# Craigslist Bike Price Prediction

Dan Weiss

# Question

Can I predict the prices of bikes for sale on Craigslist based on features present in the listing?

Bike Features Related to Price:

- Quality
- Condition
- Location

# Data

- 3000 most recent bikes listed on Craigslist Bay Area
  - Taken on 4/17/2018
  - Violated Craigslist TOS (sorry, Craig)
- Take all aspects of the listing
  - Price
  - Listing text
  - Listing attributes
  - Number of Images
  - Area

reply ✉ ☎ prohibited [?] Posted about 12 hours ago print

# 2013 Cervelo S5 Cherry Condition/Custom Paint - $1000 (willow glen / cambrian)

image 1 of 6


(google map)

condition: **excellent**

make / manufacturer: **Cervelo**

model name / number: **S5**

size / dimensions: **56cm**

2013 Cervelo S5 56cm aero road bike. Custom metallic blue sparkle paint.

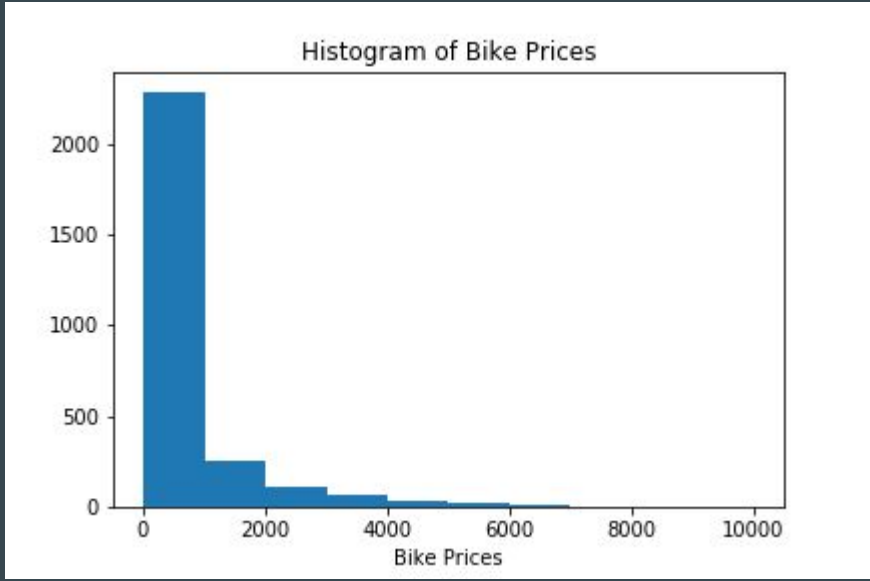
Full Ultegra 6800
3T Ergonova bar
Look Keo Blade Titanium pedals
Fulcrum Racing 5.5 wheels
Ultegra 11-25 cassette
Brand new Fizik Antares saddle
New Elite bottle cages

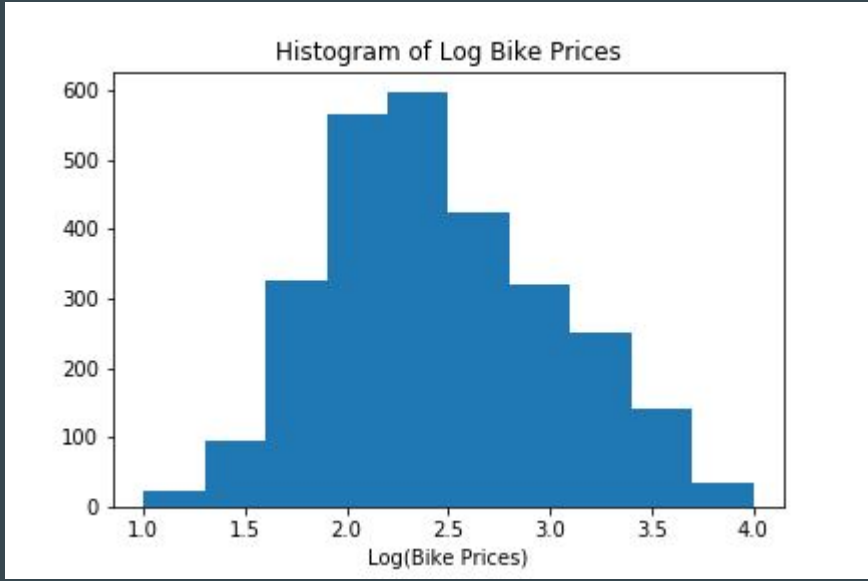
This bike rips. Stiff, super aero and uncompromisingly fast. Gorgeous custom paint by Joe's Carbon Solutions. You'll be the only person in the world with this Cervelo.

# Data Cleaning

- Filter out non- bikes (e.g. Bike racks)
- Search title and post to classify the bike as:
  - Kid's
  - Electric Bike
  - Having professional level, medium level, or entry Level components
- Calculate the length of the post, number of images

# Transform Price Variable



# Model Selection

- Linear Model
  - Log Price = B0 + B*X
- Various combinations of X
- Cross validate on 75% of data (Training Set)
- Chose model with best average $R^2$ over 5- Fold CV
- Overfitting not an issue as $R^2$ on training and test data similar

# Results

- Final Model Includes:
  - Categorical Variables:
    - Condition of the bike (New, Like New, Excellent, Good, Fair, N/A)
    - Location of the bike (North Bay, South Bay, East Bay, Peninsula, San Francisco)
    - Quality of Derailleur (Pro, Medium, Entry, N/A)
    - Kids Bike
    - Electric Bike
  - Length of the Post (in words)
  - Number of Pictures
- $R^2$ on Holdout Data: 0.41
- RMSE on Holdout Data: 0.45 ~ $2.80 (because it's log(price))