Apartment Hunting on Craigslist

Using linear regression to predict rental prices in San Francisco

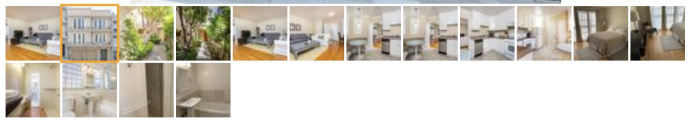
What makes an apartment great?

Price

$3,795 / 2br - 950ft^2 - Spacious 2BR,1BA PacHeights, washer dryer, Lafayette pk (pacific heights)

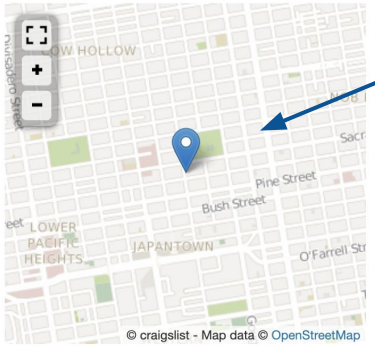
image 2 of 16

Photos are usually a good sign



Available now, nice, spacious, immaculate, FURNSHED flat (will consider unfurnishing) in best location. Ideal home for families, travel nurses, vacationers, corporate stays; two large bedrooms, bathroom with separate bathtub and shower, living Room, dining area, eat-in Kitchen, granite counters, breakfast nook, dishwasher, microwave, gas stove, office area, two large walk- in closets .Washer - dryer in this apt unit.

Location



Laguna St., # 2012 near Sacramento

(google map)

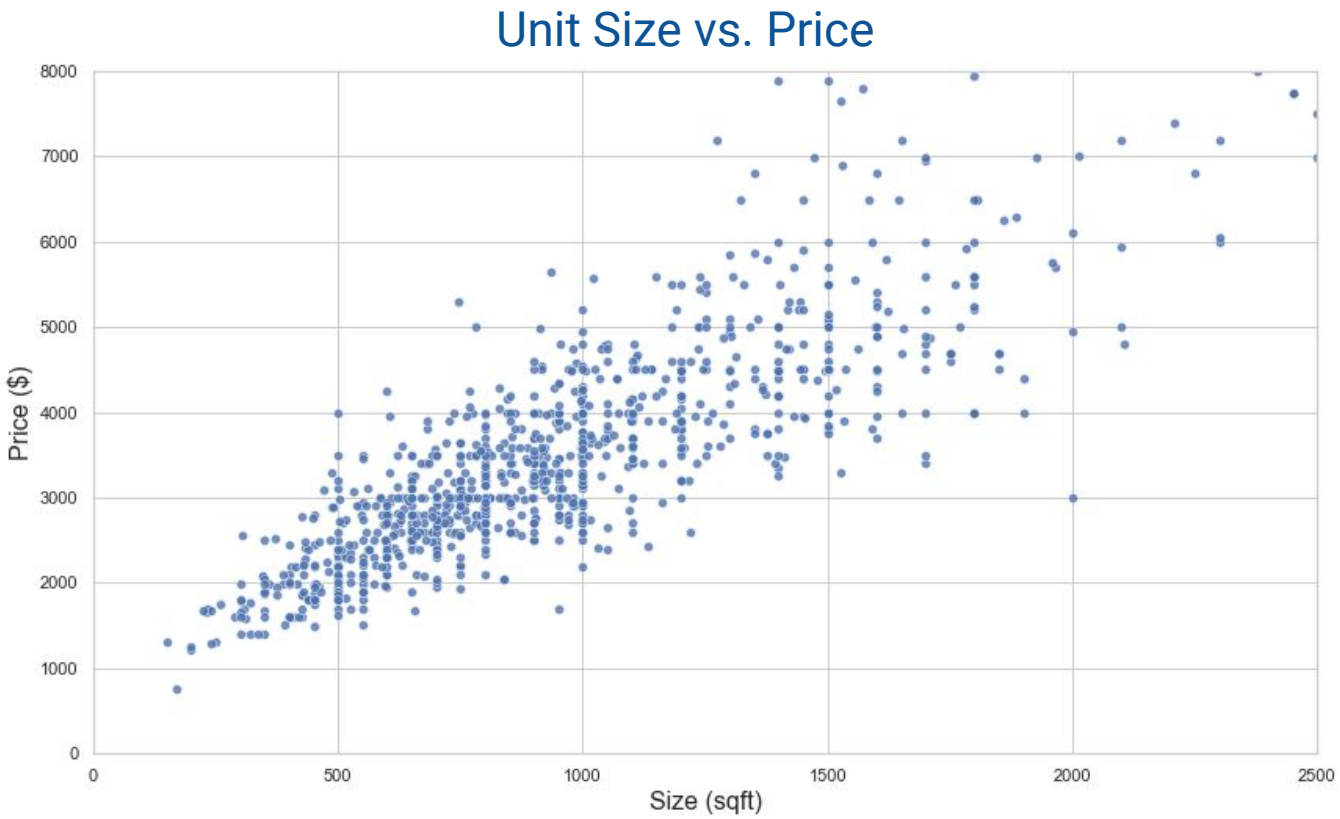
Size

2BR / 1Ba **950ft^2** available now

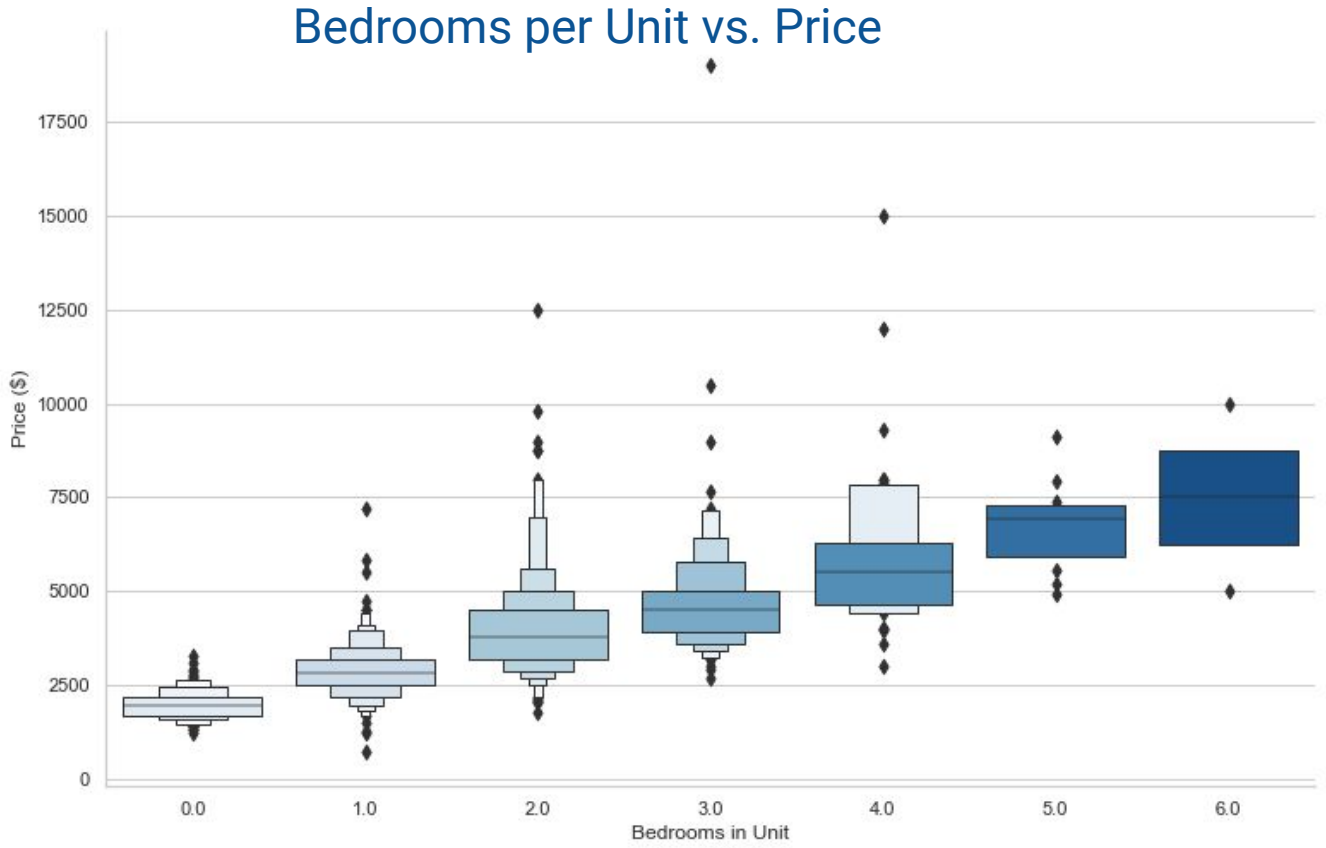
Amenities

- cats are OK - purrr
- dogs are OK - wooof
- furnished
- apartment
- w/d in unit
- no smoking
- attached garage

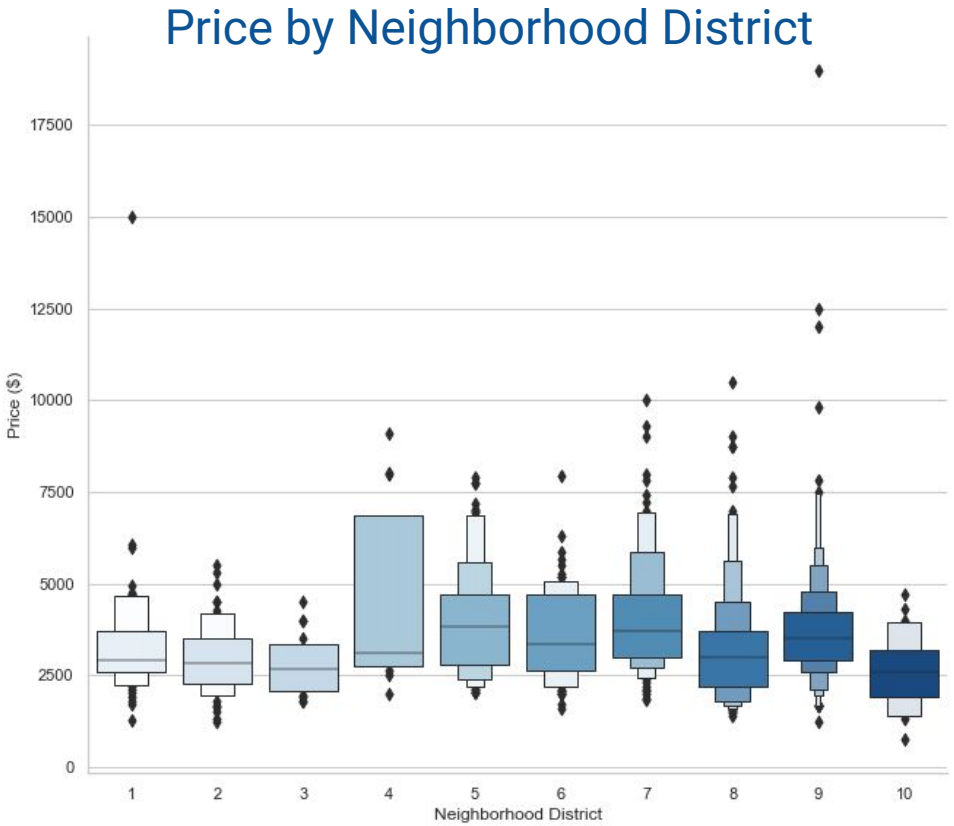
Strong relationship between unit size and price...

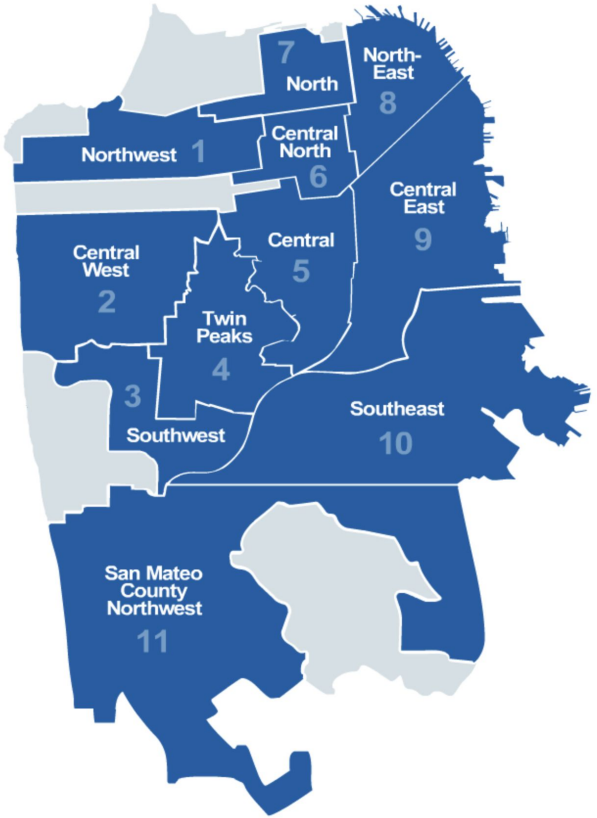


and between bedrooms and price



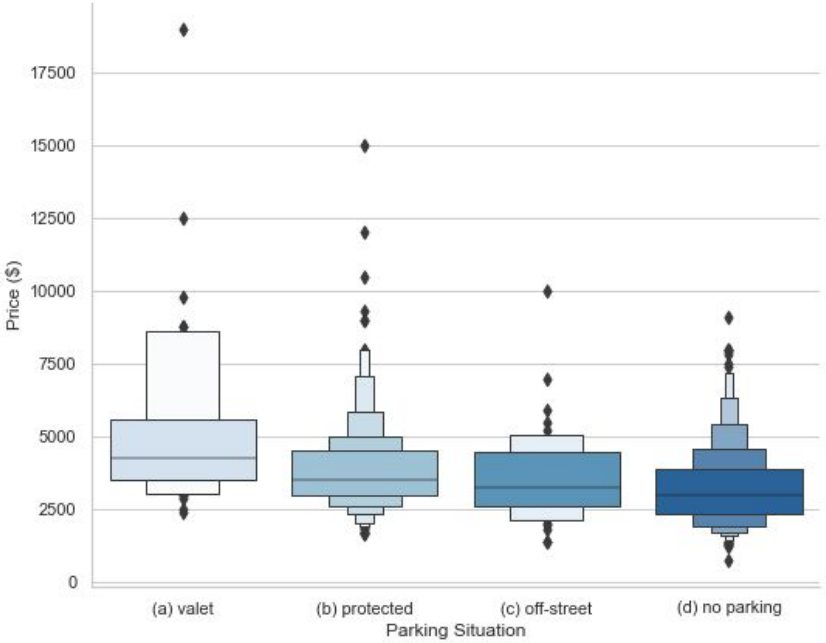
Location can affect high-end of price range



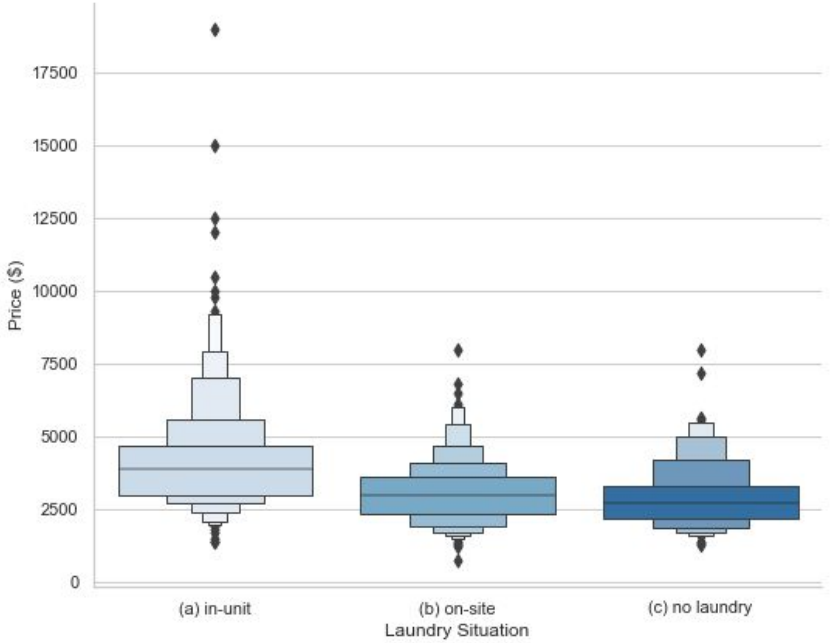


Some amenities command a premium

Price by Parking Situation



Price by Laundry Facilities



Model construction and selection

Feature Selection

Continuous Variables
- ➔ Unit size (sqft)

Discrete Variables
- ➔ Bedrooms
- ➔ Bathrooms

Ordinal Variables
- ➔ Parking
- ➔ Laundry

Categorical
- ➔ Pets
- ➔ Neighborhood

Engineering

Polynomial Features
- ➔ Degree = 2

Regularization methods tried
- ➔ Ridge (α = ~11)
- ➔ Lasso (α = ~2)

Model Selection

Ridge scored most consistent between training and validation testing
- ➔ R^2
- ➔ RMSE
- ➔ MAE

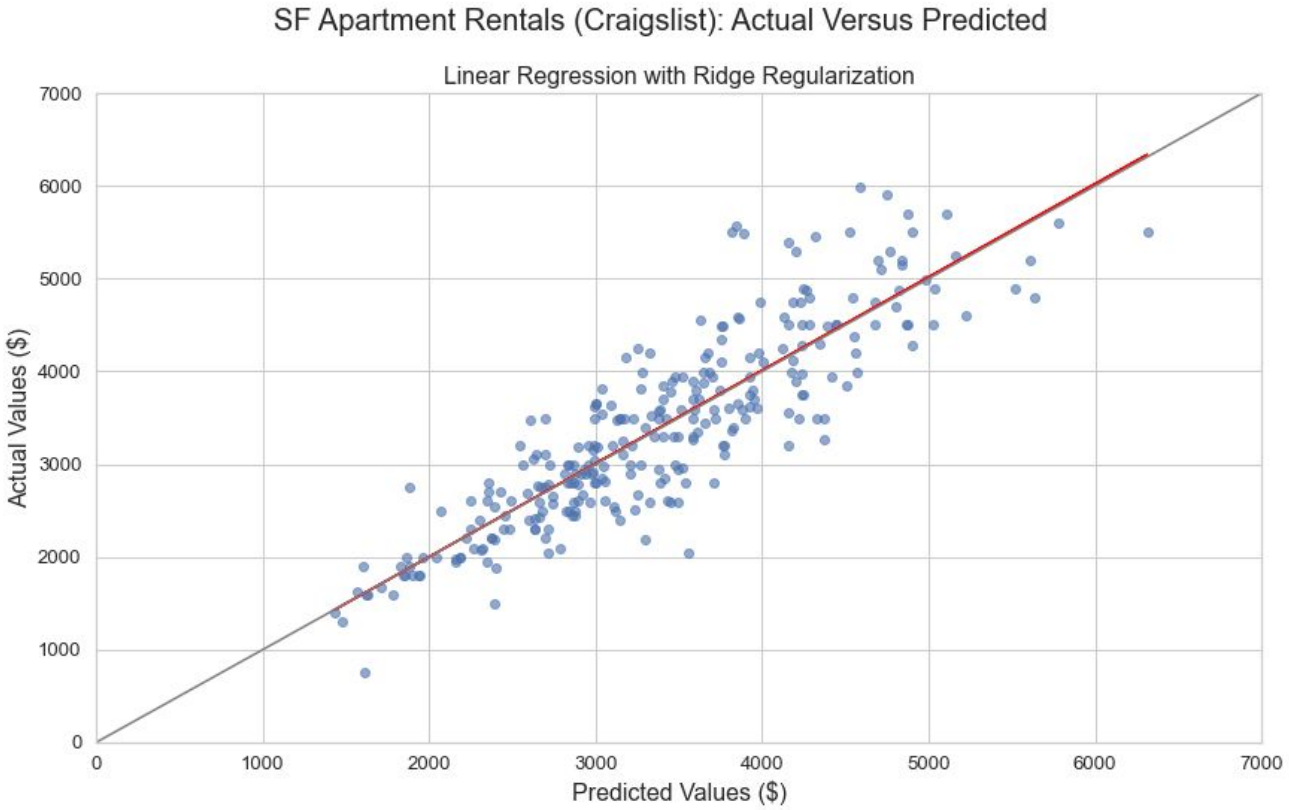
Lasso performed consistently on R^2, but showed greater variance between training and validation scoring on RMSE and MAE

Model results

R^2 Score
- ➔ Train: 79.6%
- ➔ Test: 81.7%

RMSE
- ➔ Train: ~$455
- ➔ Test: ~$465

MAE
- ➔ Train: ~$340
- ➔ Test: ~$375



Model evaluation

Tolerance of price variation (RMSE, MAE) **relative to apartment price** and **subjective to budget**

➔ ~$350-450 can be substantial for one person

➔ BUT split between 2 or more people may be okay

Model does a poor job of predicting for units > $4,000 / month

➔ Residuals plot[1] shows greater variance above $4k

➔ Most SF apartments in sample fall between $2,500 and $4,000 per month[2]

[1]Appendix A - Residuals plot (predicted versus actual prices)
[2]Appendix B - Distribution of rental prices in sample

Final thoughts

There could be other factors influencing rents

- ➔ Rent history, control
- ➔ Other amenities not standard in Craigslist posts (fitness facilities, common spaces, etc.)
- ➔ Other environmental or economic factors[1]

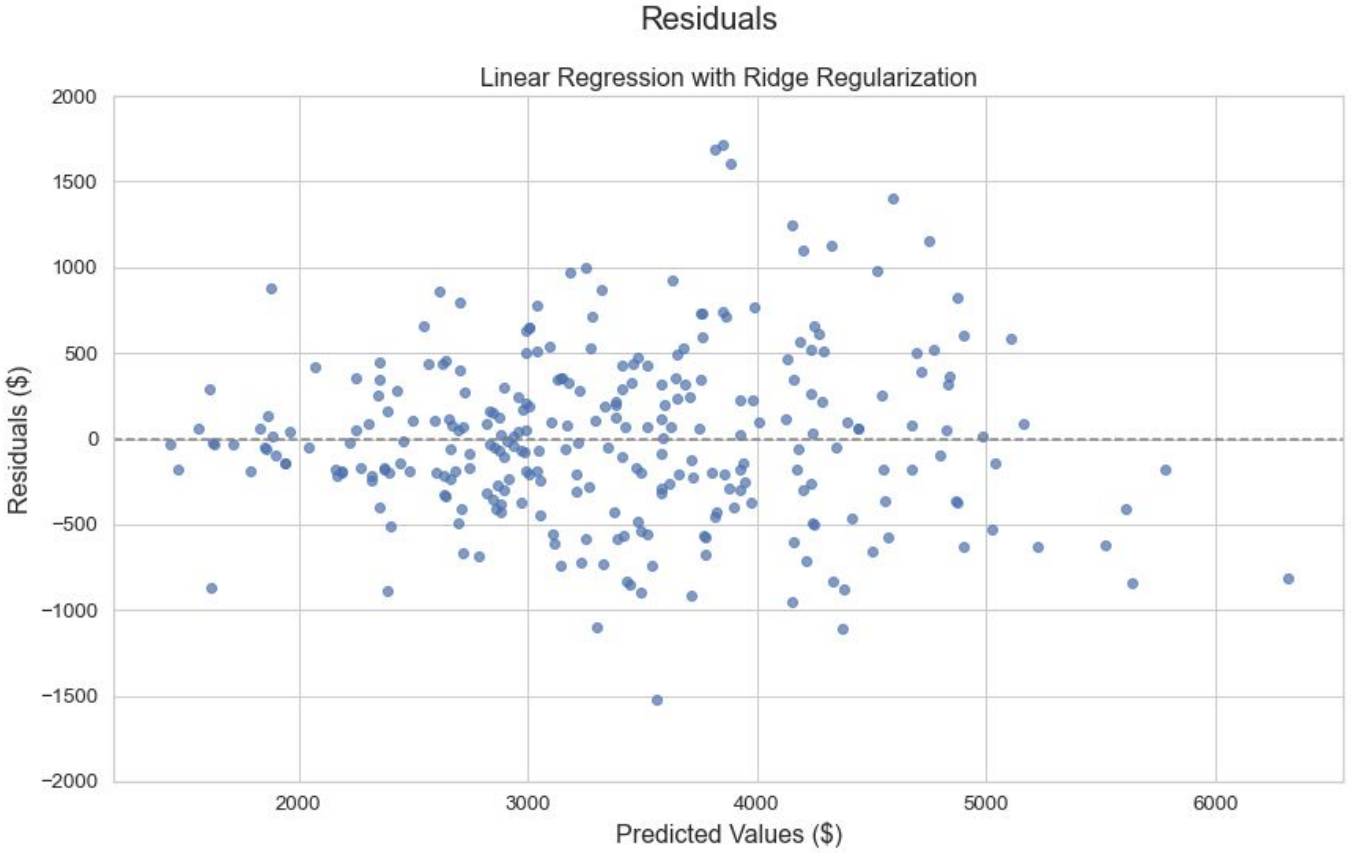
Possible future work

- ➔ Sample again when COVID's influence subsides
- ➔ Increase scope of scraping methods
  - ◆ Search text for key words like 'gym' or 'backyard' as additional amenities
  - ◆ Find information on security deposit and/or lease requirements
  - ◆ Implement image processing tools to evaluate apartment quality from photos
- ➔ Scam post identification

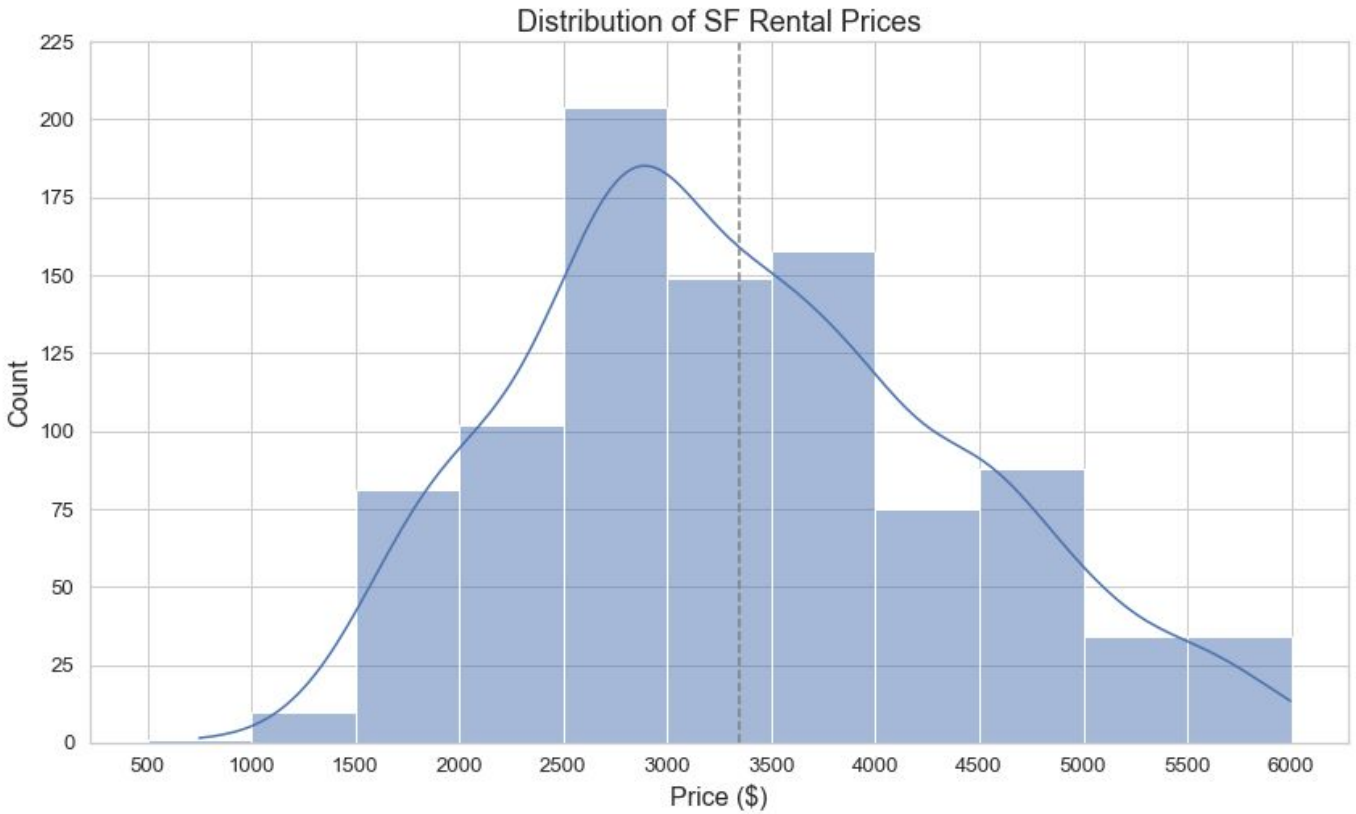
[1]Appendix C - Declining SF Rents


Thank you!
Chris Doenlen
GitHub Repository
scrapfishies@gmail.com
@scrapfishies

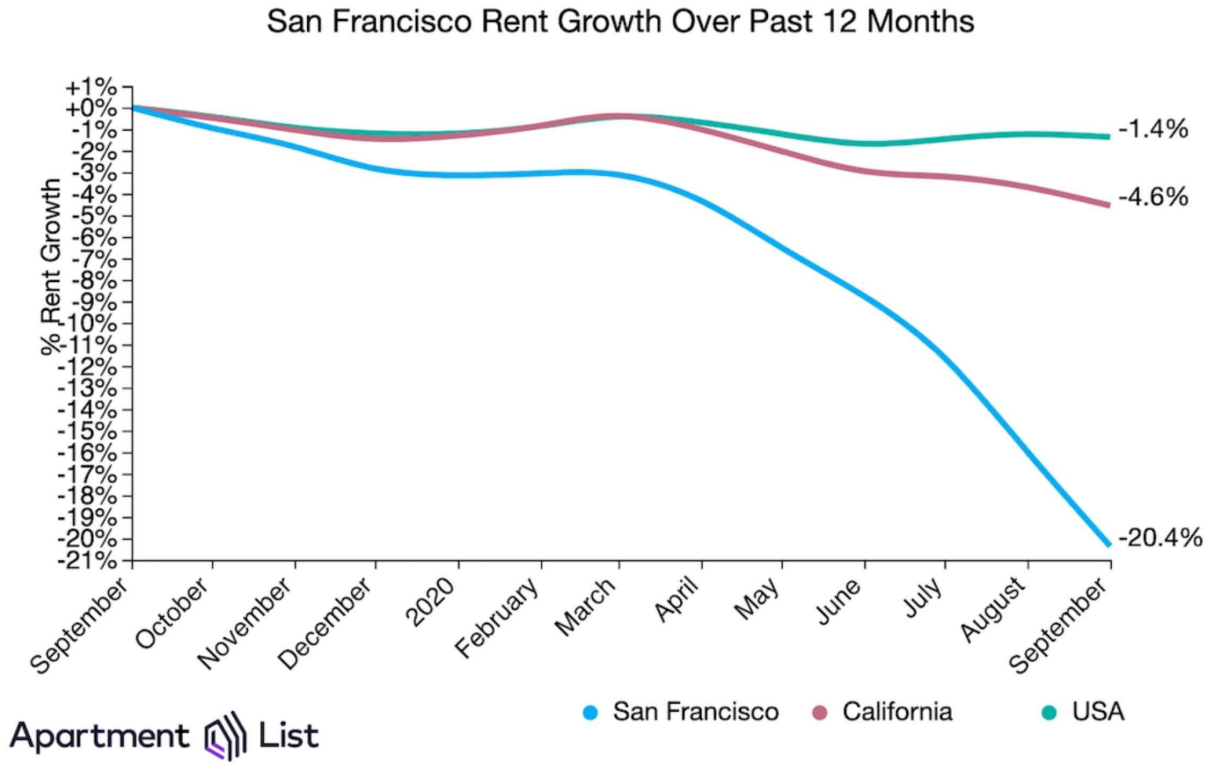
APPENDIX A - Residuals plot



APPENDIX B - Distribution of rental prices



APPENDIX C - Declining SF Rents



Source: SF Chronicle