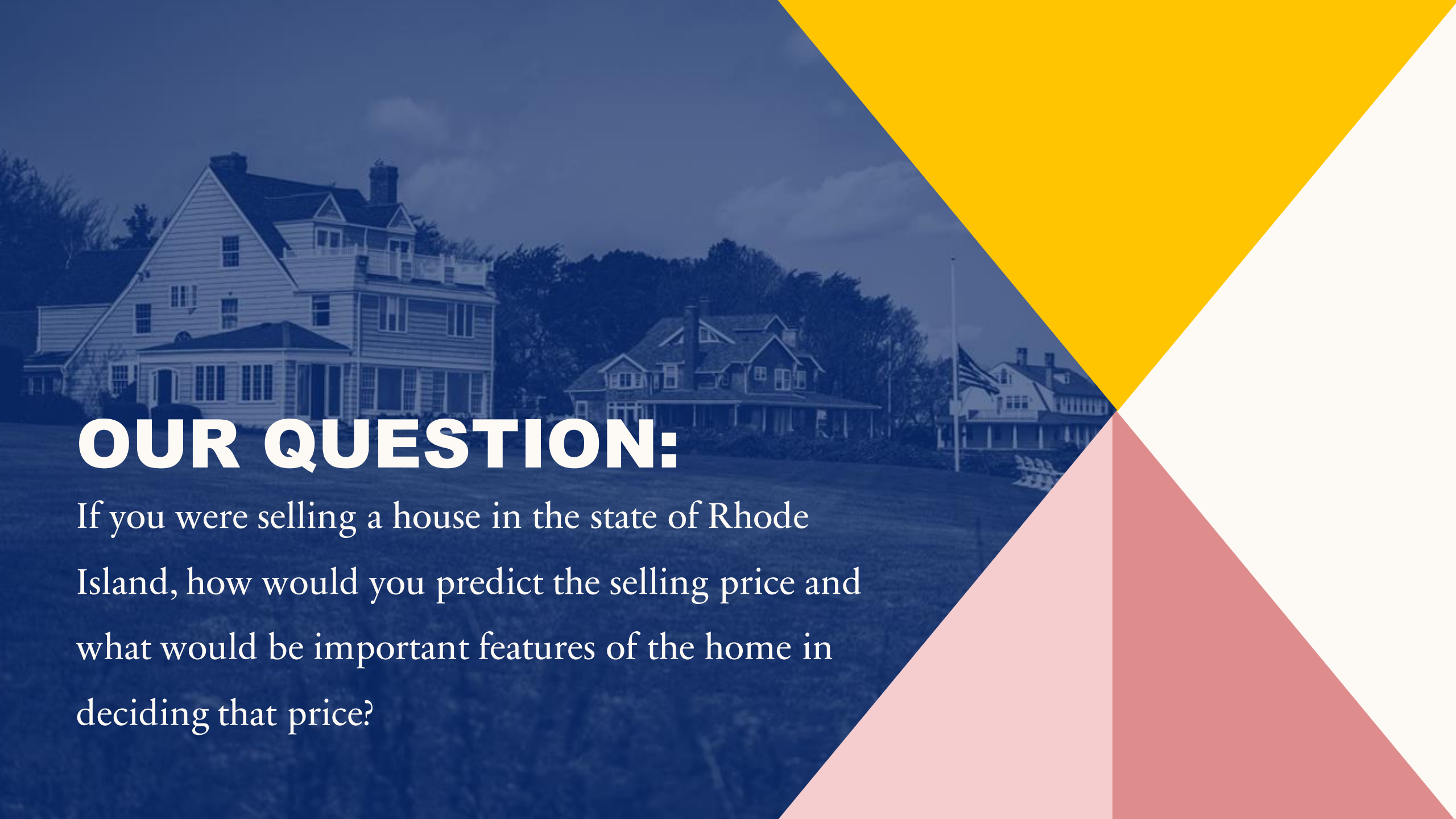
HOUSE PRICING IN RHODE ISLAND

Amara Henry
Christina LaManna
Jeremy Del Moral
Kady Epley



OUR QUESTION:

If you were selling a house in the state of Rhode Island, how would you predict the selling price and what would be important features of the home in deciding that price?

DATA SOURCE & CLEANING

- Includes selling prices of homes current on the market at the time of this project. This is not a time series dataset.
- We did not need geographic data – choosing to focus on attributes of the property such as number of rooms, bathrooms, square footage, and acreage of the lots.
- Our data is sourced from a Kaggle dataset originally from Realtor.com.
- We normalized and standardized the data prior to running any models.

MODELS CHOSEN

LINEAR REGRESSION

Iteration 1 – Including zip code as a feature
Results: R^2 = 0.3578

Iteration 2 – Excluding zip code as a feature
Results: R^2 = 0.3444

The drop in accuracy was likely due to the model having one less feature to consider

NEURAL NETWORK

This model was not as successful with our data due to the non-linearity of the model.

Results: R^2 = 0.3805

RANDOM FOREST

This model was the best fit out of the models and exceeded the minimum requirement for R^2 as defined by our project (> 0.8). We were able to achieve this by setting 100 branches.

Results: R^2 = 0.9766

FURTHER INVESTIGATION INTO RANDOM FOREST MODEL

Model	Parameters	Score
Linear Regression	*Iteration 1:* Trained 'X' Data (Bed, Bath, AcreLot, ZipCode, HouseSize) Trained 'Y' Data (Price)	R-squared Score: 35.7%
	Iteration 2: Trained 'X2' Data (Bed, Bath, AcreLot, HouseSize) *Trained 'Y2' Data (Price)*	R-squared Score: 34.4%
Neural Networks	3 Hidden Layers (64, 32, 1) Relu Activation Adam Optimizer Mean Squared Error (MSE) Loss Mean Absolute Error (MAE) Metrics Ran off of standardized X data (Bed, Bath, AcreLot, HouseSize)	R-squared Score: 38.3%
Random Forest	Random Forest Regressor Trained 'X' Data (Bed, Bath, AcreLot, ZipCode, HouseSize) Trained 'Y' Data (Price) Estimators = 100 RandomState = 42	R-squared Score: 97.6% MAE: $5,611.02 STDEV: $531,175.46 RMSE: $80,394.68

What about the model is so successful?

Utilization of decision trees in a way that addresses each feature of the property.

Data Fit

97% of the variability in the features are attributed by our regression model.

Further Analysis:

Root Mean Square Error: $80,394.68

Mean Absolute Error: $5,615.89

Standard Deviation: $531,175.46

Importance Features

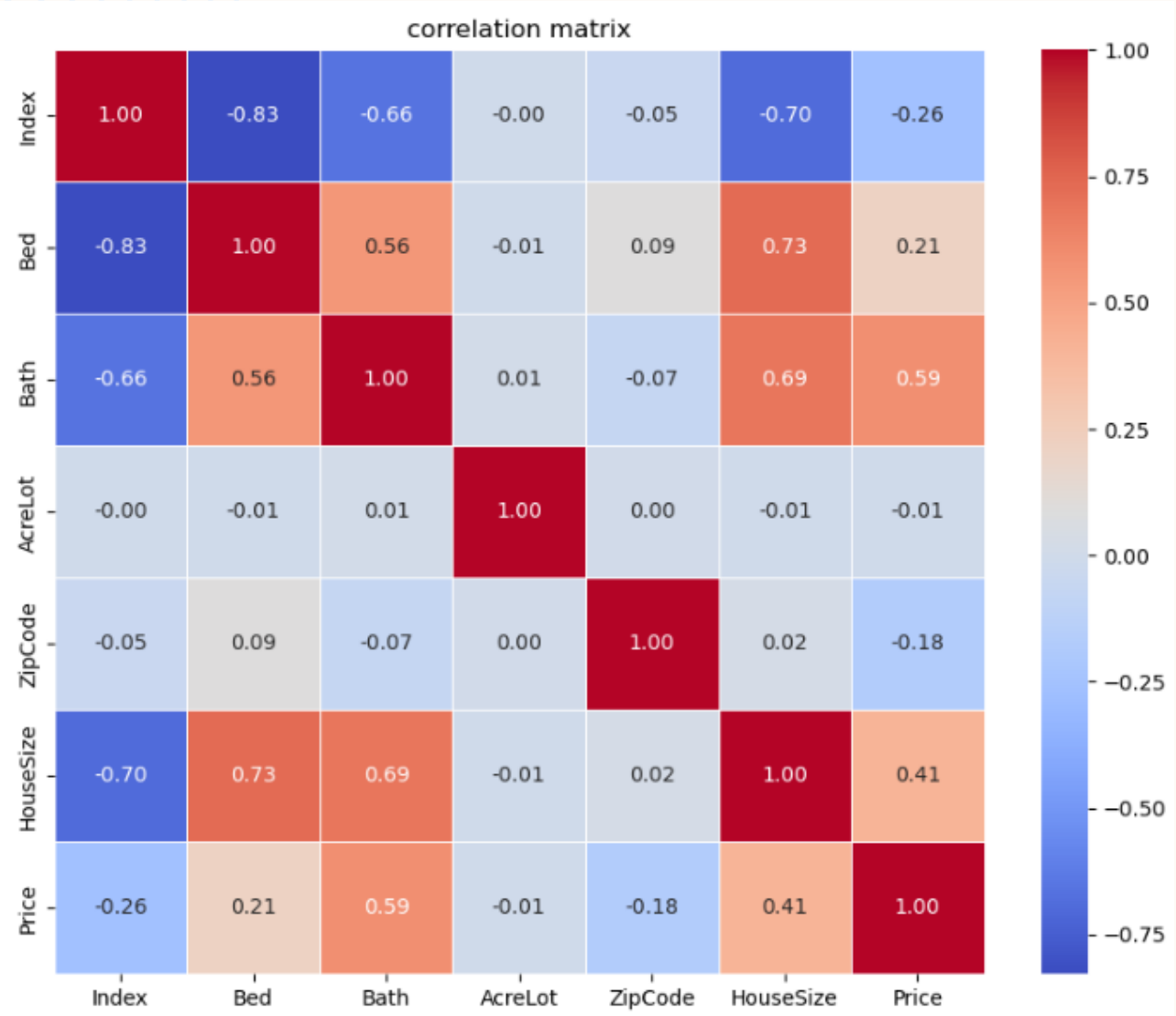
Bath: 0.437

Acreage: 0.191

House Size: 0.165

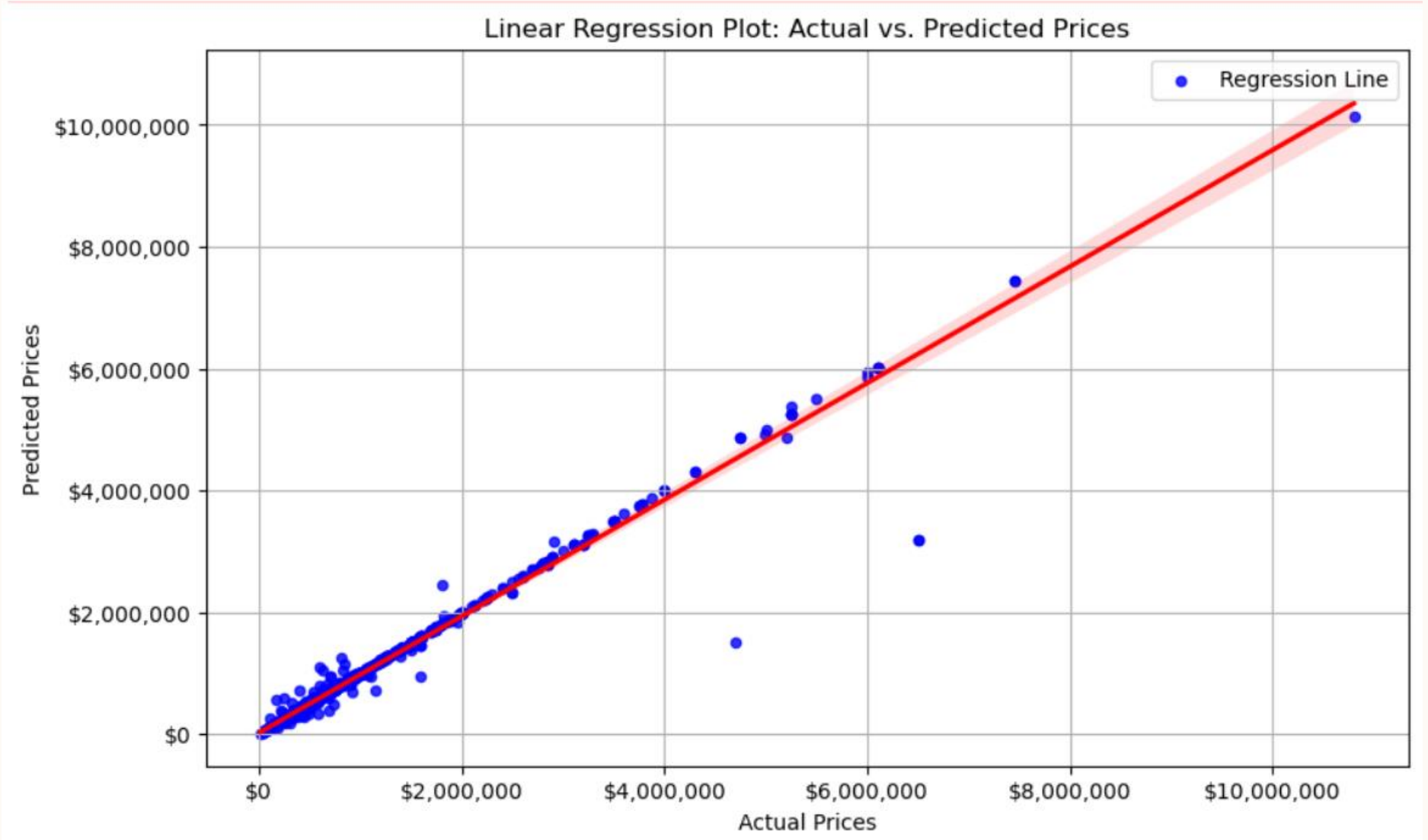
Zip Code: 0.118

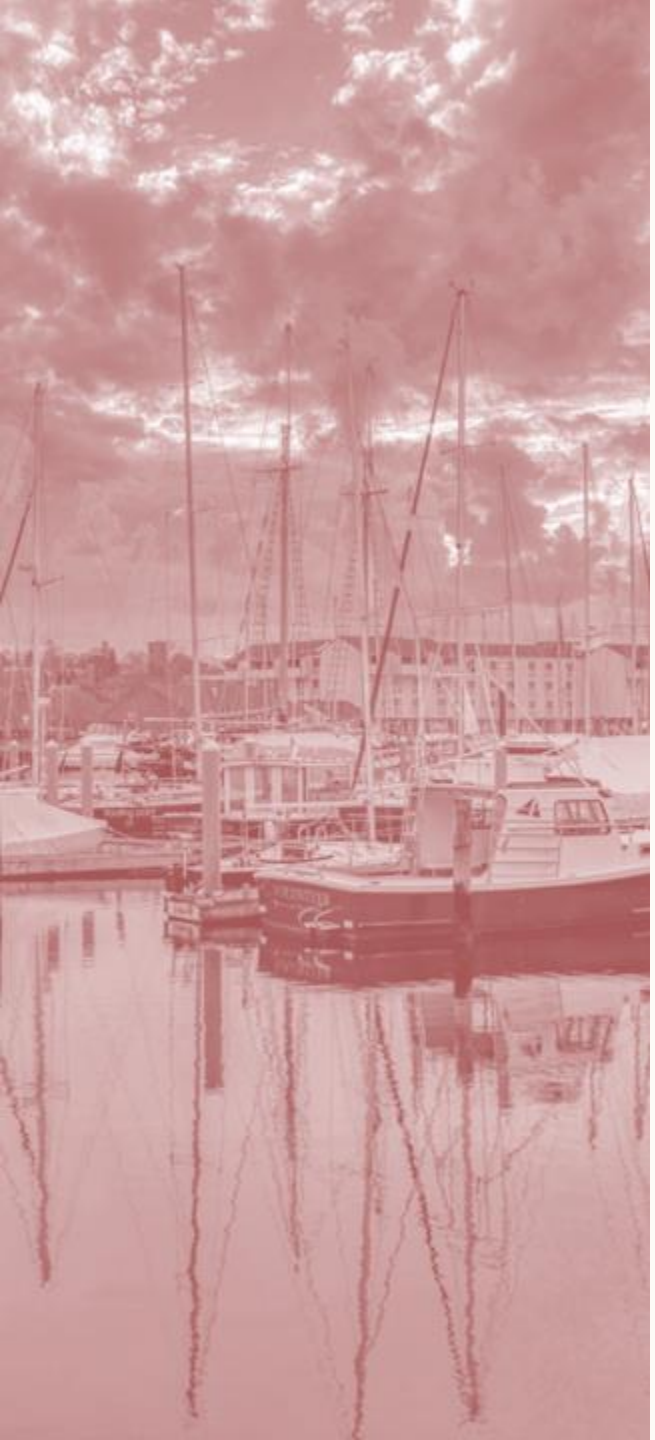
Beds: 0.089



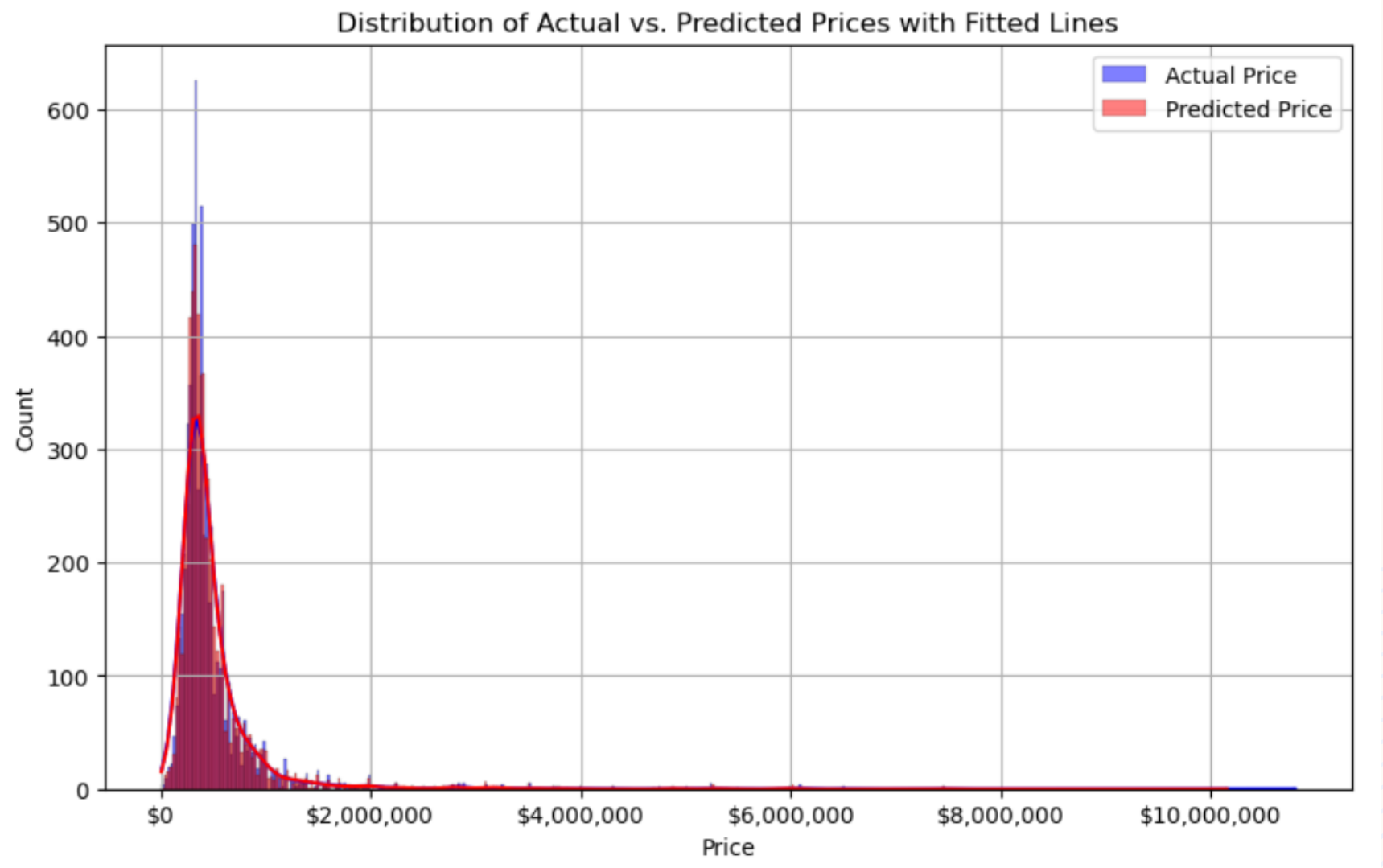
CORRELATION OF EACH FEATURE

COMPARING OUR MODEL TO ACTUAL PRICES

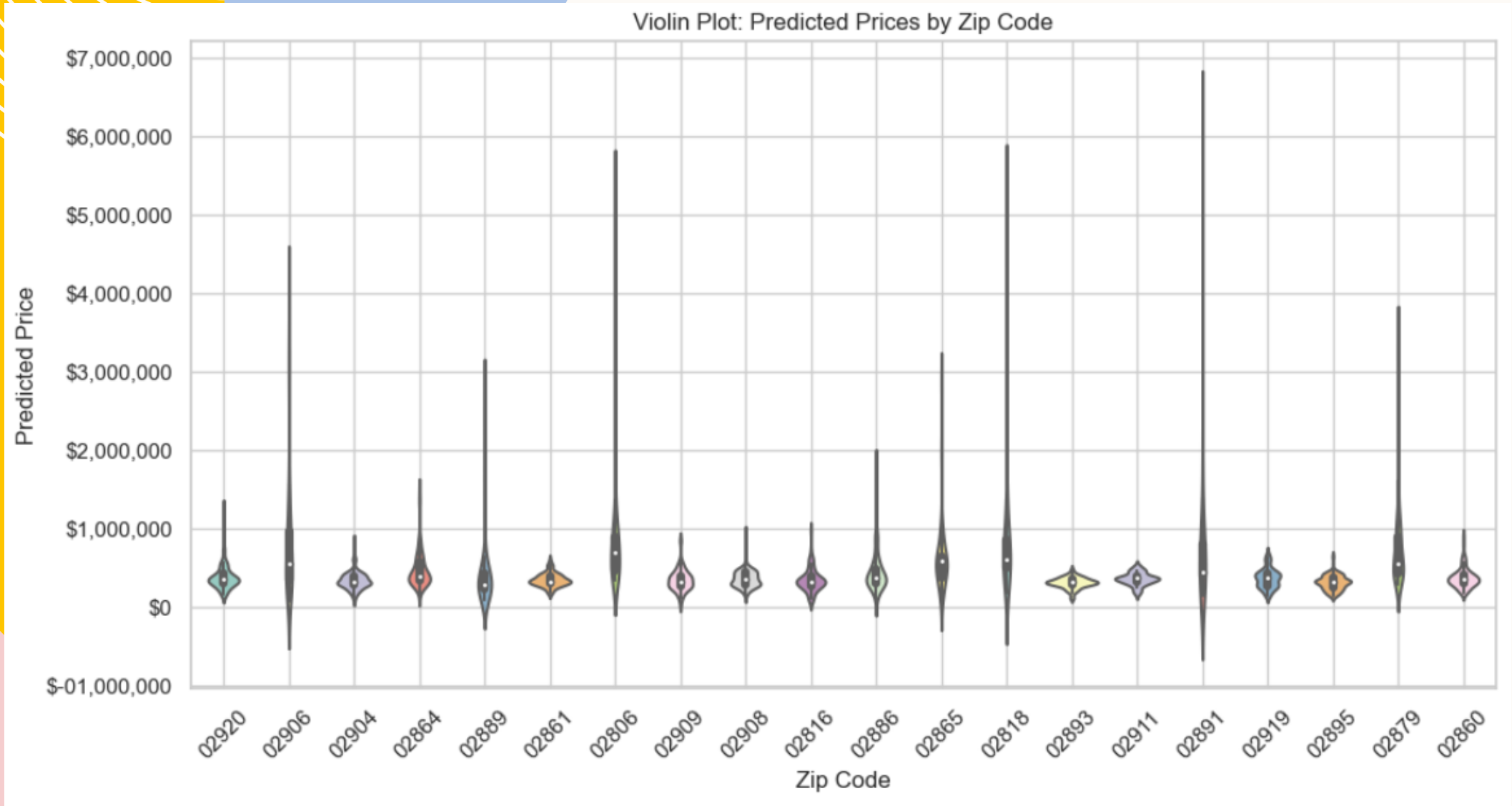




SECONDARY VIEW OF MODEL V. ACTUAL PRICES



COMPARING THE SELLING PRICES BY ZIP CODE



LIMITATIONS

- Some aspects of a home are going to be excluded from this dataset such as furnishings, recent remodels, and additional features such as pools.
- The data we have selected did have to be narrowed down due to null values such as 0 for sq. footage or no bathrooms. While we still maintained a significant number of records for current housing on the market, there have been records not accounted for.
- Our data excludes insight into foreclosures and banking institutions selling houses.
- We decided to focus solely on the attributes of the house/property. The data used to train the model was not a time series so the model solely a snap shot of the current housing climate.

SUMMARY

- Our goal was to identify the best machine learning model to answer our question of how to predict a selling price for a home and what would be the most important feature to indicate that price.
- We identified an appropriate dataset with the information we would need to evaluate, the number of bedrooms, bathrooms, acreage of the property, and square footage.
- We cleaned, standardized, and normalized this dataset to be run through two linear regressions, a neural network, and random forest models.
- Using the R^2 result from each model as an evaluation metric, we were able to determine that the random forest was the best machine learning model to predict the selling price of a home in Rhode Island.
- After further evaluation of this prediction model, we a Root Mean Square Error of $80,394.68 and a Mean Absolute Error: $5,615.89.
- The most important factors in determining the selling price of a home was be the number of bathrooms then the acreage of the property.

THANK YOU FOR YOUR TIME

Any questions?