California Housing Price Prediction

Ilya Antonov
Reuben Chorney
Rokas Skerys

Motivation & Problem Statement

Domain: Real Estate and Housing Markets

Current State/Motivation: Housing affordability and homelessness is a critical issue in California. Give policy makers the tools to improve recommendations.

Gaps to explore: Prediction of housing prices based on various factors

Contribution: This project builds a predictive model to analyze housing prices and affordability across California, helping both buyers and policymakers make informed decisions.

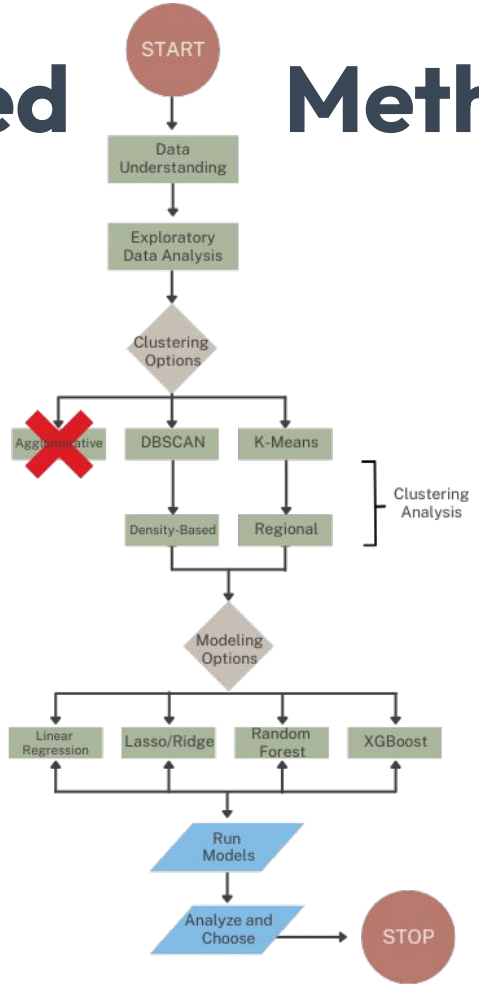
Project Introduction

I. Goals/Objective	Develop predictive model to estimate housing prices in California.
a.	Provide actionable insights for house affordability.
II. Success Indicators	Depends based on stakeholders.
a. Technological	
b. Business	Deliver visualizations that highlight key affordability trends (heatmaps, scatter plots). Real-world usability for end users through accessible and interpretable results.

Target Group	All stakeholders in the California real estate market, including buyers, sellers, real estate agencies, and policymakers. Larger group.

End user	The individuals or organizations who will directly use the predictions and affordability insights—such as homebuyers, policymakers, economists, researchers. Smaller group.

Proposed Methodology



Proposed Methodology

Data Understanding

Analyze the California housing dataset from the 1990 Census and Identify Key Variables

EDA

Visualize distributions and relationships between variables, identify and handle outliers, make key observations

Clustering

Standardize numerical variables for clustering and predictive modeling. Feature Selection. Cluster based on DBSCAN and K-Means

Identify Models

Random Forest Regressor, XGBoost, Lasso, Ridge Regression.

Run and Analyze Models

Use R^2, Mean Absolute Error (MAE), and Mean Squared Error (MSE) on validation and test sets.

Data to be used & EDA

```
data.describe()
```

	Median_House_Value	Median_Income	Median_Age	Tot_Rooms	Tot_Bedrooms	Population	Households	Latitude	Longitude	Distance_to_coast	Distance_to_LA	Distance_to_SanDiego	Distance_to_SanJose	Distance_to_SanFrancisco
count	20640.000000	20640.000000	20640.000000	20640.000000	20640.000000	20640.000000	20640.000000	20640.000000	20640.000000	20640.000000	2.064000e+04	2.064000e+04	20640.000000	20640.000000
mean	206855.816909	3.870671	28.639486	2635.763081	537.898014	1425.476744	499.539680	35.631861	-119.569704	40509.264883	2.694220e+05	3.981649e+05	349187.551219	386688.422291
std	115395.615874	1.899822	12.585558	2181.615252	421.247906	1132.462122	382.329753	2.135952	2.003532	49140.039160	2.477324e+05	2.894006e+05	217149.875026	250122.192316
min	14999.000000	0.499900	1.000000	2.000000	1.000000	3.000000	1.000000	32.540000	-124.350000	120.676447	4.205891e+02	4.849180e+02	569.448118	456.141313
25%	119600.000000	2.563400	18.000000	1447.750000	295.000000	787.000000	280.000000	33.930000	-121.800000	9079.756762	3.211125e+04	1.594264e+05	113119.928682	117395.477505
50%	179700.000000	3.534800	29.000000	2127.000000	435.000000	1166.000000	409.000000	34.260000	-118.490000	20522.019101	1.736675e+05	2.147398e+05	459758.877000	526546.661701
75%	264725.000000	4.743250	37.000000	3148.000000	647.000000	1725.000000	605.000000	37.710000	-118.010000	49830.414479	5.271562e+05	7.057954e+05	516946.490963	584552.007907
max	500001.000000	15.000100	52.000000	39320.000000	6445.000000	35682.000000	6082.000000	41.950000	-114.310000	333804.686371	1.018260e+06	1.196919e+06	836762.678210	903627.663298

```
data.isna().sum()
#We have no missing values in this dataframe
```

	0
Median_House_Value	0
Median_Income	0
Median_Age	0
Tot_Rooms	0
Tot_Bedrooms	0
Population	0
Households	0
Latitude	0
Longitude	0
Distance_to_coast	0
Distance_to_LA	0
Distance_to_SanDiego	0
Distance_to_SanJose	0
Distance_to_SanFrancisco	0

dtype: int64

```
data.shape[0]
#What is the size of dataset before dropping outliers
```

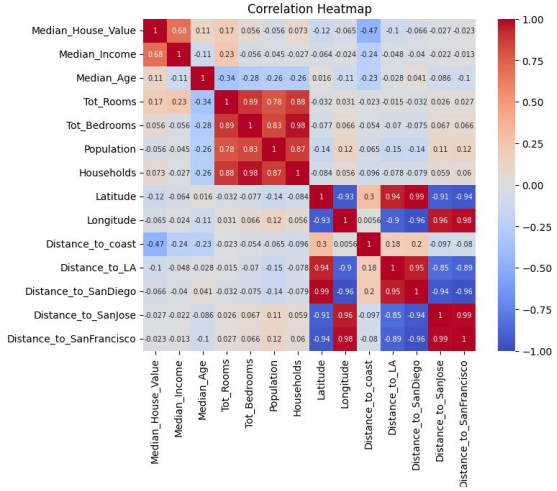
```
20640
```

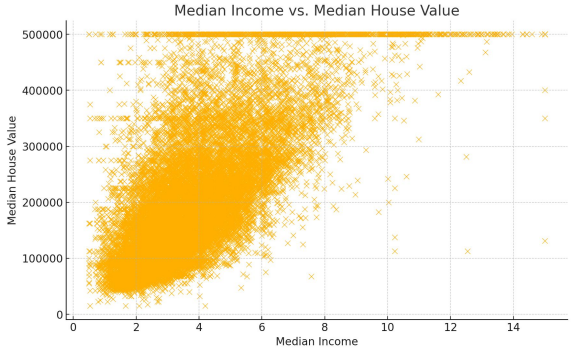
```
data_cleaned.shape[0]
#How many are left after outlier cleaning
```

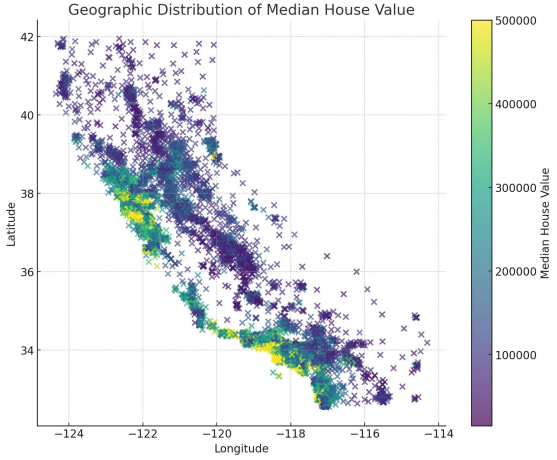
```
19608
```

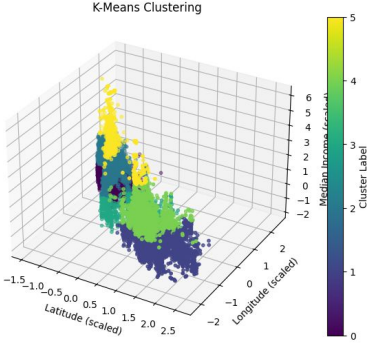
***Cleaned Data From Kaggle from 1990 US census

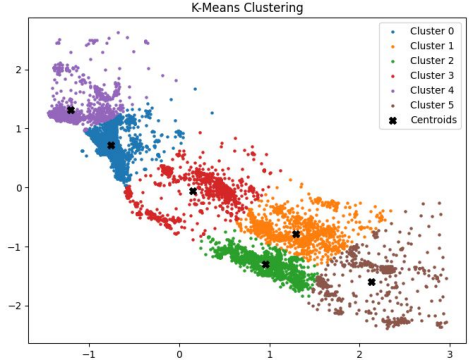
Data to be used & EDA











***Cleaned Data From Kaggle from 1990 US census

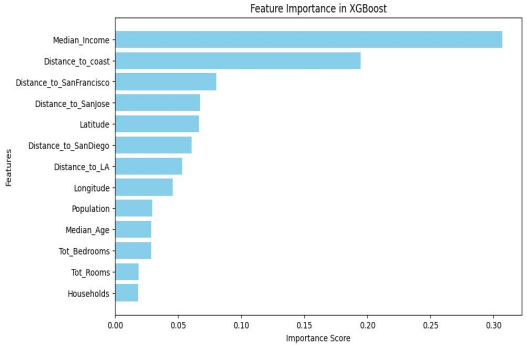
Model Comparison

1	XGBoost
2	Decision Trees
3	Linear Regression

Initial Thoughts

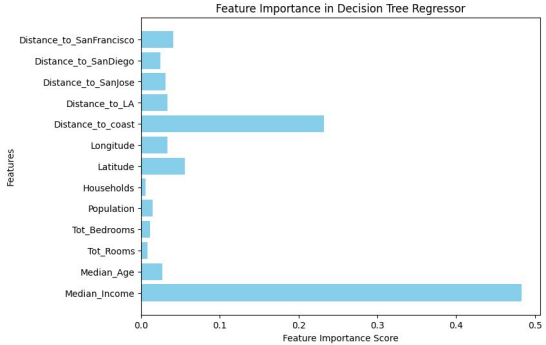
- Would be best model especially when considering multicollinearity possibility
- Would be second best model since we perceived Linear Regression as bad in comparison
- Would be worst model. Need to do analysis on which regression model we should choose.

After Traditional Feature Selection (Households Removed)



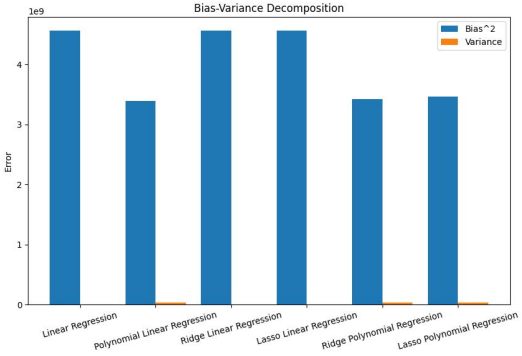
XGBoost

To boost Linear Model due to collinearity, households were removed



Decision Trees

Important features are distance to coast and Median Income. Other independent variables are not that important.



Linear

All underfitting. Choice of linear model from this.

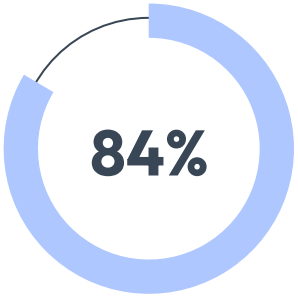
Final Model Results

Model	R^2 Score	MSE	MAE
Linear Regression (Polynomial)	0.7323	3.406×10^9	
Decision Trees	0.7285	3.453×10^9	
XGBoost	**0.8442**	**1.982×10^9**	**29,009.33**

Analysis

Our final selection is biased as at the beginning it was told that XGBoost is the best performing model. After fine-tuning all 3 models on the testing set we still get that XGBoost has the lowest MSE. Possible way to improve models: to go deeper into models, choose more parameters and fine a way to lower MSE. However, after more than 100 runs of different models on different subsets of data with different parameters it seems that improvements will not lead to a much better performance. So the only way is to wait for a completely new models.

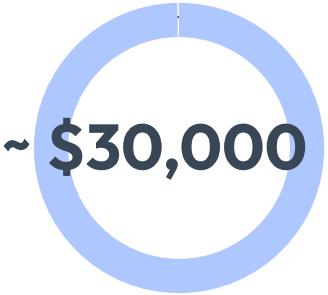
Conclusion & Future Work
XGBoost Final Model



R^2

84% R^2 has a very strong explanatory value for the variance.

Assumptions	Data is relevant. 1990 census data is relevant for today's policy makers and results can be generalizable.
Future Work	Update models on new data from California and provide more up to date recommendations for policy makers.



MAE

Relatively small average error when considering a mean of $206,000 and SD of $115,000