Chapman University
CPSC 392-02

Final Project:

What can we learn from Boston Housing Data and how can we use it?

By Julian Murillo

Link to Video Presentation:
https://www.youtube.com/watch?v=UI51TMvv434&t=1s

SPRING 2020

Data Collection

- Kaggle: https://www.kaggle.com/kyasar/boston-housing#boston_housing.csv
- Observations: 506
- Attributes/Variables: 14 (13 continuous and 1 binary)
- Missing Values: 0
- ***Attribute Information:***

1.) Crim: per capita crime rate by town

2.) Zn: proportion of residential land zoned for lots over 25,000 sq.ft.

3.) indus: proportion of non-retail business acres per town

4.) Chas : Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)

5.) Nox: nitric oxides concentration (parts per 10 million)

6.) Rm: average number of rooms per housing

7.) Age: proportion of owner-occupied units built prior to 1940

8.) Dis: weighted distances to five Boston employment centres

9.) Rad: index of accessibility to radial highways

10.) Tax: full-value property-tax rate per $10,000

11.) Ptratio: pupil-teacher ratio by town

12.) Black: 1000(Bk - 0.63)^2 where Bk is the proportion o blacks by town

13.) Lstat: % lower status of the population

14.) medv: Median value of owner-occupied homes in $1000's

From a Business Perspective:

- Geographic and demographic data is powerful

- Dense neighborhoods such as those in Boston have stacked clientele (literally)

- Finding which neighborhoods, boroughs, or districts have the most disposable income will drastically increase the likelihood of revenue growth

Methods:

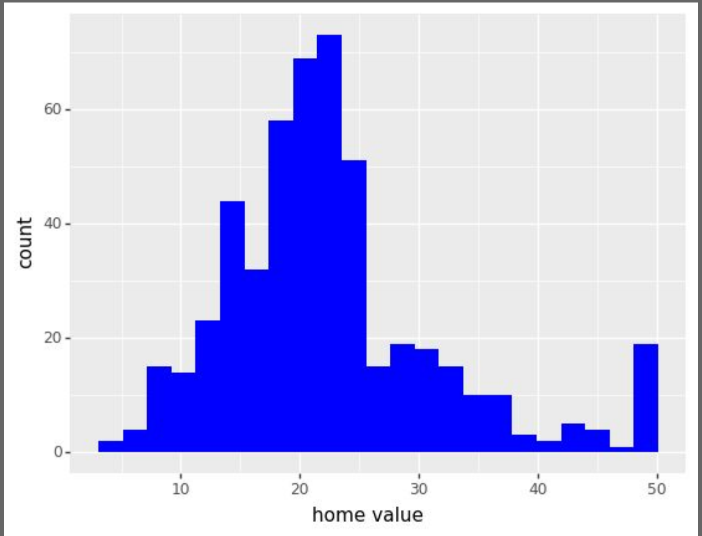
- Variables used: 12/13 (all continuous variables) for ALL models

- Standardizing all 12 predictor variables (all on different scales) for ALL models

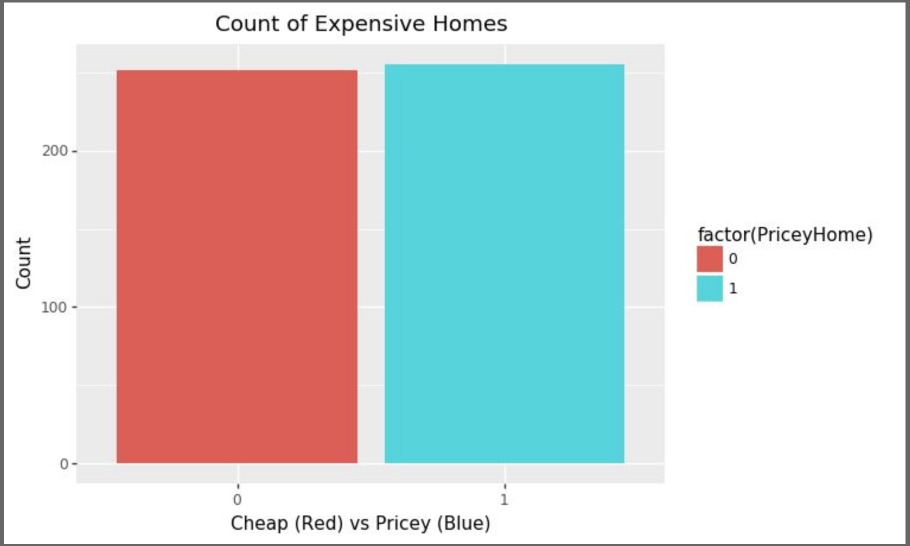
- Cross Validation: K-Fold (n_folds = 5, 20/80 split)

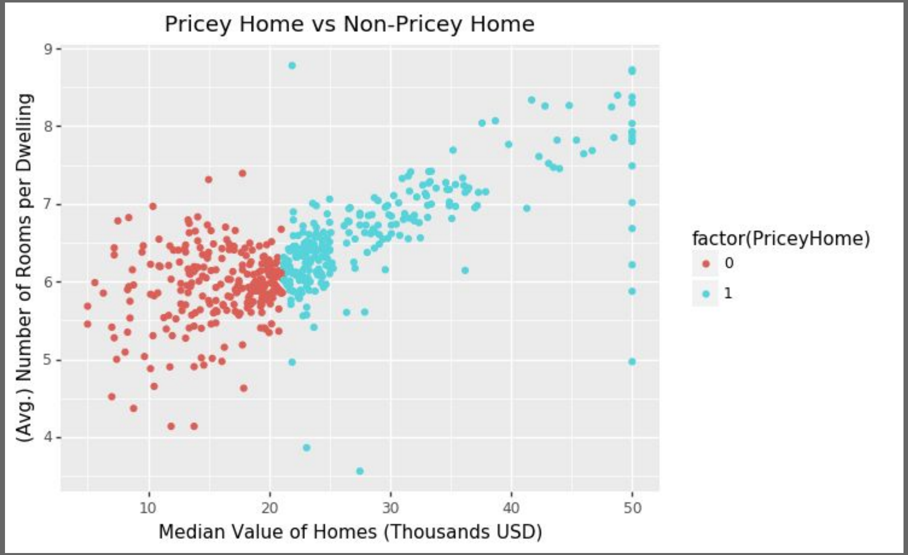
- Omit ‘Chas’ from potential predictors list

Exploring Data: Visualizations

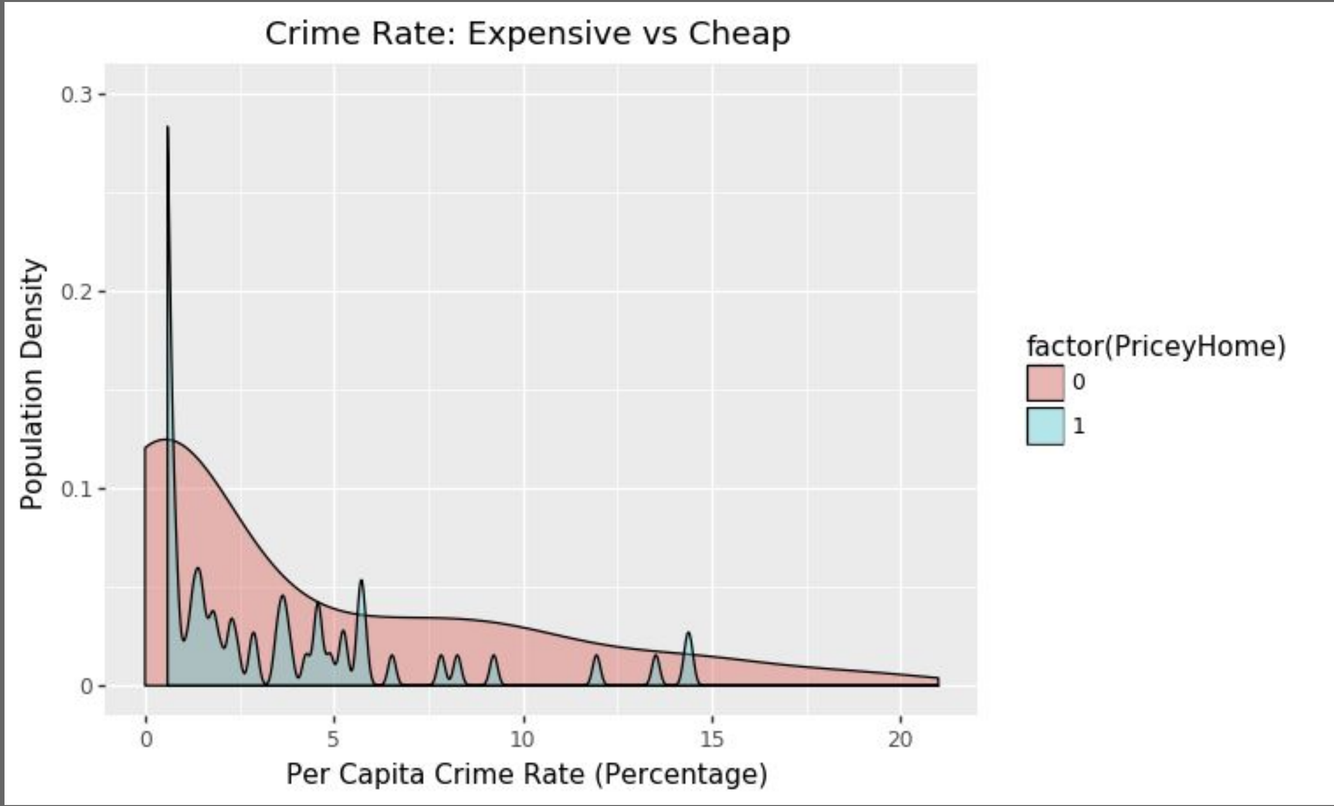


Exploring Data: Visualizations



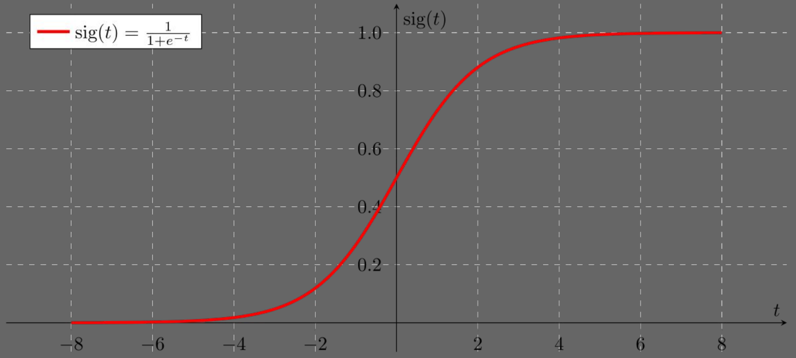


Exploring Data: Visualizations



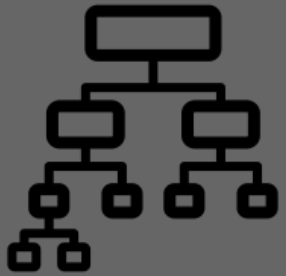
Question 1:

- **How might we go about determining whether or not a Boston home is expensive or not with the data we have available?**



- **Predict:** If home is pricey or not (0=cheap,1=pricey)

- **Analysis:** Logistic Regression and Decision Tree
  - **Why?:** Easier to predict binary variable using log-odds rather than trying to predict precise values for homes

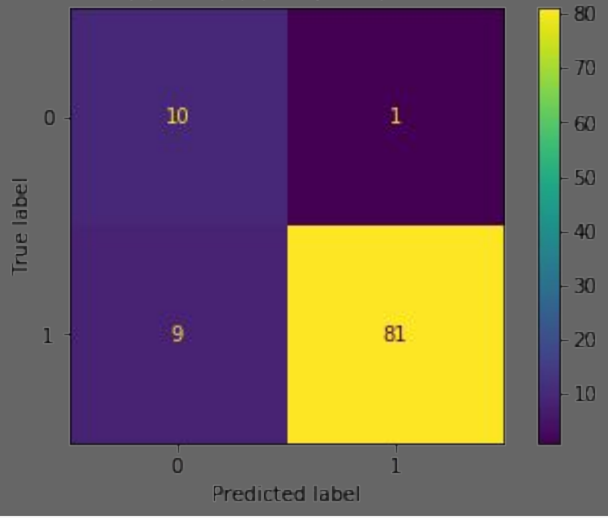


Question 1 Results: Logistic

Exponentiated Coeffs:

```
crim         0.903524
zn           1.013913
indus        1.023069
nox          0.003543
rm           4.979577
age          0.971171
dis          0.487168
rad          1.290278
tax          0.990587
ptratio      0.558602
black        1.004679
lstat        0.741257
dtype: float64
```

Best Model: C-Matrix



Accuracy Scores:

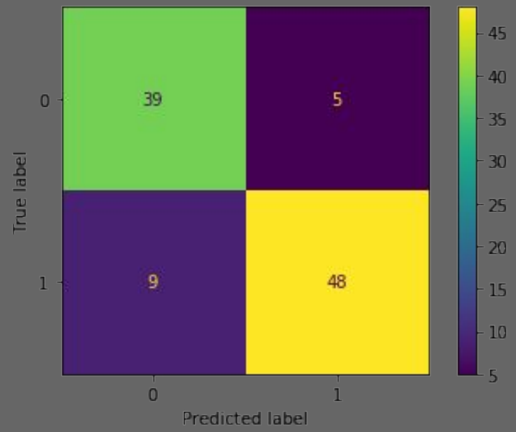
```
print(acc)
np.mean(acc)
```

```
[0.8333333333333334, 0.8910891089108911, 0.900990099009901, 0.7326732673267327, 0.7425742574257426]

0.8201320132013201
```

```
Logistic Regression: ROC AUC for last K-fold Model = 0.669
```

Question 1 Results: Decision Tree




```
Logistic Regression: ROC AUC for last K-fold Model = 0.853
```

Accuracy Scores:

```
[0.8235294117647058, 0.8316831683168316, 0.8316831683168316, 0.8613861386138614, 0.8514851485148515]
0.8399534071054164
```

Question 2:

- **In the city of Boston, is there a higher crime rate when the median values of homes are lower? How might we go about predicting crime rate?**
- **Predict:** Crime Rate (crim)
- **Analysis:** Ridge Regression and Lasso

- **Why?:** Variable importance

Question 2 Results: Ridge/Lasso

Ridge Regression:

```
TRAIN MAE:  2.923988878738928
TEST MAE:  2.322512797946774

TRAIN R2:  0.4384582782576133
TEST R2:  0.4039410999038191

 Alpha = 10.0
Our Model is Underfit
```

	Coef	Name
0	0.764239	zn
1	-0.669253	indus
2	-0.578999	nox
3	0.110325	rm
4	-0.202880	age
5	-1.391240	dis
6	4.165649	rad
7	0.460953	tax
8	-0.066871	ptratio
9	-0.649449	black
10	2.298568	lstat

LASSO:

```
TRAIN MAE:  2.883268714779122
TEST MAE:  2.1230343749597784

TRAIN R2:  0.4152944168019205
TEST R2:  0.6899058347438429

 Alpha = 0.051952642702813225
Our Model is Underfit
```

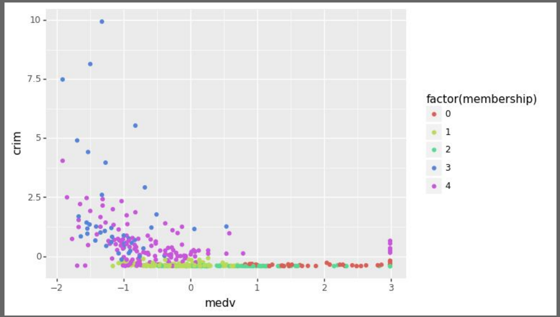
	Coef	Name
0	0.741308	zn
1	-0.281694	indus
2	-0.701652	nox
3	-0.000000	rm
4	0.000000	age
5	-1.230479	dis
6	4.453919	rad
7	-0.000000	tax
8	-0.101011	ptratio
9	-1.090720	black
10	1.677738	lstat

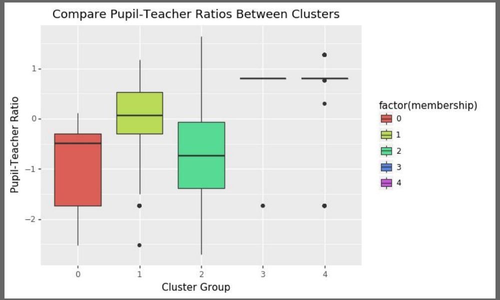
Question 3:

- **Does the number of rooms per house increase or decrease with an increase in the parent-teacher ratio in Boston? What factors determine the parent-teacher ratio?**
- **Predict:** Nothing, Cluster!
- **Analysis:** Gaussian Mixture Model and K-Means
- **Why?:** Distinguish neighborhoods and select best to set up shop

Question 3 Results:

K-Means:





K = 6 yielded a score of: .2545

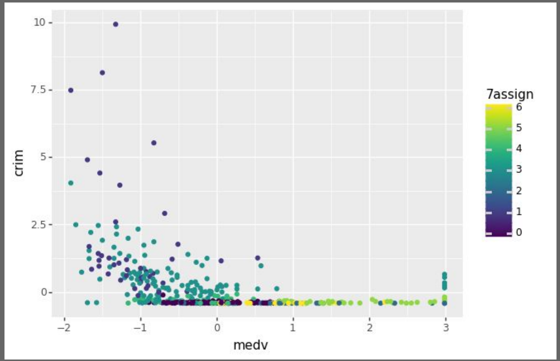
K = 5 yielded a score of: .3345

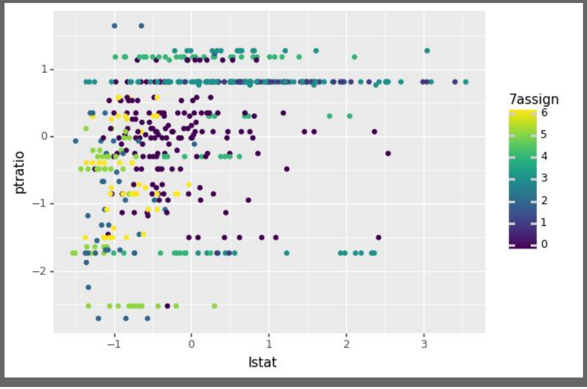
K = 4 yielded a score of: .3075

K = 3 yielded a score of: .2903

K = 2 yielded a score of: .2812

GMM:






```
[0.4351476018326517, 0.3484041386144893, 0.38652448285606467, 0.43850060560309195, 0.4564911676410142, 0.518672076144
5247]
```

Conclusion:

- Limitations:
  - Small data set with few variables
  - Small data set, small number of features (but clean data)
  - Only on one city
- Future Revisions:
  - Cross Validate Ridge and LASSO
  - Bruesch Pagan Test for Ridge and LASSO
  - Try running Elastic Net model
  - Include factored “Chas” variable
  - Narrow-down data set to just top 10% of expensive homes
- Business Recommendation:
  - Set up shop in neighborhood with good access to highways
  - Low p-t ratios
  - High property value
  - Less nitric oxides
  - Less crime
  - More bedrooms