are left with the below attributes.

Region: Region a city belongs to

Name: Name of the city

StateName: State of the city.

Date: Date on which data has been collected

Population: Total number of population as on specified date

RegionID: ID for the region.

SizeRank: Rank of the city as per its population.

RegionType: Region type a city placed in.

The data model of our project is provided in figure 5. It Shows the overall view of the entire data structure. It consists of the entities and their relationships with the datasets. It gives a One-to-Many relationship. Because the dataset includes denormalized data, the normalization method was used to have one fact and one dimension table.

3.2 Data Cleaning and Exploration

The data for this project is collected from Kaggle.com and USCensus.com. Datasets are thoroughly examined for feature extraction and data quality check. The raw datasets were uncleaned and unprepared so they needed a lot of cleaning and preparation. Categorizing linear data into distinct categories and maintaining categorical variables are examples of data preparation processes. The housing market dataset was collected from kaggle.com, and the Population dataset is collected from UScensus. The housing market dataset has a lot of missing values, null values, and unnecessary fields. Modification and cleaning will be required because the datasets contain textual, numerical, null, and zero-valued records. So for data cleaning and data preparation we have used Tableau Prep Builder, JupyterLab, and AWS Glue. In the first stage, we have used Tableau Prep Builder for cleaning, transforming, splitting, and merging the datasets. Tableau is a software program that makes data preparation simple and straightforward. Tableau Prep Builder is a tool for aggregating, organizing, and classifying data in preparation for analysis. At this stage Primarily, We have split the region field it has city name and state_id combined so we split it into two separate columns, then cleaned the housing dataset, then transformed datatypes of the fields, and gave the geographical roles to the city and state fields. Next, we have pivoted the Population dataset using the column to row transformation and allotted geographical roles to the city and state fields and city and state/region. Lastly, we have

merged both files on common fields city, date, and state. The data preparation and merging flow of the two datasets are provided in

Figure 2. The figure was taken from the Tableau Prep Builder.

Figure 1

ER Diagram

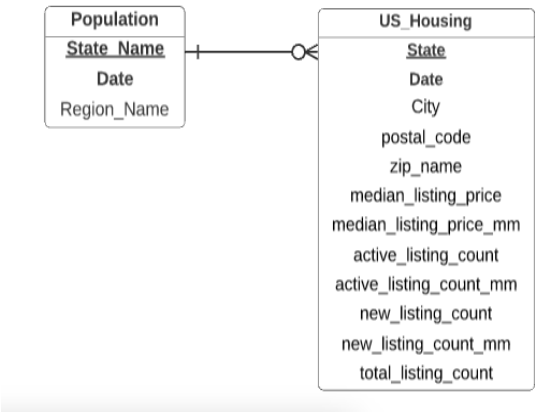
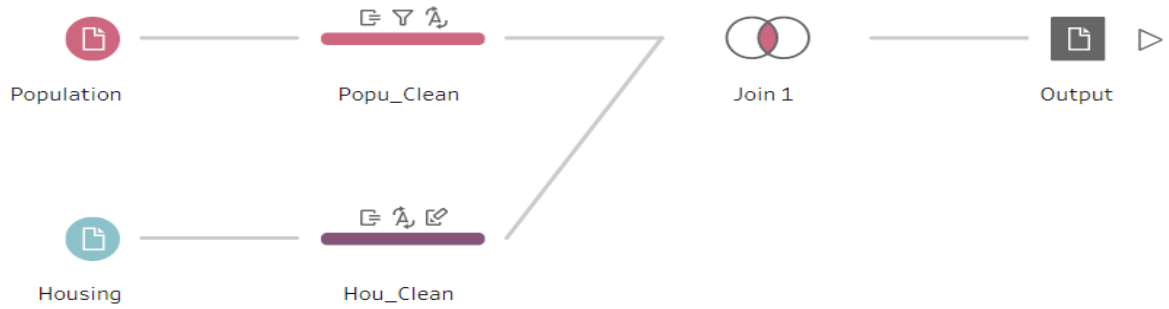


Figure 2

Data preparation and merging flow of the two datasets



In the second stage, we have used the Jupyter Notebook for Exploratory Data Analysis and Data Preparation. At this stage, we have done data analysis using some visualizations like histograms and scatter plots, dropped unnecessary and repeated columns, and replaced null values with mean, median, and standard deviation. We have dropped fields like Pending

ratio_mm, Pending ratio_yy, City-1, State-1, Date-1, and Zip name because they are not useful for our analysis. The fields in which we have replaced null values with mean, median, and standard deviation are Median_listing_price_mm, Median_listing_price_yy, Active_listing_count_mm, and Active_listing_count_yy. Also, we have tried to find the ratios between the increase and decrease in prices since the previous month and year. Result being that compared to previous years, there has been an increase in prices by 65.38%.

Figure 3

Pair Plots to see the Relationship Between Two Attributes

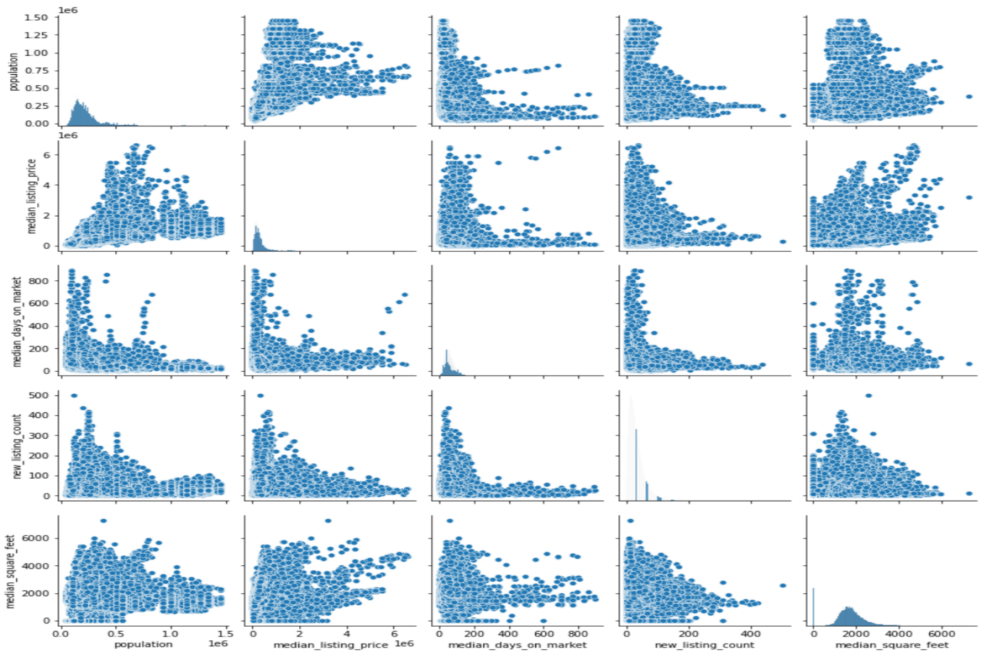
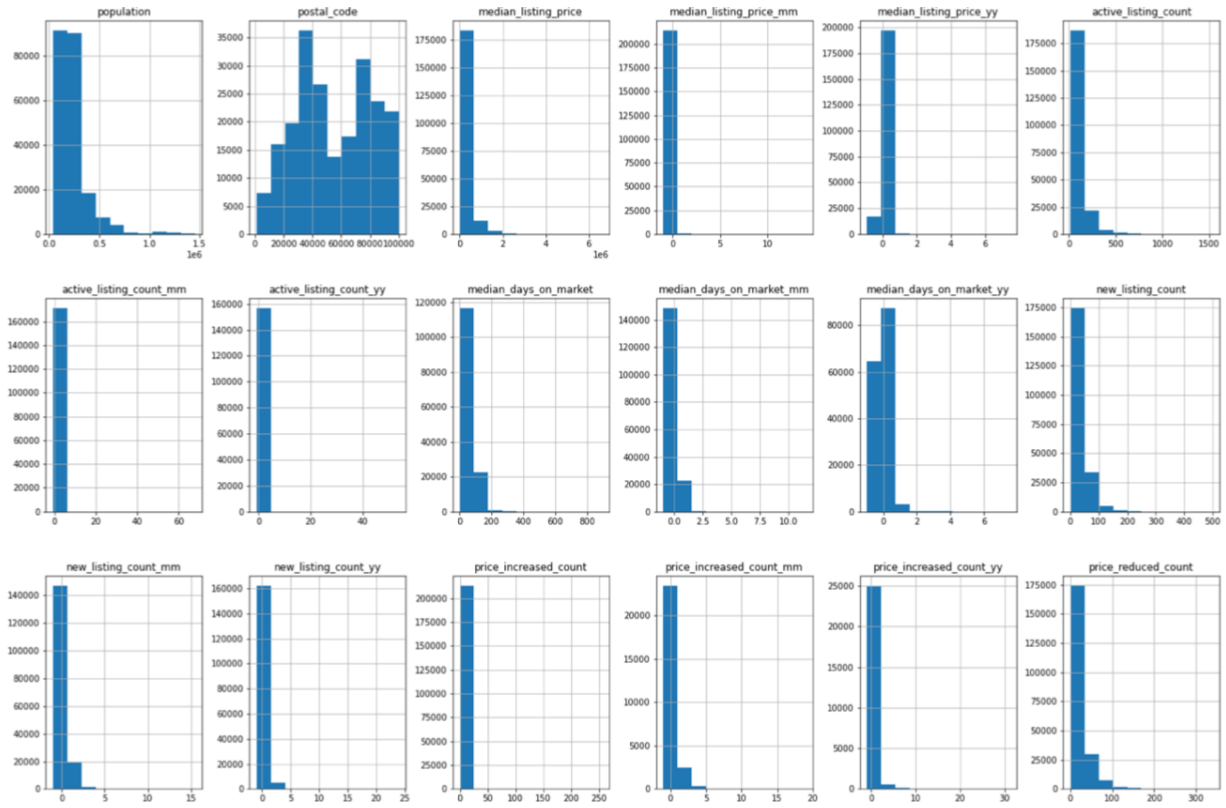


Figure 4

Histograms of Individual Attribute



The figure 5 shows the box plot graph of state housing price range comparison. As you can observe Massachusetts state has the highest variance housing price range whereas Delaware has the lowest. It can also be observed that California boasts max house prices. Apart from that it can also be observed that the states with maximum urbanisation boast the top house prices in the United States.

Figure 5

States housing price range comparison

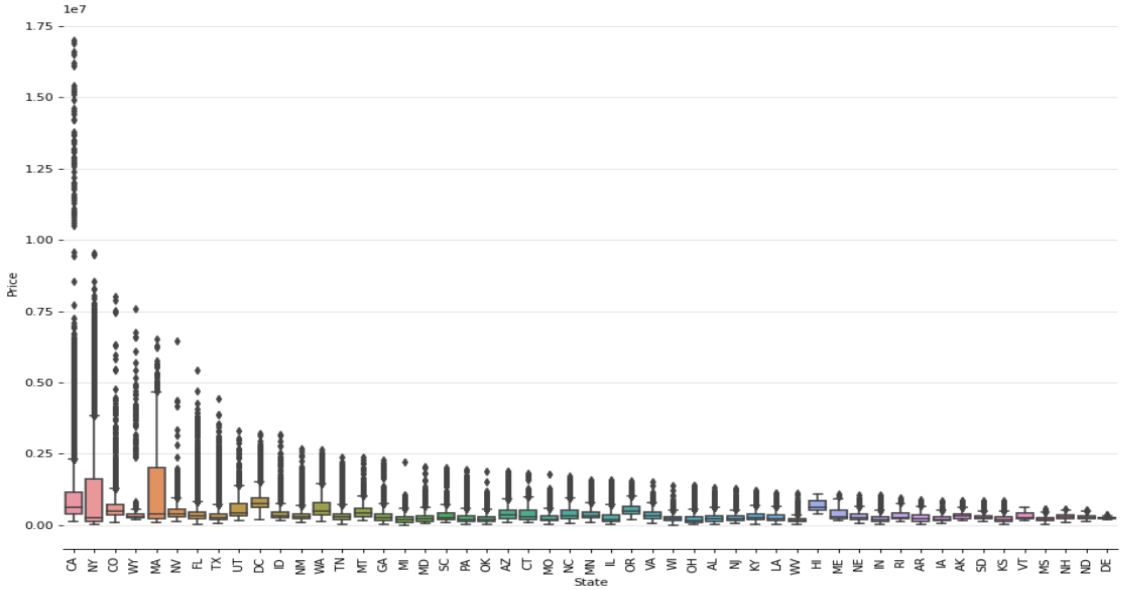
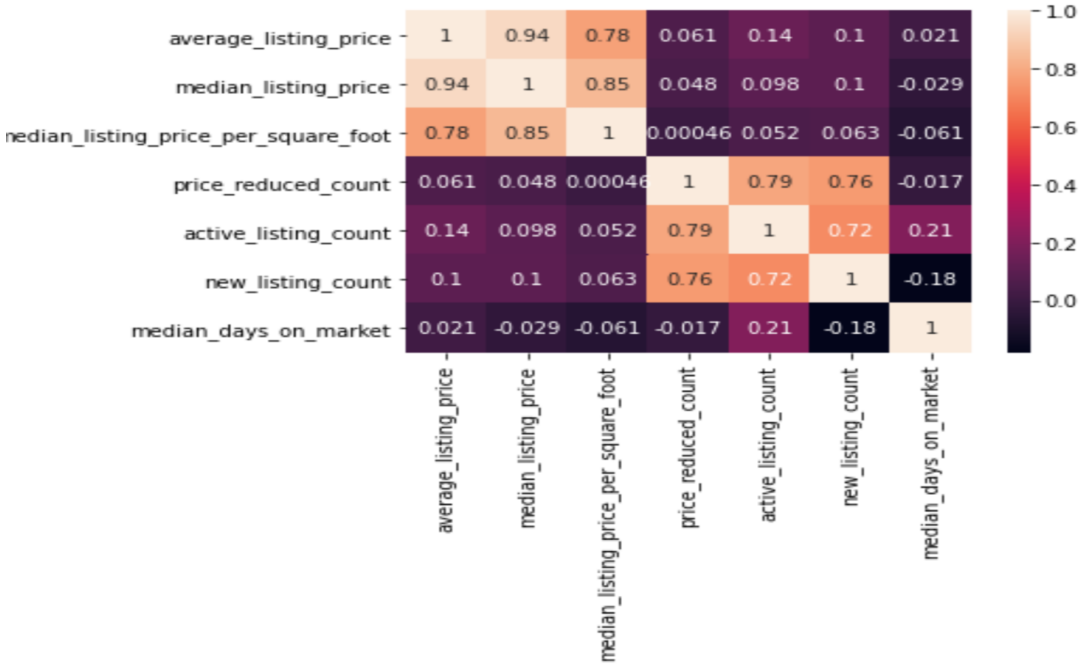


Figure 6

Heatmap to Determine the Correlation Between the Attributes

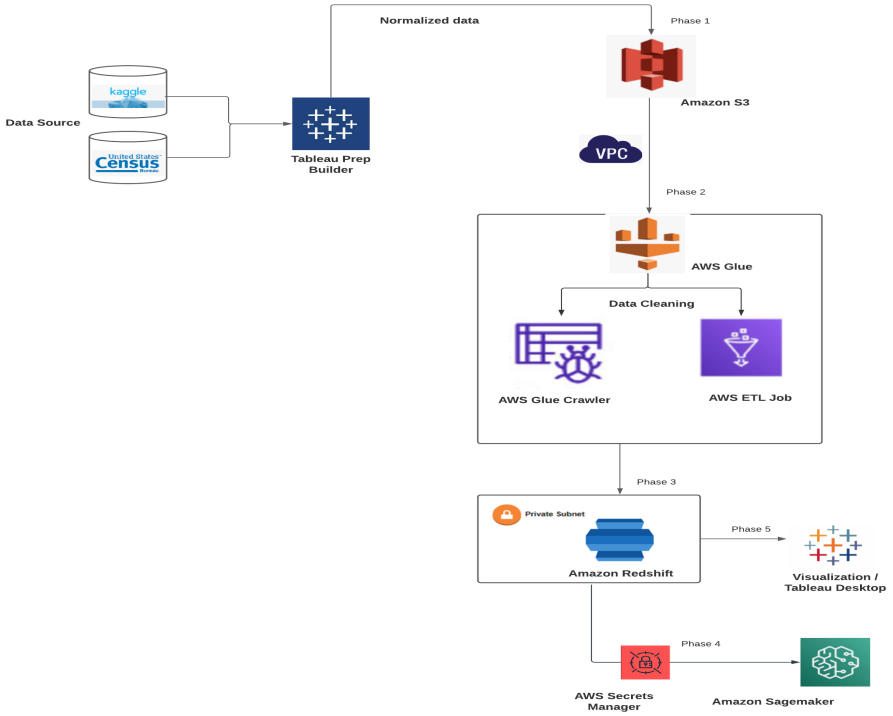


4. Architecture

Figure 7 shows the data flow diagram on AWS of our project. It consists of the different stages of the project starting from the data collection to the visualization. In the first step, we have collected the data from kaggle.com and USCensus.com and have used Tableau Prep Builder to normalize and clean the data sets. Next, the prepared data is uploaded to an AWS storage S3 bucket. Next, Crawlers are created with the needed IAM roles, to eliminate the null values, edit and convert the data using ETL Glue processes. Subsequently, the data is uploaded to Redshift using AWS Glue, where the files' data is combined in the format required. The majority of the data analysis is completed using Sagemaker and Tableau (Data Visualization), which are both connected to Redshift.

Figure 7

Architecture Diagram



4.1 AWS S3

S3 storage is an AWS cloud storage solution that allows you to access various data sources from any path. When building the buckets for deployment, this path is chosen. AWS provides ‘pay as you go’ services where the pricing is based on the amount of data stored and the network transmitting capacity. The attributes and authorization of every object uploaded to an S3 bucket are unique. The data stored in the S3 buckets are used for further analysis. These buckets also consist of an object that helps in storing backup and disaster recovery. To avoid any setbacks, both the production and backup data would be subjected to the same security restrictions, such as encryptions, permissions, and audit requirements. S3 buckets are also used for application hosting and software delivery.

4.2 AWS Glue

AWS Glue is a serverless data integration technology that enables data discovery, preparation, and combination for analytics, machine learning, and application development. AWS Glue comes with all of the data integration options you'll need to get started with data analysis. When fresh data arrives, AWS Glue may perform the ETL operations. An AWS Lambda function is used to execute the ETL jobs as soon as fresh data in Amazon S3 becomes available. As part of the ETL processes, a new dataset is registered in the AWS Glue Data Catalog. It scans all datasets, identifies data types, and suggests data storage configurations. Glue automatically develops the program required to perform the data processing and loading activities. It allows to conduct and manage ETL processes, as well as integrate and duplicate data across numerous data stores, with ease.

4.3 AWS Redshift

Amazon Redshift employs SQL to investigate structured and semi-structured data in data warehouses, operational databases, and data lakes, combining AWS technologies to give the most cost-effective pricing at all scales. Amazon Redshift, the most common cloud data warehouse, is distinctively positioned to benefit from all three trends and can be a great place to start monetizing or extracting great insights from data.

4.4 AWS Sagemaker

Amazon SageMaker is a highly scalable service that allows data scientists and programmers to quickly design, test, and execute machine learning (ML) models. By removing the hard lifting from every stage in a machine learning process, SageMaker makes it simple to build high-quality models. SageMaker unifies all machine learning components into a single toolset, allowing models to be deployed more quickly and for less money.

4.5 Tableau

Tableau is a data visualization application that is often used for Business Intelligence, but its capabilities are not restricted to that. It aids in the construction of dynamic graphs and charts in the dashboards and spreadsheets with the purpose of gaining market intelligence. Tableau can integrate information from a multitude of heterogeneous data sources interface for dynamic analysis, including databases, spreadsheets, big data, and cloud data. It is capable of recovering data from any location. may fetch data from a rudimentary database.

5. Security Management Services

AWS provides services to protect data, identities, and applications against security breaches. The following are some of the services that are used:

5.1 Identity and Access Management

AWS Identity and Access Management (IAM) allows for perfectly alright access management throughout the whole AWS infrastructure. You can control who has accessibility to which resources and services, and under what conditions, using IAM. You manage permissions to your workforce and systems with IAM policies and procedures to ensure lowest access.By using this service there is no overhead for tracking the authorized and unauthorised users, hence the data and the services are secured using the IAM service.

5.2 Virtual Private Cloud

Amazon VPC allows you complete control throughout the virtual networking environment, covering resource placement, connection, and security. It is a logical container that separates resources that we create from other Amazon cloud customers. Inside each container different services are placed with essential IAM roles to interact with the other AWS services. But it hides the usage of the cloud services from one customer to the other by setting a boundary between the customers. With Amazon VPC, you can have even more exact control over your cloud network, adding an extra layer of security for your processes and data. A user may manage network gateways and subnets as well as configure network settings such as IP address ranges, route tables, network gateways, and subnets.

5.3 AWS Secret Manager

The secret manager is used for hiding the critical and non-exposable credentials from the customers or other users. The database passwords are highly confidential credentials as anyone may tamper the data from the database which will greatly affect our analysis. The secret manager stores the values in the form of a key-value pair which is encrypted and kept secret in the secret manager service. These values are retrieved through an API which decrypts the values and makes a connection with the needed service. In this project, we have encrypted the database name, username, password, port number, server where the database is hosted. Other types of secrets, such as API keys and OAuth tokens, can also be handled by the service.

5.4 Endpoint

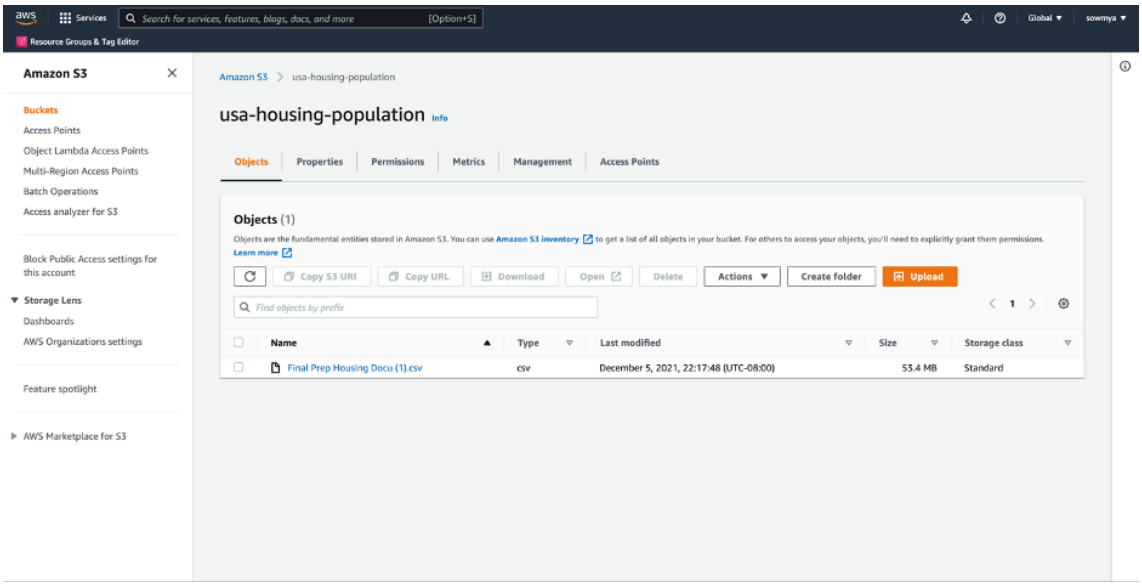
The communication between the services residing inside the VPC happened with the help of endpoints. They make it possible to communicate between VPC time and products while avoiding glitches. Users are distributed to a variety of devices in the digital community using a gateway load balancer. Auditing, adherence, cover tampering, and other community products can all be protected with these devices. When you create an interface VPC endpoint, an Elastic Network Interface with a private IP address is deployed in your subnet. An Amazon EC2 instance in the VPC can communicate with an Amazon S3 bucket via the ENI and AWS network. Using the interface endpoint, apps on your on-premises network infrastructure may access S3 buckets through AWS Direct Connect or Site-to-Site VPN. The interface endpoint supports an ever-expanding list of AWS services.

6. Implementation

Firstly, the data that is collected from kaggle.com and USCensus.com is merged using Tableau Prep Builder. Here, the two datasets are merged by common attributes like State, City and Date. The final .csv final has been generated with the name ‘Final prep Housing Docu’. This data is then uploaded to the Amazon S3 buckets for further analysis. This S3 bucket has been named as ‘us-housing-population’ as shown in figure 9.

Figure 9

S3 bucket loaded with the data file



Next, a Virtual Private Cloud is created with many endpoints in order to give access to the amazon buckets. A security group was made to connect the AWS glue to JDBC. We'll use the standard VPC for all the services to interact with one another, and build inbound rules for connecting to JDBC for Redshift. Create an IAM role after building a glue crawler, selecting a source type and adding a data store. Finally, build a database to store the records before running the crawler. Then, before verifying the connection, create a cluster in Amazon Redshift and assign the role that was established in the previous stage. After the connection to Redshift is

established, create a new crawler in the data store, pick a connection, specify the role, select the database, and begin the crawler. Figures 10 and 11 are examples of this.

Figure 10

AWS ‘project - crawler’ created

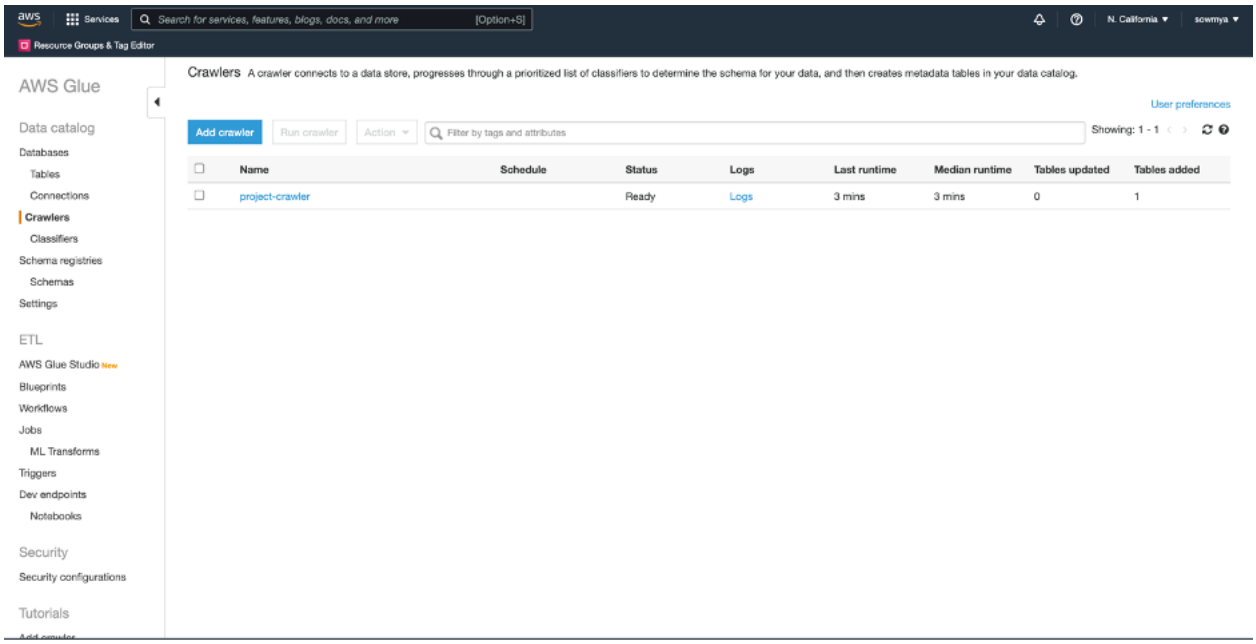
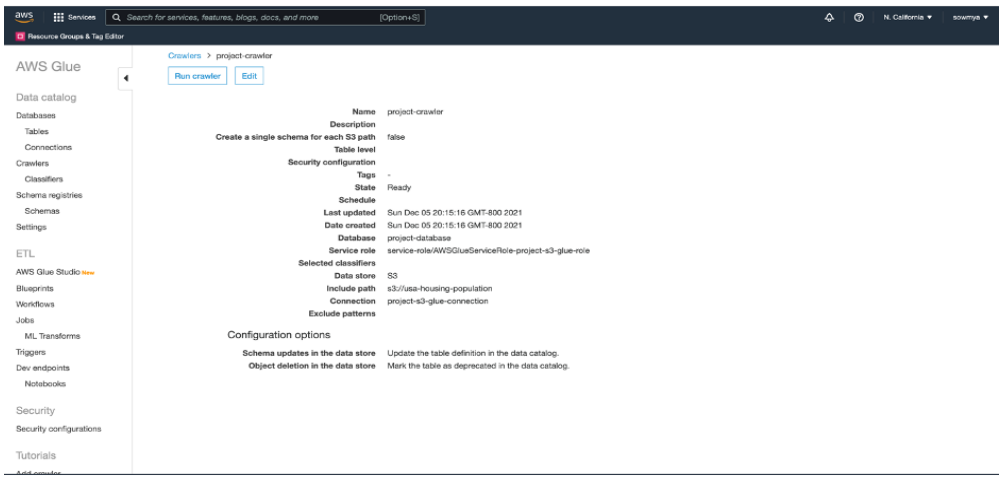


Figure 11

Configurations of the AWS crawler



Further, to create the table structure for the data residing in the S3 bucket, a crawler has been created. To construct an instance job, use either the source or destination connection. Two new

connections called ‘ project-s3-glue-connection’ and ‘ project-s3-glue-redshift-job-connection’ are created for connecting the AWS glue to S3 and Redshift. After the connection has been made and the crawler is run, the AWS Glue tables metadata is automatically created. These steps have been shown in figures 12, 13 and 14 respectively.

Figure 12

AWS Glue Connections to Redshift and S3

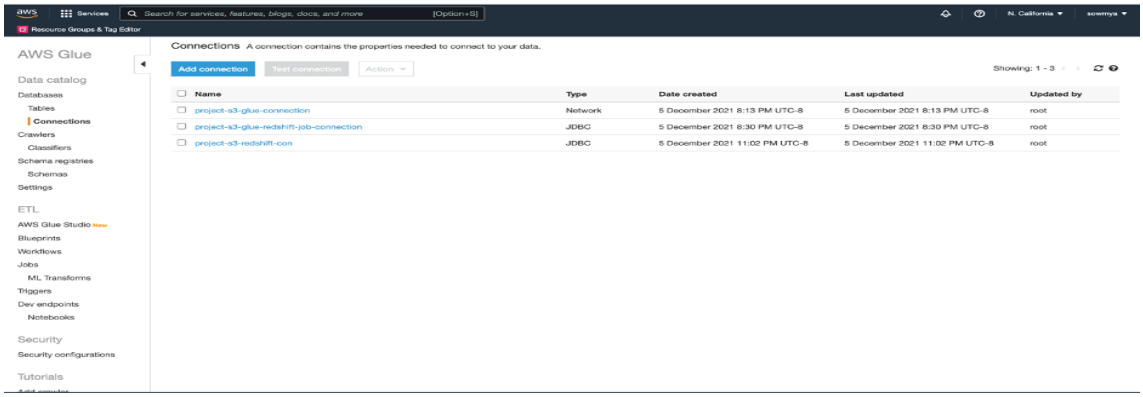


Figure 13

AWS Glue Tables

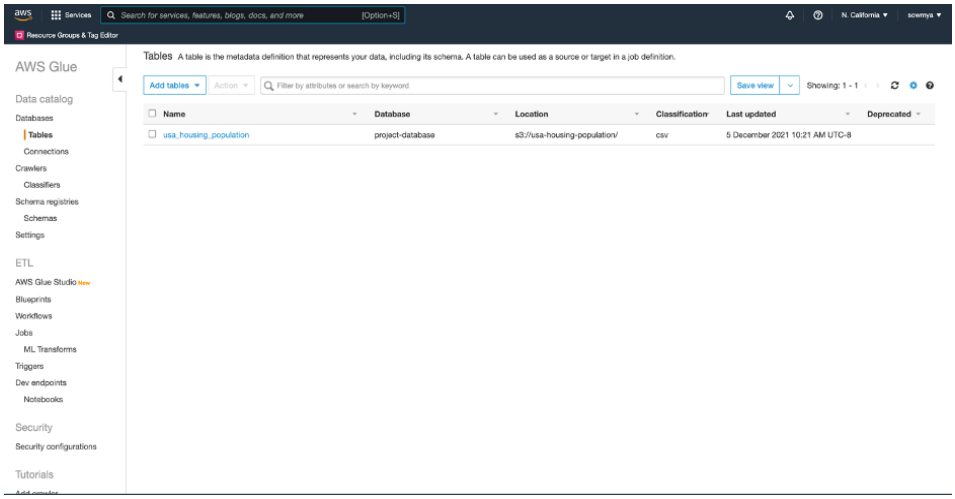
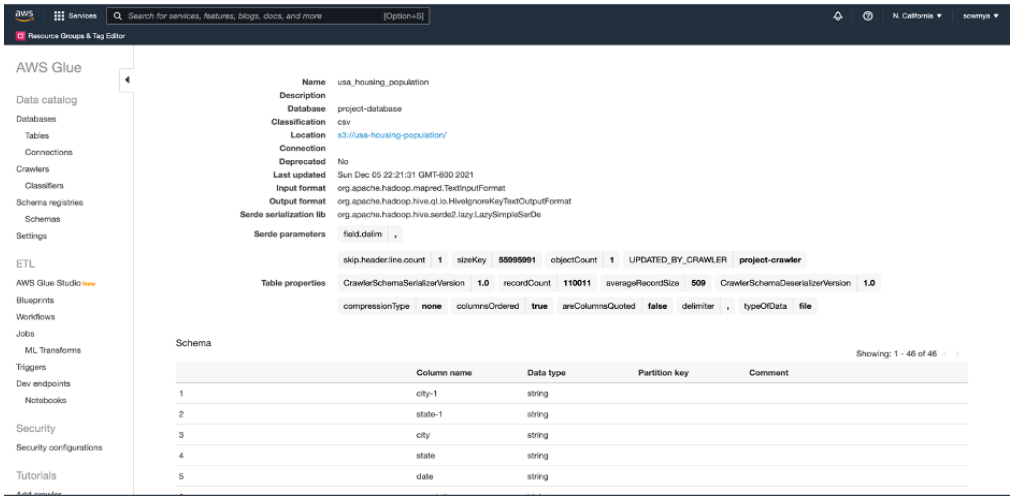


Figure 14

Table Metadata



Now, run the AWS job for transferring the data from S3 bucket to Redshift. This will populate the table with its corresponding data. A pyspark script is run to modify and clean the data. Here, we have removed nulls, unwanted data and renamed some of the columns. As shown in figures 15 and 16.

Figure 15

AWS Glue Job to transfer data from S3 to Redshift

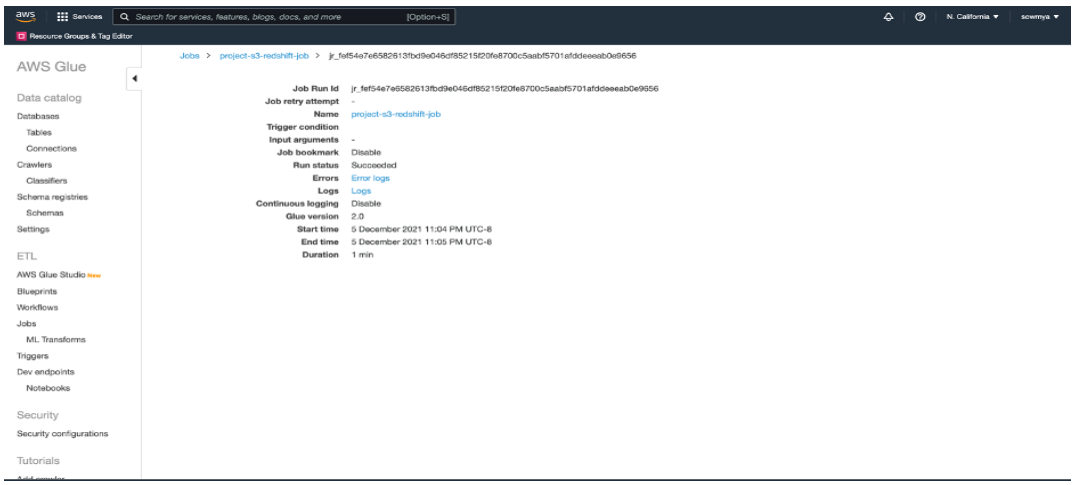
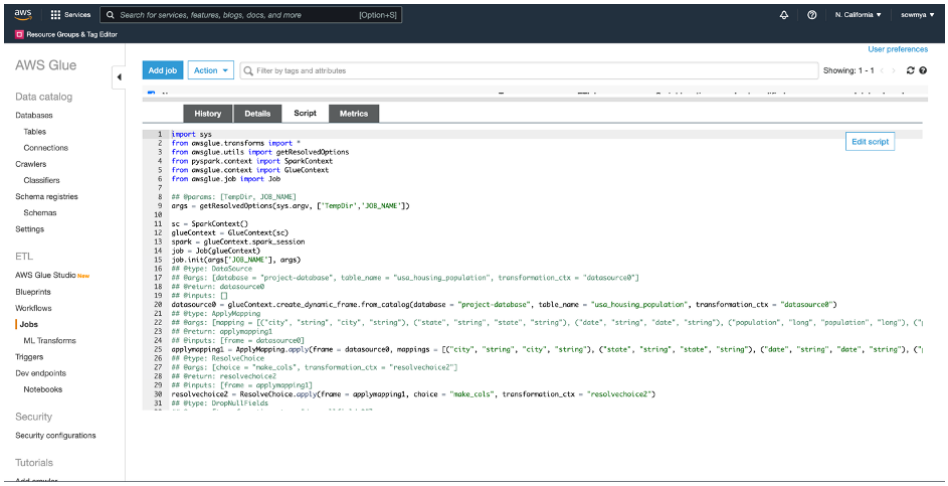


Figure 16

Pyspark Script to remove the null values, unwanted columns and renaming few columns



Now, as shown in figure 17 and 18, AWS Redshift service is used to create a cluster with the name ‘project-cluster’. AWS Redshift is a data warehouse where it has a unit called cluster containing two components: compute node and leader node. The Compute node has its own CPU, memory and disk space. Once the job is run , the data is transferred to the ‘dev’ database with username ‘awsuser’ and password.

Figure 17

Redshift Cluster

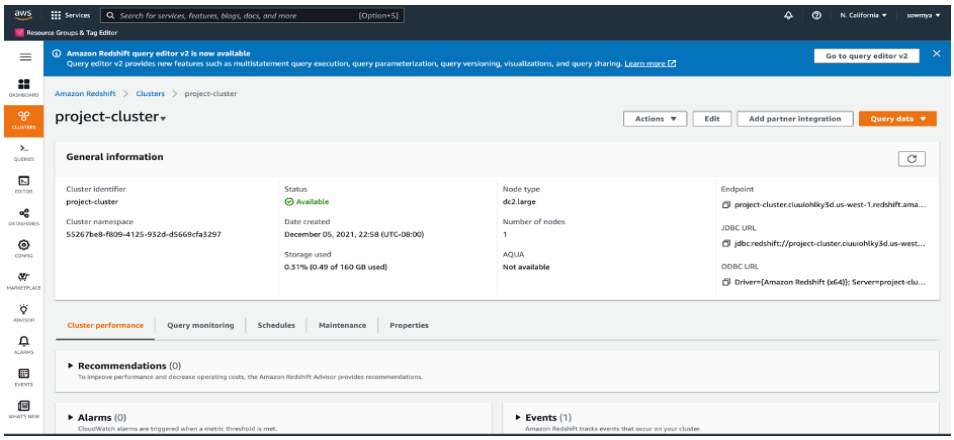
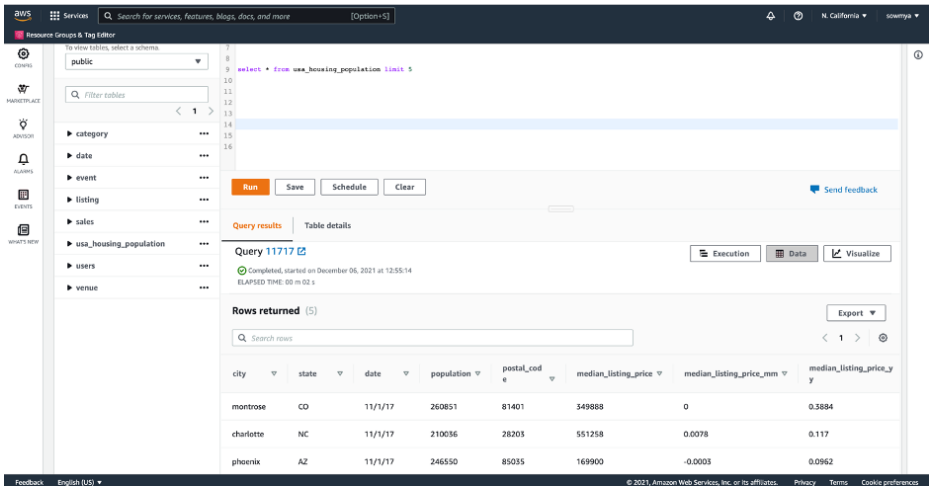


Figure 18

Redshift database with US housing population table



After the database has been created, AWS Sagemaker jupyter notebook is used for Exploratory Data Analysis. This helps in extracting the insights of data. The sagemaker and the Redshift were connected securely using the Secret Manager. This is shown in figures 19, 20 and 21.

Figure 19

Sagemaker Notebook

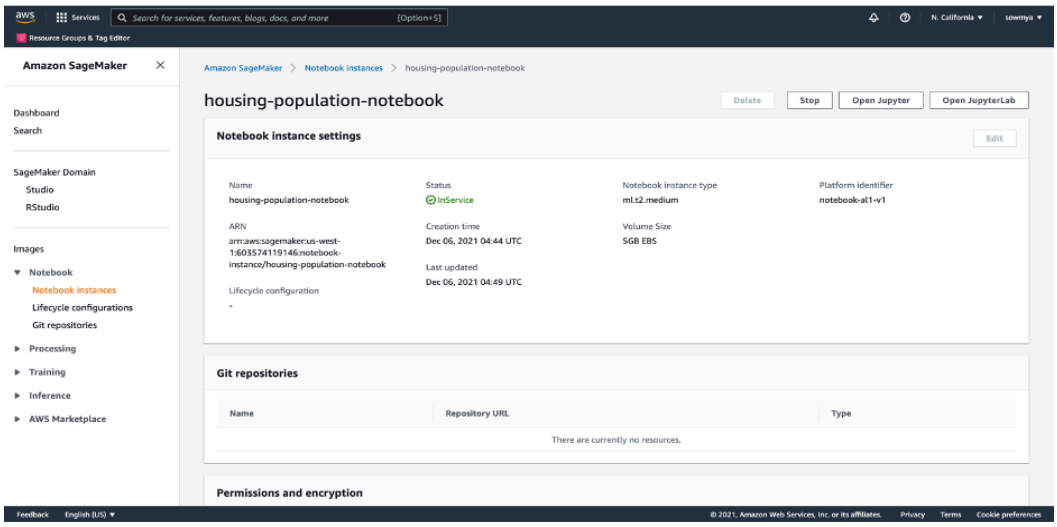


Figure 20

Data Exploration and EDA

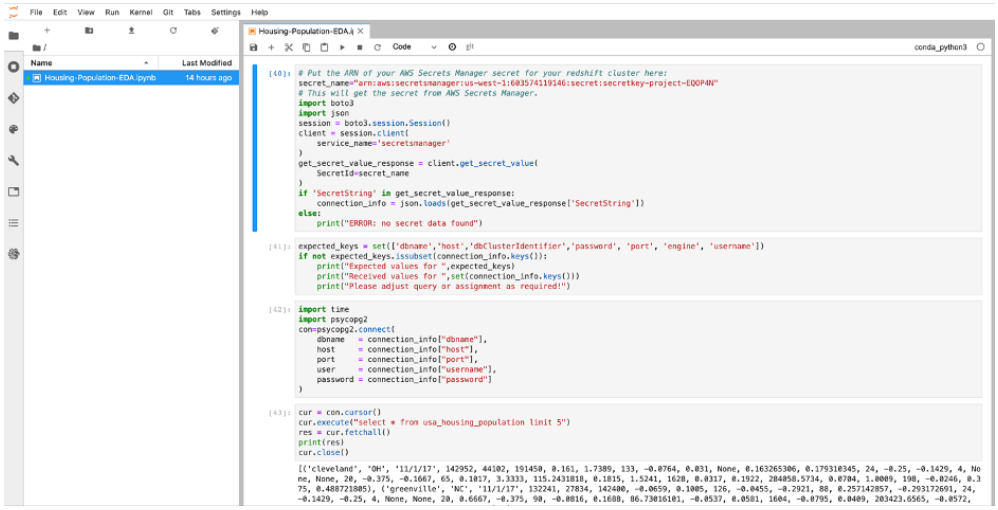
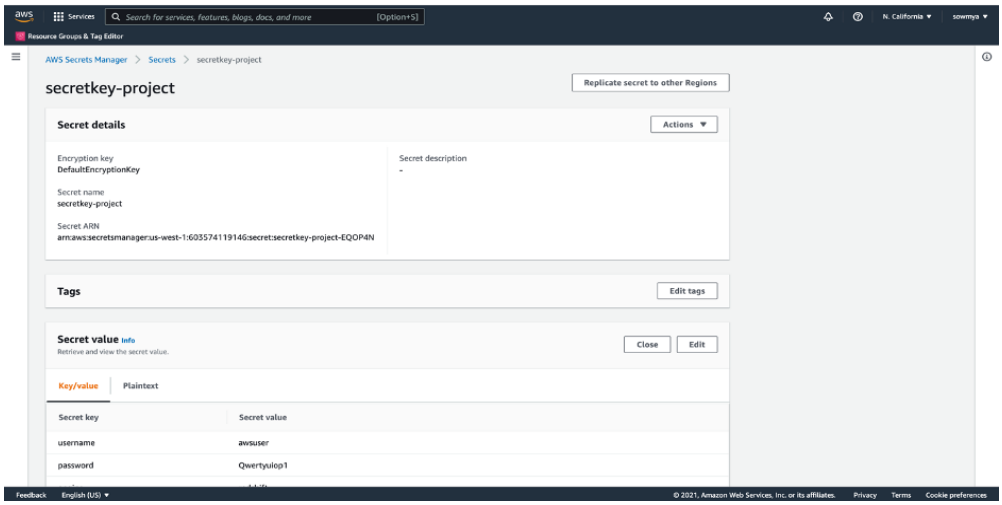


Figure 21

Secret Key Manager



Finally, we have created visualization in Tableau by connecting Tableau with Redshift with the help of database name, server name, port numbers, username and password.

7. Data Visualization

At this stage, we have created some visualizations using our dataset to show our objectives in a graphical representation. It is a way to summarize findings and represent them in a form that helps interpret and identify outliers, trends, and patterns. So for visualizing our findings we have used Tableau Desktop. It is a data visualization software. So we have connected the tableau desktop with a cleaned and Processed dataset that was already stored in the AWS Redshift with the help of the Endpoint URL of the Redshift. It requires database name, port number, and Password to access the datasets available in Redshift to ensure not to compromise the security and integrity of the dataset. We have provided all the details and accessed the dataset in tableau from Redshift. After that, we made sure that the format and data types of all fields are correct. Finally, we have created multiple charts and graphs to get excellent insights from the dataset which are provided below.

Figure 22

Connecting dataset in Redshift with Tableau Desktop

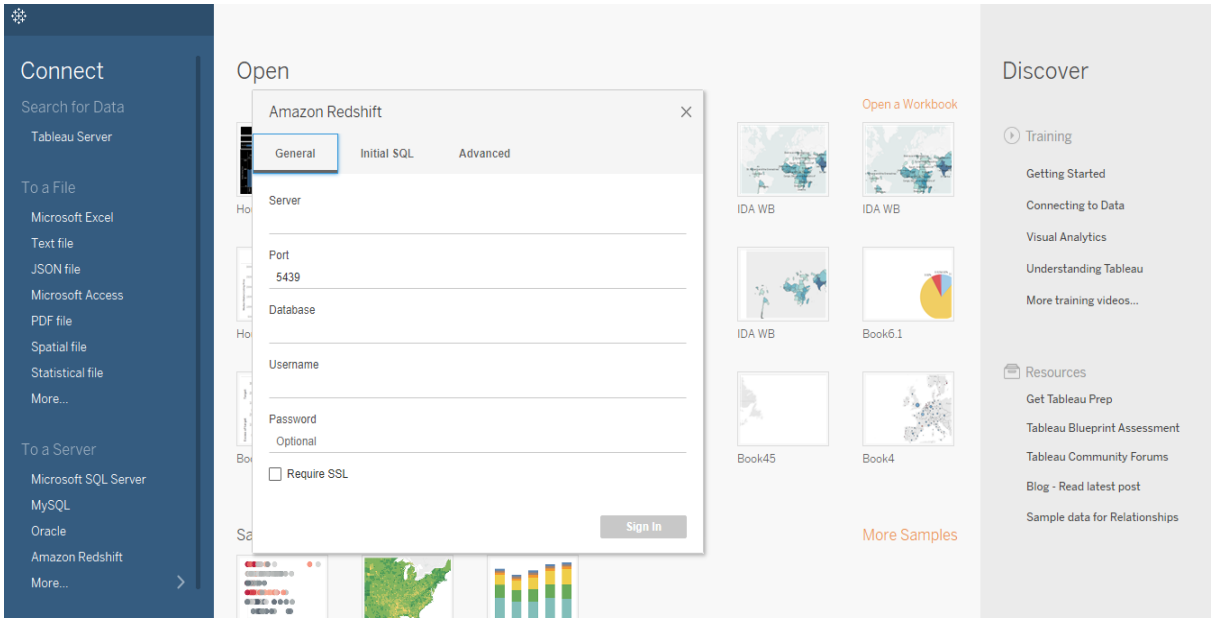


Figure 23

Sample of the dataset accessed from Redshift

Sort fields | Data source order | Show aliases | Show hidden fields | 1,000 rows

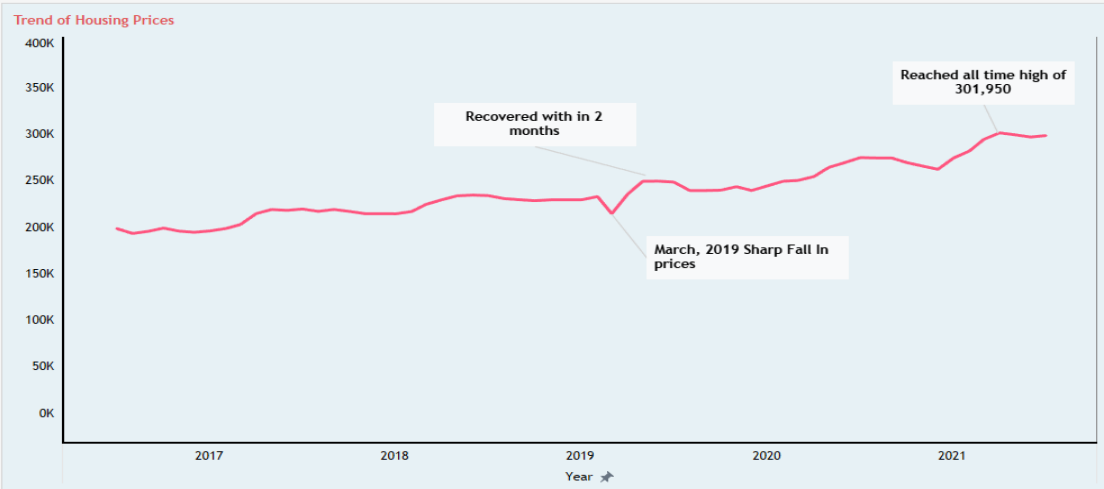
City	State	Date	Population	Postal Code	Median Listing Price	Median Listing Pri...	Median Listing Pri...	Active Listing Count
las vegas	NV	01-05-2018	270,257	89117	494,000	-0.212700	0.10390	
san diego	CA	01-05-2018	582,978	92128	672,000	-0.006700	0.00450	
washington	DC	01-05-2018	417,463	20009	624,950	-0.056500	0.04640	
atlanta	GA	01-05-2018	222,001	30317	509,950	0.004800	0.19780	
chicago	IL	01-05-2018	237,627	60622	620,000	0.056200	0.02990	
san antonio	TX	01-05-2018	198,079	78255	487,400	-0.009200	0.01580	
phoenix	AZ	01-05-2018	256,784	85044	366,950	0.054500	-0.02020	
new orleans	LA	01-05-2018	201,392	70130	474,950	-0.040500	-0.10220	

7.1 Charts

Some of the visualizations done for our project are:

Figure 24

Trend of Housing Prices



The visualization shown in the Figure 24, is the trend line chart for the Median House Listing Prices in the United States, It indicated that there is a sharp fall in the month of March, 2019 but the prices have recovered within 2 months that we have annotated in the chart and then

the prices have reached all time high in the month of April, 2021. Although there are fluctuations in the price the overall trend of House Prices is upwards.

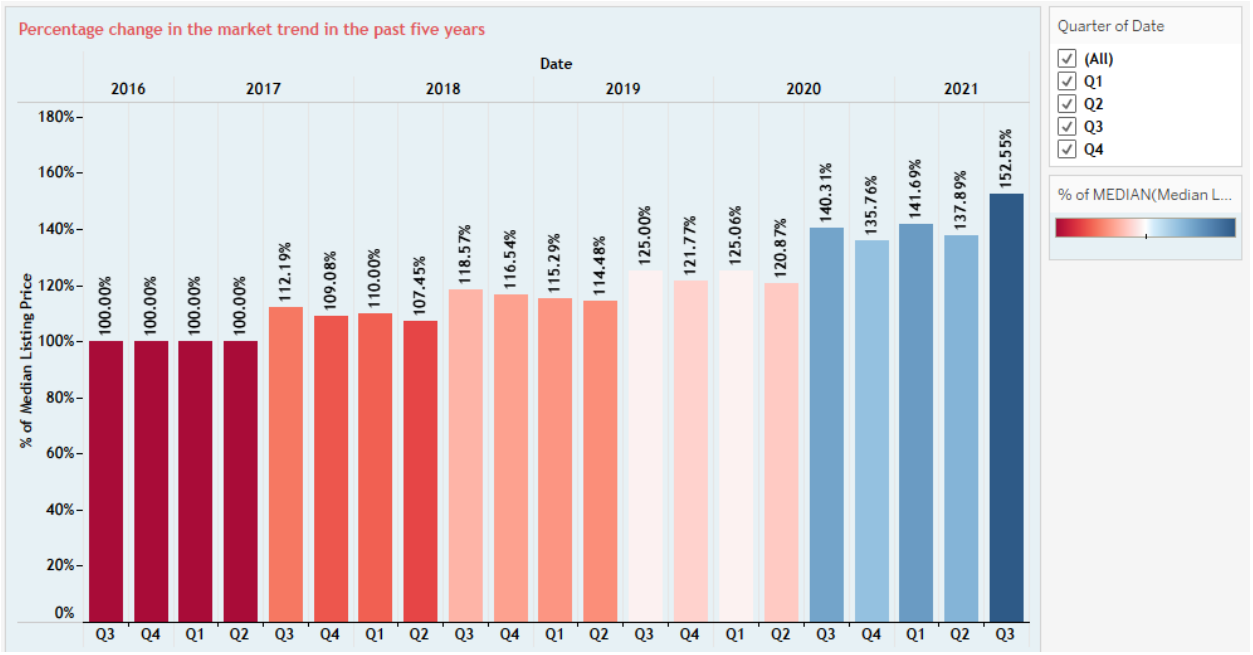
Figure 25

Which are the best months to buy the house?

The Bar chart in the figure 25, indicated that average prices of houses in the US aggregated monthly wise. So we have noticed an important pattern that the average house prices in the middle of the year , May and June, are higher compared to the start and end of the year. The lowest monthly average house prices are seen in the month of December, and highest in the month of May.

Figure 26

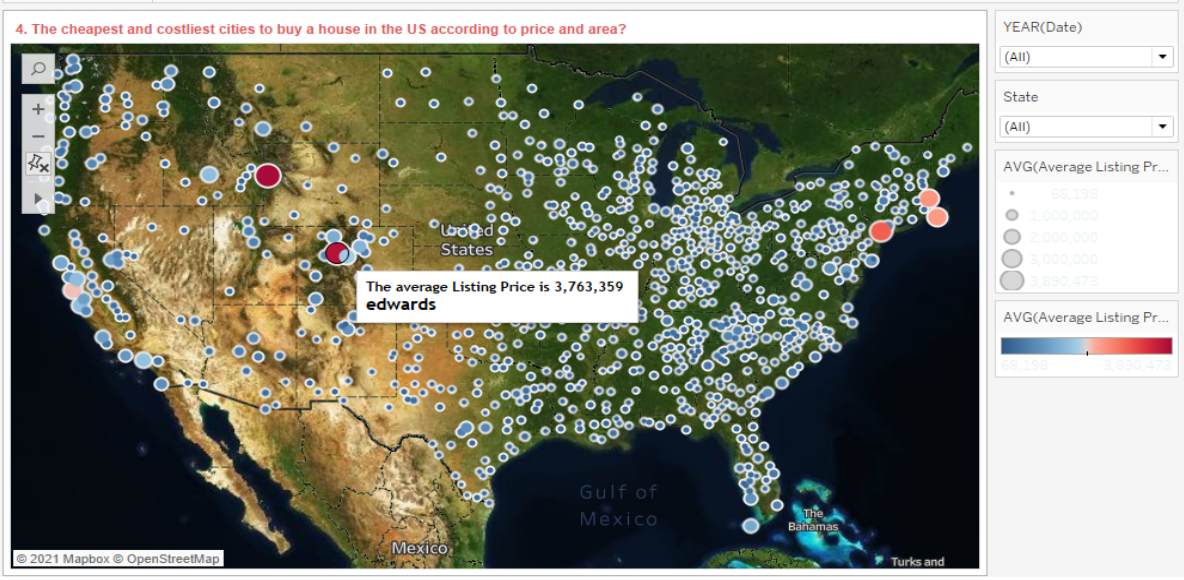
Percentage change in the market trend in the past five years



The coloured bar chart in Figure 26 represents that, the quarterly growth of the median house prices in the US, We have taken Quarter 3 of 2016 as a base quarter. The red color represents minimum growth and the Blue color represents the maximum growth. We can notice that there is a 52.55% increase in the prices from quarter 3, 2016 to quarter 3, 2021. Also the prices never went below the prices of the base quarter.

Figure 27

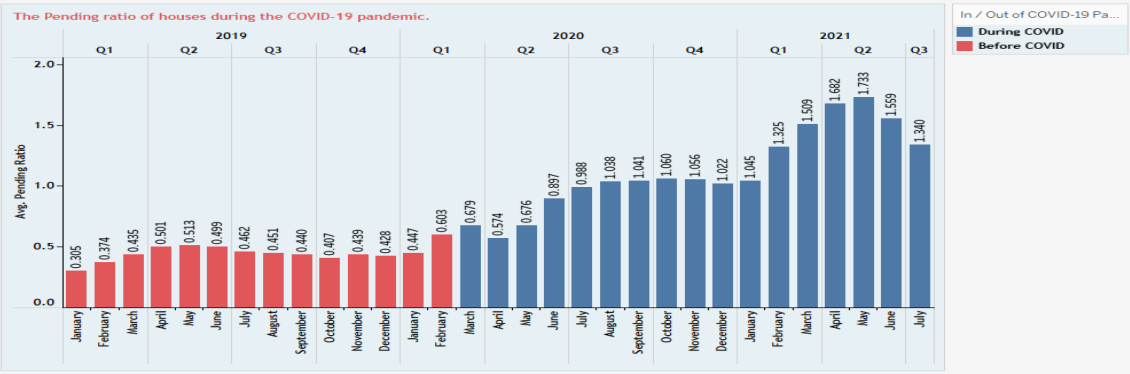
The cheapest and costliest cities to buy a house in the US according to price and area?



The Bubble map in Figure 27 represents all the cities of the US, Whereas the size and color of the Bubbles represents the average listing prices of the houses in those cities. We can note that Edwards has the highest average listing price in the US with over Three Million Dollar for a house.

Figure 28

The Pending ratio of houses during the COVID-19 pandemic.



The bar chart in Figure 28 gives information about the ratio of pending houses with respect to the total number of listed houses across all the cities in the United States. In this

visualization we want to see the pending hoses trend just before and during the COVID 19 pandemic. So we have taken data from 2019 to 2021 and analyzed. The red colored bars represent before COVID and Blue colored bars represent during the COVID 19. We have noticed an interesting trend that the number of pending houses has increased tremendously during the COVID 19 pandemic.

Figure 29

Top 10 cities

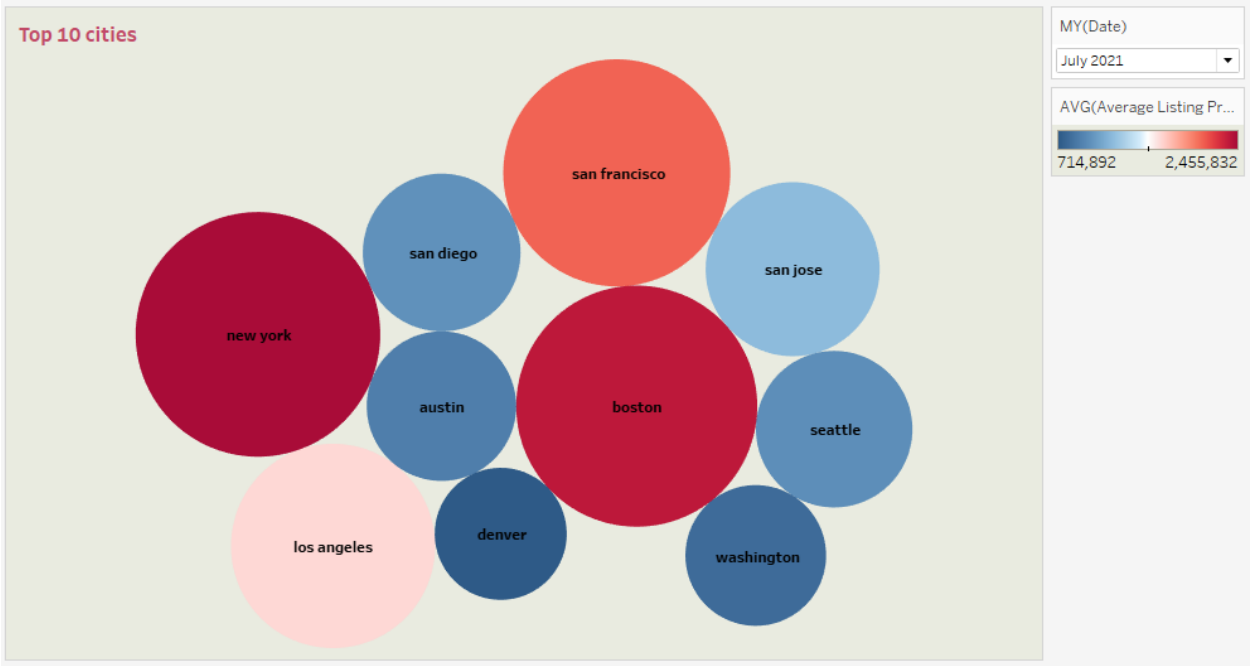
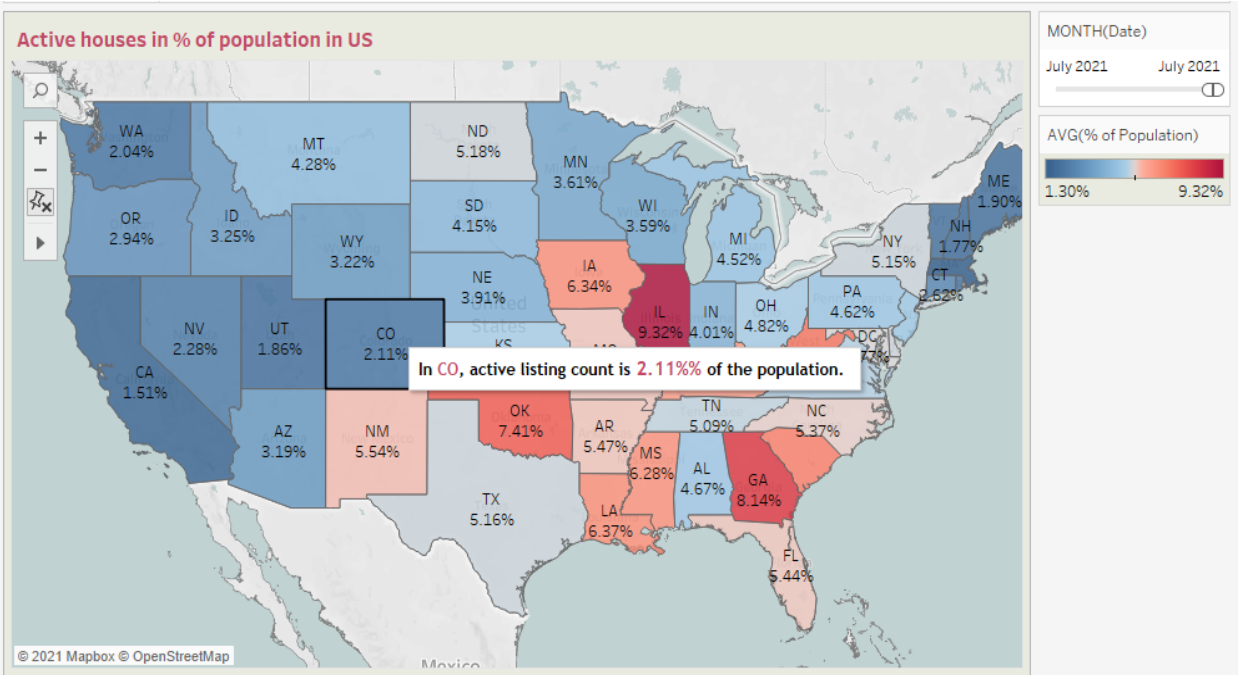


Figure 29 represents the top 10 costliest cities in the United States. It states that Newyork is the costliest city in the US following Boston.

Figure 30

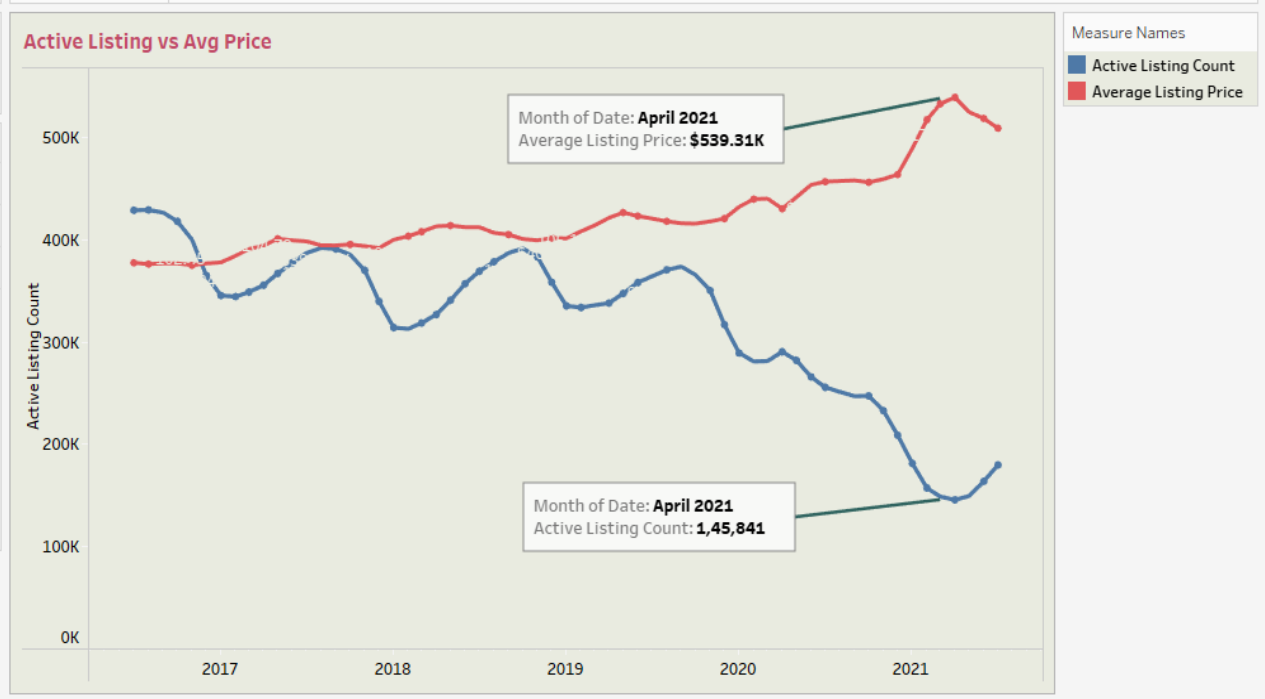
Active houses in % of population in US



At this stage we wanted to analyze the number of houses for sale in the market and relate that number with the population of the states. So the Choropleth map in Figure 30 represents that the active listing houses the percentage of a state's population. We have noticed that the Illinois state has scored the highest percentage. It means the active houses for sale in Illinois is 9.32% of its population. In this way we can analyze data for any specified month of given years.

Figure 31

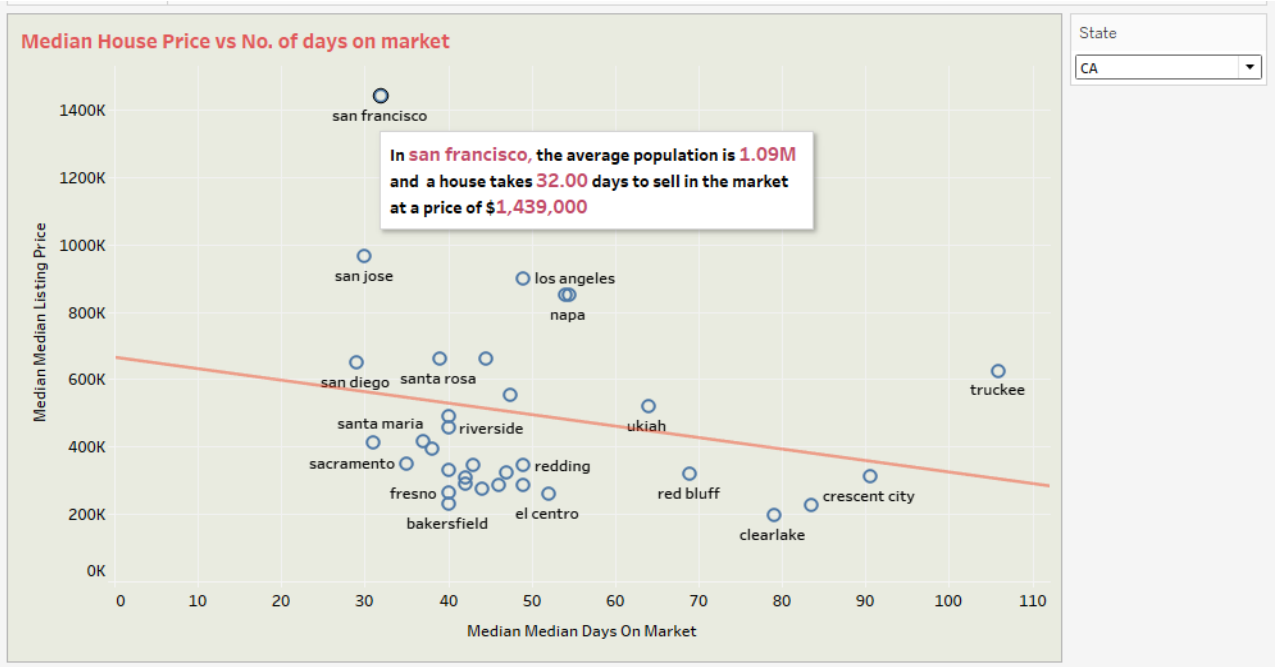
Active Listing vs Avg Price



The dual line chart in Figure 31 consists of information about the average line price per a house and the Average number of houses listed. The blue line represents the total number of actively listed houses for sale in the market whereas the red line represents the average listing price pre house over the period. So the total number of actively listed houses is decreasing over the period and average listing price per house is increasing. They both seem inversely proportional to each other. That means less number of houses for sale means more price per house, It follows simple supply and demand rules.

Figure 32

Median House Price vs No. of days on market



In Figure 32 we have analyzed the correlation between house price and number of days on an average a house will be there in the market to get sold. We have drawn a linear regression line and noticed the negative relationship between them. Here we have shown for California, You can see for any particular or combination of states.

Figure 33

Population VS Price per Sq foot

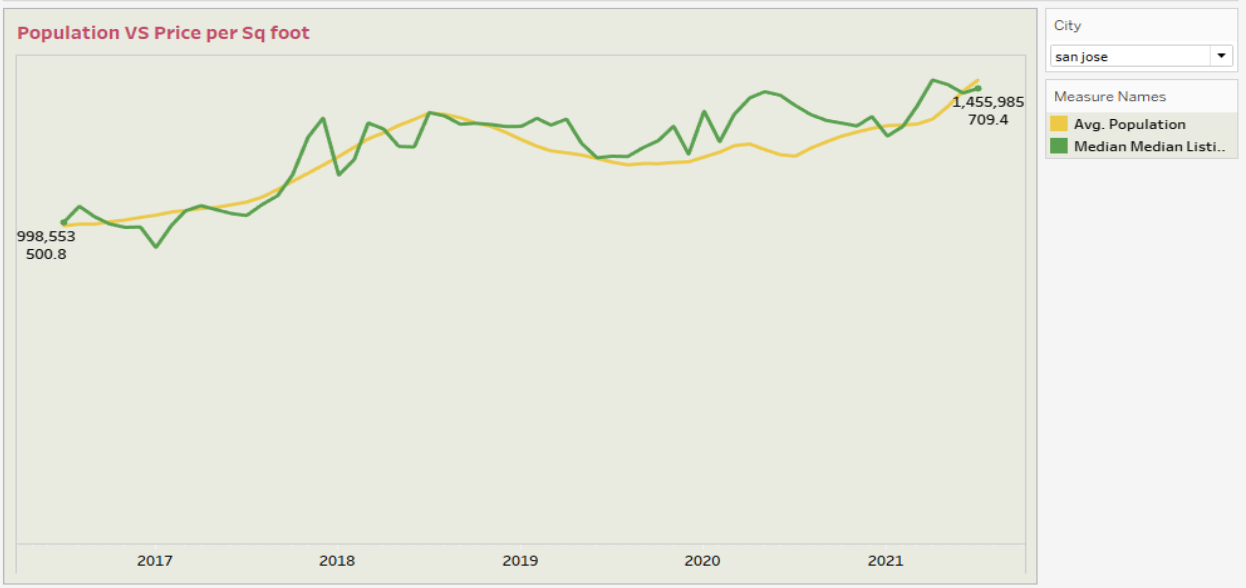
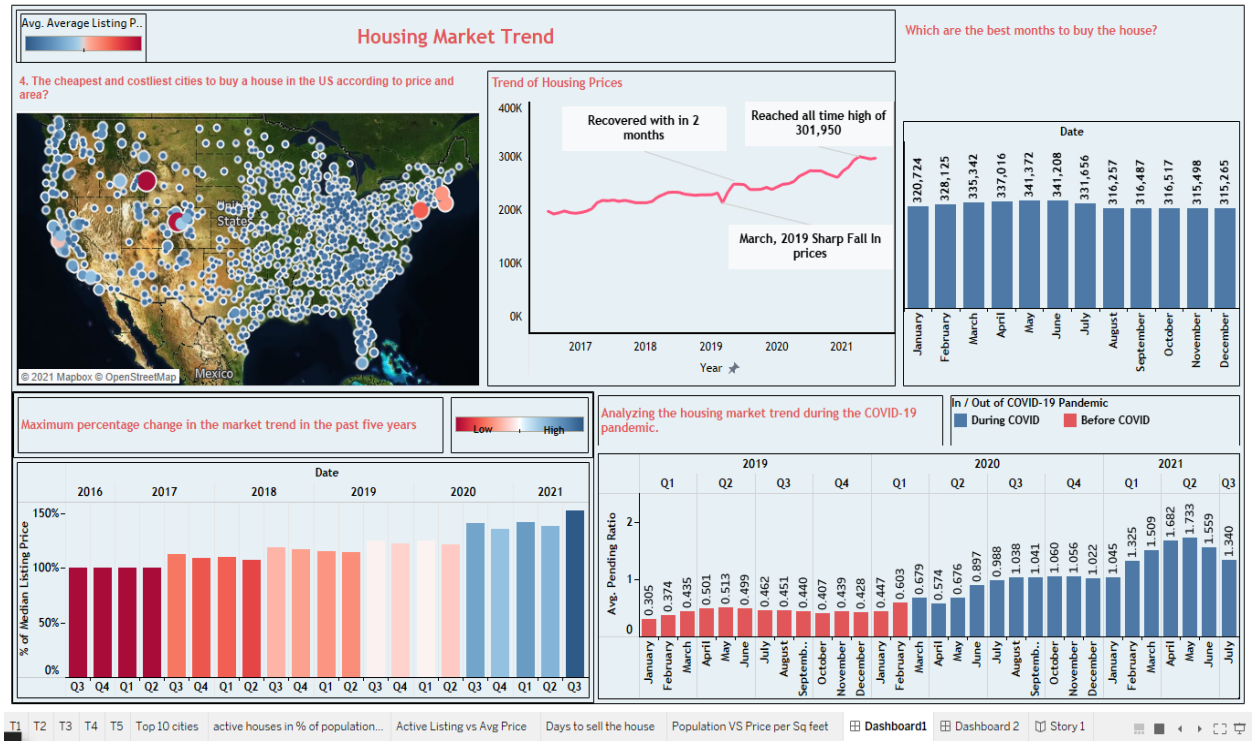


Figure 33 represents the merged Line charts, where the green line represents the median listing price per square foot in San Jose, and the red line represents the average population of San Jose. We have drawn a trend line for both parameters and noticed that the Price per Square Foot is increasing with respect to the Population of the city. Like that we can select individual cities and analyse the trend.

Figure 34

Dashboard 1



So combining all the visualization graphs we have created two Dashboards for customers for their own analysis and understanding purpose. The Figure 34 represents the Housing market trend dashboard, it is a combination of charts provided in figures from 24 to 28. From this dashboard customer can analyse information like Trend of Housing Prices, which are the best months to buy the house, In the last five years, the quarterly percentage change in the market trend, What are the cheapest and most expensive cities in the United States to buy a property in terms of price and area?, and The Pending ratio of houses during the COVID-19 pandemic.

Figure 35

Dashboard 2

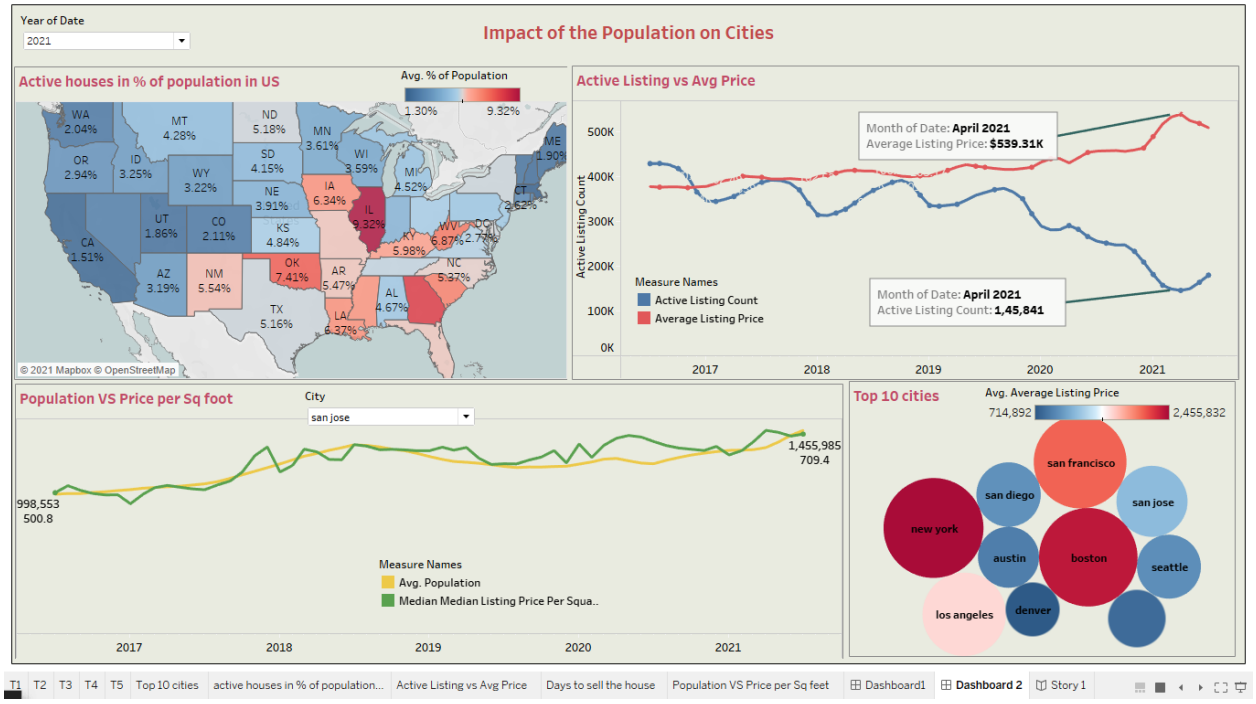


Figure 35 represents our second Dashboard. Impact of the population of cities on the housing prices and the demand for a house in the specified city is the major analysis that can be done using this dashboard. It consists of visualizations like Population VS Price per Sq foot. Top 10 cities in the US, Active Listing vs Avg Price, and Active houses in % of population in the US.

8. Conclusion

We have attempted to break down the US market information to check and see the changes in the cost of the houses over the most recent five years by building an information investigation pipeline. Diverse AWS administrations were utilized for ingestion, readiness, stockpiling, and perceptions. Endeavors were made to comprehend the clump information, to make the various patterns that could be trailed by financial backers prior to trading the houses. We have observed that the market is developing consistently and it has been seen that the general normal development of the US real estate market is 41% from 2016 to 2021. Our investigation and representations can be utilized by new and old financial backers and furthermore any financial backer or broker who is hoping to put resources into various urban communities of the US.

9. Future work

Check the factors affecting the price of the houses like employment,and salary growth.This project can be further extended by implementing Machine learning algorithms to forecast future prices in various regions across the United States in order to aid smarter investment decisions.The architecture will be enhanced to make it easier to handle streaming data and provide faster insight into real-time data.

10. References

Bureau, U. S. C. (2021, December 9). Census.gov. Retrieved December 12, 2021, from https://www.census.gov/#:~:text=The%20U.S.%20Census%20Bureau%20has,Census%20population%20of%20331.4%20million.

Kouwenberg, R. R. P., & Zwinkels, R. C. J. (2011). Chasing Trends in the U.S. Housing Market. SSRN Electronic Journal. Published. https://doi.org/10.2139/ssrn.1539475

Maftouni, M. (2021, August 26). US Real Estate Market Trends from 2016 to 2021. Kaggle. Retrieved December 12, 2021, from https://www.kaggle.com/maedemaftouni/real-estate-market-trends.

United States Housing Prices & Market - Redfin. (n.d.). Retrieved December 12, 2021, from *https://www.redfin.com/us-housing-market.*