California Housing Market Application

A Project Report

Presented to

DATA-228-21

By

Nupur Pathak, Revathi Boopathi, Sree Divya Cheerla, Vani Bhat

11/23/2022

ABSTRACT

The surge in demand for the US housing market and shortage of inventory has made it difficult for people to find their dream home with all the features they desire with ease. Hybrid work culture because of the pandemic is fueling the demand in the housing market. There are various parameters that are impacting the property values such as school proximity, migration of people impacting the population density, crime rate, job opportunities, new constructions, recreational centers (restaurants, malls, etc), medical care facilities, access to public transportation, and others. In this project, we plan to analyze house price listings and show how the housing market trends have been over the years in California. We also plan to bring out the idea of how house/property values are influenced by the presence of school districts and how migration has affected the house prices. We are redefining the housing market by predicting the house prices based on the customer's preference, to fulfill their desire to buy their dream house. Big data has enabled the handling of large-scale data involving millions of records seamlessly for an application like housing. AWS has served as a market leader because of its improved scalability, reliability, security, etc. over its competitors. We will leverage big data applications by using AWS cloud services to analyze and predict house listings / property records according to the customer's preferences.

Keywords: Big Data Application, AWS S3, AWS Glue, AWS Cloud Formation, AWS SageMaker, California Housing Market, House price prediction.

Acknowledgements

The authors are deeply indebted to Professor Andrew H. Bond and TA Pravallika Vallapuri for their invaluable comments and guidance in the development of this project throughout the process.

Table of Contents

List of Figures

List of Tables

Chapter 1. Introduction

1.1 Project goals and objectives

In this project, we are building a California housing market application to ease the process of housing property search for the customers based on their preferred buying criteria. We also plan to bring out the idea of how house/property values are influenced by the presence of public schools with higher ranking in the region.

1.2 Problem and motivation

The surge in demand for the US housing market and shortage of inventory has made it difficult for people to find their dream home with all the features they desire with ease. Hybrid work culture resulting from the pandemic is fueling the demand in the housing market. There are various parameters that are impacting the property values such as school proximity, migration of people impacting the population density, crime rate, job opportunities, new constructions, recreational centers (restaurants, malls, etc.), medical care facilities, access to public transportation, and others. Therefore, it is becoming challenging to find a suitable property.

1.3 Project application and impact

The housing market application will be useful for the residents to identify the right property based on their preferences. With the collaboration of the school data, it will also aid in identifying the public and private schools in the proximity of the house.

1.4 Project results and expected deliverables

We are developing a web platform for the California housing market application. This web platform is developed on a cloud-based web development platform wix. This web platform can be used to find a property listing based on various filters such as city, zip code, number of beds, number of baths and price range. In addition to this, there are sections to detail the listings as well as school in the region of interest.

1.5 Project report structure

The key sections covered in this report are background and related work in the field of US housing market, system requirements and analysis, system design, system implementation, system testing and experiment, and conclusion and future work. The details on each section are covered in the following sections.

Chapter 2 Background and Related Work

2.1 Background and used technologies

The surge in demand for the US housing market and shortage of inventory has made it difficult for people to find their dream home with all the features they desire with ease. Hybrid work culture resulting from the pandemic is fueling the demand in the housing market. There are various parameters that are impacting the property values such as school proximity, migration of people impacting the population density, crime rate, job opportunities, new constructions, recreational centers (restaurants, malls, etc.), medical care facilities, access to public transportation, and others. The five major characteristics of big data like volume, variety, velocity, value, veracity, and variability have enabled the handling of large-scale data involving millions of records seamlessly in any application.

In our project, we will leverage big data applications by using AWS cloud services to analyze and predict house listings / property records according to the customer's preferences. Using this application, the housing data is combined with school data to analyze how the house prices are influenced by the presence of highly ranked schools within the vicinity. Housing market data is huge and volatile in nature and demands significantly high computing capacity. Pay-as-you-go cloud computing models like AWS support this characteristic of big data workloads for efficient processing. As the data stored in the cloud is backed up and stored using encryption algorithms, it is safer and can be accessed across the team. The house listings data along with school is stored in Amazon S3 (Simple Storage Service) which is an object storage service provided by AWS to store data in the form of objects in buckets offering scalability, security, availability, and high performance. The data stored on Amazon S3 can be optimized and accessed to meet specific business requirements. For building any application, having clean and transformed data that fits the application requirements is essential. In our project, to clean and transform the house listings data and school data, AWS Glue which is a serverless data integration service offered by AWS is used. The extract, transform, and load (ETL) jobs are orchestrated on the data to make it easier to integrate with other sources for further analytics and application development. The transformed data in S3 is moved to AWS SageMaker where a linear regression model is implemented with features like bed, bath, average school ranking, etc. The house price is further predicted and assessed; the data is then moved to AWS S3. The output of the SageMaker is then given as input to Amazon Athena which helps to query the data logs inside Amazon S3 bucket using standard SQL and generates the view of the data. The results of the query are configured to store in the output S3 bucket which is then imported into Tableau to gain insights.

2.2 Literature **survey**

Mishra et al. (2021) in their paper talks about recent trends in the big data applications using Amazon Web Services (AWS) cloud platform. The five important characteristics of big data like volume, velocity, variety, veracity, and value makes it easy to identify patterns and insights to make better business decisions. Big data contains data both in structured and unstructured form in huge volumes making it difficult for processing using traditional techniques or algorithms. This paved path for cloud platforms which provide features like storage, security, networking, software, analytics with low operating costs and flexibility to pay only for the services used making it easy to scale and run the business efficiently and with high performance. The authors talk about different cloud computing types available like public, private and hybrid clouds and explain the difference between each of them. They also discuss various cloud services available like IaaS, PaaS, SaaS and serverless and how they can be employed.

As the volume of the data increases, the computing gets complex and demands an environment which can handle and provide high computing capacity making pay-as-you-go-cloud computing models ideal for this. These models make it easy to scale the business up and down as per the business requirements without any additional hardware installments. The authors then discuss different AWS components available for big data collection, processing, storing, and analyzing like Amazon S3, Amazon Glue, Amazon Machine Learning, Amazon Lambda, Amazon Elastic MapReduce, Amazon QuickSight, and others. They focus mainly on Amazon Athena and Amazon QuickSight by explaining the architecture, advantages, and disadvantages of them by explaining a case study.

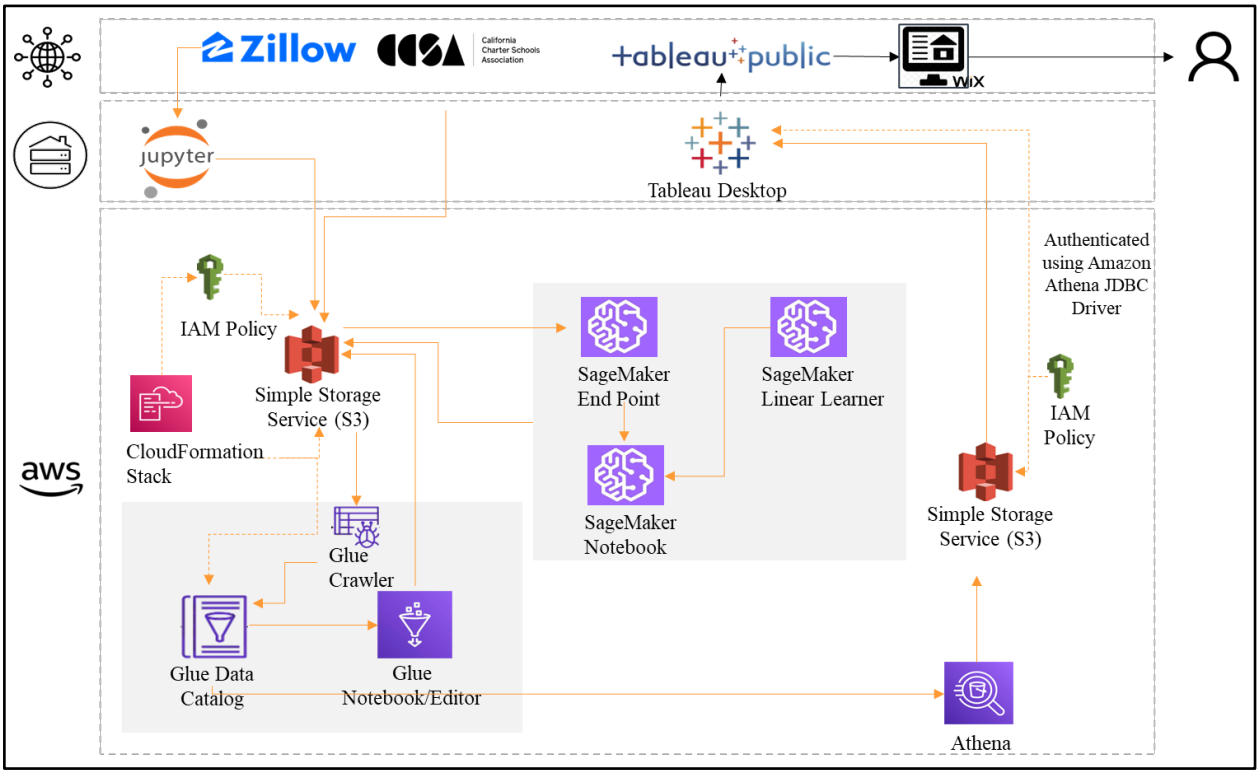
Chapter 3 System Design

3.1 System architecture design

The system architecture design diagram with the tools and technologies involved can be seen in Figure 1.

Figure 1

System Architecture Design



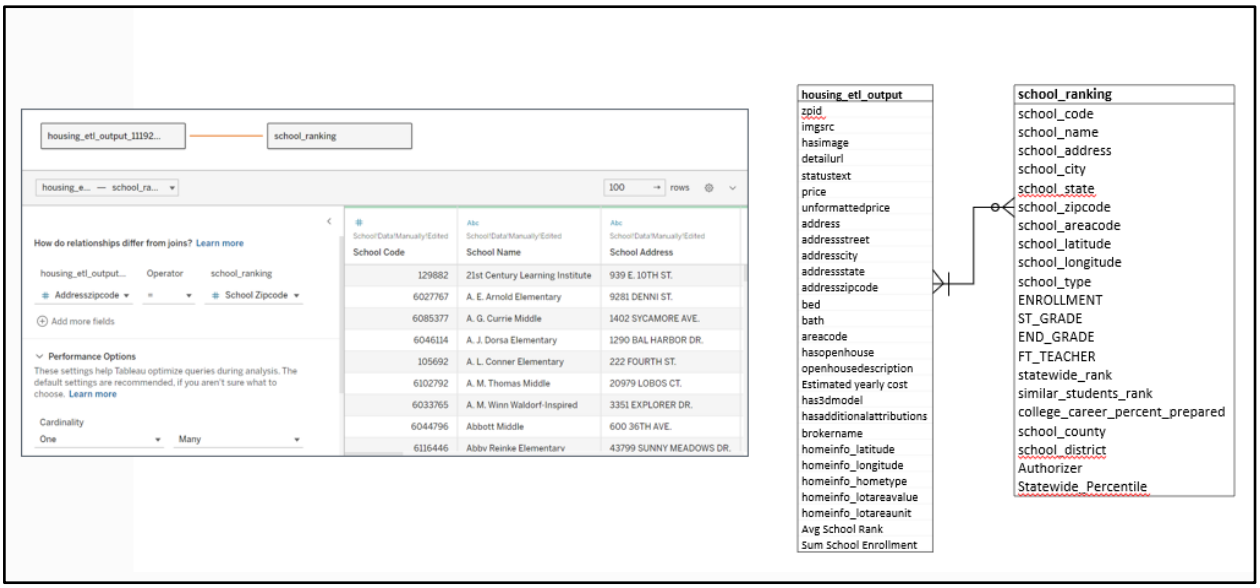
3.2 System data and database design

The Housing Data (7,490 listings) is collected with the help of web scraper. The data is scraped for cities in California iterating listings on each page for the search results. Listings sourced from the site are collated in Jupyter notebook. The School Ranking Data (10,576 schools) is collected from California Charter Schools Association which provides the rating for public and private schools in California. The school state-wide rankings are sourced from CCSA for 2019. Each school is ranked on a scale of 10. The two datasets are connected using the zip code. The

addresszipcode field in the housing dataset is linked with school_zipcode field in the school ranking dataset.

Figure 2

Data Model



3.3 System interface and connectivity design

3.3.1 Development of connectivity within AWS

Cloud formation stack is used to establish interface and connectivity between AWS S3 and AWS GLUE Data Catalog. In this process we load a .yaml file with details on AWS Glue Database, AWS IAM roles like ‘iamgetrole’ and ‘iampassrole’, the schema for the data tables and the new S3 bucket to be created. The established connectivity helps to automatically get the data from the data catalog table to apply various transformations in AWS GLUE on the data which in turn is stored back to S3.

For the modeling in the AWS SageMaker, the connection is established between AWS S3 and an instance of SageMaker endpoint notebook to import data using BOTO which integrates the Python application with AWS S3. The data and the results of the model stored in the S3 is

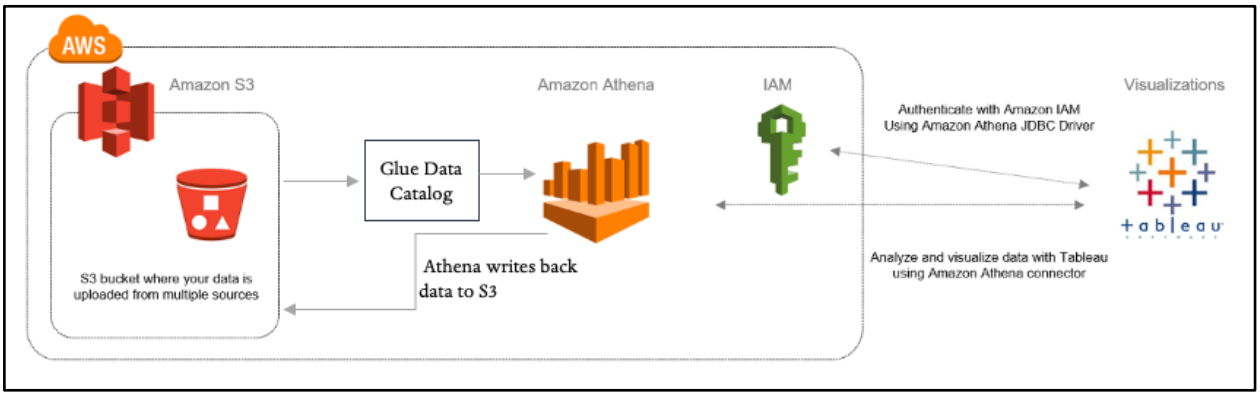
crawled to Glue tables using a crawler which serves as a data platform for third party applications.

3.3.2 External interfaces to third-party systems/components

The house data, school ranking data and predictions data from the glue table is interfaced to the third-party application Tableau using AWS Athena which creates a view for the data in the glue table. Amazon Athena is interfaced with Tableau to visualize the data using Amazon Athena JDBC driver. This would require authentication using Amazon IAM. Below diagram summarizes the authentication.

Figure 3

Third Party Interface Design



3.4 Design problems, solutions, and patterns

There are a few design trade-off decisions that were taken in order to cope up with the system constraints. Given the scope of the application and cost constraints, we have stored the data from SageMaker and AWS Glue back to AWS S3. The data is further moved to AWS Data Catalog as per any resource requirement. This will save the cost of Glue development endpoint being active consistently throughout the application run. Also, EC2 Spot instances are leveraged in order to cut down the cost while implementing the SageMaker models.Tableau is leveraged as

a visualization tool as compared to Power BI as it was observed that it provides free service initially in order to publish the dashboard in a website.

Chapter 4 System Implementation

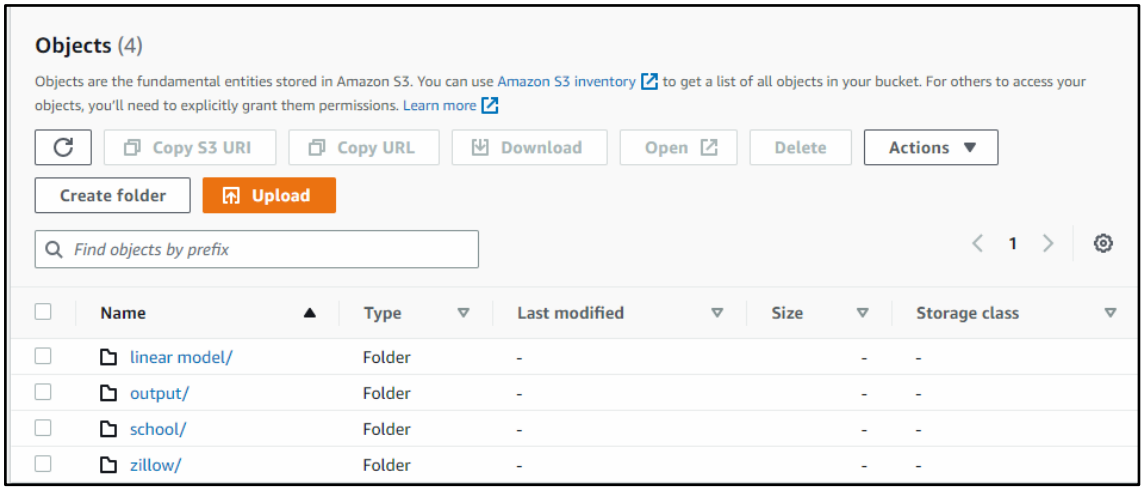
4.1 System implementation summary

The team has scraped house data from Zillow and used US School data from CCSA to build the housing application. Various AWS Services like AWS S3, AWS Cloud Formation, AWS Glue, AWS SageMaker, AWS Athena and Tableau are used which has led to successful implementation of the California Housing Market Application.

The two datasets in .CSV formats are loaded into two AWS S3 buckets which serves as an input.

Figure 4

S3 Bucket



AWS Cloud Formation is used which creates collection of related AWS and third-party resources, IAM roles, database and tables.A .yaml file is configured with the meta data of two of the input datasets, necessary IAM roles required and also the path for the input files which upon

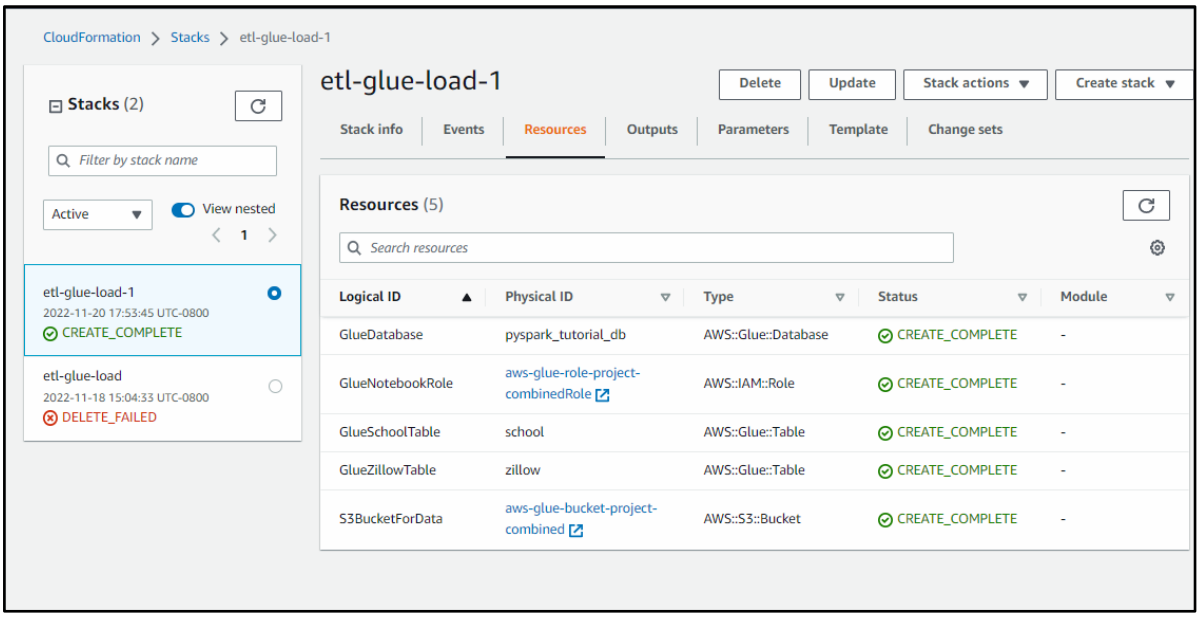
load creates schemas for the input, loads necessary IAM roles needed for AWS Glue and also fetches data from S3 to AWS Glue tables.

Figure 5

Sample. yaml File Structure

Figure 6

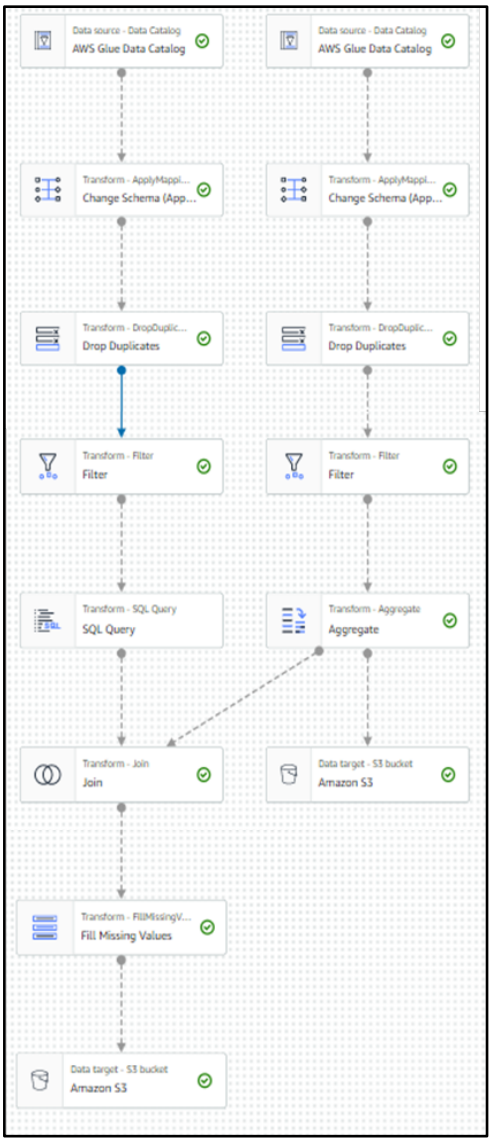
CloudFormation based Resource Creation



Various data transformation techniques like handling missing and null values, removing duplicates and unwanted values, filtering data for the ‘CA’ state, left joining the house and school datasets are performed in the Glue visual editor and are saved back to the AWS S3 output folder. The visual editor flow to transform data in AWS Glue is as follows.

Figure 7

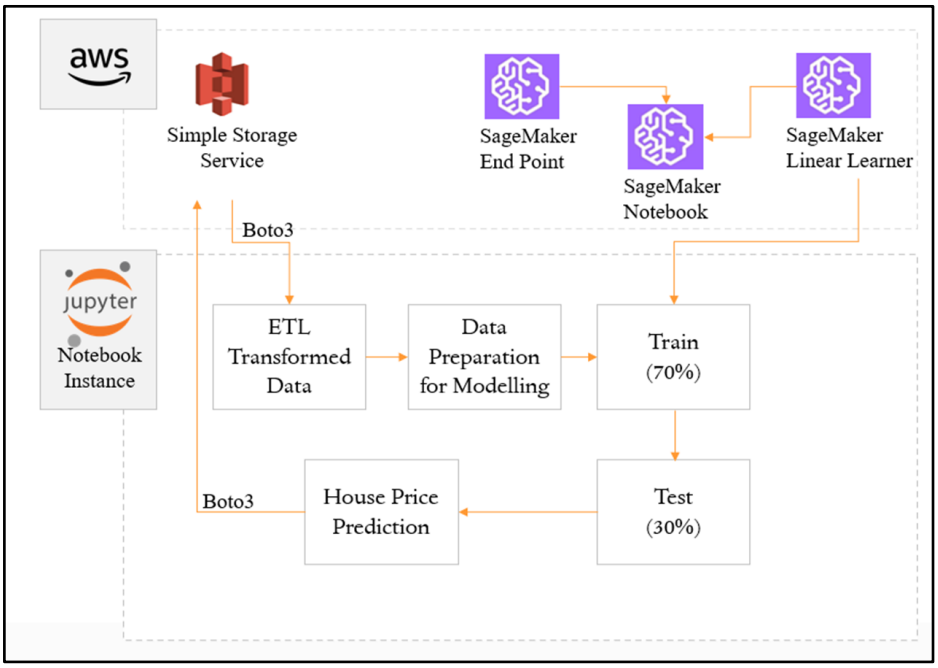
AWS Glue ETL Transformations



Transformed data from S3 is imported to AWS SageMaker to make house price prediction using Linear Regression. The application prediction is called ‘Bestimate’ which ‘California Housing Market Application’ predicts for its home buyers.

Figure 8

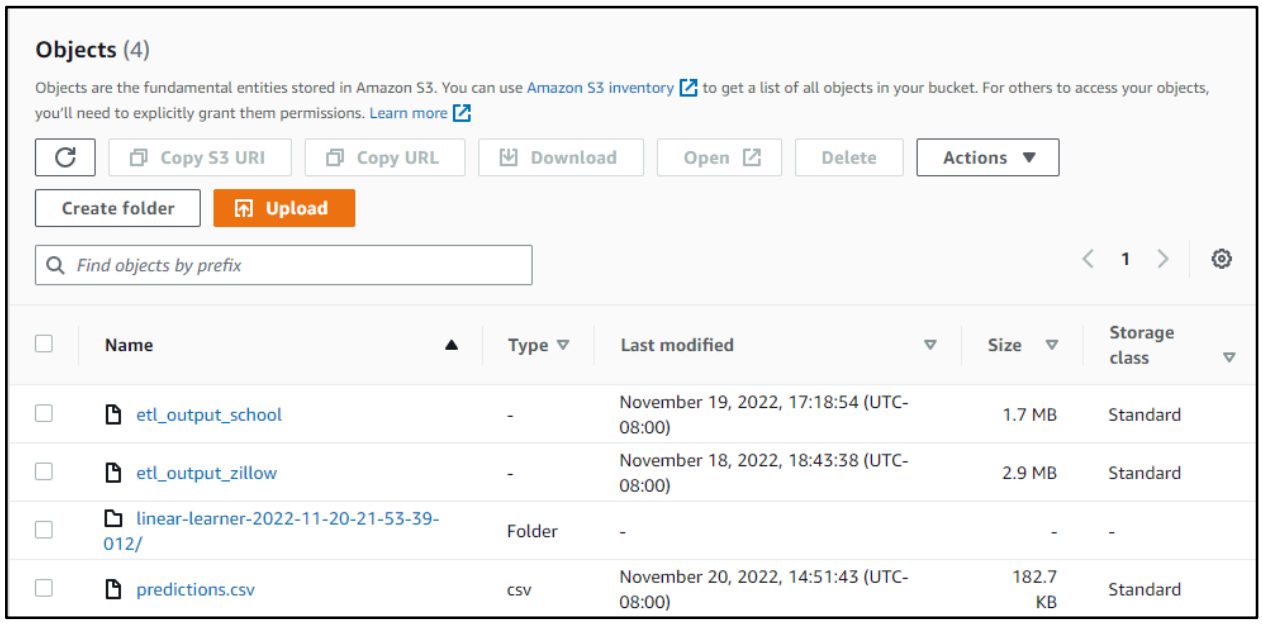
SageMaker Architecture



The predictions are stored into an AWS S3 bucket output folder which is then retrieved into glue tables using crawler jobs.

Figure 9

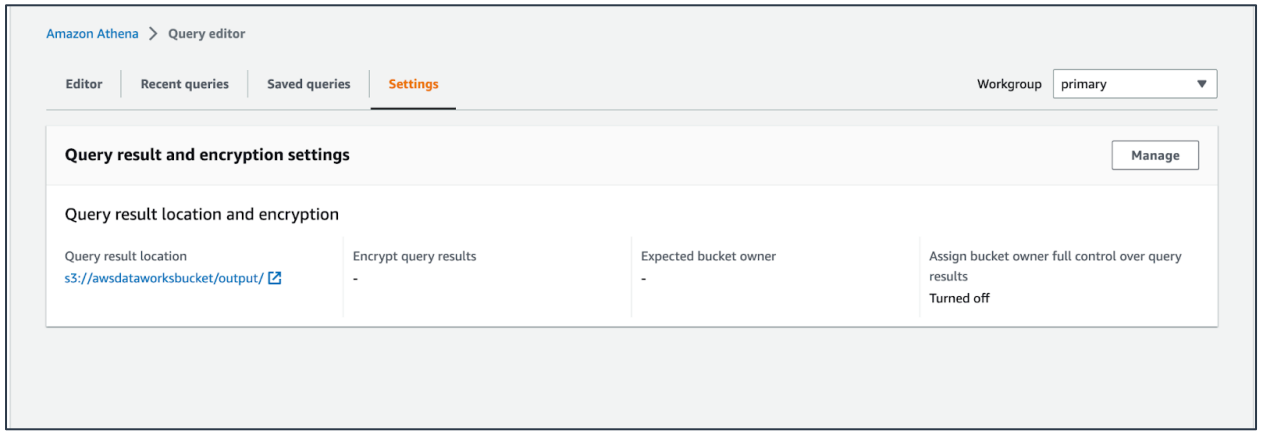
Output folder in AWS S3



The predicted data from AWS Glue is then fetched into AWS Athena for querying and creating a view of the data and is written back to S3 for storage.

Figure 10

Athena Query Result Location



The final data from S3 is then imported into Tableau by establishing a data connection from S3 bucket to Tableau by installing the necessary drivers and setting up the environment by creating Access keys.

Figure 11

IAM Access Key Credentials

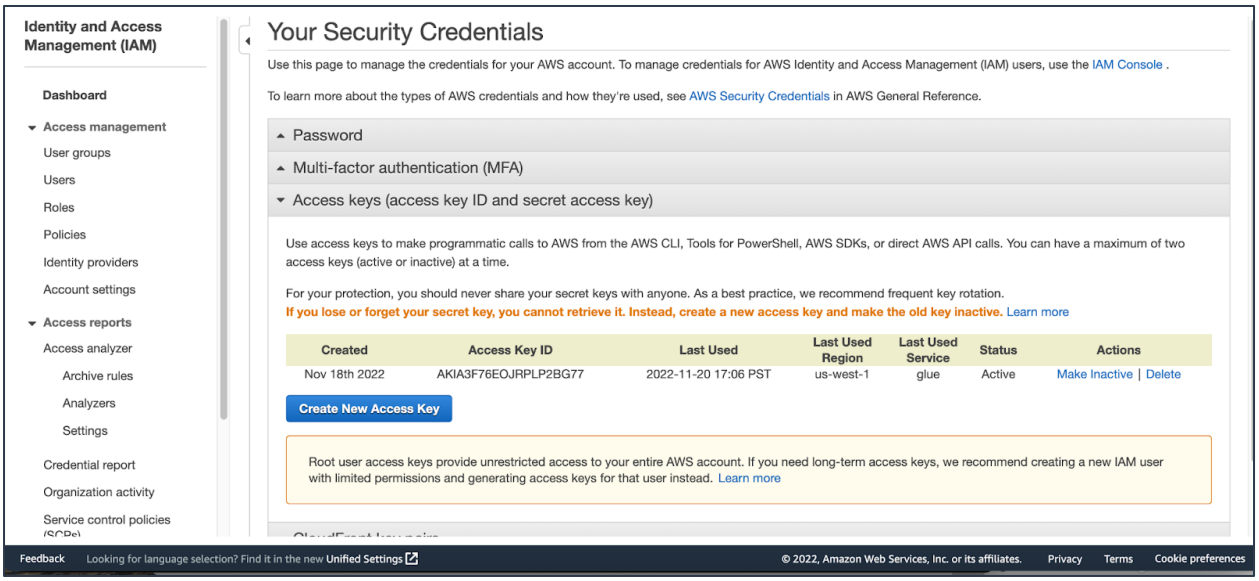
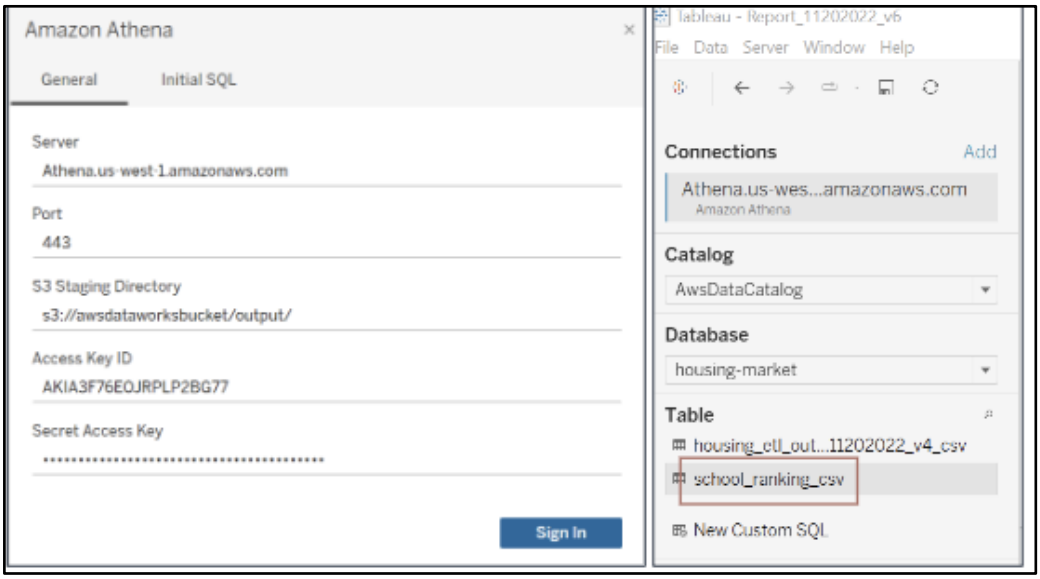


Figure 12

Data Pipeline Connection from S3 to Tableau



In Tableau the data is used to analyze various trends and insights on the house and school data. The implemented visuals show the trend in the housing market in California, also shows how house prices vary due to the presence of schools nearby.

Figure 13

Home Page Dashboard

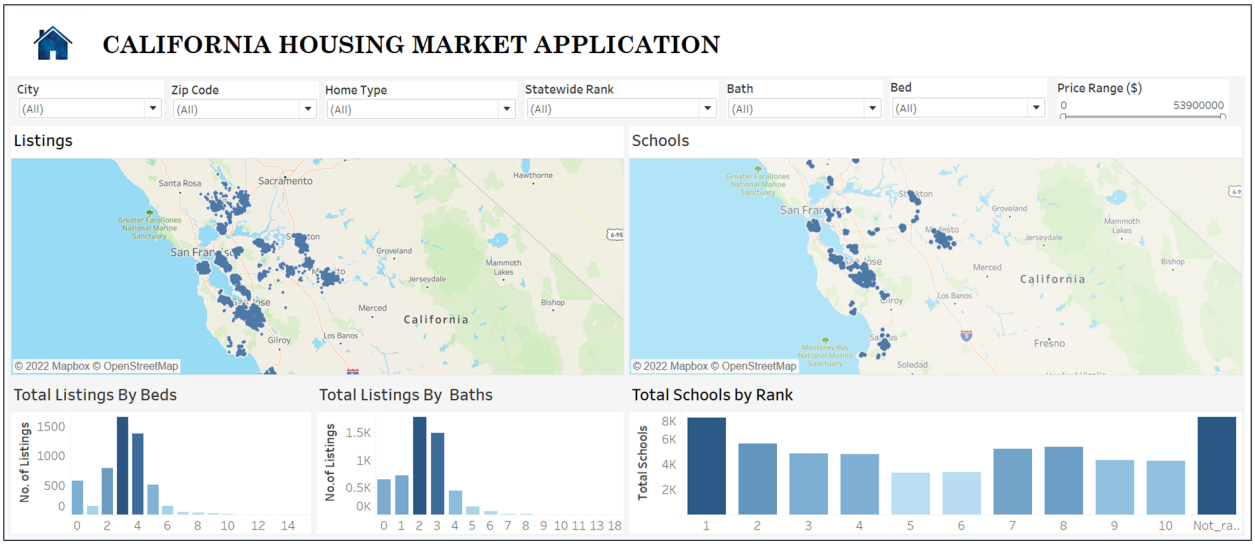


Figure 14

Listings Details Dashboard

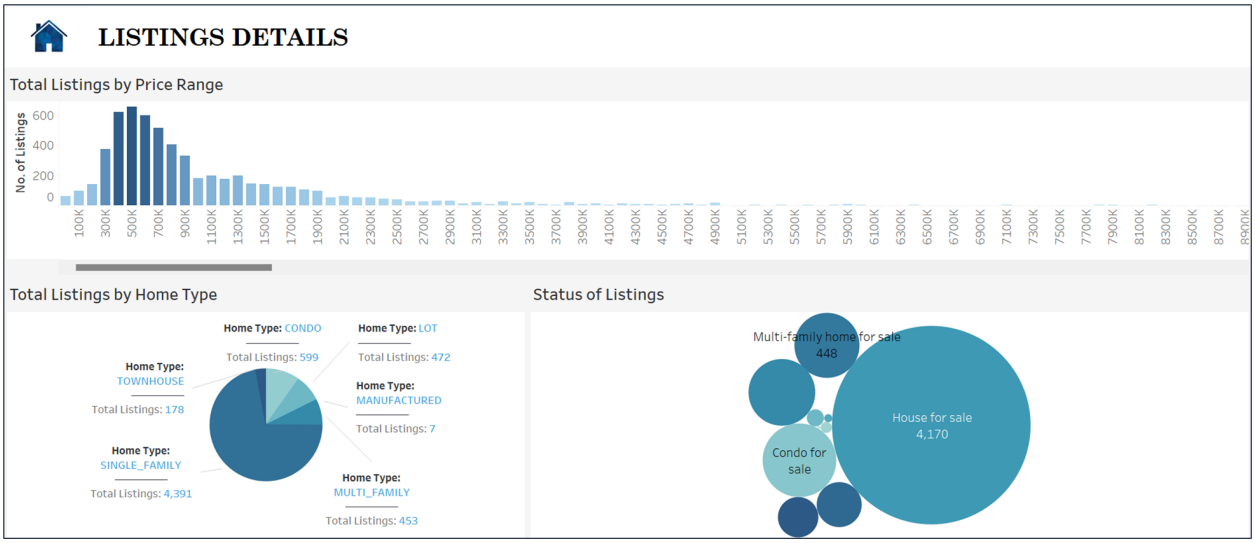
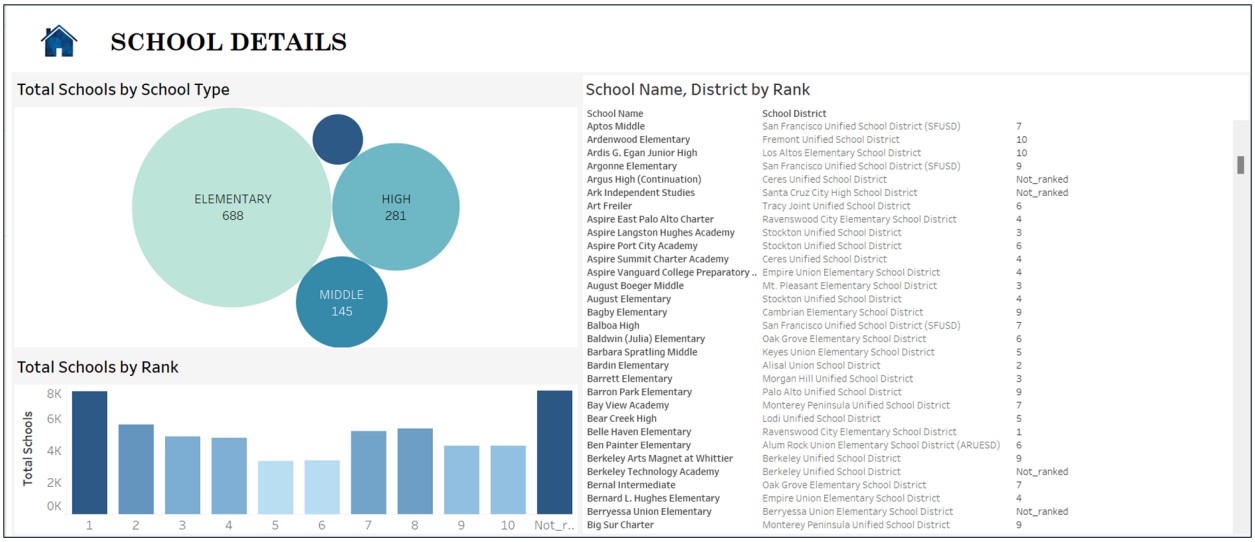


Figure 15

School Details Dashboard



The tableau dashboards are then embedded into the web page in the Wix website platform.

Figure 16

California Housing Market Application Site - I

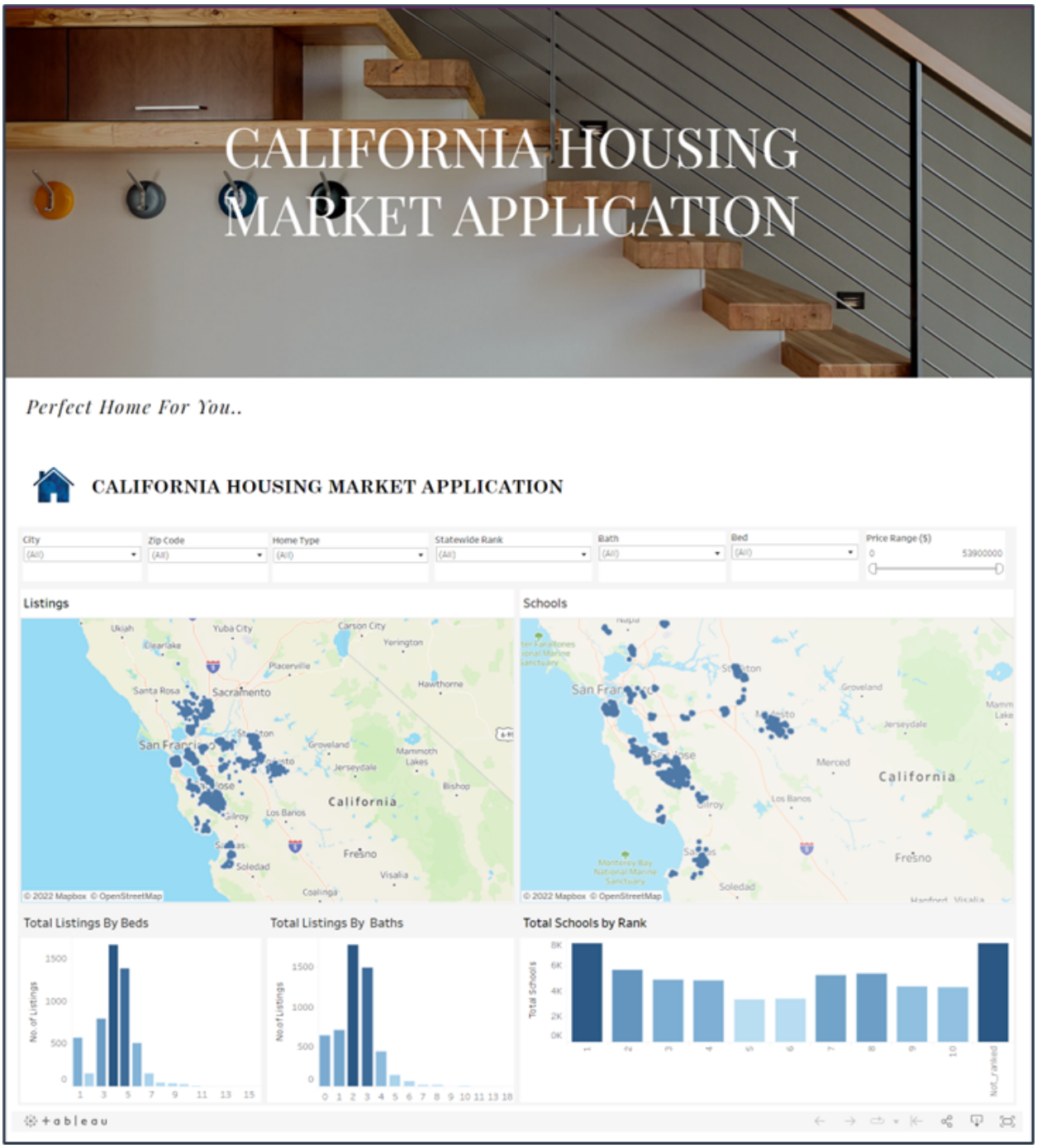
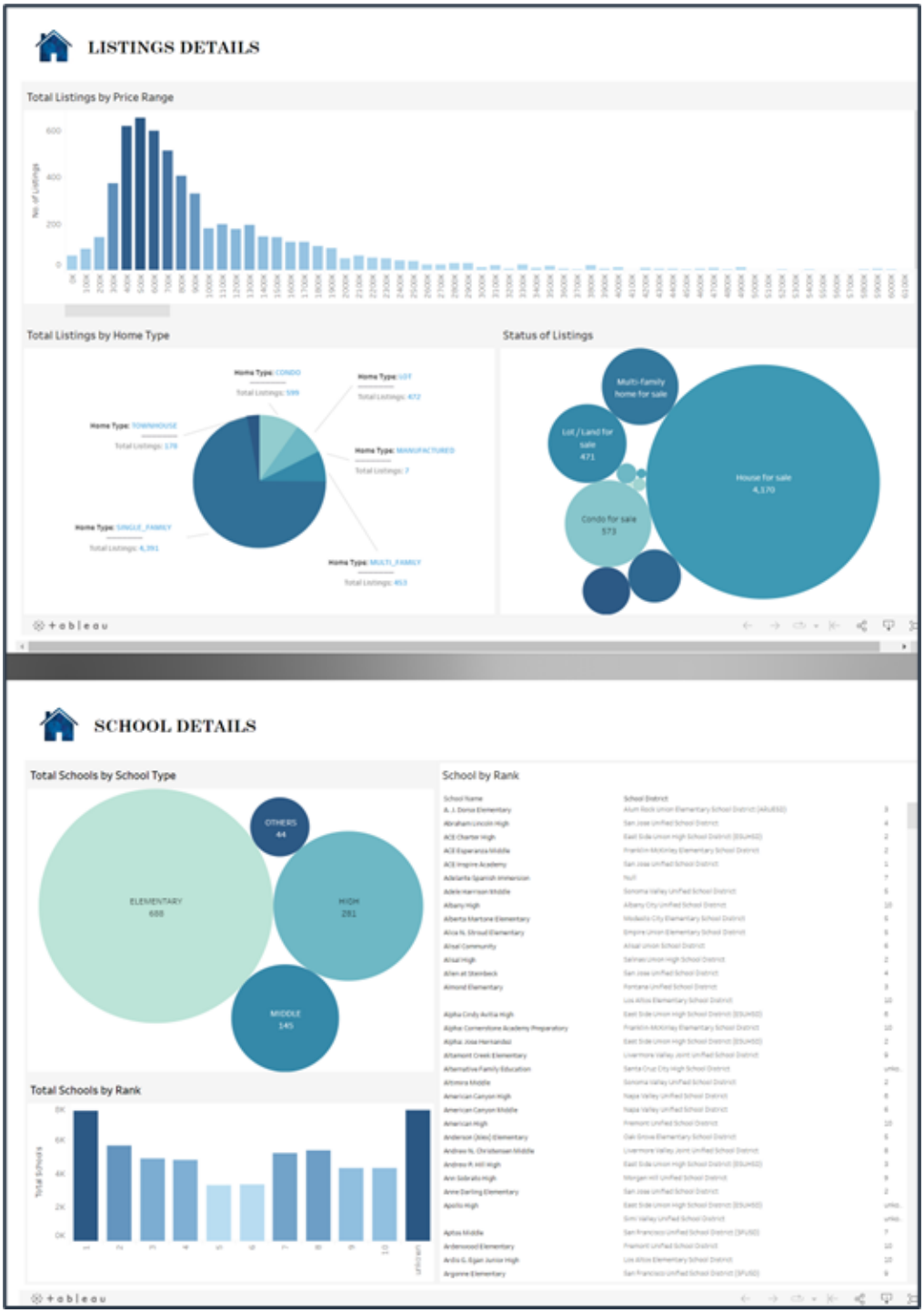


Figure 17

California Housing Market Application Site- II



4.2 System implementation issues and resolutions

The biggest hurdle at the beginning phase of the project was the data collection.Either the house data available online consisted of few 100's of records only or websites charged considerably for the data. Finally, the issue was resolved after the team made use of a web scraper to scrape the data from zillow. Another issue which was faced during implementation is creation of multiple IAM roles and policies. Failing to create a role/policy or even a small change leads to errors while using AWS Glue to transform data. This was fixed using AWS Cloud formation where the team used a .yaml file to load IAM roles, database and tables and fetch data from AWS S3 to AWS Glue table automatically.

4.3 Used technologies and tools

Below table shows various tools and technologies the team has used throughout the developmental process to build the big data application.

Table 1

Tools and Technologies Used

Tools and Technologies	**Usage**
Python,Jupyter Notebook (Web Scraper)	To scrape house listings data from Zillow
AWS S3	AWS S3 stores the extracted house data and US school ranking data
AWS Glue	● Database to store data from S3 ● Tables to store data from S3 which lies inside database ● Data Transformation,ETL (Extract, Transform and Load) and store back to S3

Tools and Technologies	Usage
AWS SageMaker	To implement Linear Regression Machine Learning Model to predict the house prices for the extracted listings correlated with US school ranking data
AWS Athena	To query the data from the AWS Glue tables and connect to Tableau
Tableau	Dashboard creation to display visuals
Website	Frontend UI where the tableau dashboard is hosted

Chapter 5 System Testing and Experiment

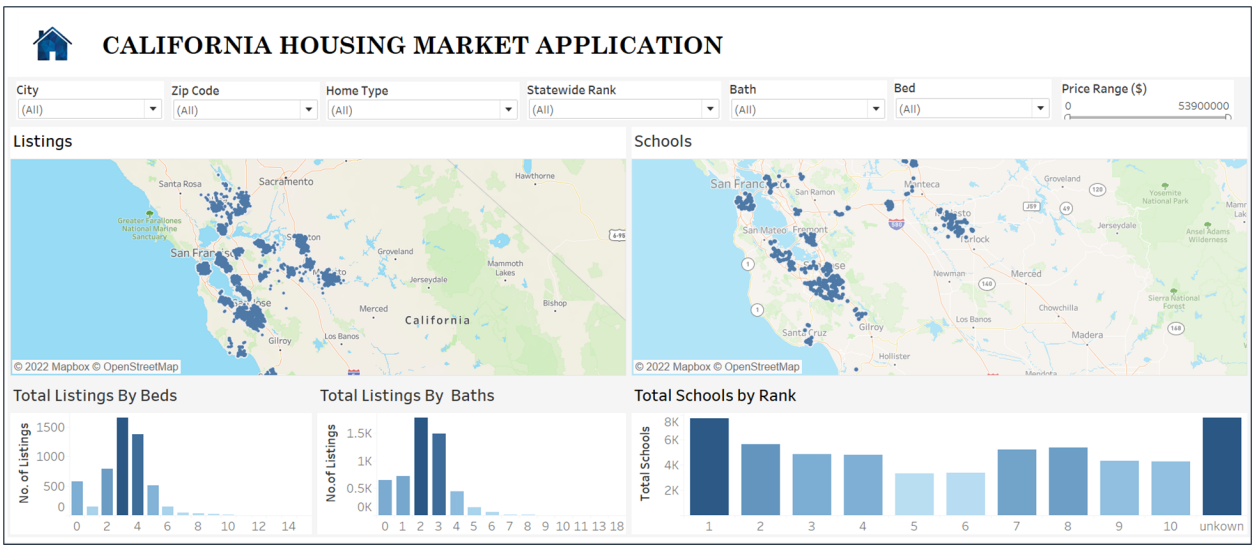
5.1 Case study

The final stage of the architecture involves building interactive dashboards and interface for customers to help find the perfect home as per their requirements. For our project, we have employed Tableau, a data visualization and analytical tool as our medium to build visuals. We have created multiple dashboards as per the application requirements which helps people filter them as per their desire making it easy to one stop destination to find their dream home.

The predicted data is pulled into Tableau desktop from AWS S3 buckets through Amazon Athena for visualizations. To establish connection between AWS S3 buckets where the final processed data is stored and Tableau through Amazon Athena, necessary drivers are installed, and a data pipeline is established. This allows Tableau to get the processed data directly through the pipeline

Figure 18

Home Page Dashboard



This dashboard gives an overview of all the house listings available in California and their attributes like location, city, zip code, number of bedrooms, number of baths, their price etc. It also gives information about the list of schools available in California and their location making it easier for customers to know the house's details along with the nearby school's information. The customer by adding the features they are looking for using the filters can see the location of houses that fit the parameters, schools nearby, number of beds and baths available and the ranking of the school available.

Figure 19

Search Filters

The filter section has filters based on the features of the house like city, zip code, home type, bath, bed, and price range making it easy for customers to filter their search based on their requirements. The city filter has a list of all the cities present in California while zip code filters

provide the city zip codes to enable customers to search the houses using either of the options. The home type filter has a list of all the home types available like apartments, condos, single-family, multi-family etc. The bed and bath filters have information about the number of bedrooms and bathrooms available respectively which can be used to limit the search further. The price range filter has information about the price range available in USD which makes it easy for customers to narrow down their search. In addition to all these filters, we have also added a Statewide rank filter which contains the rank of all schools present in the California state based on which one can filter the houses which are nearer to the schools that are highly ranked.

Figure 20

Listings by Map Visual

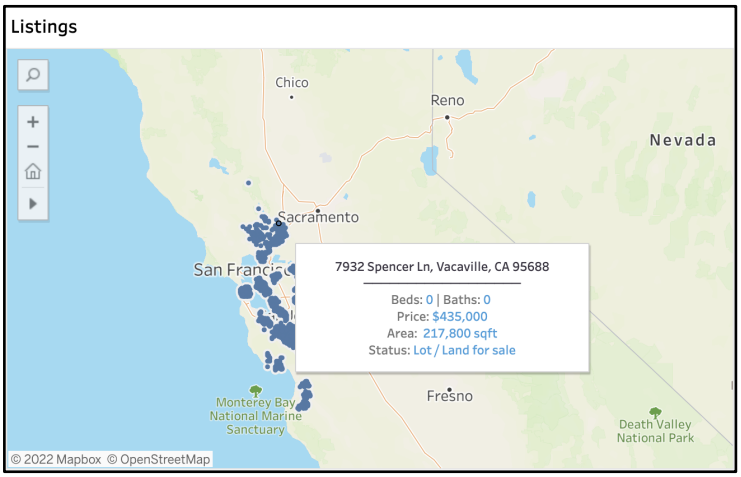
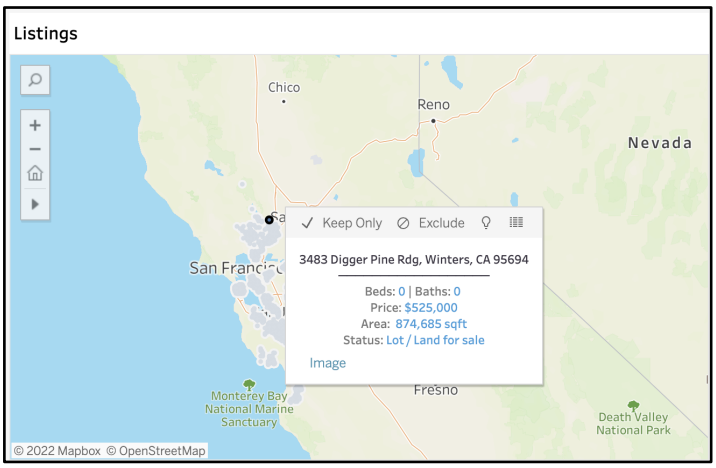


Figure 21

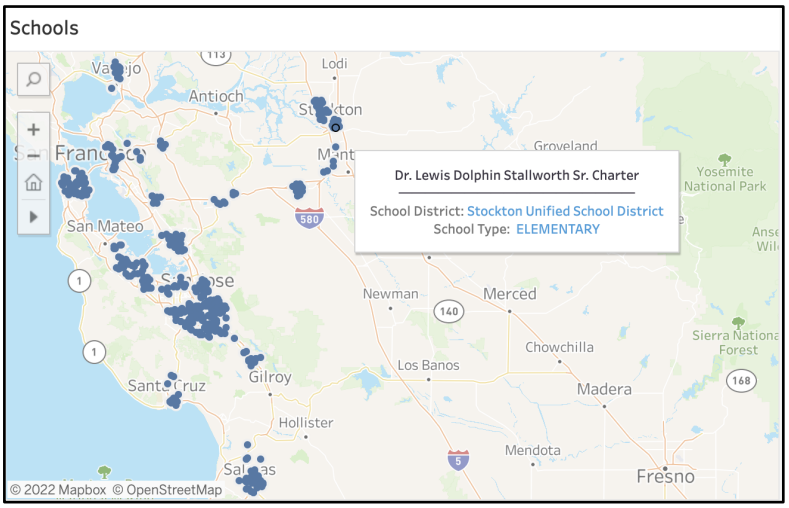
Listings by Map Visual with Image URL in Tooltip



In this Map visual, we can find the location of house listings spread across the state of California. When we hover on each data point, we can see the address of the house, number of beds and baths available, the price of the house, the area of the house and the type of the house details in the tooltip. We can also see the image of the house by clicking on the image option available in the tooltip.

Figure 22

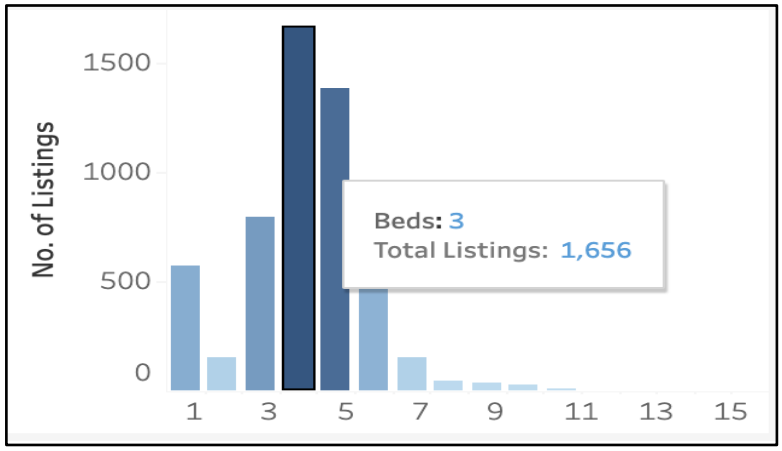
Schools by Map Visual



Similarly, in the map visual, we can see the location of schools present in California along with the name of the school and which school district it falls under in the tooltip. When the customer selects any particular house in the Listings map, he/she can also see the schools available within the vicinity of the house and along with their rank from the total schools by rank visual all at once place which helps them gain more insight about the house.

Figure 23

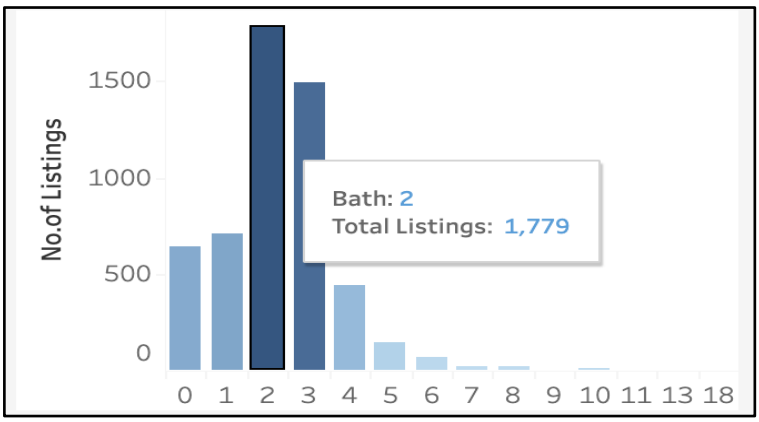
Total Listings by Beds Visual



In this Bar chart, we can see the distribution of house listings based on the number of bedrooms present. We can see that currently there are more house listings with four and five bedrooms and very few listings that have more than eight bedrooms.

Figure 24

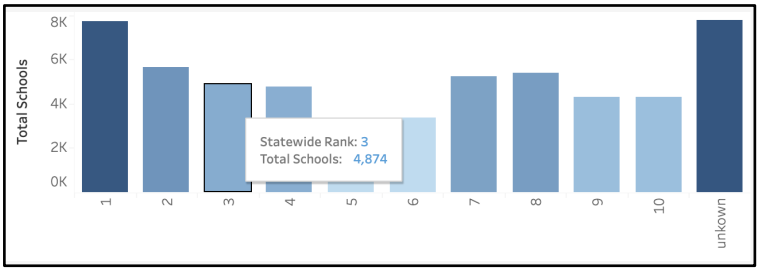
Total Listings by Baths Visual



In this Bar chart, the distribution of house listings based on the number of bathrooms available can be seen. We can see that currently the number of listings with two and three bathrooms are more compared to listings with more than six bathrooms.

Figure 25

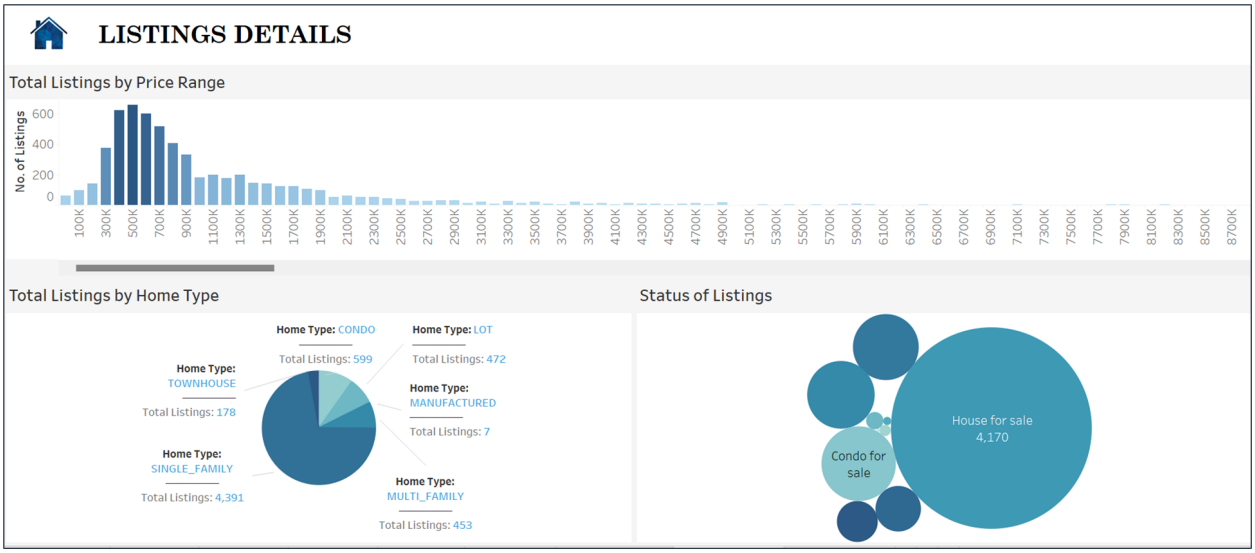
Total Schools by Rank Visual



In this Bar chart, the distribution of schools can be seen based on the statewide rank achieved and can be seen that though there are more schools that are highly ranked, there are few schools that are not yet ranked as they might be newly constructed schools which are not yet evaluated.

Figure 26

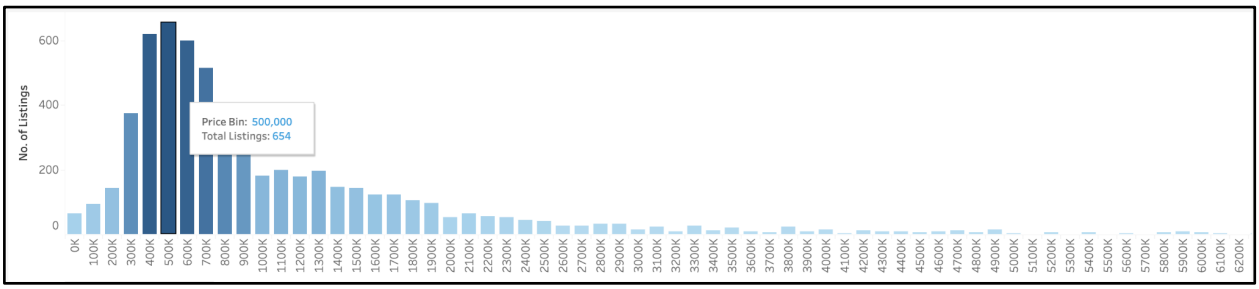
Listings Details Dashboard



This Dashboard gives an overview of details of the house listings like how the houses listed are distributed based on the price range, type of the home like condo, apartment, single-family, multi-family etc., and status of listings like whether it's a new construction or land for sale or condo for sale, and others. The customer can select any price bin and see what the distribution of houses in that price range is based on their house type and the status of the listings.

Figure 27

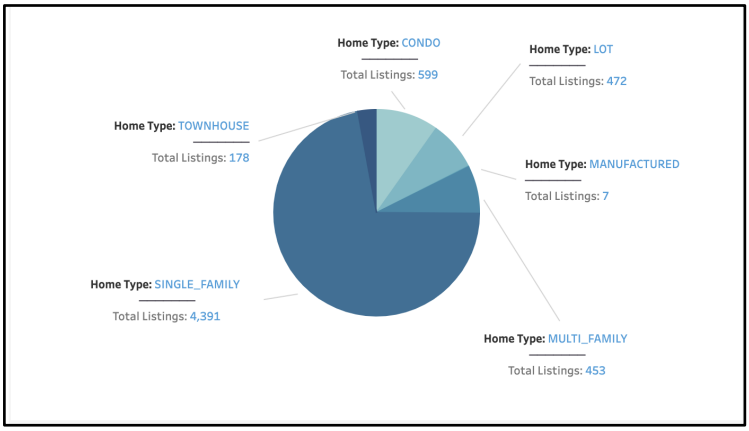
Total Listings by Price Range Visual



This Bar chart explains the distribution of listings based on their price. The prices are split in the form of bins with 100K each making it easier to see the number of listings falling within a specific price range. From the visual, we can see that currently there are a greater number of houses listed within the 400k – 700k price range. As the price range increases, there are fewer listings as we can see from the chart.

Figure 28

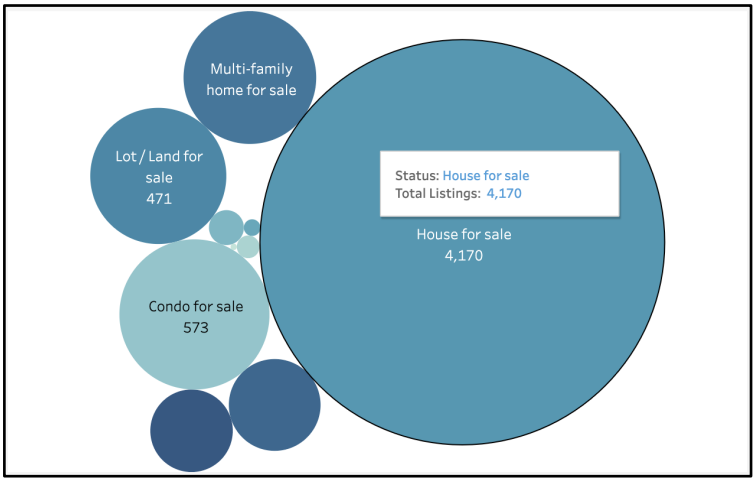
Total Listings by Home Type Visual



The pie chart contains details about the number of listings based on the type of home like Townhouse, condo, single-family, multi-family etc. we can see from the chart that single-family houses dominate the market followed by condo.

Figure 29

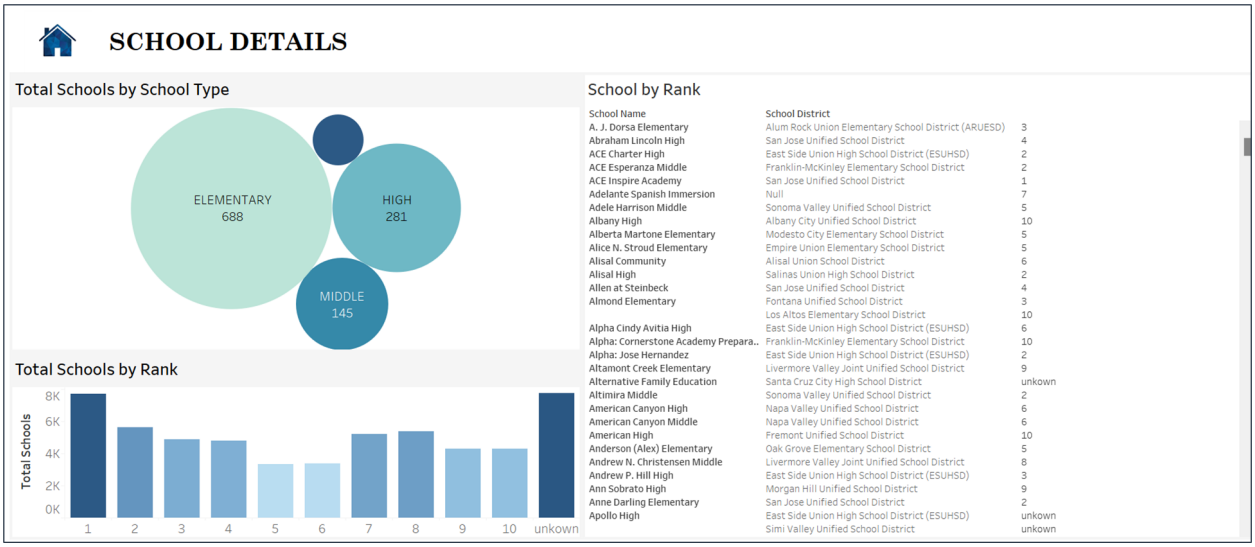
Status of Listings Visual



The bubble chart talks about the distribution of listings based on the status of listings like house for sale, land for sale, condo for sale etc., We can see from the chart that number of listings for House for sale category dominates the graph followed by condo for sale category.

Figure 30

School Details Dashboard

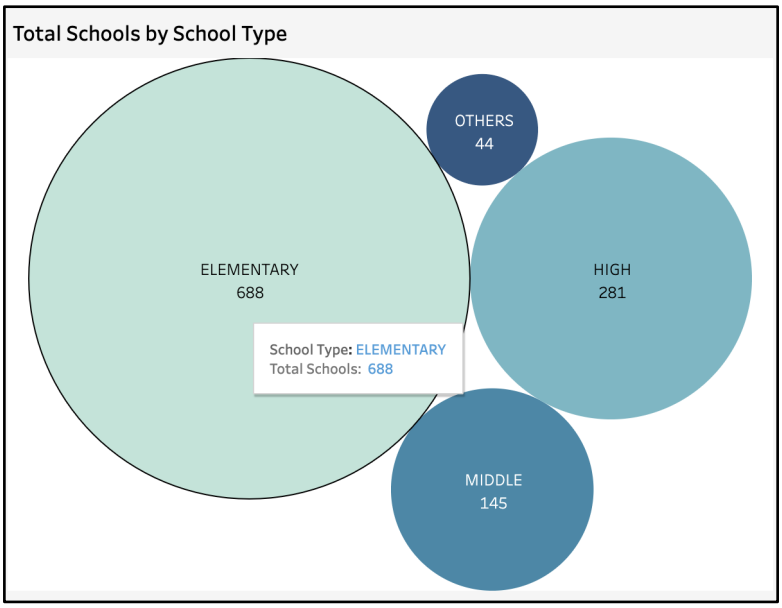


This dashboard gives an overview of details of schools like type of school, the school district they fall under, their rank and the number of schools available with the same rank. Any

customer can get information about the schools, by selecting any school name and know the type of the school, which districts it falls under and its statewide rank.

Figure 31

Total Schools by School Type Visual



The bubble chart gives information about the type of the school like elementary, middle etc., We can see from the chart that currently there are more elementary schools followed by high schools in California state.

Figure 32

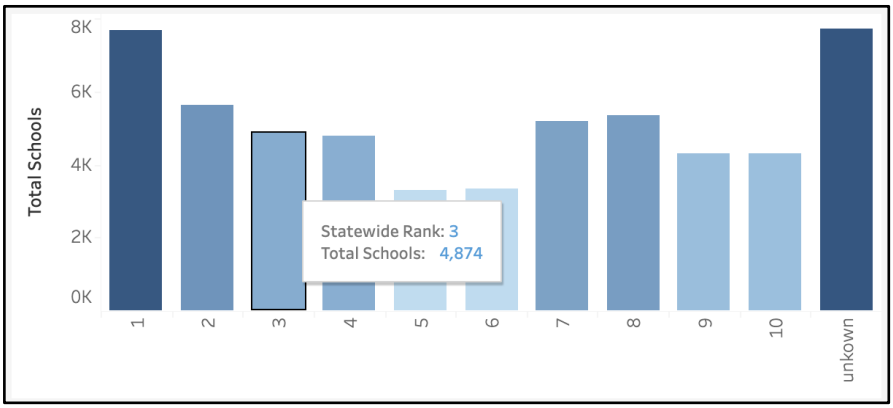
School Name, District by Rank Visual

School Name	School District	
A. J. Dorsa Elementary	Alum Rock Union Elementary School District (ARUESD)	3
Abraham Lincoln High	San Jose Unified School District	4
ACE Charter High	East Side Union High School District (ESUHSD)	2
ACE Esperanza Middle	Franklin-McKinley Elementary School District	2
ACE Inspire Academy	San Jose Unified School District	1
Adelante Spanish Immersion	Null	7
Adele Harrison Middle	Sonoma Valley Unified School District	5
Albany High	Albany City Unified School District	10
Alberta Martone Elementary	Modesto City Elementary School District	5
Alice N. Stroud Elementary	Empire Union Elementary School District	5
Alisal Community	Alisal Union School District	6
Alisal High	Salinas Union High School District	2
Allen at Steinbeck	San Jose Unified School District	4
Almond Elementary	Fontana Unified School District	3
	Los Altos Elementary School District	10

This table contains details of schools like name of the school, the school district it falls under and its statewide rank. Note that for some schools there are multiple ranks in the table as some schools have multiple branches in different school districts with the same name.

Figure 33

Total Schools by Rank Visual



In this Bar chart, the distribution of schools can be seen based on the statewide rank achieved and can be seen that though there are more schools that are highly ranked, there are few schools that are not yet ranked as they might be newly constructed schools.

Chapter 6 Conclusion and Future Work

6.1 Project summary

The demand for the housing market in California and absence of an ideal search platform has made it difficult for the people to find the houses of their choice. Hence, to ease this process the team has built a big data application 'California Housing Market Application' which can scale large amounts of housing data where home buyers can look for the desired houses in the area of their choice. Throughout this project, the team has made use of various AWS services like Amazon S3, Amazon Glue, Amazon SageMaker, Amazon Athena and Tableau to scale, store, transform, predict and visualize the house and school data. The application gives home buyers the insights on the housing trends in California, flexibility to search houses in the desired

area, the house price prediction and impact on house prices due to the presence of schools. This can help buyers make informed decisions.

6.2 Future work

As a future work, there is a scope for automation of the entire process of the application building by extending the scope of the Cloud Formation template from creating resources required for ETL to creating all resources required for system implementation. An automated event triggering through Lambda function will serve as an extension that would enable a seamless connectivity across the AWS components. Also, stream processing through AWS Kinesis for web scraping from Zillow would ensure live capture and analytics of data. Combination of all these steps would make the housing application an end-to-end product that can further be hosted as a service for usage of the general public.

References

- https://docs.aws.amazon.com/AmazonS3/latest/userguide/Welcome.html
- https://www.ijert.org/research/big-data-analytics-on-aws-cloud-IJERTV10IS040325.pdf
- https://docs.aws.amazon.com/AWSCloudFormation/latest/UserGuide/cfn-whatis-howdoesitwork.html
- https://datadiverse.com/connect-s3-data-using-athena-with-tableau-desktop/
- https://www.google.com/imgres?imgurl=https%3A%2F%2Fwww.pngitem.com%2Fpimgs%2Fm%2F192-1925872_insurance-agent-house-real-estate-business-home-insurance.png&imgrefurl=https%3A%2F%2Fwww.pngitem.com%2Fmiddle%2Fihbibm T_insurance-agent-house-real-estate-business-home-insurance%2F&tbnid=N4OeOh_PV-bgkM&vet=12ahUKEwi8wpX8xL77AhXQlGoFHYP6B_0QMygCegUIARCDAg..i&docid=bVEOCGh5KWAjQM&w=860&h=625&q=home%20logo%20HD&ved=2ahUKEwi8wpX8xL77AhXQlGoFHYP6B_0QMygCegUIARCDAg
- https://www.dreamstime.com/illustration/california-houses.html
- https://templates.office.com/en-us/presentations
- https://homeshots.us/
- https://www.ccsa.org/what-we-do/student-success
- https://www.zillow.com/homes/for_sale/
- https://www.tableau.com/
- https://public.tableau.com/app/discover
- https://www.zillow.com/https://jupyter.org/
- https://www.wix.com/

- https://www.dreamstime.com/illustration/california-houses.html
- https://templates.office.com/en-us/presentations

- https://homeshots.us/
- https://docs.aws.amazon.com/AWSCloudFormation/latest/UserGuide/cfn-whatis-howdoesitwork.html

Appendices

Table 2

Appendix A – Hardware/Software Standards

Criteria	Remarks
Used Big Data concepts from this Course DATA 228	Yes
This criterion is linked to a Learning Outcome Significance to the real world	Yes
Code Walkthrough	Yes https://github.com/DivyaCheerla/Data228_Datworks_Project (github)
Version Control Use of Git / GitHub or equivalent; must be publicly accessible	Yes https://github.com/DivyaCheerla/Data228_Datworks_Project (github)
This criterion is linked to a Learning Outcome Discussion / Q&A	Yes

Lessons learned Included in the report and presentation?	Yes
This criterion is linked to a Learning Outcome Innovation	Yes Used AWS features
This criterion is linked to a Learning Outcome Teamwork	Yes
This criterion is linked to a Learning Outcome Technical difficulty	The data consisted of null values, duplicates and inconsistent data which resulted in inefficient results. To solve the problem and improve the efficiency of the analytical queries we used AWS Glue, as an ETL tool to clean and transform our dataset.
AWS Tools	S3, AWS Glue, Amazon Athena, Amazon SageMaker
Used ETL tool	Python, AWS Glue
Visualization Tool	Tableau, Wix Website