California Housing Price Predictions

Aarya Khanna, Audrey Huang, Ella Cruz, Heidi Yu, Richard Xu, Siddharth Singh, Jordyn Fuchs

Contents

01
Data Collection

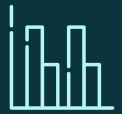
Data Collection

- Obtaining Data
  - www.kaggle.com/datasets/camnugent/california-housing-prices/data
- Dataset
  - 20640 observations x 10 variables
- Variables
  - Longitude, latitude, housing median age, total rooms, total bedrooms, population, households, median income, median house value, ocean proximity
- Goals
  - Create visualizations displaying relevant features correlated to the housing prices in California

sklearn dataset: fetch_california_housing

8 features

Including average number of household members, average number of rooms, median income, and location



1 target

Target variable is median home value for each California district in hundreds of thousands of dollars

20,640 samples

Containing all numerical features for each data sample

https://scikit-learn.org/stable/modules/generated/sklearn.datasets.fetch_california_housing.html

02

Data Cleaning

Data Cleaning with R

- Cleaned data to exclude housing values where total bedrooms had NA values
- 20,640 to 20,433 → eliminated 207 values

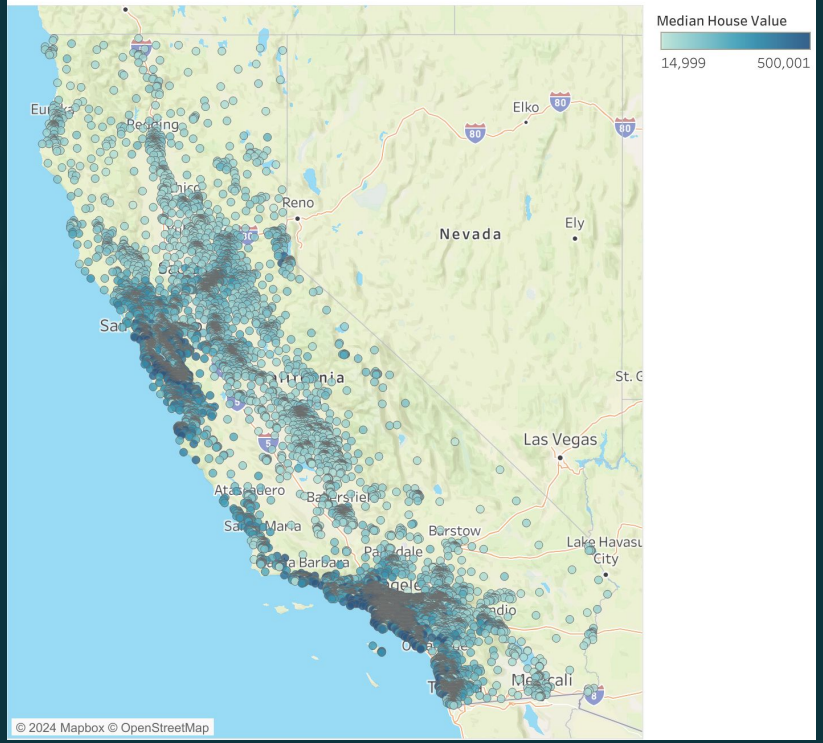
```r
house_without_ocean <- housing[-c(10)]
house_without_ocean[which(is.na(total_bedrooms)), ] <- -1

filtered_data <- subset(house_without_ocean, total_bedrooms > 0)
filtered_data
```

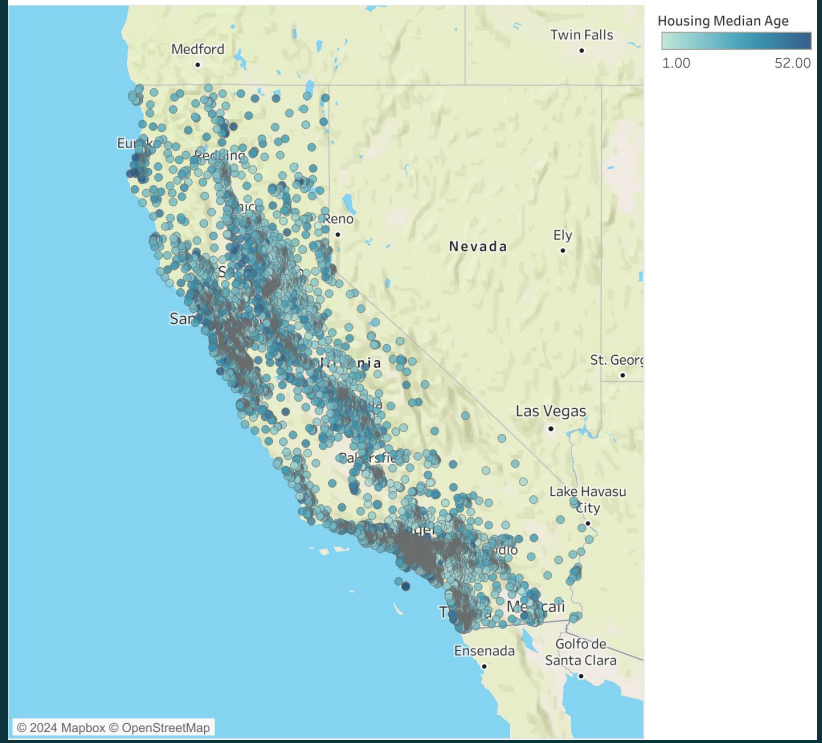
03
Visualization

Visualizations with Tableau

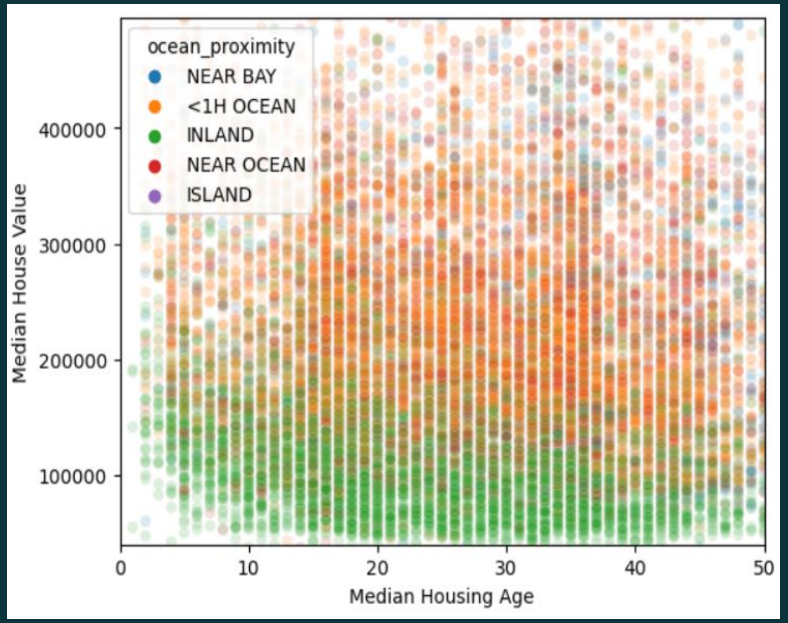
Median House Value Map

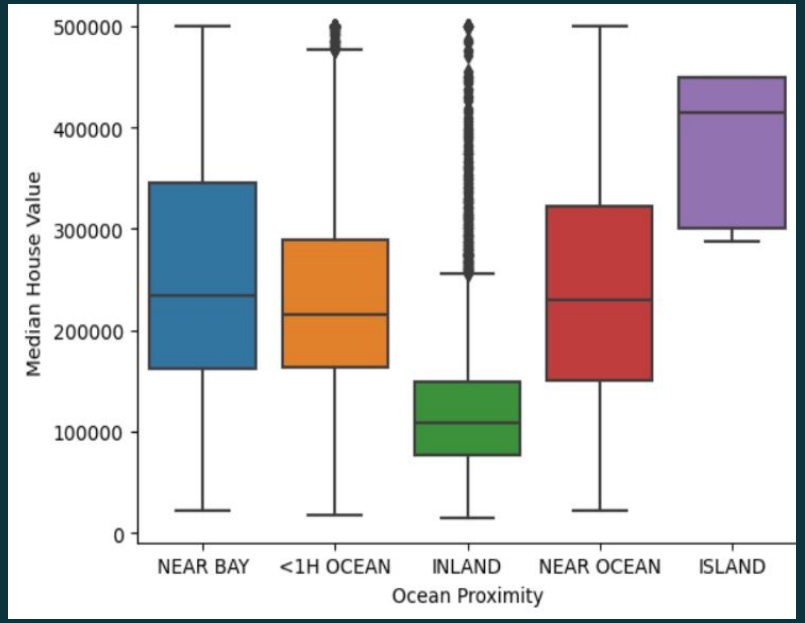


Median Age Map

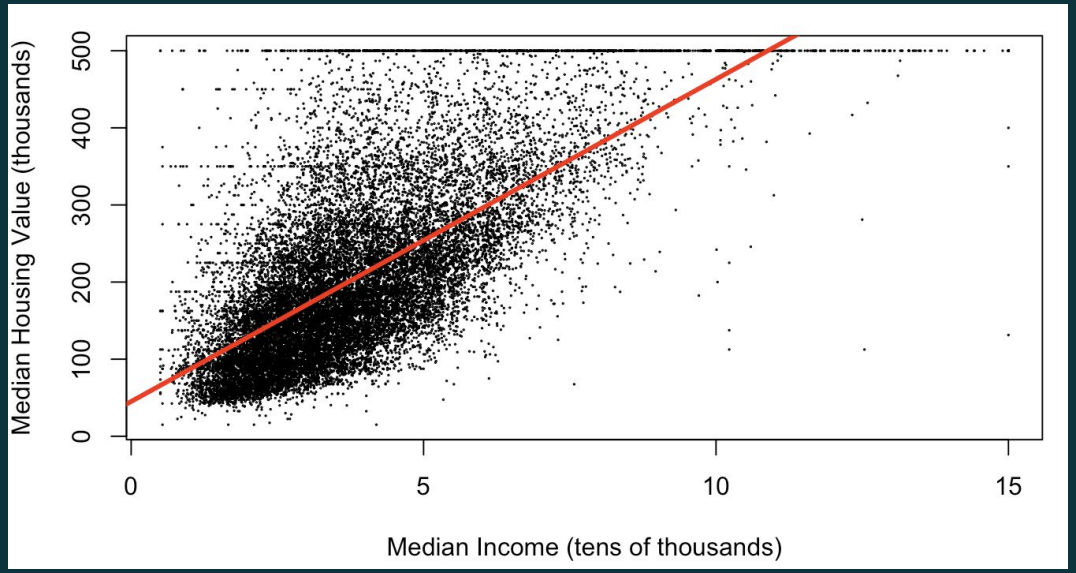


Visualizations

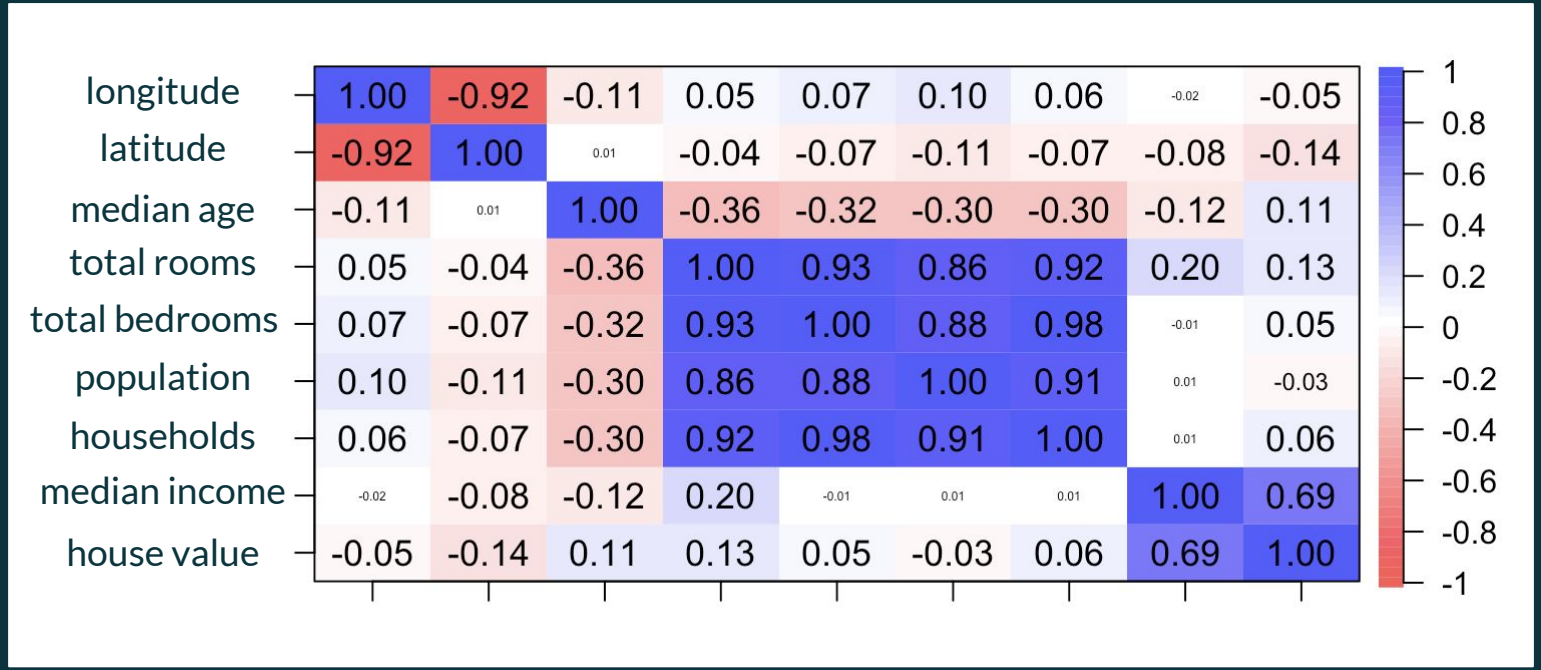




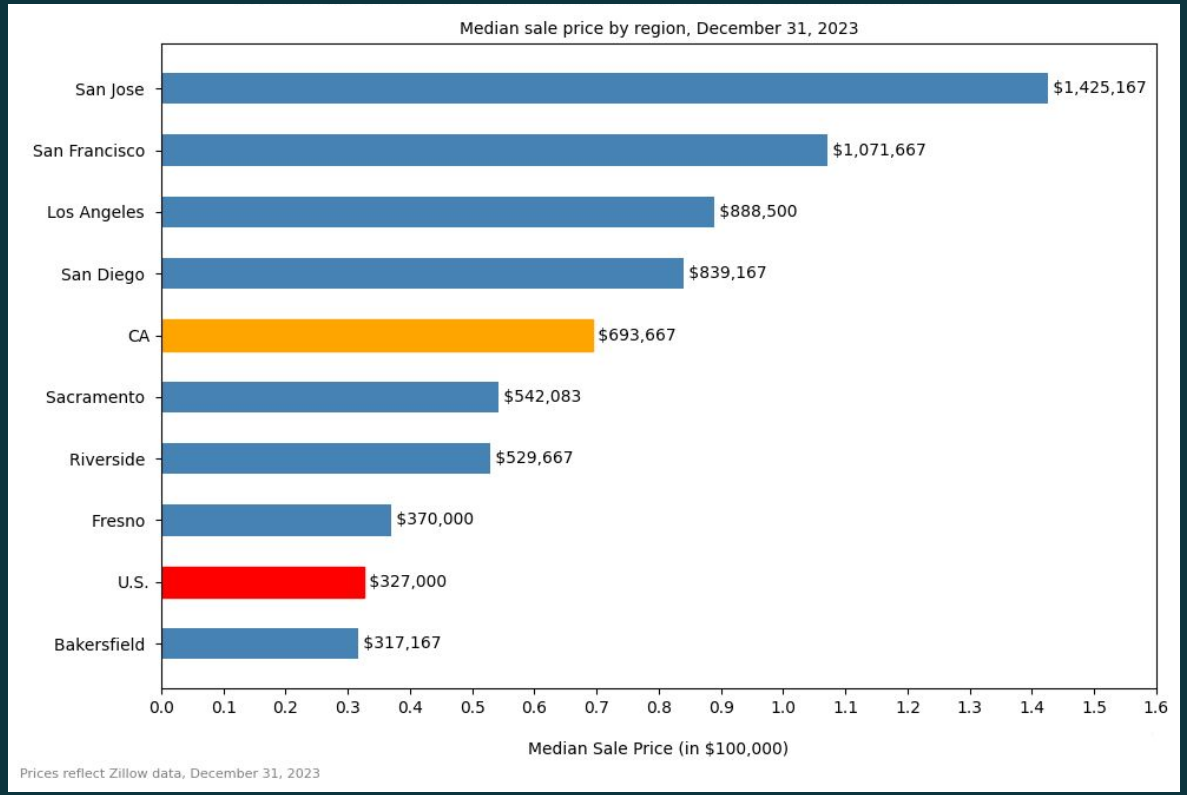
Linear Regression



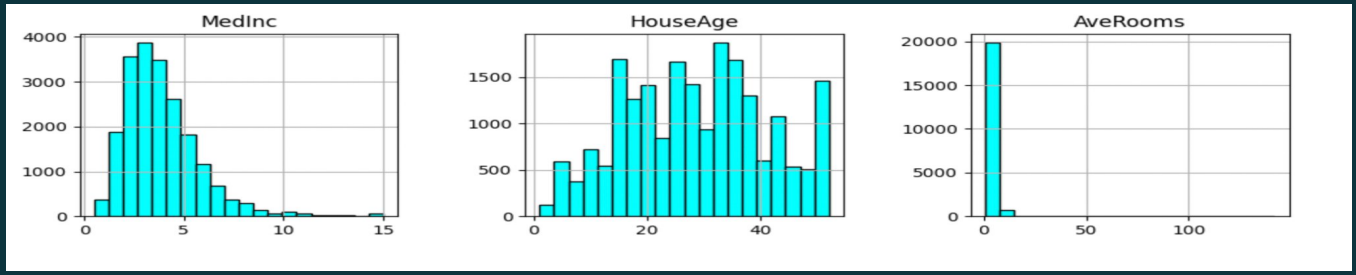
Correlation Matrix

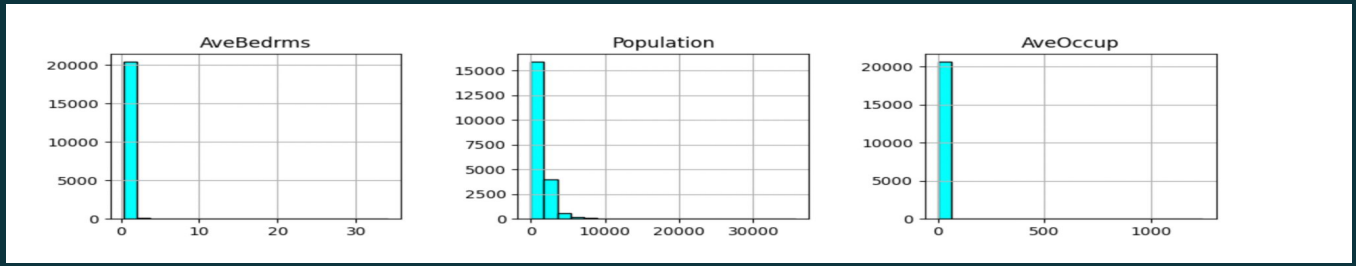


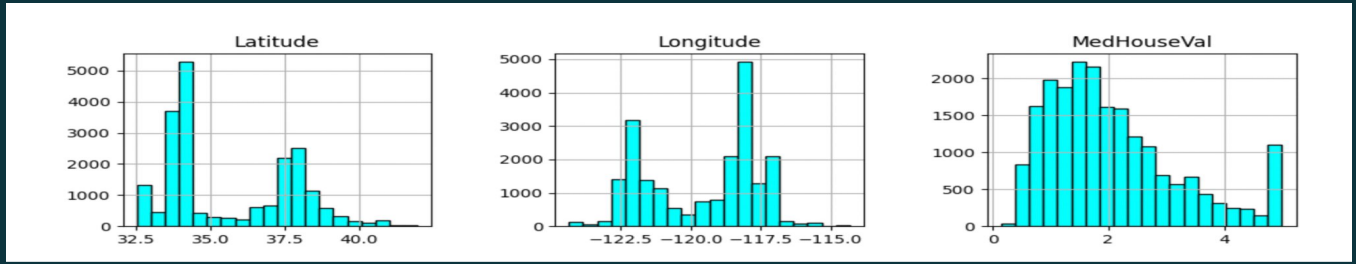
Horizontal Bar Chart



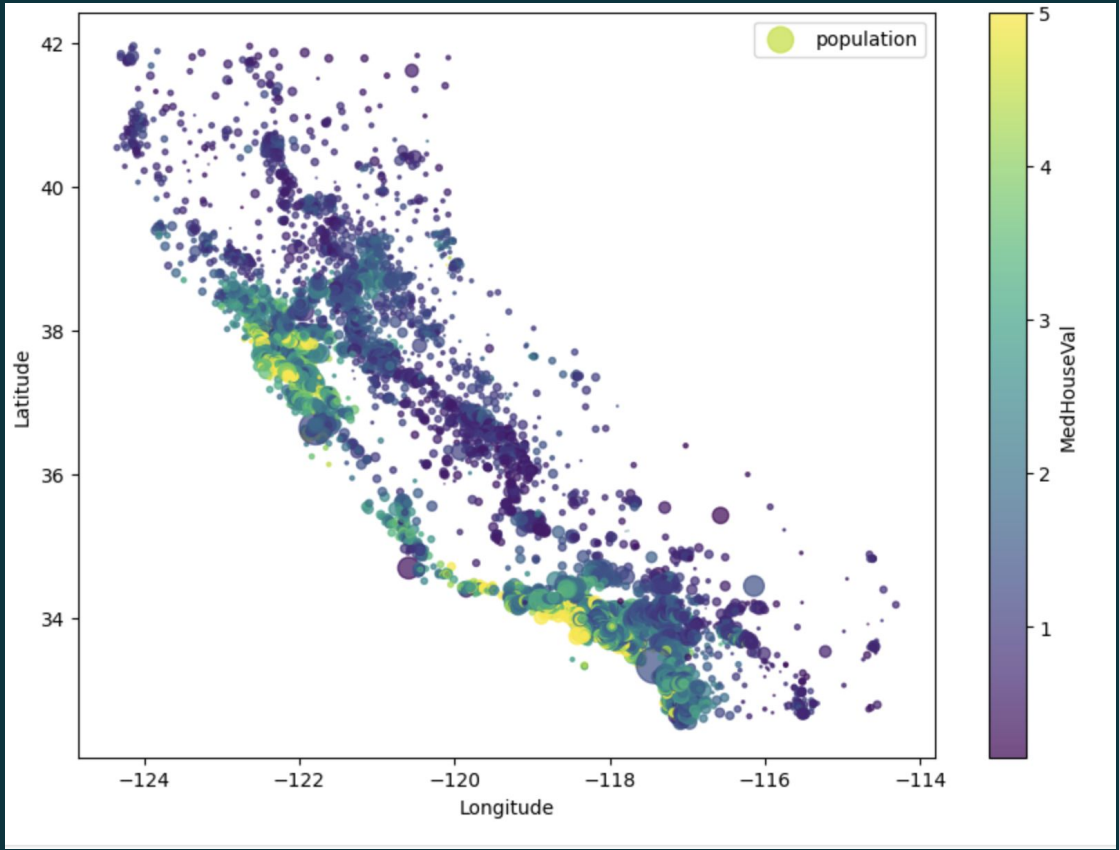
Feature Distributions



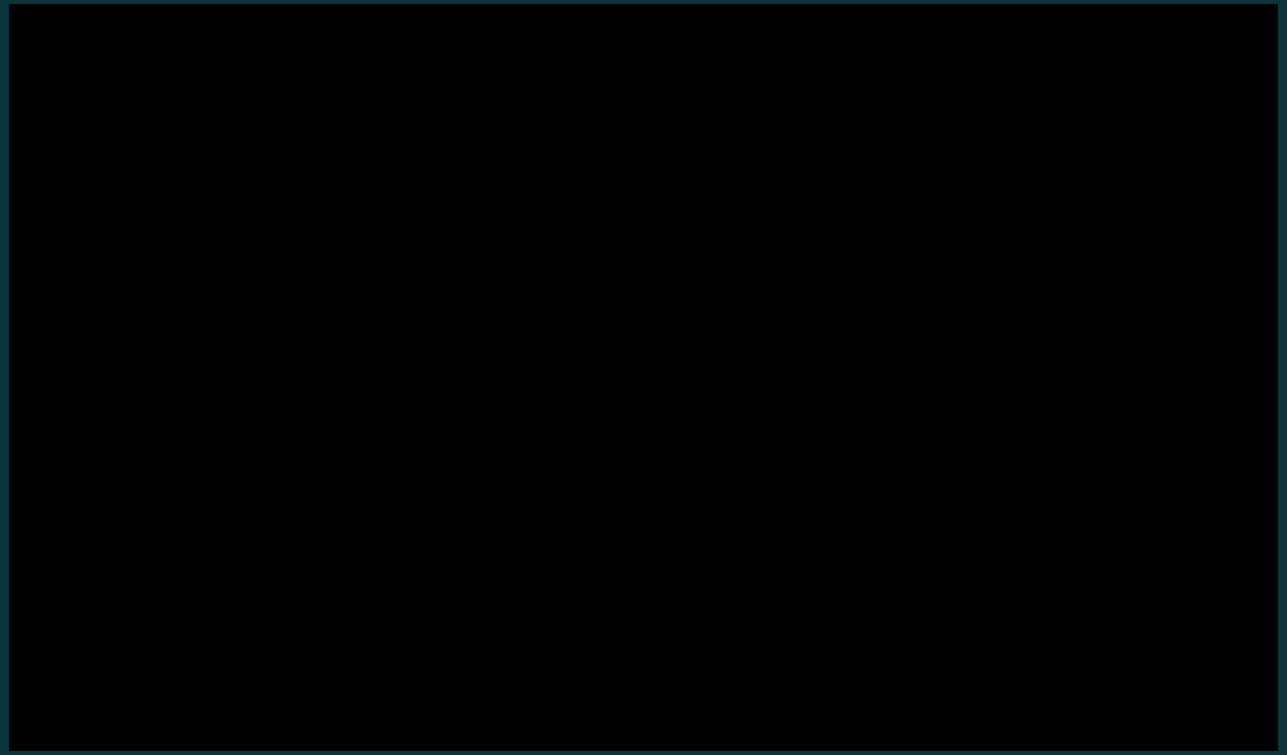




Value, Population, Location



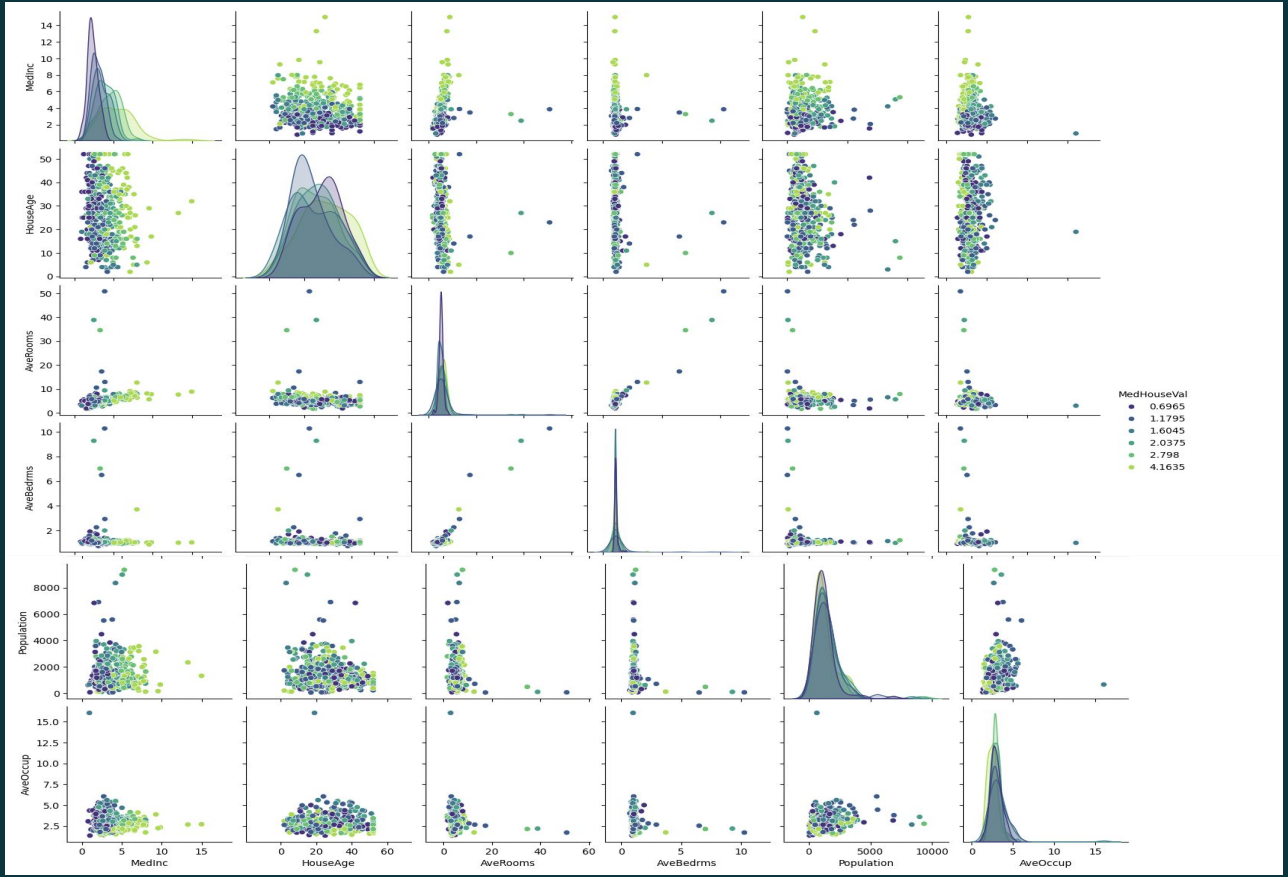
Power BI Dashboard



04

Machine Learning

Feature Correlation



Multivariable Regression

```
from sklearn import linear_model

X = df[['MedInc', 'HouseAge', 'Population']]
y = df['target']

reg = linear_model.LinearRegression()
reg.fit(X, y)

#predict the cost of a home for a buyer with a higher-end income looking for a newer home in a big city
predicted_price = reg.predict([[9.5, 15, 10000]])
print(predicted_price)
print(reg.coef_)
print(reg.intercept_)
```

```
[4.52345725]
[4.32273443e-01 1.82943214e-02 3.16069821e-05]
-0.1736250996619492 8
```

05
Conclusions

Findings

Location

Coastal locations and highly populated cities tend to have higher housing prices

Home Features

The age of the home and number of rooms also has significant impact on the price

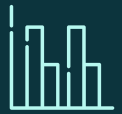
06

Future Ideas

Next Steps

Tune Parameters

Continue analysis on current parameters and try other types of regression to improve predictions



Clustering

Try a clustering/k-nearest neighbor model to group houses with similar features together

New Data

Consider other influences, including the stock market, employment, politics, etc.