Boston House Prices

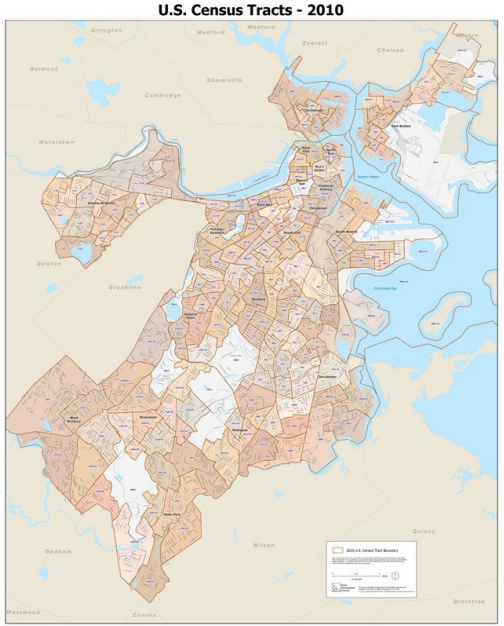
Alex Jaimes Sandoval, Roshan Sanyal, and Megan Schaeb

Rationale

- As of 2022 the housing sector accounts for 16.4% of the total US GDP. As such, we look at Boston housing data as its is one of the largest housing markets in the US.

Summary

- Question of interest: what factors impact the median value of Boston houses?
  - Focusing on environment, social, and architectural factors
  - Excluding location
- Created models to predict median house value by census tract in Boston, MA
- Interactions greatly increase accuracy
- The final model has great predictive power by including interactions between the variables

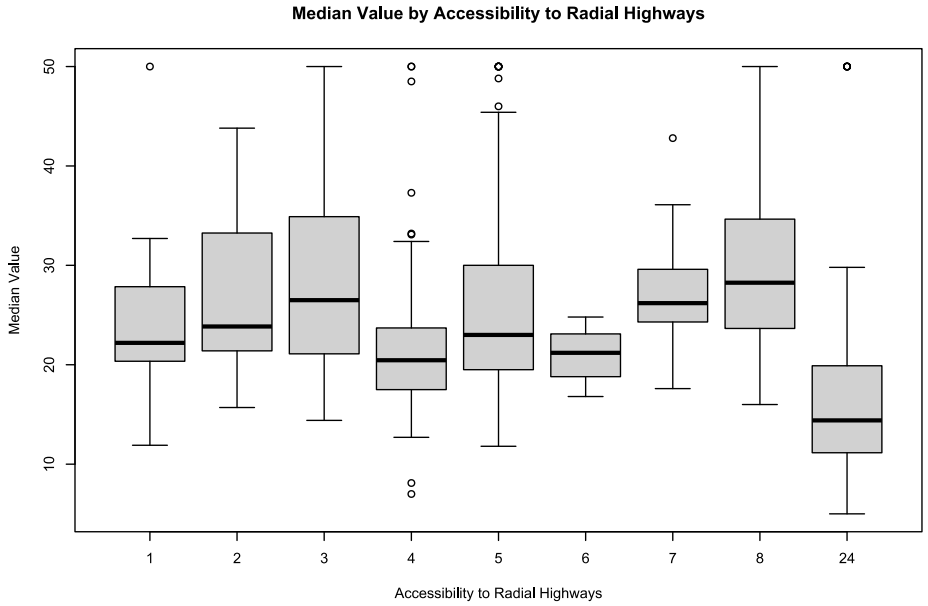


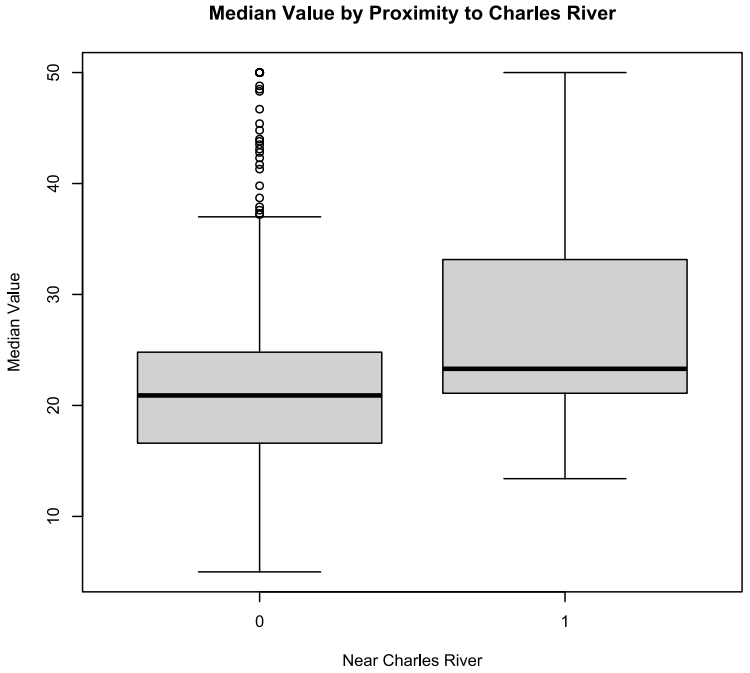
Our Data

- Data set consists of 506 observations and 13 potential predictor variables
  - 11 quantitative variables, 2 categorical variables
- Data sourced from Kaggle
- Used 90% of the data for training and 10% for testing

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV
1	0.00632	18.0	2.31	0	0.5380	6.575	65.2	4.0900	1	296	15.3	396.90	4.98	24.0
2	0.02731	0.0	7.07	0	0.4690	6.421	78.9	4.9671	2	242	17.8	396.90	9.14	21.6
3	0.02729	0.0	7.07	0	0.4690	7.185	61.1	4.9671	2	242	17.8	392.83	4.03	34.7
4	0.03237	0.0	2.18	0	0.4580	6.998	45.8	6.0622	3	222	18.7	394.63	2.94	33.4
5	0.06905	0.0	2.18	0	0.4580	7.147	54.2	6.0622	3	222	18.7	396.90	5.33	36.2

Exploratory Analysis

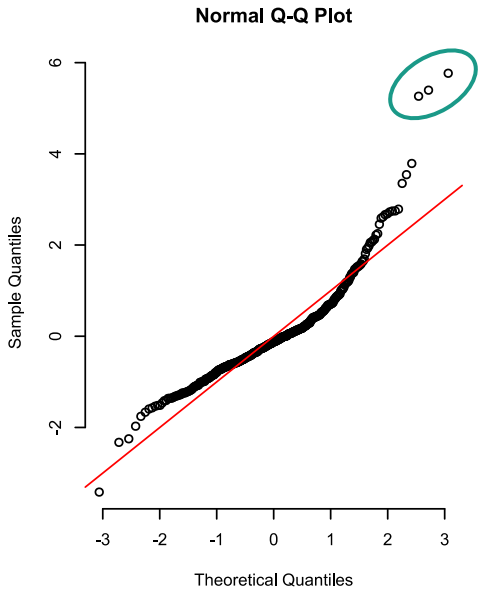




Preliminary Models and Adjustments to Data

Issues with full model that included all 13 predictor variables:

- Exhibited multicollinearity
  - Identified RAD variable as the source and removed the variable from future models
- Outliers
  - Removed from data set
- Non-normality and non-constant variance among residuals



Multicollinearity

- RAD and TAX both have a high VIF
- Removed the RAD variable from future models

	VIF
CRIM	1.861919
ZN	2.405411
INDUS	4.542974
CHAS	1.099373
NOX	4.694575
RM	2.011739
AGE	3.158231
DIS	4.101432
RAD	20.049527
TAX	10.660453
PTRATIO	2.227791
B	1.326824
LSTAT	3.100999

	VIF
CRIM	1.709995
ZN	2.211126
INDUS	3.658050
CHAS	1.046958
NOX	4.353862
RM	1.896570
AGE	3.067205
DIS	3.923741
TAX	3.584171
PTRATIO	1.730388
B	1.310980
LSTAT	3.031051

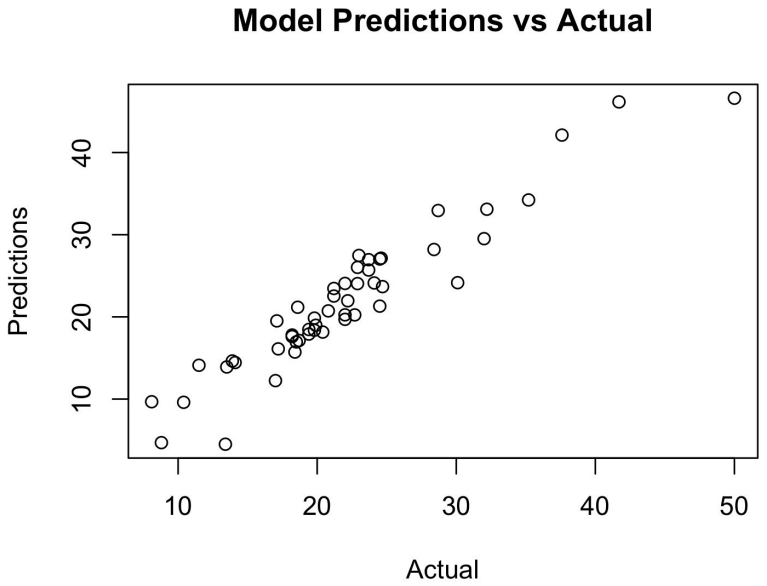
Model Selection

Fit several models and compared them using AIC, BIC, and adjusted R^2

	AIC	BIC	Adjusted R^2
Full model	1375.4480	1428.955	0.7551746
Full model with interactions	975.1649	1300.320	0.9110525
Backwards step selection with interactions	929.3969	1065.221	0.9122839
Mixed selection with interactions	1038.4237	1112.510	0.8848868
Forward step selection with interactions	1041.1643	1123.482	0.8846762

Final Model

- Final model: backward step selection with interactions
- 31 predictors in model + intercept
  - 12 individual predictors
  - 19 interaction terms
- Out of sample MSE was the lowest of all models we tested
- Correlation between actual and predicted values = 0.95



Final Model

- 12 individual predictors
- 19 interaction terms

	Estimate
(Intercept)	-111.5172643
CRIM	0.2477158
ZN	-0.0671455
INDUS	-0.9606610
CHAS1	64.7696488
NOX	15.4165173
RM	26.0201766
AGE	0.3556474
DIS	-0.9727279
TAX	-0.0090777
PTRATIO	0.6913679
B	0.0566791
LSTAT	2.7099459
CRIM:CHAS1	2.0652894
CRIM:NOX	-2.5024018
CRIM:RM	0.1526244
CRIM:LSTAT	0.0184398
ZN:TAX	0.0002658
INDUS:NOX	1.9968596
INDUS:RM	0.2173042
INDUS:PTRATIO	-0.0801020
CHAS1:NOX	-45.9441967
CHAS1:RM	-6.6636860
CHAS1:AGE	0.1385749
CHAS1:LSTAT	-0.7043011
NOX:AGE	-0.5086090
RM:TAX	-0.0229653
RM:PTRATIO	-0.5208911
RM:LSTAT	-0.3529709
AGE:TAX	0.0002082
AGE:B	-0.0006168
TAX:PTRATIO	0.0083652
TAX:LSTAT	-0.0022495

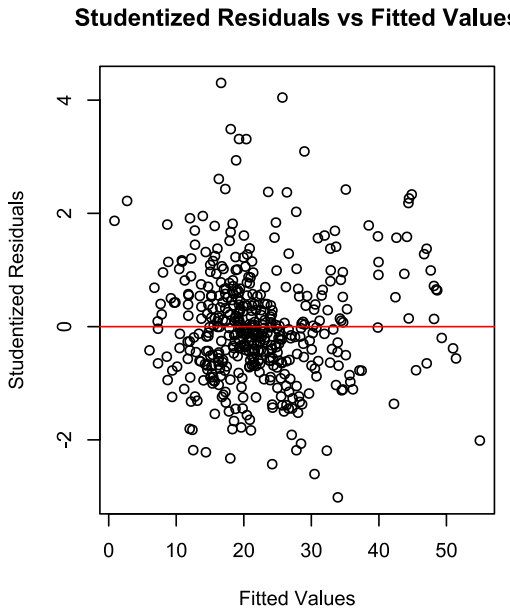
Model Assumptions

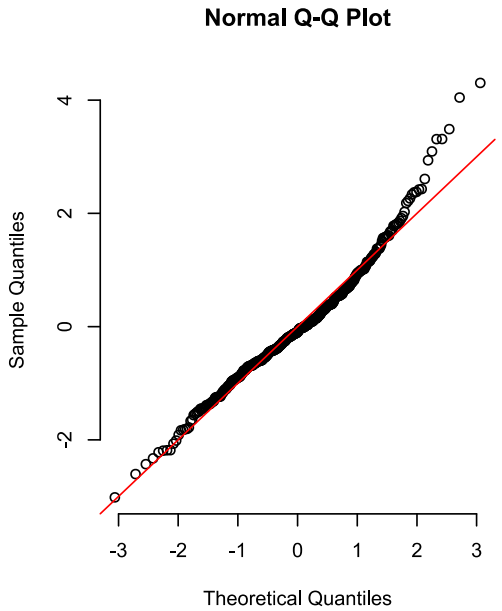
Normality does not hold:

- Shapiro-Wilks Test: p = 1.279e-7

Homoscedasticity holds:

- Breusch-Pagan Test: p = 0.05375





Conclusion

- Considered models with and without interaction terms
- Removed outliers and column with multicollinearity
- Recommendations
    - Could use k-folds cross validation to assess model accuracy.