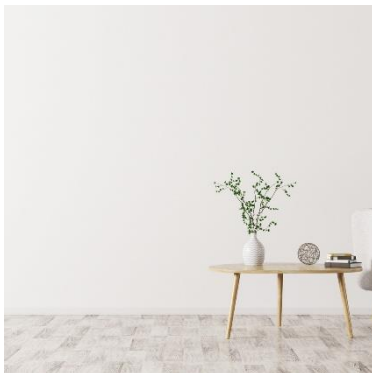


# California Housing Market Application



# Agenda



## I. Introduction

- a. Problem and Motivation
- b. Project Goals and Objective

## I. System Design

- a. Architecture Design
- b. Data Collection
- c. Data Model

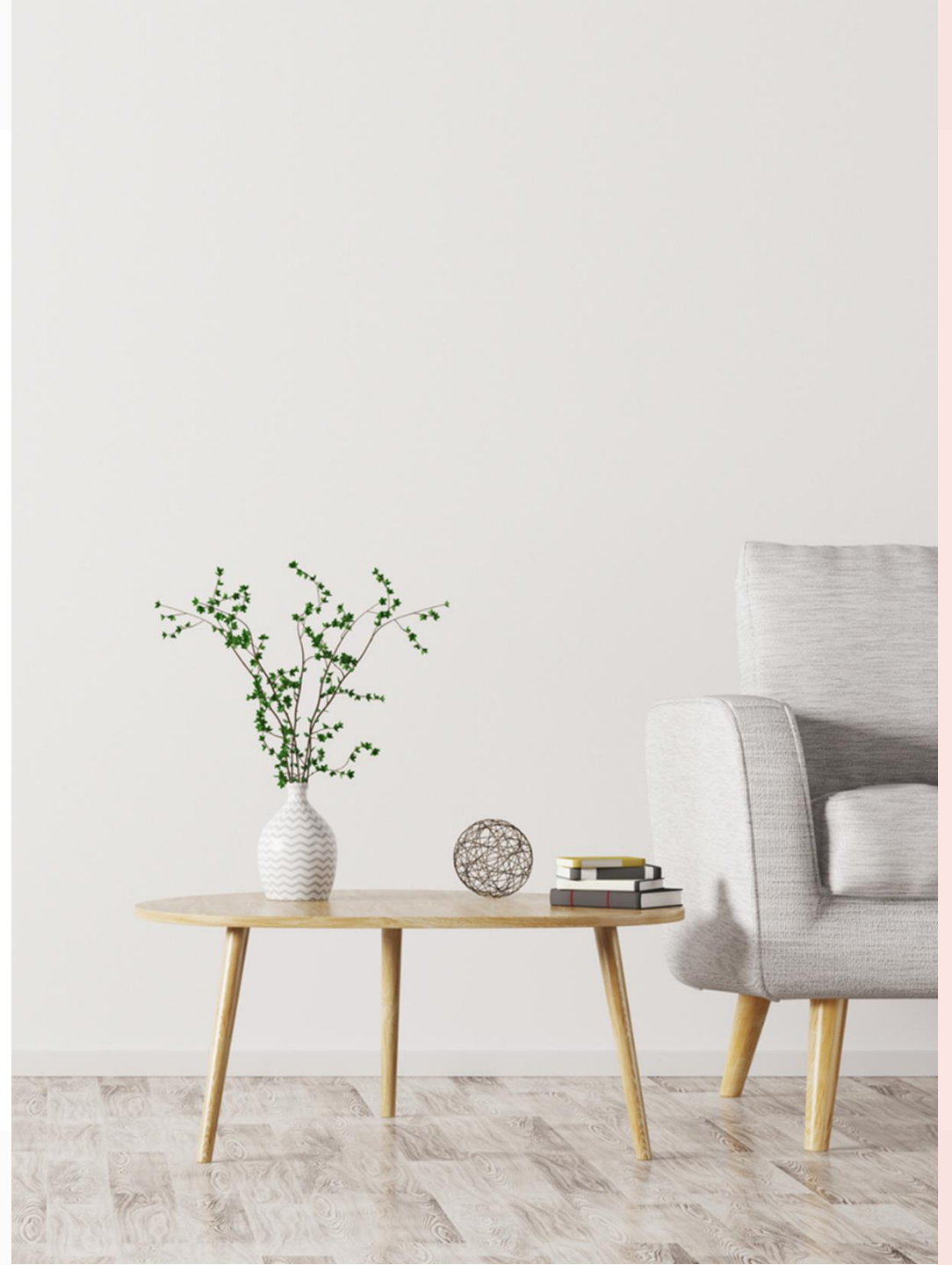
## II. System Implementation

- a. Setup AWS resources
- b. Data Storage
- c. Extract, Transform, and Load (ETL)
- d. Modeling
- e. Analytics and Visualization

## III. Conclusion and Future Work

# I. Introduction

- a. Problem and Motivation
- b. Project Goals and Objective



# 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 property values such as:

- School proximity
- Migration of people impacting the population density
- Crime rate
- Job opportunities
- New construction
- Recreational centers
- Medical care facilities
- Access to public transportation, and others





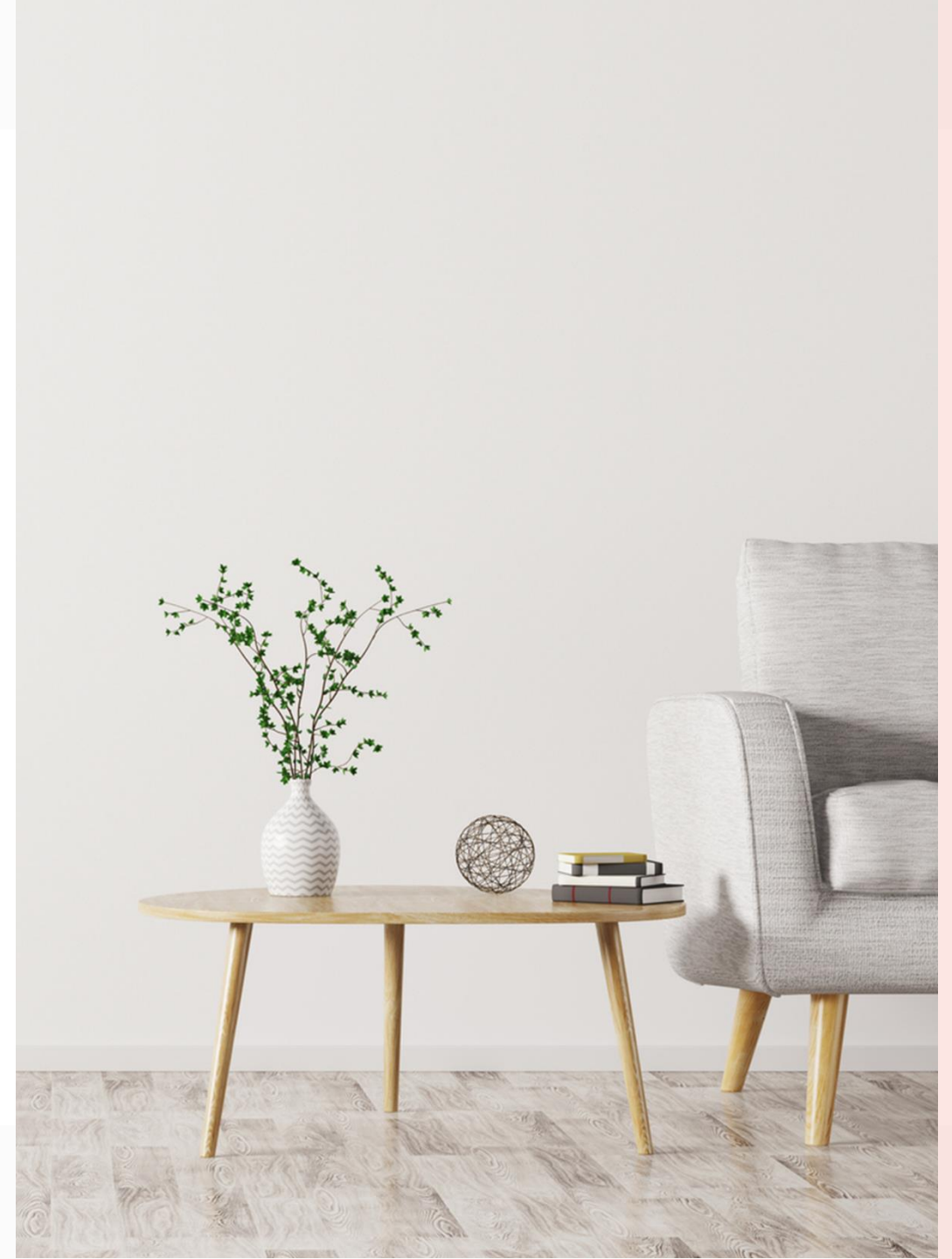


# Project Goals and Objectives

- Build California housing market application to ease the process of housing property search for the customers based on their preferred buying criteria
- Influence on house/property values by the presence of schools with higher ranking in the region

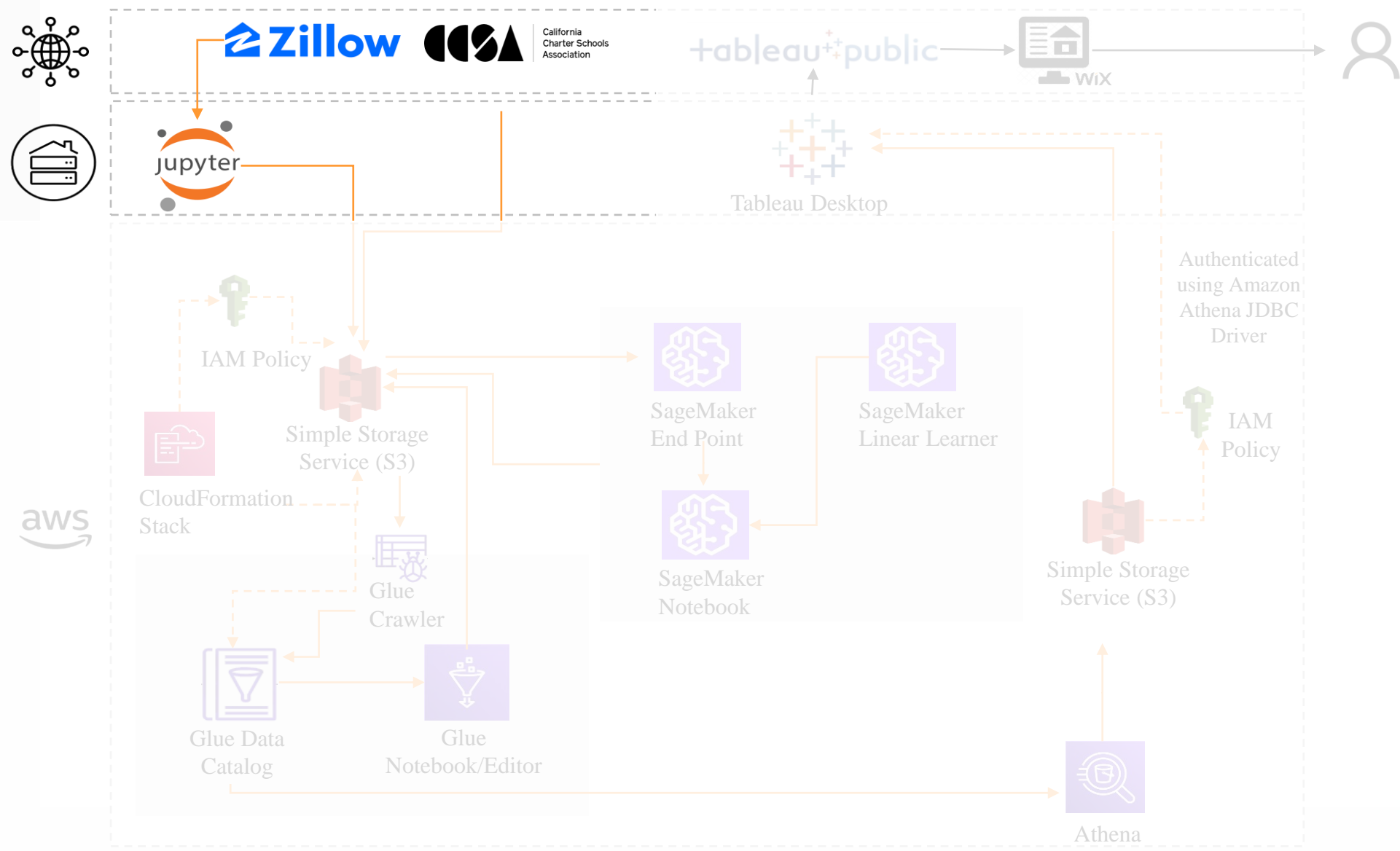
## II. System Design

- a. Architecture Design
- b. Data Collection
- c. Data Model



**SJSU** SAN JOSÉ STATE UNIVERSITY

# Data Collection





# Data Collection: Housing Data + School Ranking



## Housing Data:

- Data on the house 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



## School Ranking Data:

- California Charter Schools Association 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.

housing\_etl\_output\_11192...

school\_ranking

housing\_e... — school\_ra...

100 → rows ⚙️ ▼

How do relationships differ from joins? [Learn more](#)

housing\_etl\_output...

Operator

school\_ranking

# Addresszipcode ▼

= ▼

# School Zipcode ▼

⊕ Add more fields

▼ Performance Options

These settings help Tableau optimize queries during analysis. The default settings are recommended, if you aren't sure what to choose. [Learn more](#)

Cardinality

One ▼

Many ▼

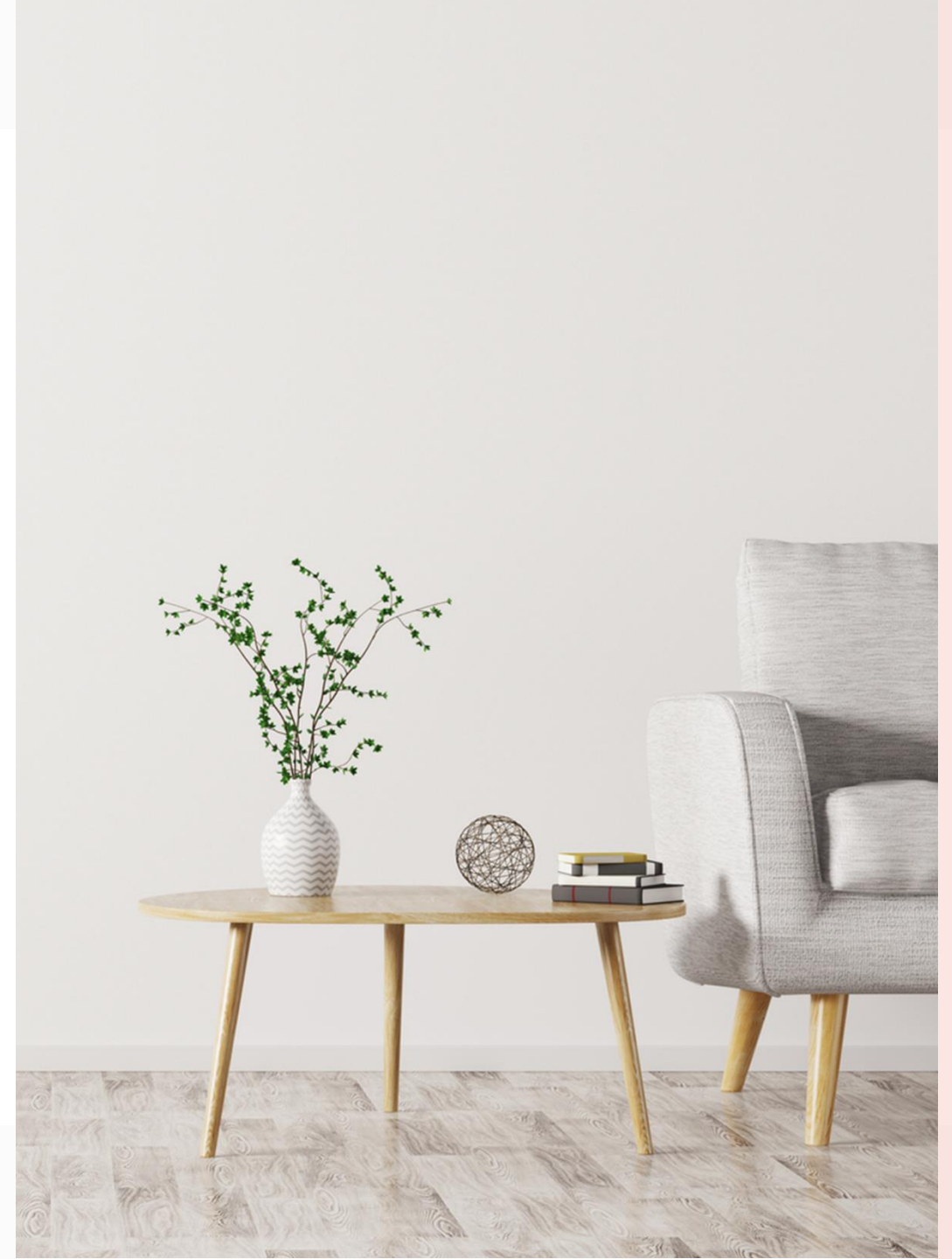
#	Abc	Abc
School!Data!Manually!Edited	School!Data!Manually!Edited	School!Data!Manually!Edited
School Code	School Name	School Address
129882	21st Century Learning Institute	939 E. 10TH ST.
6027767	A. E. Arnold Elementary	9281 DENNI ST.
6085377	A. G. Currie Middle	1402 SYCAMORE AVE.
6046114	A. J. Dorsa Elementary	1290 BAL HARBOR DR.
105692	A. L. Conner Elementary	222 FOURTH ST.
6102792	A. M. Thomas Middle	20979 LOBOS CT.
6033765	A. M. Winn Waldorf-Inspired	3351 EXPLORER DR.
6044796	Abbott Middle	600 36TH AVE.
6116446	Abbv Reinke Elementary	43799 SUNNY MEADOWS DR.

housing_etl_output
zpid
imgsrc
hasimage
detailurl
statustext
price
unformattedprice
address
addressstreet
addresscity
addressstate
addresszipcode
bed
bath
areacode
hasopenhouse
openhousedescription
Estimated yearly cost
has3dmodel
hasadditionalattributions
brokername
homeinfo_latitude
homeinfo_longitude
homeinfo_hometype
homeinfo_lotareavalue
homeinfo_lotareaunit
Avg School Rank
Sum School Enrollment

school_ranking
school_code
school_name
school_address
school_city
school_state
school_zipcode
school_areacode
school_latitude
school_longitude
school_type
ENROLLMENT
ST_GRADE
END_GRADE
FT_TEACHER
statewide_rank
similar_students_rank
college_career_percent_prepared
school_county
school_district
Authorizer
Statewide_Percentile

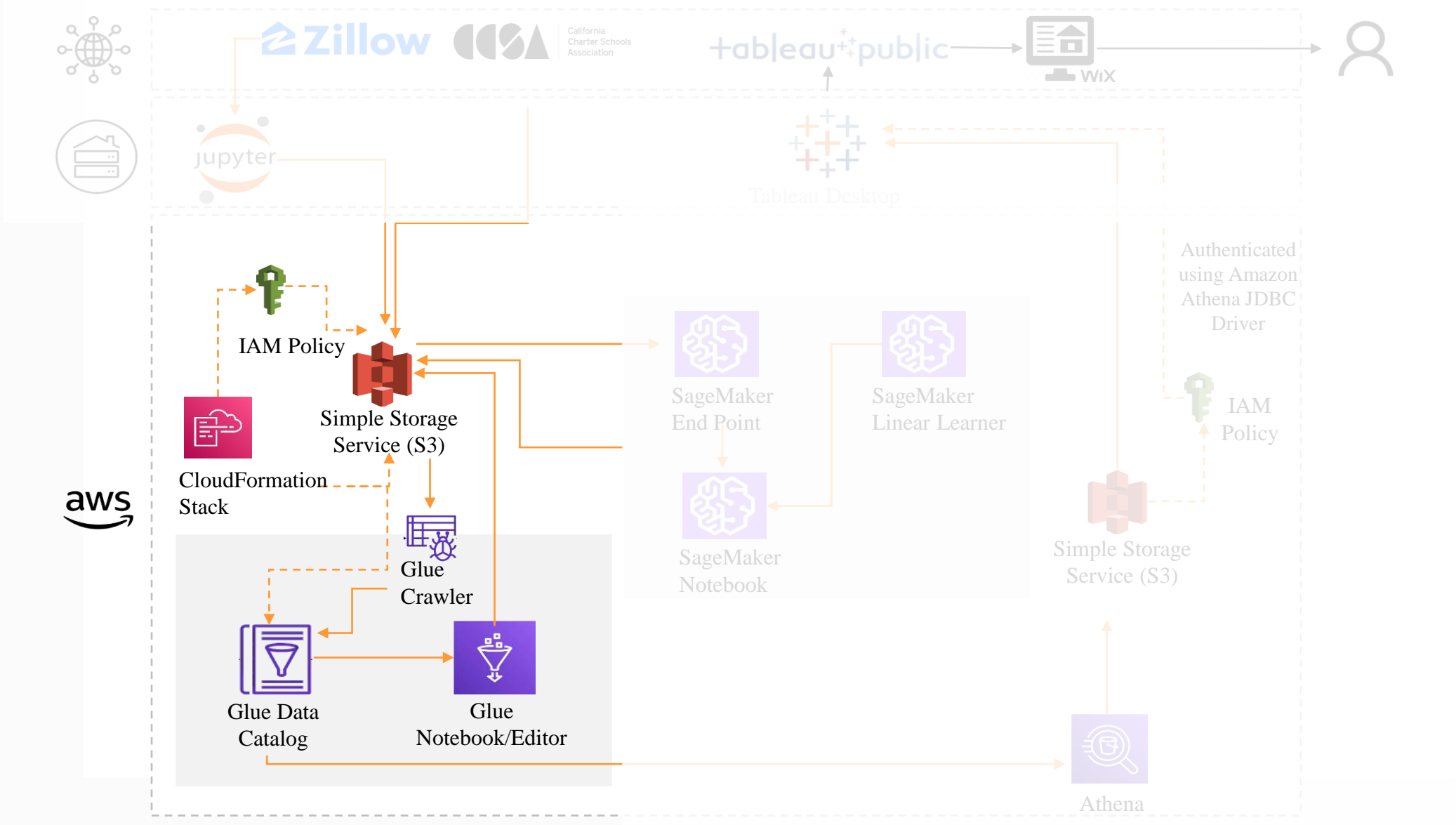
# III. System Implementation

- a. Setup AWS resources
- b. Data Storage
- c. Data Transformation (ETL)
- d. Modeling
- e. Analytics and Visualization



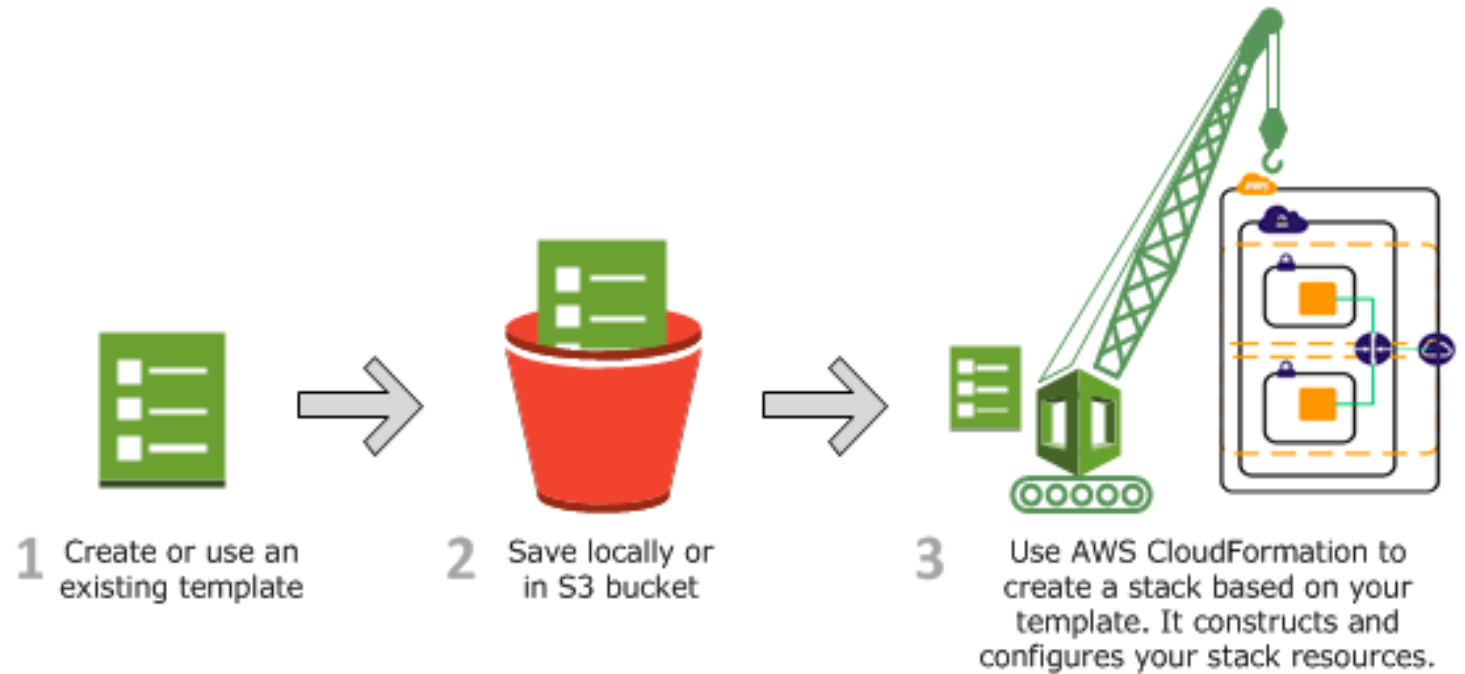


# Setup AWS resources, Data Storage, ETL



# Setup AWS resources: AWS Cloud Formation

Automate resource  
creation for AWS  
GLUE



# Setup AWS resources: AWS Cloud Formation

.yaml file to configure the resources

```

school_zillow_combined.yaml

- Effect: Allow
  Action:
    - "glue:*"
    - "s3:*"
    - "ec2:DescribeVpcEndpoints"
    - "ec2:DescribeRouteTables"
    - "ec2:CreateNetworkInterface"
    - "ec2:DeleteNetworkInterface"
    - "ec2:DescribeNetworkInterfaces"
    - "ec2:DescribeSecurityGroups"
    - "ec2:DescribeSubnets"
    - "ec2:DescribeVpcAttribute"
    - "iam:ListRolePolicies"
    - "iam:GetRole"
    - "iam:GetRolePolicy"
    - "cloudwatch:PutMetricData"
    - "ec2:CreateTags"
    - "ec2:DeleteTags"
    - "logs:CreateLogGroup"
    - "logs:CreateLogStream"
    - "logs:PutLogEvents"
  Resource: "*"
- Effect: Allow
  Action:
    - "s3:*"
  Resource:
    - "arn:aws:s3:::{S3PySparkBucketName}/*"
    - "arn:aws:s3:::{S3PySparkBucketName}/"
- Effect: Allow
  Action:
    - "iam:GetRole"
    - "iam:PassRole"

```

Loads database, IAM roles, data tables with schema, S3 bucket

CloudFormation > Stacks > etl-glue-load-1

etl-glue-load-1 Delete Update Stack actions Create stack

Stack info | Events | **Resources** | Outputs | Parameters | Template | Change sets

Stacks (2) Filter by stack name

Active View nested < 1 >

etl-glue-load-1  
2022-11-20 17:53:45 UTC-0800  
CREATE\_COMPLETE

etl-glue-load  
2022-11-18 15:04:33 UTC-0800  
DELETE\_FAILED

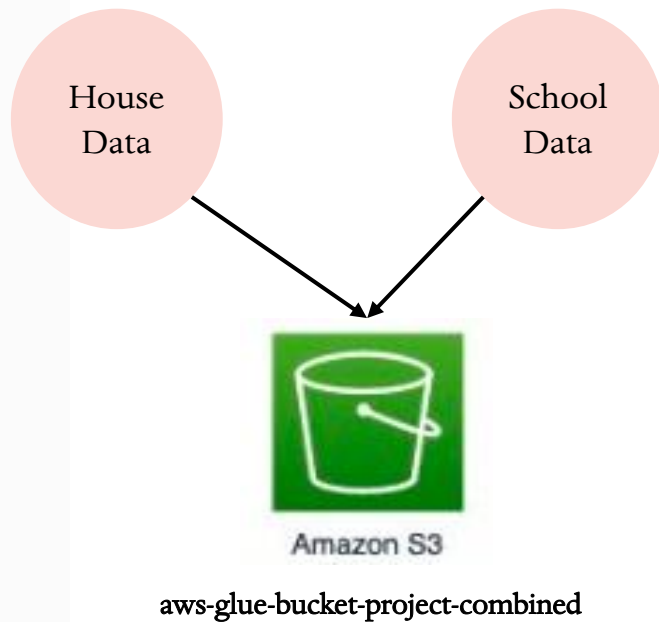
Resources (5) Search resources

Logical ID	Physical ID	Type	Status	Module
GlueDatabase	pyspark_tutorial_db	AWS::Glue::Database	CREATE_COMPLETE	-
GlueNotebookRole	aws-glue-role-project-combinedRole	AWS::IAM::Role	CREATE_COMPLETE	-
GlueSchoolTable	school	AWS::Glue::Table	CREATE_COMPLETE	-
GlueZillowTable	zillow	AWS::Glue::Table	CREATE_COMPLETE	-
S3BucketForData	aws-glue-bucket-project-combined	AWS::S3::Bucket	CREATE_COMPLETE	-



# Data Storage: AWS S3

Add school and house data into S3 bucket



**Objects (1)**

Objects are the fundamental entities stored in Amazon S3. You can use [Amazon S3 inventory](#) to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. [Learn more](#)

[Refresh](#) [Copy S3 URI](#) [Copy URL](#) [Download](#) [Open](#) [Delete](#) [Actions](#) [Create folder](#)

[Upload](#)

<input type="checkbox"/>	Name	Type	Last modified	Size	Storage class
<input type="checkbox"/>	<a href="#">House_Data_ForETL.csv</a>	csv	November 21, 2022, 00:16:37 (UTC-08:00)	11.4 MB	Standard

**Objects (1)**

Objects are the fundamental entities stored in Amazon S3. You can use [Amazon S3 inventory](#) to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. [Learn more](#)

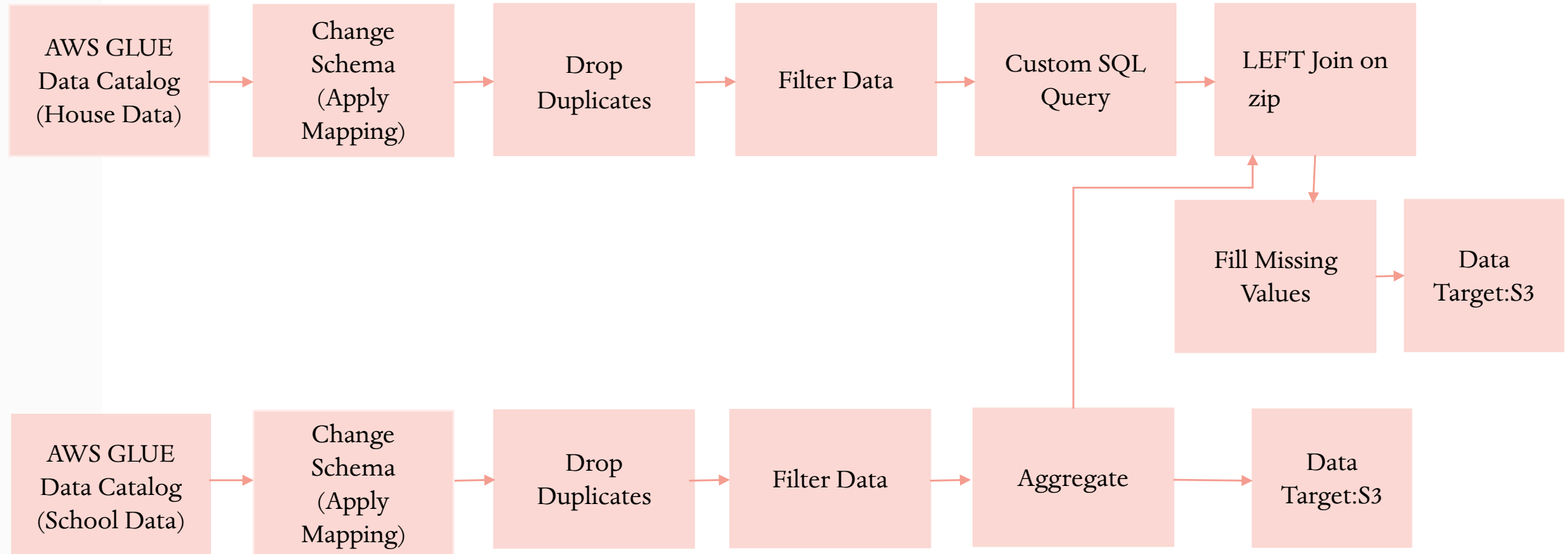
[Refresh](#) [Copy S3 URI](#) [Copy URL](#) [Download](#) [Open](#) [Delete](#) [Actions](#) [Create folder](#)

[Upload](#)

<input type="checkbox"/>	Name	Type	Last modified	Size	Storage class
<input type="checkbox"/>	<a href="#">School_Ranking_ForETL.csv</a>	csv	November 20, 2022, 17:59:58 (UTC-08:00)	2.6 MB	Standard

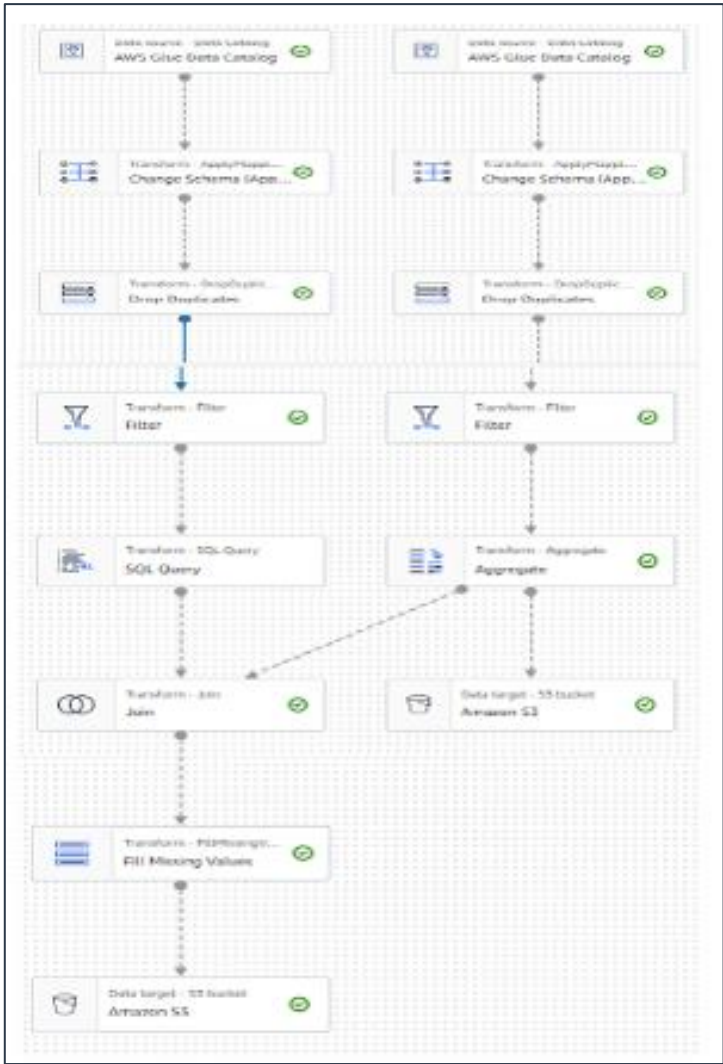
# Data Transformation (ETL): AWS GLUE

AWS GLUE visual editor flow to transform the datasets and load to data target S3

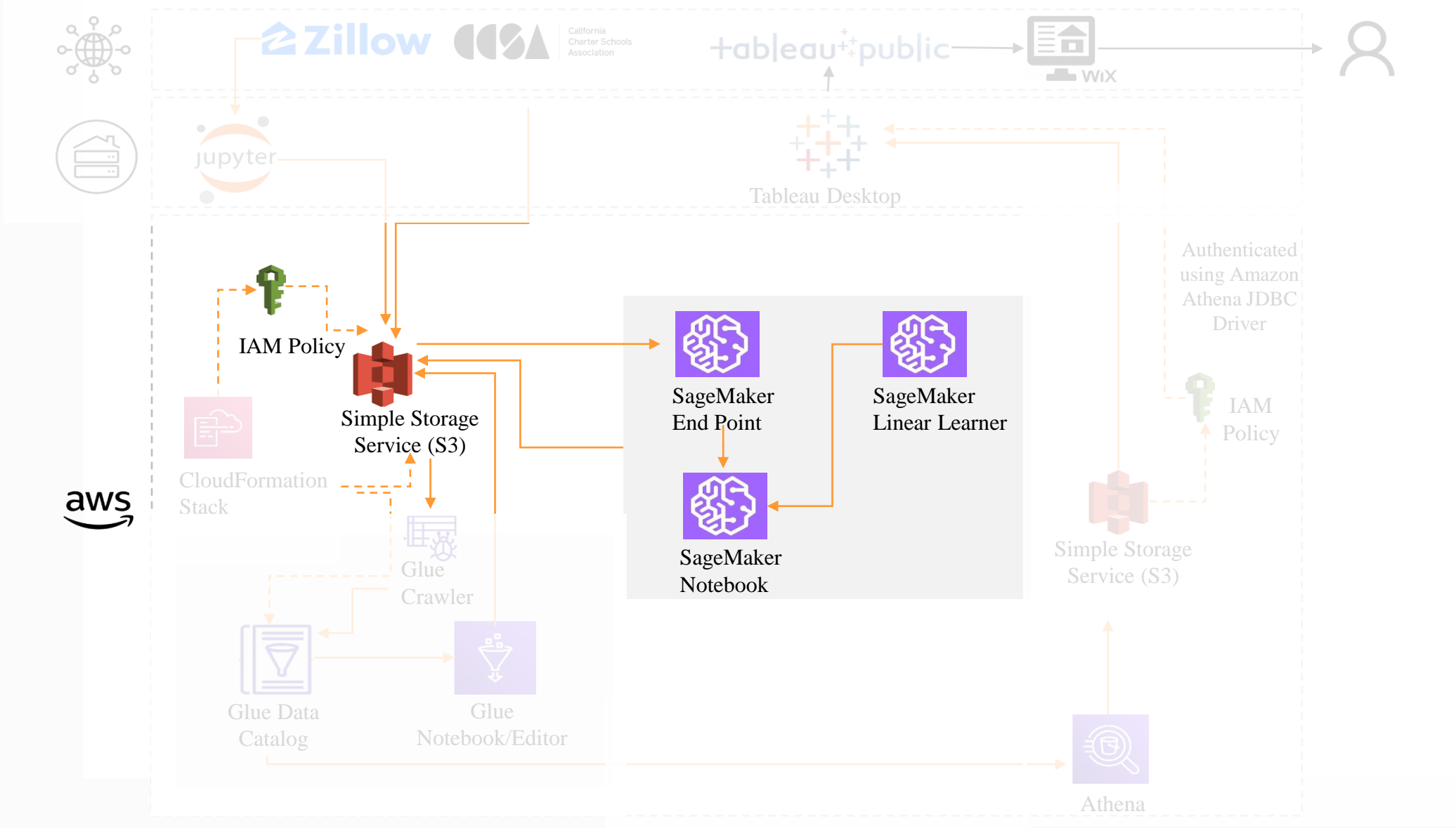


# Data Transformation (ETL): AWS GLUE Visual Editor Flow Diagram

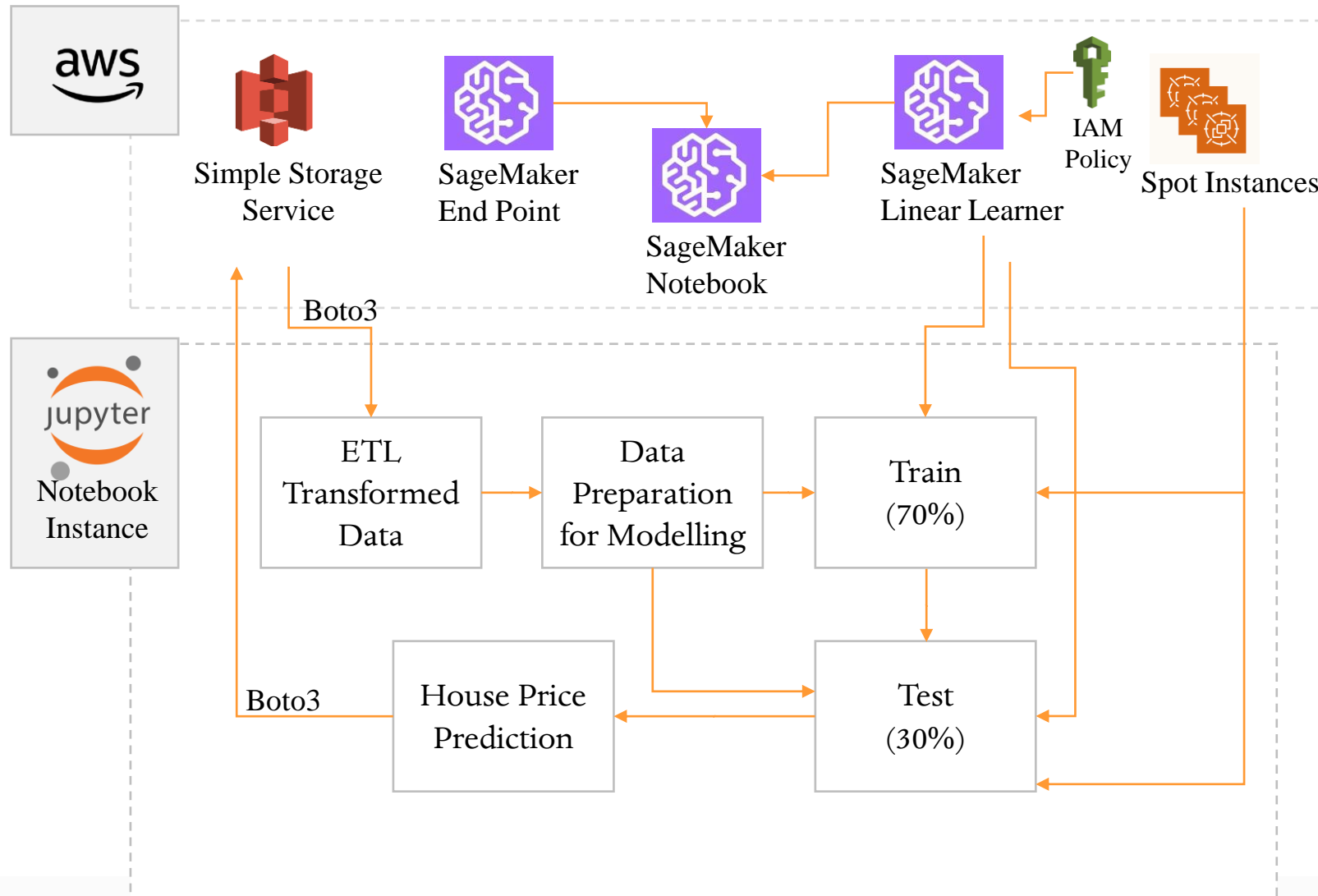
Transforming data from AWS  
GLUE Data Catalogue using AWS  
Glue's ETL Job







# Modeling: SageMaker Architecture and Process Flow



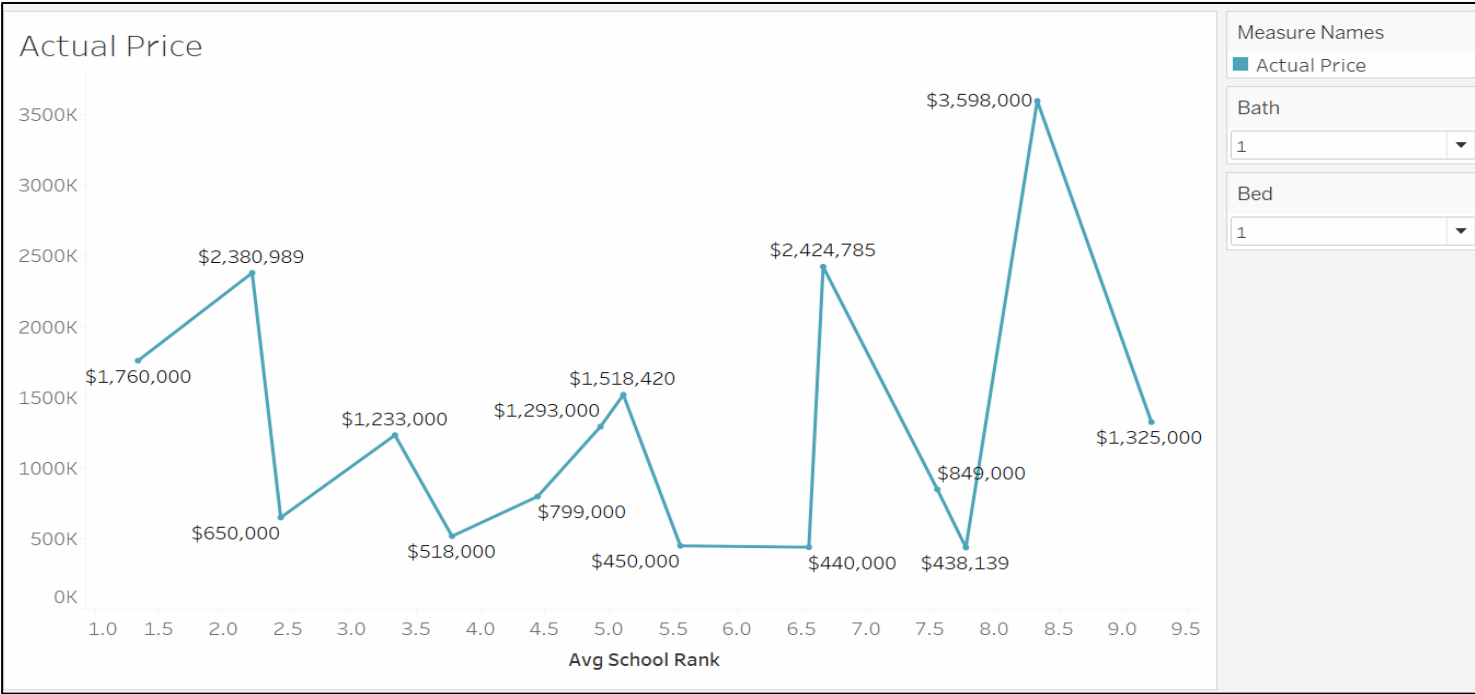
# Modeling:

## Model Features and Evaluation metrics

Features

Bed	Bath	Additional Attributes	Home type	Price per square feet	Avg School Rank	School Enrollment	Home price
2	1	False	Townhouse	\$ 687	8.9	600	980 k
3	1	True	Single Family	\$ 300	6	300	1.1 M
1	1	True	Condo	\$ 200	2.3	250	200 k
...	...	...	...	...	...	...	...

House price by School Ranking



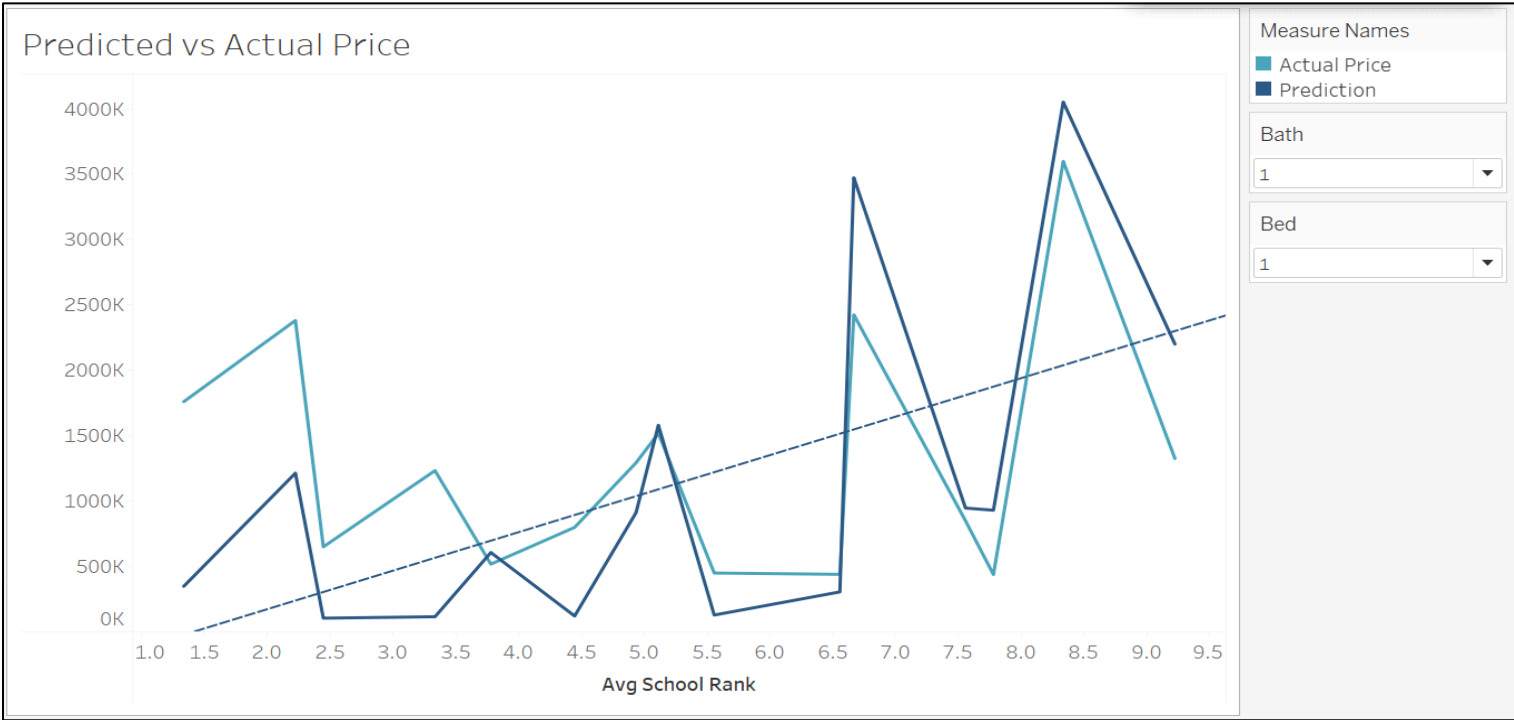


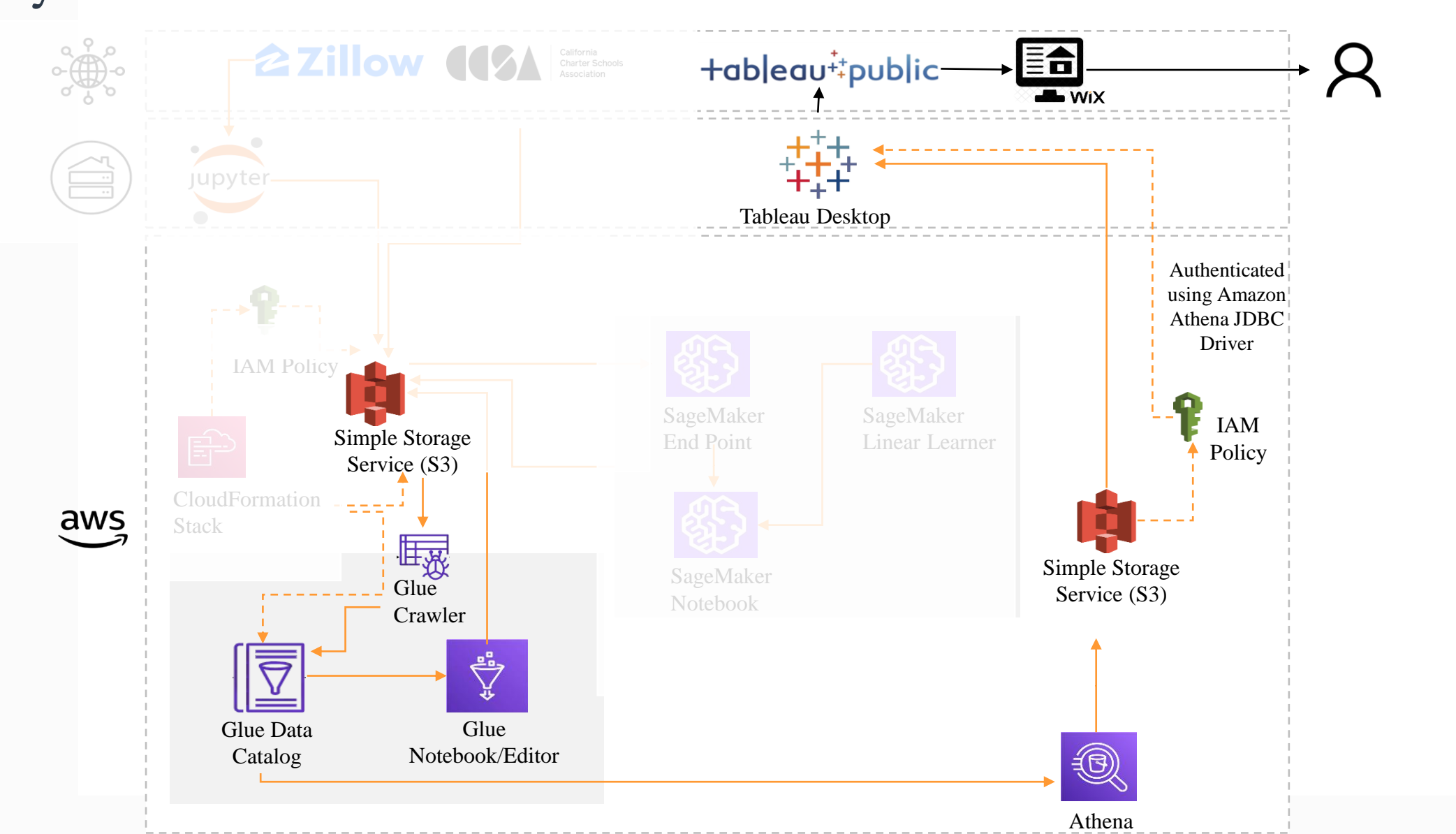
# Modeling: Predictions

Features	Bed	Bath	Additional Attributes	Home type	Price per square feet	Avg School Rank	School Enrollment	Home price	Predictions
	2	1	False	Townhouse	\$ 687	8.9	600	980 k	900 k
	3	1	True	Single Family	\$ 300	6	300	1.1 M	1 M
	1	1	True	Condo	\$ 200	2.3	250	200 k	250 k
	...	...	...	...	...	...	...	...	

## Evaluation Metrics and visual inspection

- $R^2 = 0.6$
- Adjusted  $R^2 = 0.59$
- Explained Variance Score – 0.6





# Analytics and Visualization: Amazon Athena to Tableau Dashboard

Amazon Athena > Query editor

EditorRecent queriesSaved queriesSettings

Workgroupprimary

Query result and encryption settings

Manage

Query result location and encryption

Query result location

s3://awsdataworksbucket/output/

Encrypt query results

-

Expected bucket owner

-

Assign bucket owner full control over query results

Turned off

Identity and Access Management (IAM)

Dashboard

Access management

User groups

Users

Roles

Policies

Identity providers

Account settings

Access reports

Access analyzer

Archive rules

Analysts

Settings

Credential report

Organization activity

Service control policies

Your Security Credentials

Use this page to manage the credentials for your AWS account. To manage credentials for AWS Identity and Access Management (IAM) users, use the IAM Console .

To learn more about the types of AWS credentials and how they're used, see AWS Security Credentials in AWS General Reference.

Password

Multi-factor authentication (MFA)

Access keys (access key ID and secret access key)

Use access keys to make programmatic calls to AWS from the AWS CLI, Tools for PowerShell, AWS SDKs, or direct AWS API calls. You can have a maximum of two access keys (active or inactive) at a time.

For your protection, you should never share your secret keys with anyone. As a best practice, we recommend frequent key rotation.

If you lose or forget your secret key, you cannot retrieve it. Instead, create a new access key and make the old key inactive. Learn more

Created	Access Key ID	Last Used	Last Used Region	Last Used Service	Status	Actions
Nov 18th 2022	AKIA3F76EOJRPLP2BG77	2022-11-20 17:06 PST	us-west-1	glue	Active	Make Inactive   Delete

Create New Access Key

Root user access keys provide unrestricted access to your entire AWS account. If you need long-term access keys, we recommend creating a new IAM user with limited permissions and generating access keys for that user instead. Learn more



Amazon Athena

GeneralInitial SQL

Server

Athena.us-west-1.amazonaws.com

Port

443

S3 Staging Directory

s3://awsdataworksbucket/output/

Access Key ID

AKIA3F76EOJRPLP2BG77

Secret Access Key

.....

Sign In

Tableau - Report\_11202022\_v6

FileDataServerWindowHelp

Connections

Add

Athena.us-wes...amazonaws.com

Amazon Athena

Catalog

AwsDataCatalog

Database

housing-market

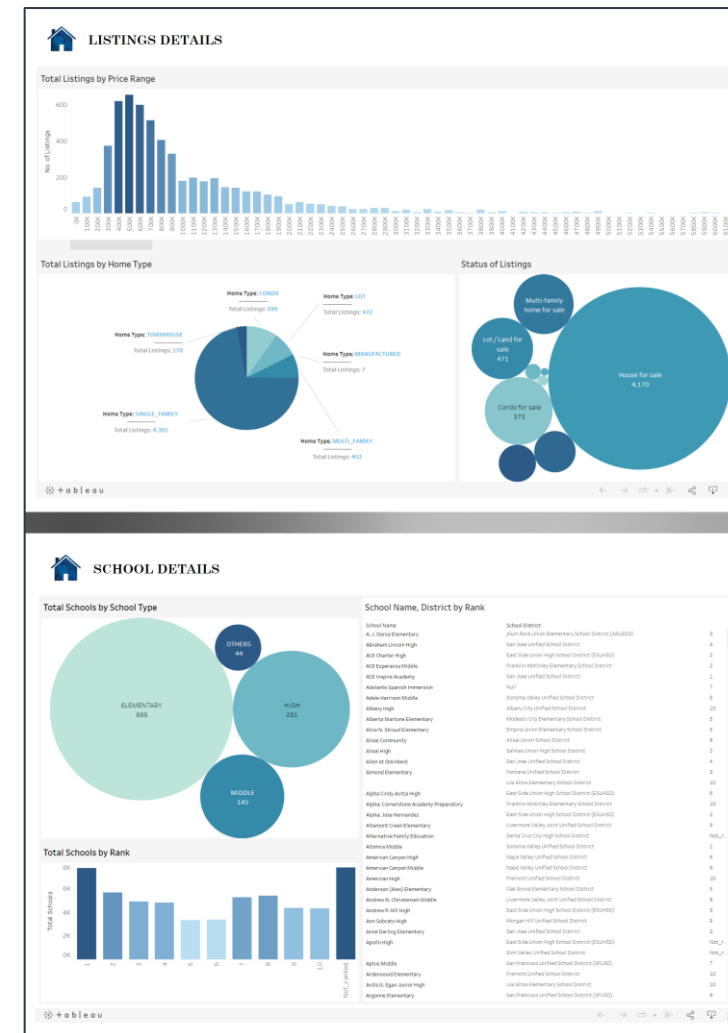
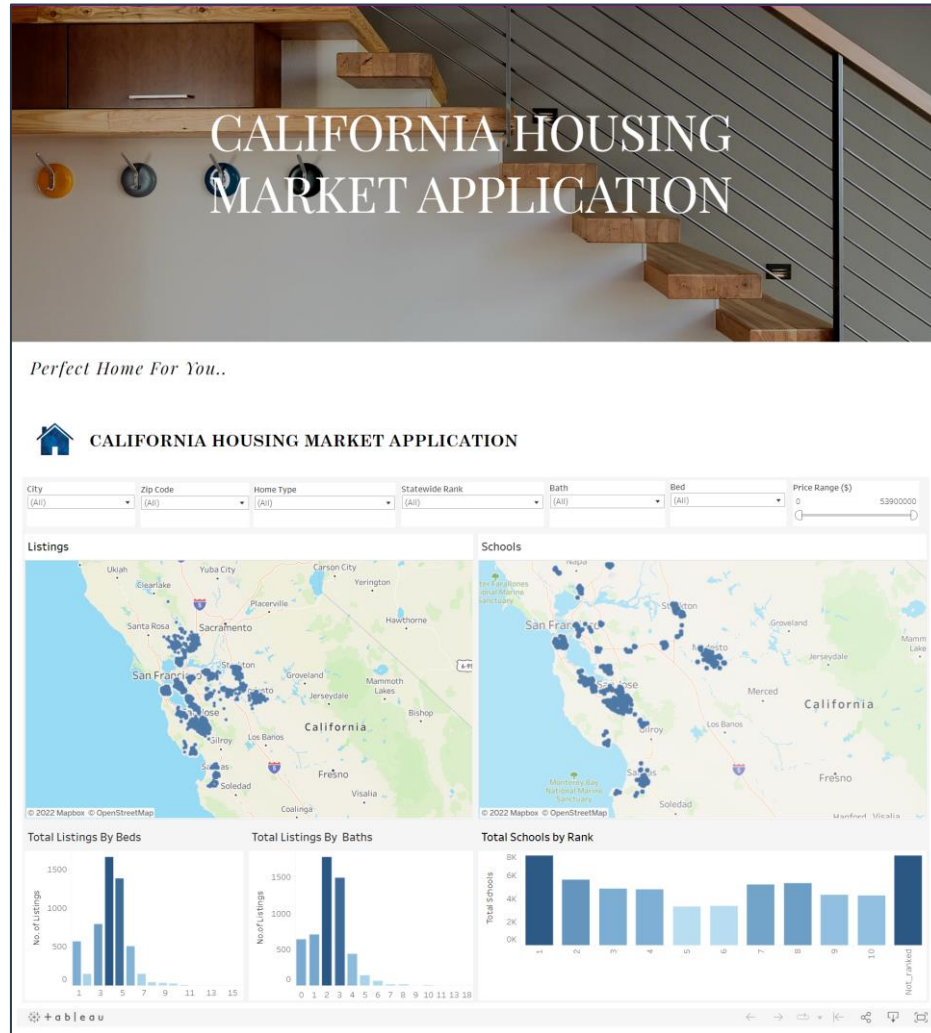
Table

housing\_etl\_out...11202022\_v4\_csv

school\_ranking\_csv

New Custom SQL

# California Housing Market Application Site







CALIFORNIA HOUSING MARKET APPLICATION

City

(All)

Zip Code

(All)

Home Type

(All)

Statewide Rank

(All)

Bath

(All)

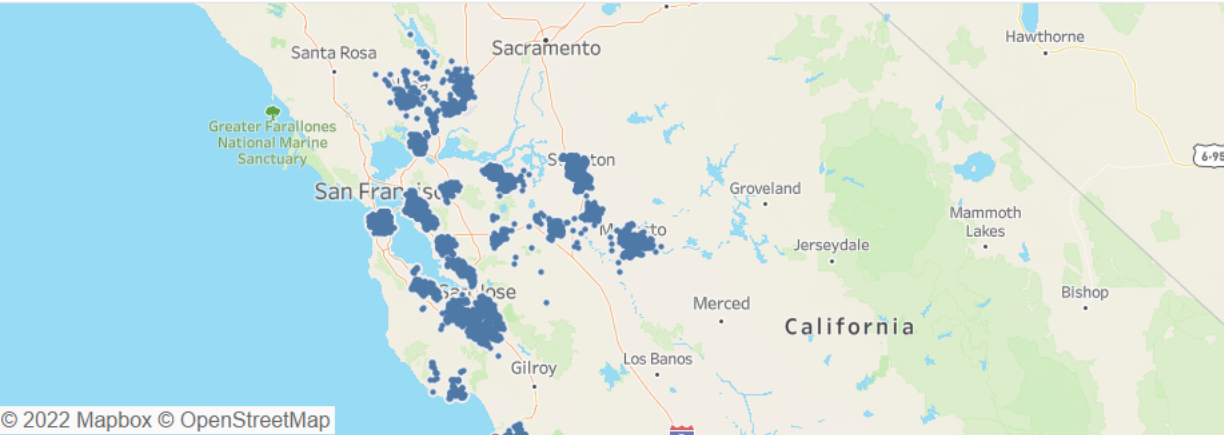
Bed

(All)

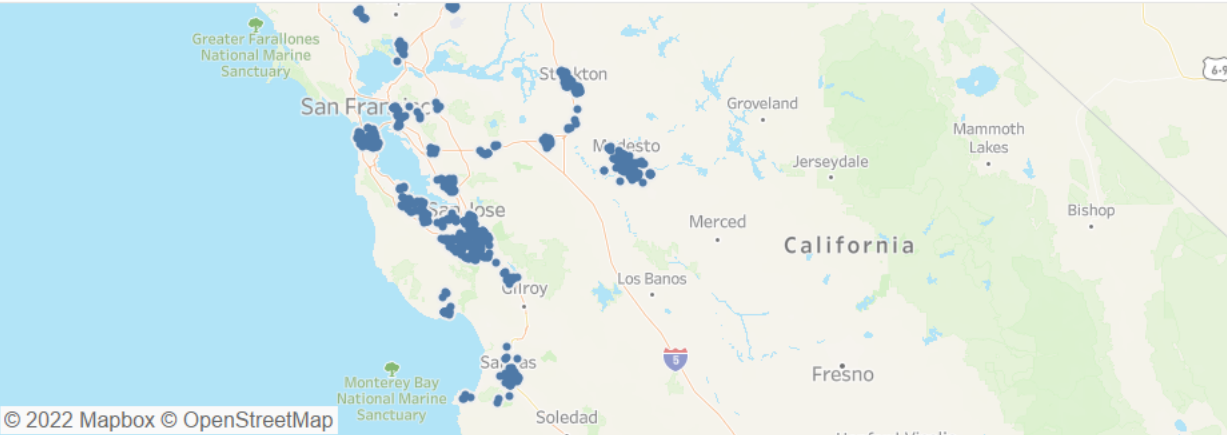
Price Range (\$)

053900000

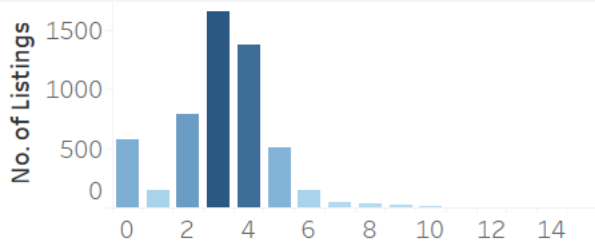
Listings



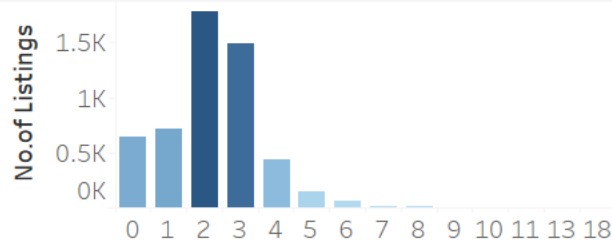
Schools



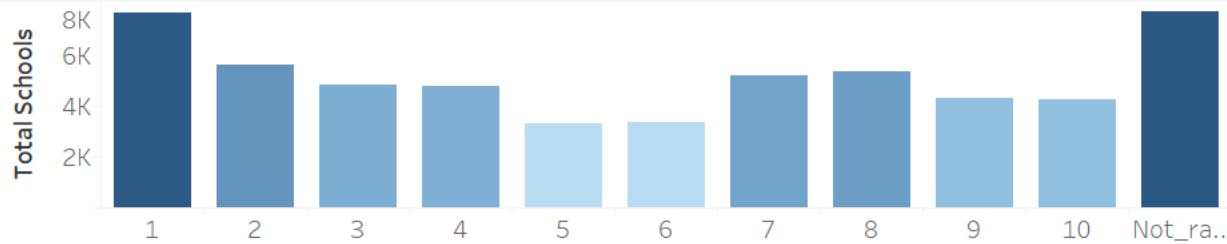
Total Listings By Beds



Total Listings By Baths



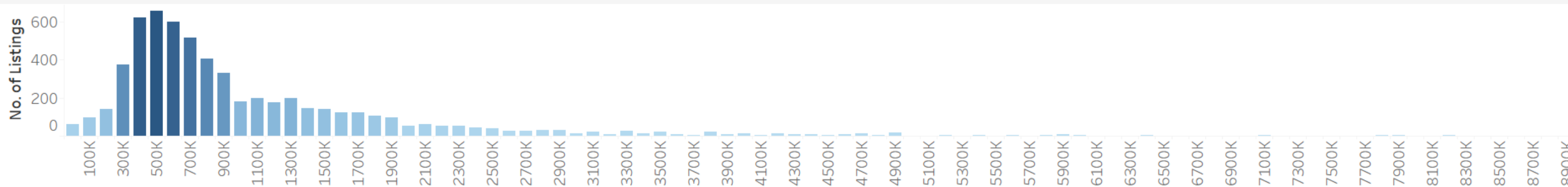
Total Schools by Rank



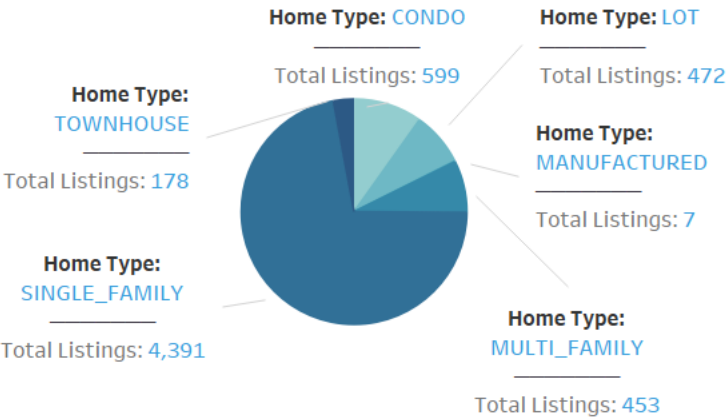


## LISTINGS DETAILS

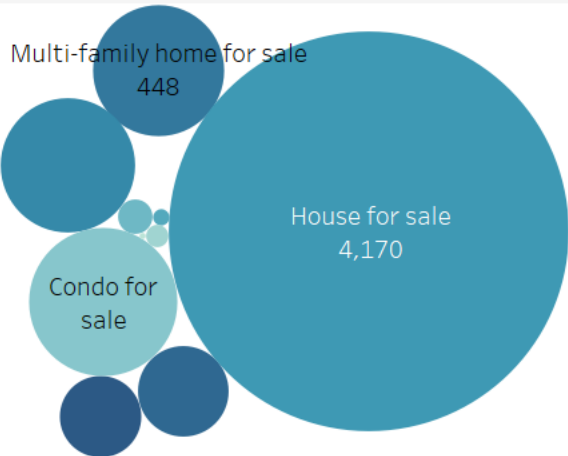
Total Listings by Price Range



Total Listings by Home Type



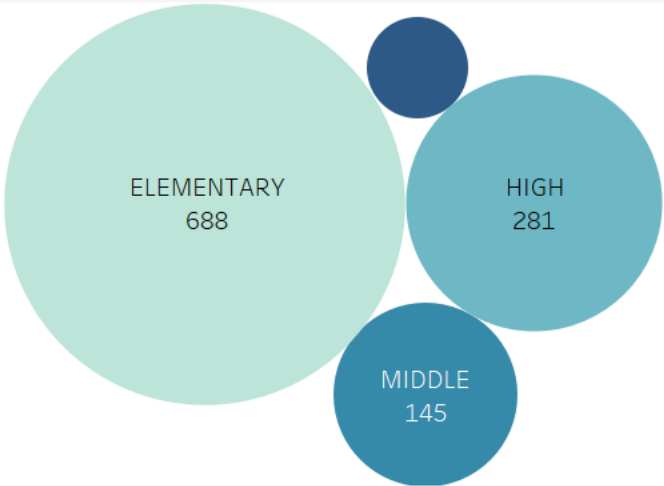
Status of Listings



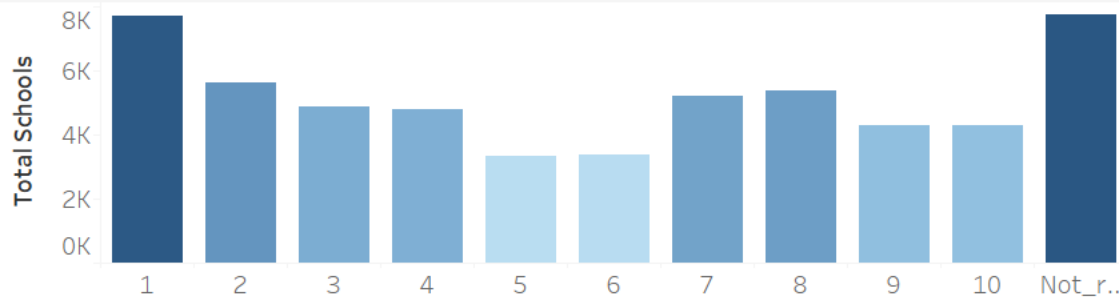


## SCHOOL DETAILS

Total Schools by School Type



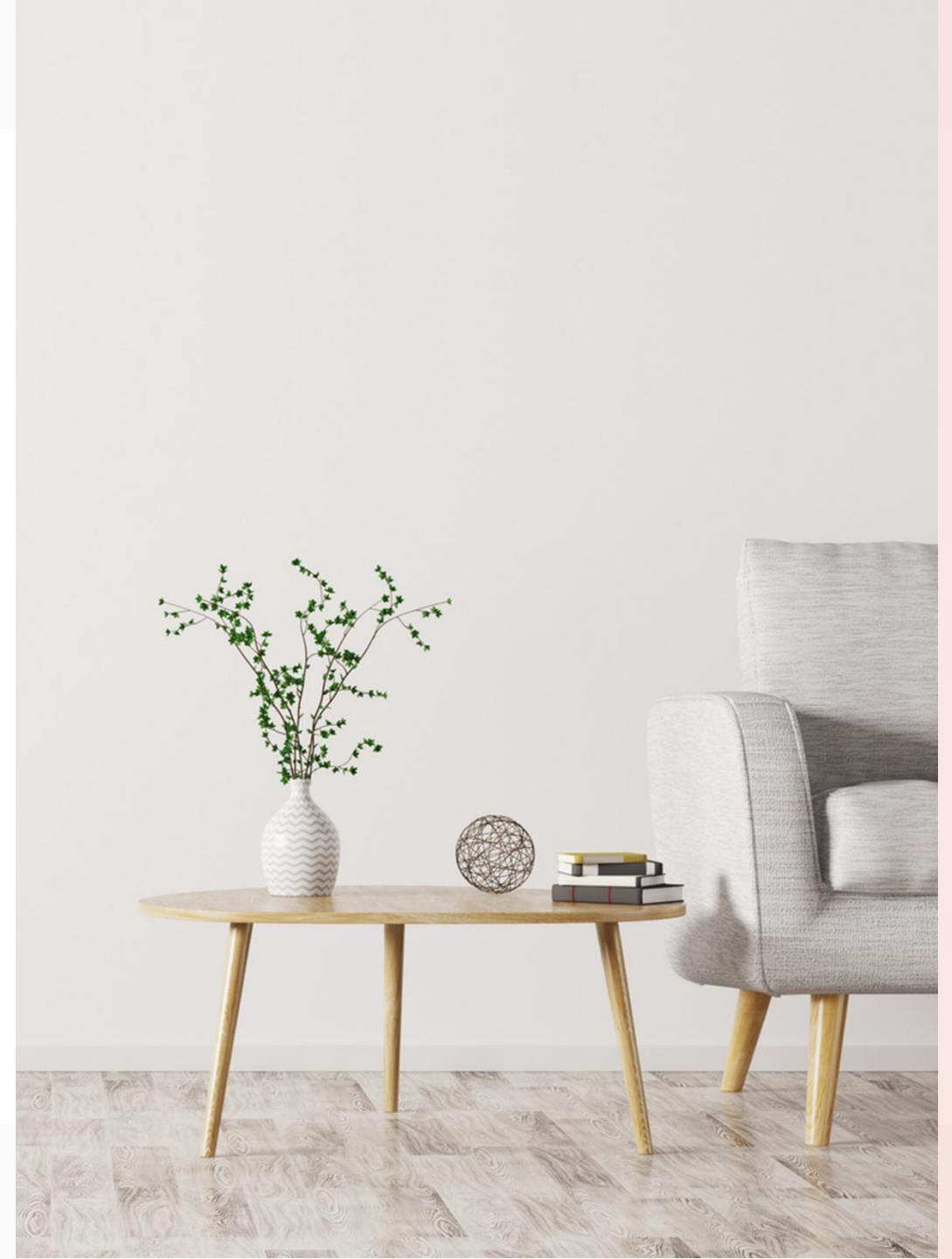
Total Schools by Rank



School Name, District by Rank

School Name	School District	
Aptos Middle	San Francisco Unified School District (SFUSD)	7
Ardenwood Elementary	Fremont Unified School District	10
Ardis G. Egan Junior High	Los Altos Elementary School District	10
Argonne Elementary	San Francisco Unified School District (SFUSD)	9
Argus High (Continuation)	Ceres Unified School District	Not_ranked
Ark Independent Studies	Santa Cruz City High School District	Not_ranked
Art Freiler	Tracy Joint Unified School District	6
Aspire East Palo Alto Charter	Ravenswood City Elementary School District	4
Aspire Langston Hughes Academy	Stockton Unified School District	3
Aspire Port City Academy	Stockton Unified School District	6
Aspire Summit Charter Academy	Ceres Unified School District	4
Aspire Vanguard College Preparatory ..	Empire Union Elementary School District	4
August Boeger Middle	Mt. Pleasant Elementary School District	3
August Elementary	Stockton Unified School District	4
Bagby Elementary	Cambrian Elementary School District	9
Balboa High	San Francisco Unified School District (SFUSD)	7
Baldwin (Julia) Elementary	Oak Grove Elementary School District	6
Barbara Spratling Middle	Keyes Union Elementary School District	5
Bardin Elementary	Alisal Union School District	2
Barrett Elementary	Morgan Hill Unified School District	3
Barron Park Elementary	Palo Alto Unified School District	9
Bay View Academy	Monterey Peninsula Unified School District	7
Bear Creek High	Lodi Unified School District	5
Belle Haven Elementary	Ravenswood City Elementary School District	1
Ben Painter Elementary	Alum Rock Union Elementary School District (ARUESD)	6
Berkeley Arts Magnet at Whittier	Berkeley Unified School District	9
Berkeley Technology Academy	Berkeley Unified School District	Not_ranked
Bernal Intermediate	Oak Grove Elementary School District	7
Bernard L. Hughes Elementary	Empire Union Elementary School District	4
Berryessa Union Elementary	Berryessa Union Elementary School District	Not_ranked
Big Sur Charter	Monterey Peninsula Unified School District	9

## IV. Conclusion and Future Work





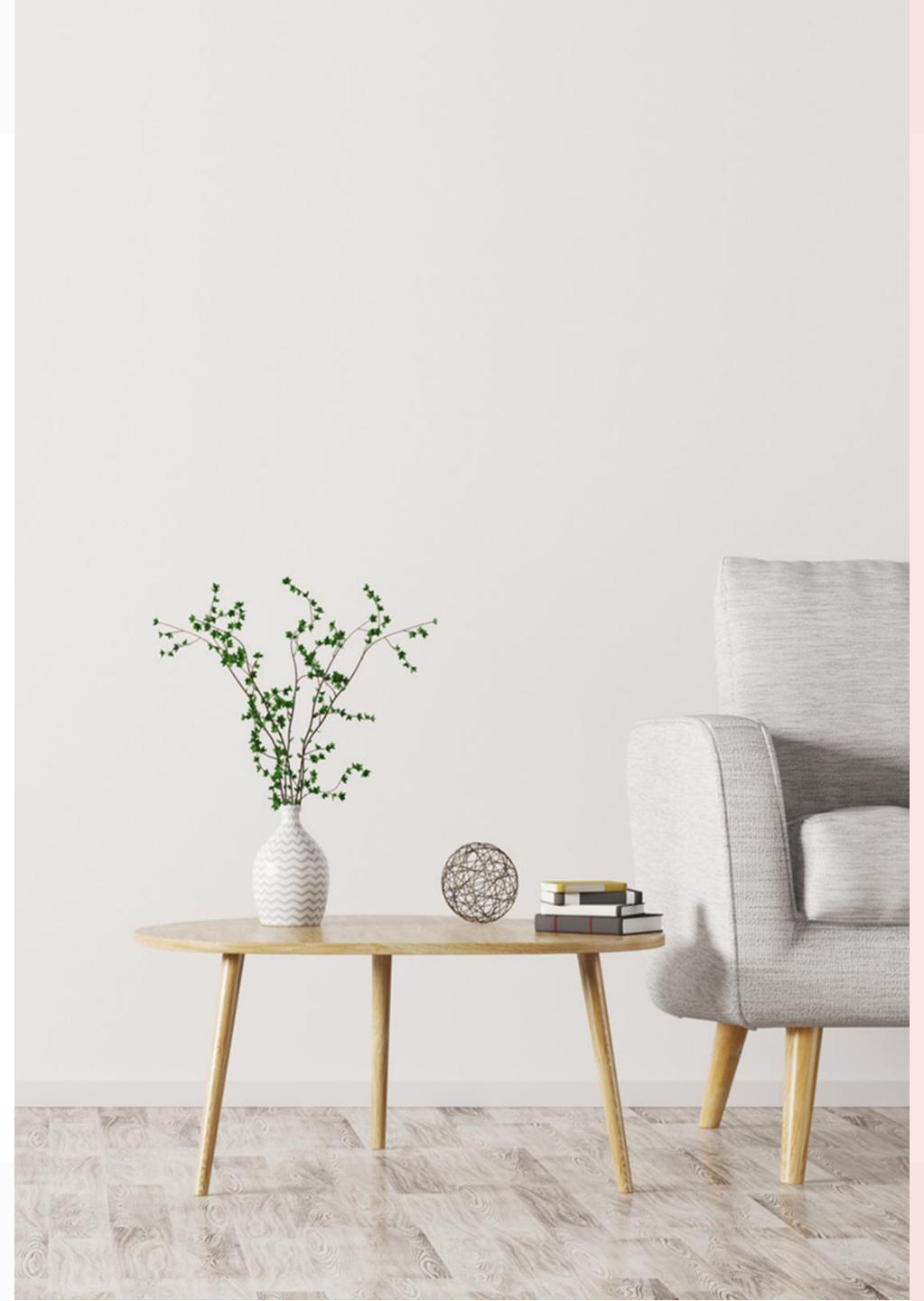
# Conclusion and Future Work

Overall,

- Built a scalable 'California Housing Market' big data application leveraging various services in AWS cloud platform.
- It serves as a search platform for the home buyers to view the properties, get the best price quote and insights about house price variation due to the presence of schools

As a future work,

- The application can be improvised to use AWS Kinesis to live capture the data and analytics
- Event triggering through AWS Lambda to connect various AWS components and data pipeline for continuous deployment and execution in a single click
- Employ Amazon simple notification Service (SNS) to notify the customers about the new listings via email/text message.



# References

- <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.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-what-is-how-does-it-work.html>



# Thank You

Wednesday, November 30, 2022