SF Bay Area Housing Dataset

By Averi Tanlimco and Sania Bandekar

Dataset: SF Bay Area Housing

- **Goal:** Predict the Price of Bay Area Houses

	Address	City	State	Zip	Price	Beds	Baths	Home size	Lot size	Latitude	Longitude	SF time	PA time	School score	Commute time
0	2412 Palmer Ave	Belmont	CA	94002	1459000	3	2.0	1360.0	5001.0	37.516781	-122.304623	63	33	77.9	33
1	1909 Hillman Ave	Belmont	CA	94002	1595000	4	2.0	2220.0	3999.0	37.521972	-122.294079	63	33	77.9	33
2	641 Waltermire St	Belmont	CA	94002	899999	2	1.0	840.0	4234.0	37.520233	-122.273144	63	33	77.9	33
3	2706 Sequoia Way	Belmont	CA	94002	1588000	3	2.0	1860.0	5210.0	37.520192	-122.309437	63	33	77.9	33
4	1568 Winding Way	Belmont	CA	94002	1999000	4	3.5	2900.0	16117.2	37.524280	-122.291241	63	33	77.9	33
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
7140	The Davis	Mountain House	CA	95391	603990	5	3.0	2327.0	NaN	37.756444	-121.547719	120	125	65.3	120
7141	The Berkeley	Mountain House	CA	95391	619990	5	4.0	2410.0	NaN	37.756444	-121.547719	120	125	65.3	120
7142	Geranium	Mountain House	CA	95391	666340	5	4.0	2486.0	NaN	37.764721	-121.537761	120	125	65.3	120
7143	The Pepperdine	Mountain House	CA	95391	659990	5	4.0	2856.0	NaN	37.756444	-121.547719	120	125	65.3	120
7144	The Stanford	Mountain House	CA	95391	644990	5	4.0	2679.0	NaN	37.756444	-121.547719	120	125	65.3	120

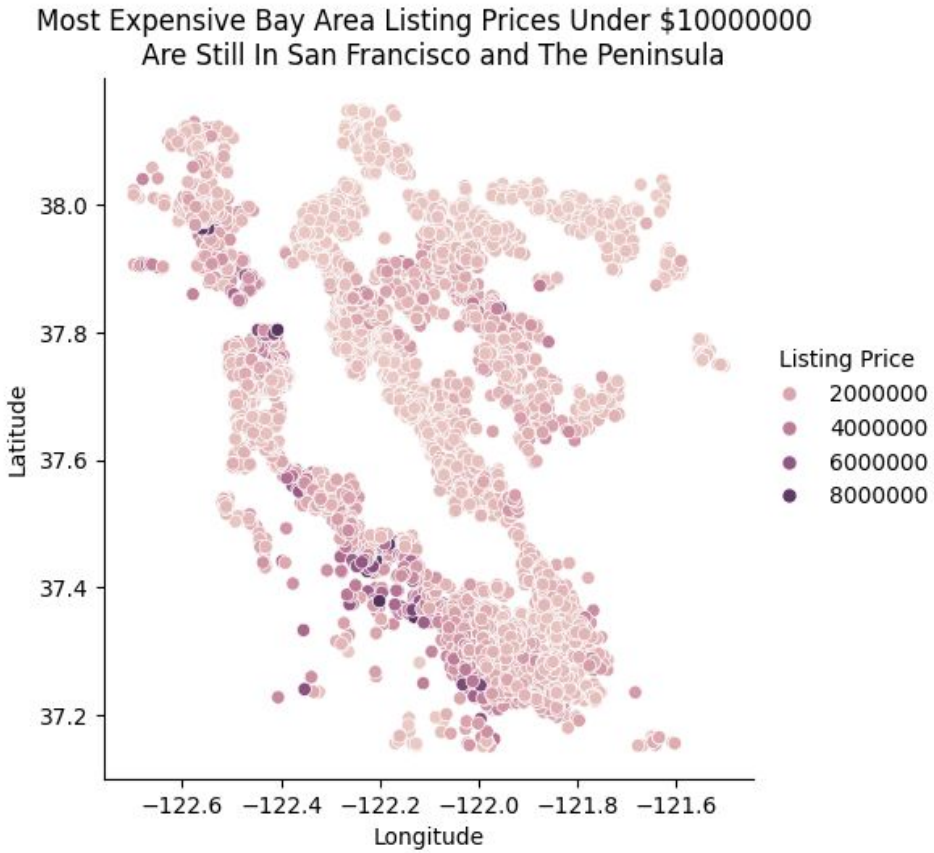
7145 rows × 15 columns

- `Address` - the adddress of the house
- `City` - the city the house is at
- `State` - California, this data set is from the Bay Area
- `Zip` - postal zip code
- `Price` - listing price of the house
- `Beds` - number of bedrooms
- `Baths` - number of bathrooms
- `Home size` - the square footage of the house

- `Lot size` - the square footage of the lot
- `Latitude` - latitude coordinate
- `Longitude` - longitude coodinate
- `SF time` - the commute time by car at 8 AM to San Francisco
- `PA time` - the commute time by car at 8 AM to commute to Palo Alto
- `School score` - the quality of the schools in the neighborhood
- `Commute time` - the commute time by car at 8 AM to the general Bay Area.

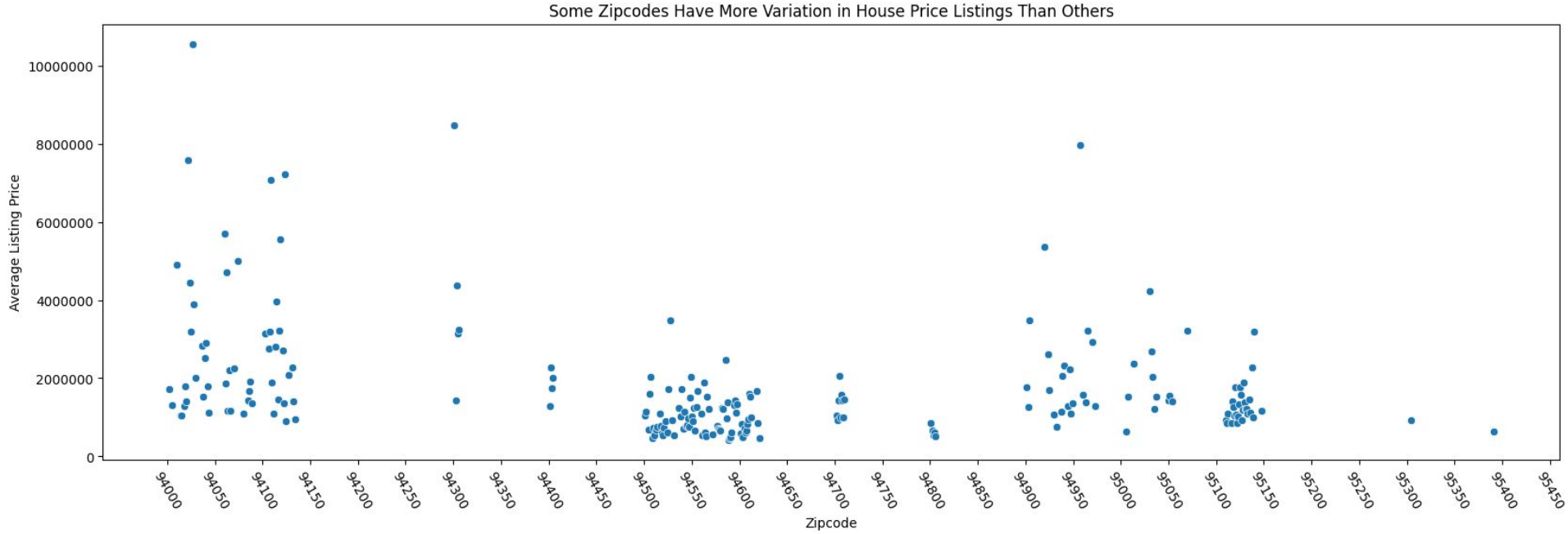
Questions / Visualizations

Q1: How is a house's coordinates related to its price?



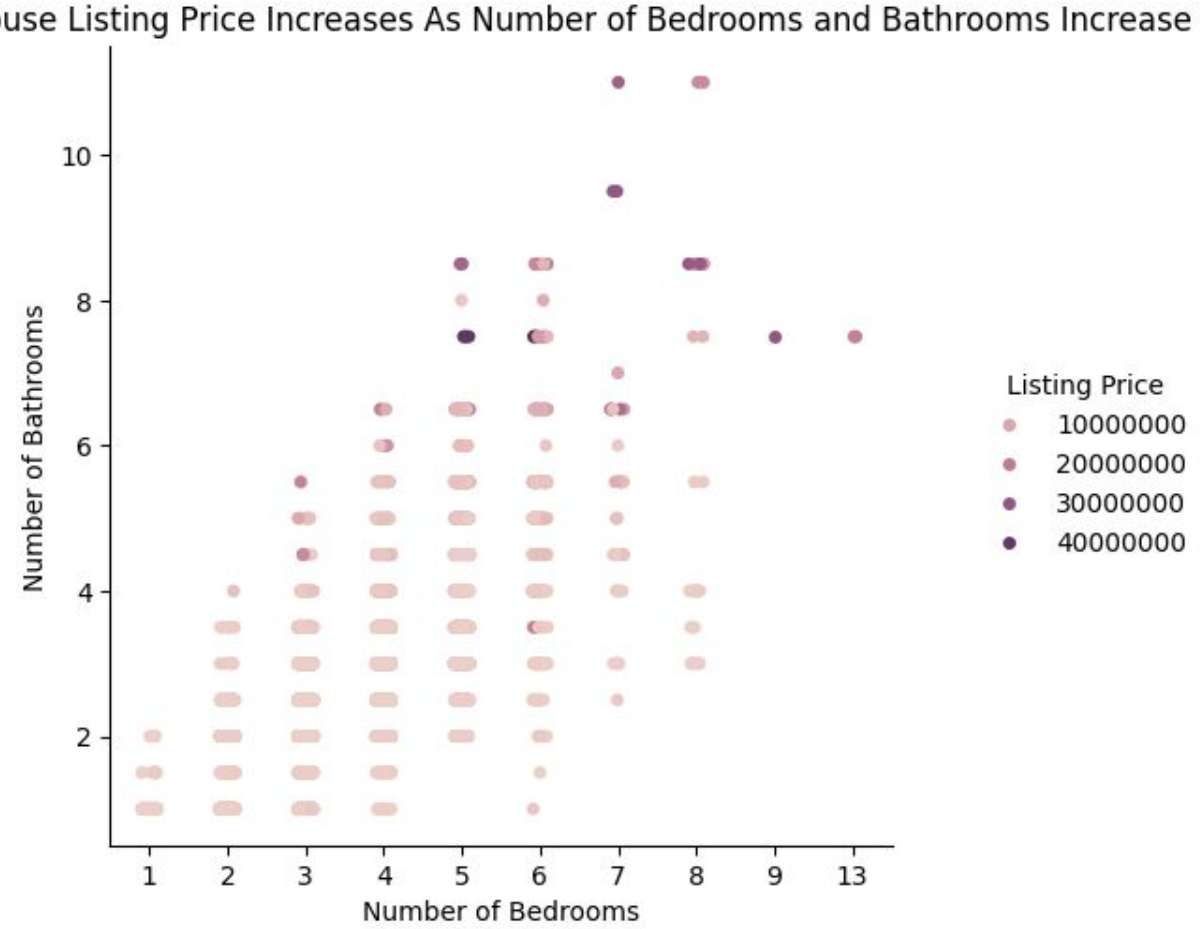
Q2: What is the average price of houses and average school quality in different cities?

Q3: What is the relationship between zip codes and price?

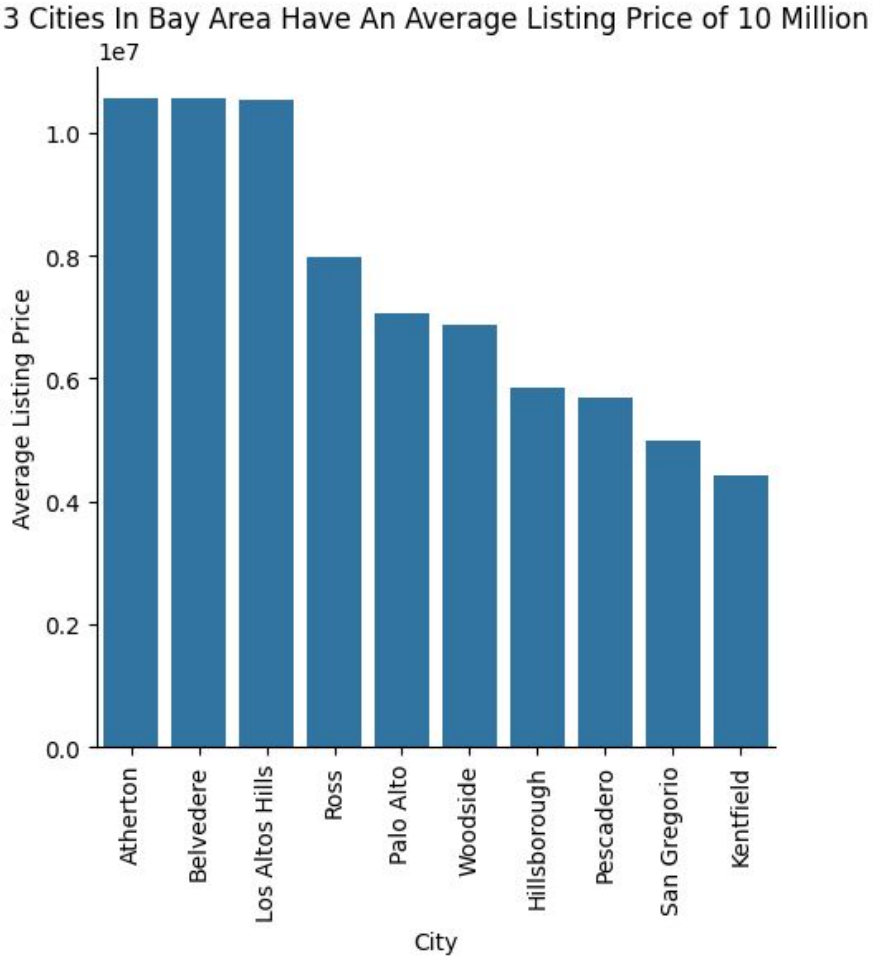


Q4: How do lot size and home size relate to the house price?

Q5: Does having more rooms (bedrooms, bathrooms) increase the price?



Q6: What are the 10 most expensive cities to live in?

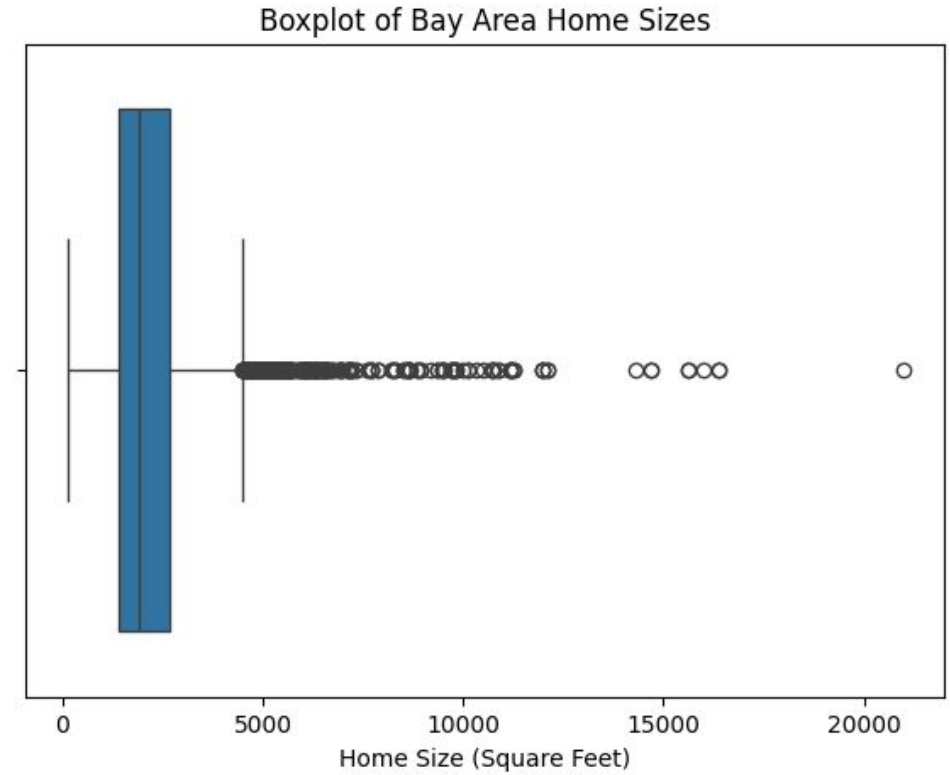


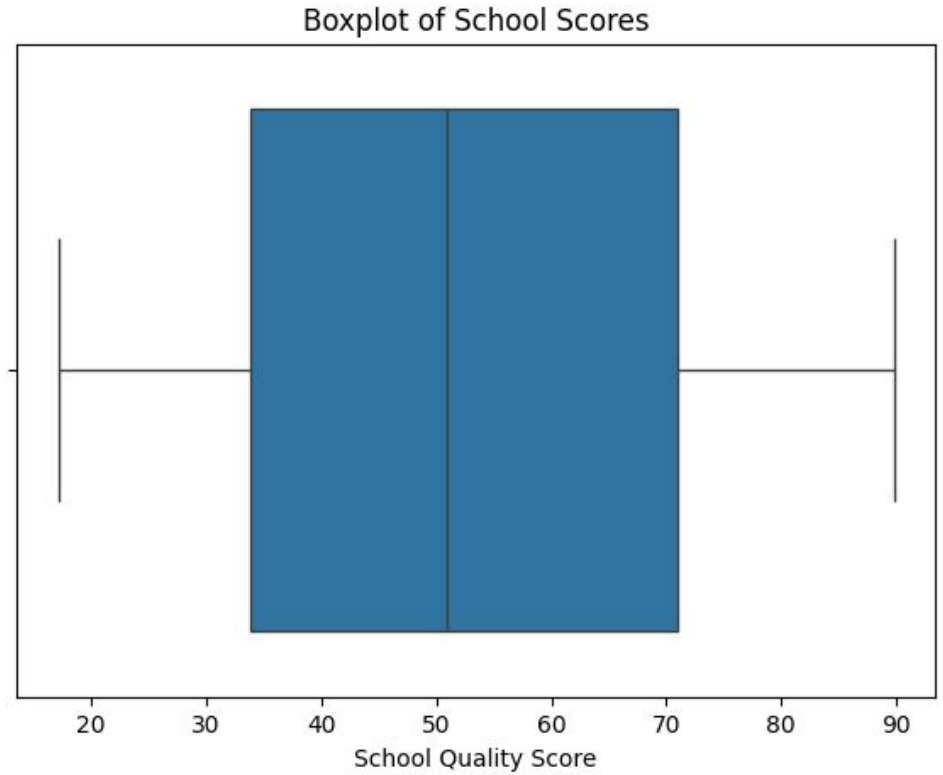
Preparing Data for Machine Learning

How Data Visualization Helps

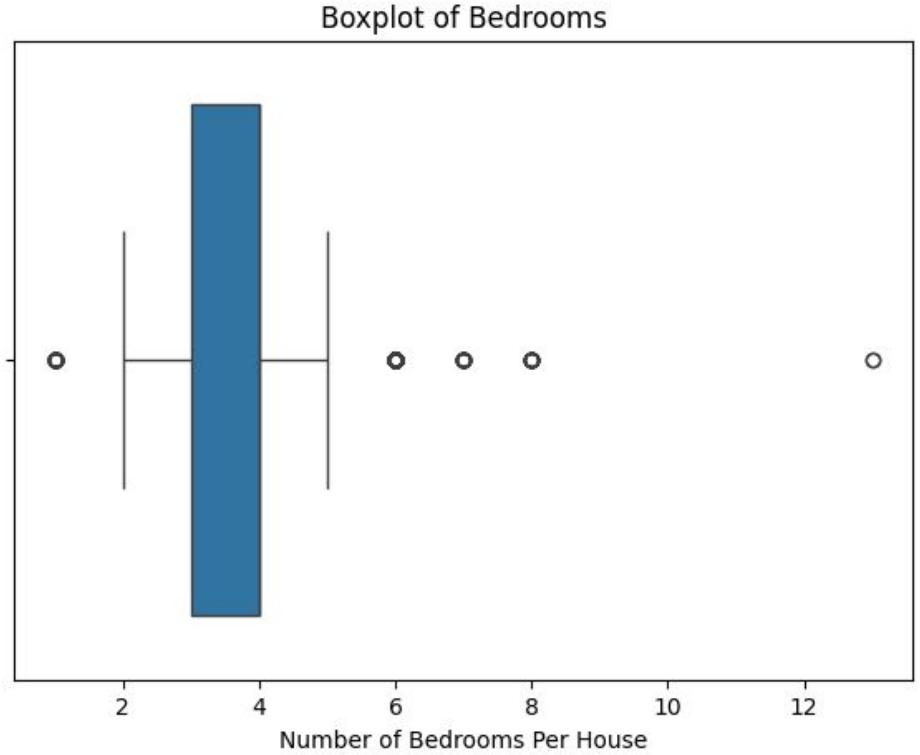
ML Goal: Predict the Price of Bay Area Houses

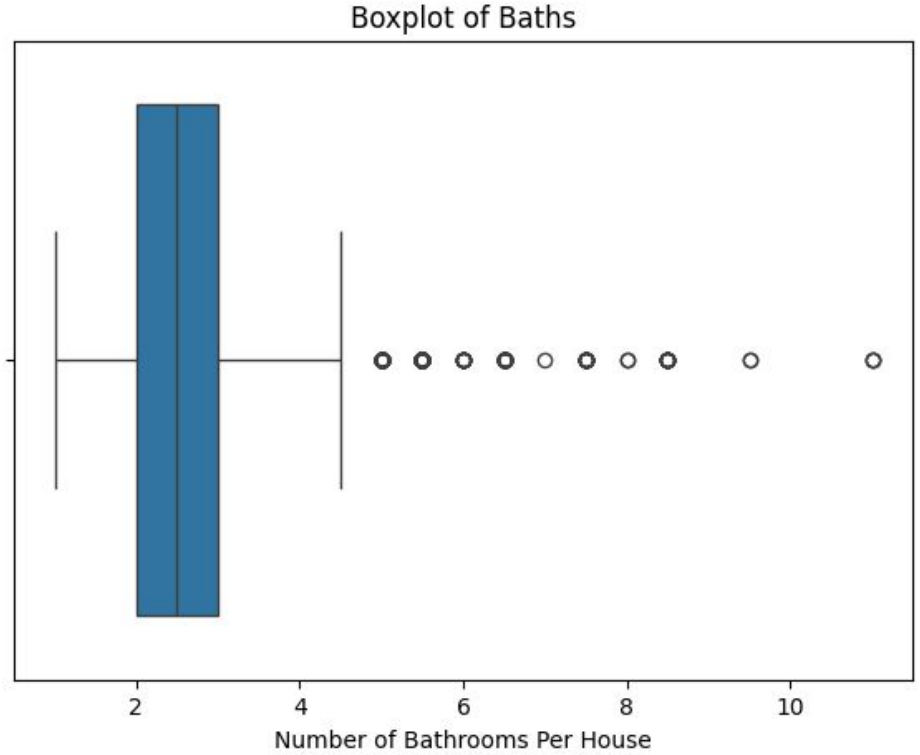
Visualize Chosen Features as Box Plots to Identify Outliers





Visualize Chosen Features as Box Plots to Identify Outliers





Machine Learning

Split Dataset Into Training and Testing

```
# The variables we want to use to predict 'Price' are 'Home size', 'School score', 'Beds', 'Baths'
# since they seem to be positively correlated with price
influencing_vars = no_outlier_cleaned_df[['Home size', 'School score', "Beds", "Baths"]]

# The variable we want to predict is 'Price'
target_var = no_outlier_cleaned_df['Price']

# results in four lists:
#the list of influencing values for the training dataset,
#the list of influencing values for the testing dataset,
#the list of target values for the training dataset,
#the list of target values for the testing dataset
#the test_size=0.2 means that the test dataset will be 20% of the no_outliers_cleaned_df dataset
influencing_train, influencing_test, target_train, target_test = train_test_split(influencing_vars, target_var, test_size=0.2, random_state=123)
```

Performance of Machine Learning Models

Linear Regression:

```
The Cross Validation Scores for Mean Squared Error for Linear Regression are:
 [-5.51714173e+11 -7.09608971e+11 -8.03281009e+11 -4.88320956e+11
 -9.07720868e+11 -4.50231442e+11 -3.62407081e+11]
The Average Cross Validation Score for Mean Squared Error for Linear Regression is: -610469214259.3729
```

Random Forest Regression:

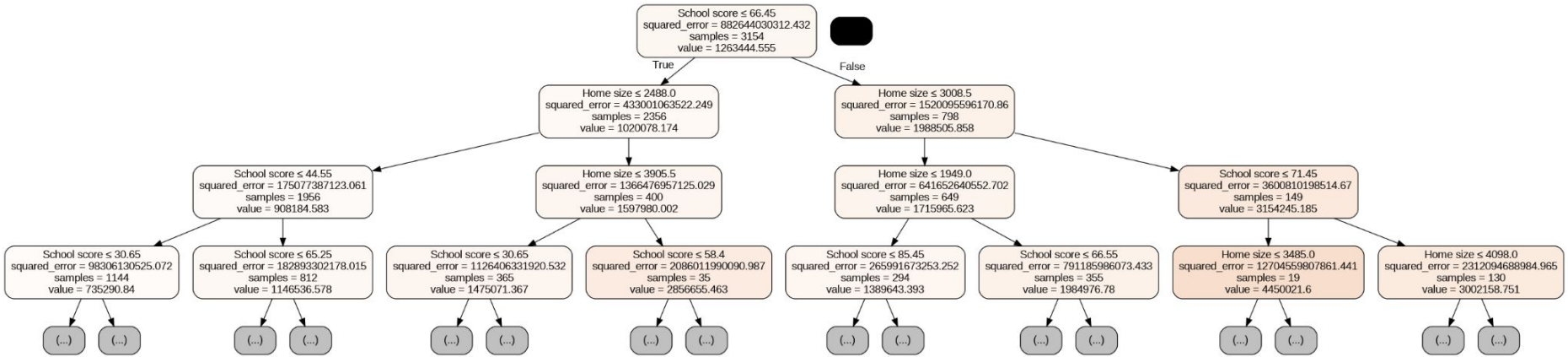
```
The Cross Validation Scores for Mean Squared Error for Random Forest Regression are:
 [-3.61014664e+11 -4.08881507e+11 -2.65584029e+11 -2.08203740e+11
 -3.79941745e+11 -1.52289277e+11 -2.03543329e+11]
The Average Cross Validation Score for Mean Squared Error for Random Forest Regression is: -282779755705.6136
```

Decision Tree Regression:

```
The Cross Validation Scores for Mean Squared Error for Decision Tree Regression are:
 [-7.75505599e+11 -7.04026031e+11 -3.78308424e+11 -6.16014559e+11
 -3.88924450e+11 -3.72126527e+11 -2.95942325e+11]
The Average Cross Validation Score for Mean Squared Error for Decision Tree Regression is: -504406845010.9193
```

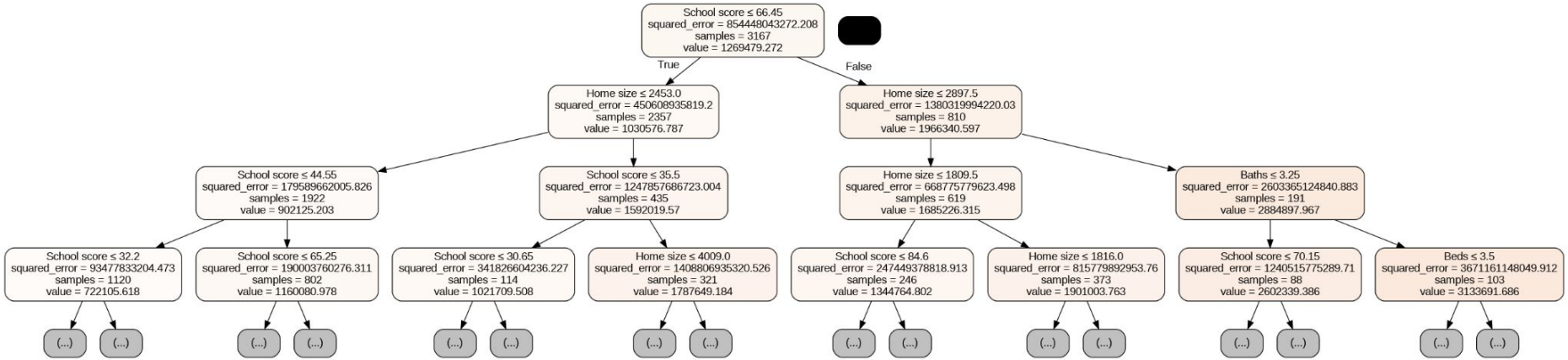
Visualization of Best ML Model (Random Forest)

The First Four Levels of The Last Decision Tree (100th) In The Random Forest:



Visualization of Refined ML Model

The First Four Levels of The Last Decision Tree (1000th) In The Random Forest:



Refine Best Model Using Hyperparameters and Training Data

Refine Best Model Using Hyperparameters:

```
#create random forests model with hyperparameters
refined_ran_for_model = RandomForestRegressor(n_estimators=1000, random_state=123)
```

Cross Validation Performance Using Refined Model And Training Data:

```
(Refined) The Cross Validation Scores for Mean Squared Error for Random Forest Regression are:
 [-3.44720357e+11 -4.12880593e+11 -2.58059880e+11 -2.07597085e+11
 -3.88819504e+11 -1.49108848e+11 -2.07336419e+11]
(Refined) The Average Cross Validation Score for Mean Squared Error for Random Forest Regression is: -281217526678.1075
```

Performance of Refined ML Model Using Testing Data

```
refined_ran_for_model.fit(influencing_train, target_train)
predictions = refined_ran_for_model.predict(influencing_test)
mse = mean_squared_error(target_test, predictions)
print("Testing Data MSE:",mse)
```

```
Testing Data MSE: 178182015031.42554
```

Challenges

Challenges

- Deciding what hyperparameters to use
- Visualizing the Random Forest Model
  - Getting the list of estimators used in the Random Forest model only works if the model is fitted first
  - Initial print outs of the graph visualization contained too many nodes, had to restrict using max_depth parameter

Thank You For Listening!