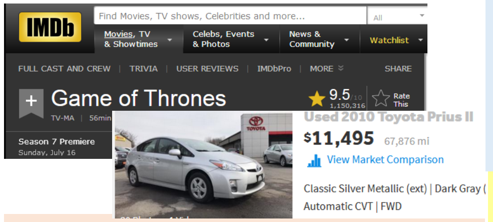
DYOD – Download Your Own Data

Robin Lock, Ivan Ramler, Choong-Soo Lee



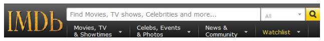
Goal

Provide ways for students to get datasets that are

- Easily downloadable
- Individual and personalized
- Real and readily understood
- Common, predictable structure

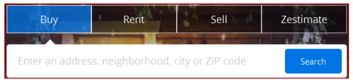
Three Data Sources

- IMDb.com Ratings of TV Series



- Used Car Prices from cars.com

- House prices from zillow.com



Easily Downloadable:

- Requests are made through web apps
- Each dataset obtained by web scraping
- Each app allows local save as .csv file

All three apps are available at :

mystlawu.edu/~clee/datasets

Individual and personalized:

- Students choose their own car model or TV series
- Students choose a zip code from which to sample cars or houses.

Main webpage contains links to zip code locators

Real and readily understood:

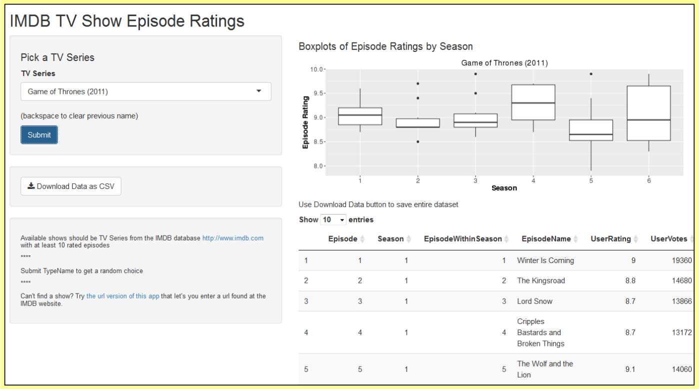
- All data are scraped live from the websites
- TV ratings, car/house prices, characteristics of cars (age, miles) and houses (bedrooms, baths, …) are all familiar to students.

Common, predictable structure:

- All students get data on the same variables
- Relationships can be anticipated, should be similar between datasets, but not identical

IMDb Ratings: *shiny.stlawu.edu:3838/TVSeries* (Shiny App)

Student specifies: Name of a TV series



TV Series: Game of Thrones (2011)

(backspace to clear previous name)

Available shows should be TV Series from the IMDB database http://www.imdb.com with at least 10 rated episodes

****

Submit TypeName to get a random choice

****

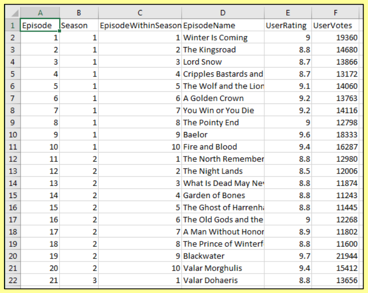
Can't find a show? Try the url version of this app that let's you enter a url found at the IMDB website.

Use Download Data button to save entire dataset

Show 10 entries

	Episode	Season	EpisodeWithinSeason	EpisodeName	UserRating	UserVotes
1	1	1	1	Winter Is Coming	9	19360
2	2	1	2	The Kingsroad	8.8	14680
3	3	1	3	Lord Snow	8.7	13866
4	4	1	4	Cripples Bastards and Broken Things	8.7	13172
5	5	1	5	The Wolf and the Lion	9.1	14060

Output dataset

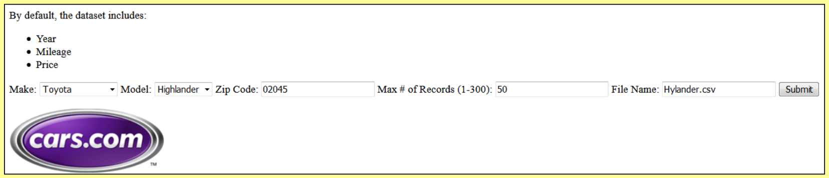


	A	B	C	D	E	F
1	Episode	Season	EpisodeWithinSeason	EpisodeName	UserRating	UserVotes
2	1	1	1	Winter Is Coming	9	19360
3	2	1	2	The Kingsroad	8.8	14680
4	3	1	3	Lord Snow	8.7	13866
5	4	1	4	Cripples Bastards and	8.7	13172
6	5	1	5	The Wolf and the Lion	9.1	14060
7	6	1	6	A Golden Crown	9.2	13763
8	7	1	7	You Win or You Die	9.2	14116
9	8	1	8	The Pointy End	9	12798
10	9	1	9	Baelor	9.6	18333
11	10	1	10	Fire and Blood	9.4	16287
12	11	2	1	The North Remember	8.8	12980
13	12	2	2	The Night Lands	8.5	12006
14	13	2	3	What Is Dead May Nev	8.8	11874
15	14	2	4	Garden of Bones	8.8	11243
16	15	2	5	The Ghost of Harrenha	8.8	11445
17	16	2	6	The Old Gods and the	9	12268
18	17	2	7	A Man Without Honor	8.9	11802
19	18	2	8	The Prince of Winterf	8.8	11600
20	19	2	9	Blackwater	9.7	21944
21	20	2	10	Valar Morghulis	9.4	15412
22	21	3	1	Valar Dohaeris	8.8	13656

Episode	Episode name
Season	Average User Rating
Episode within Season	Number of Raters

Car Prices: *myslu.stlawu.edu/~clee/dataset/cars* (Javascript)

Student specifies: Car make/model, zip code

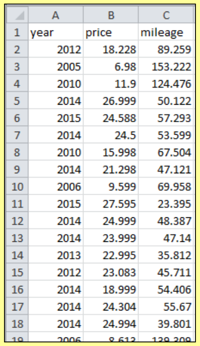


By default, the dataset includes:

- Year
- Mileage
- Price

Make: Toyota Model: Highlander Zip Code: 02045 Max # of Records (1-300): 50 File Name: Hylander.csv Submit

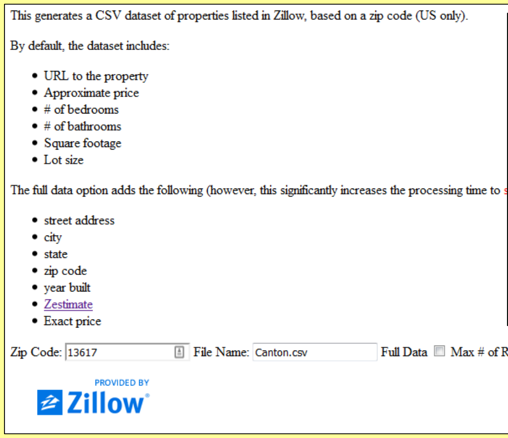
Output dataset



	A	B	C
1	year	price	mileage
2	2012	18.228	89.259
3	2005	6.98	153.222
4	2010	11.9	124.476
5	2014	26.999	50.122
6	2015	24.588	57.293
7	2014	24.5	53.599
8	2010	15.998	67.504
9	2014	21.298	47.121
10	2006	9.599	69.958
11	2015	27.595	23.395
12	2014	24.999	48.387
13	2014	23.999	47.14
14	2013	22.995	35.812
15	2012	23.083	45.711
16	2014	18.999	54.406
17	2014	24.304	55.67
18	2014	24.994	39.801
19	2006	8.613	139.209

House Prices: *myslu.stlawu.edu/~clee/dataset/zillow* (Javascript)

Student specifies: zip code



This generates a CSV dataset of properties listed in Zillow, based on a zip code (US only).

By default, the dataset includes:

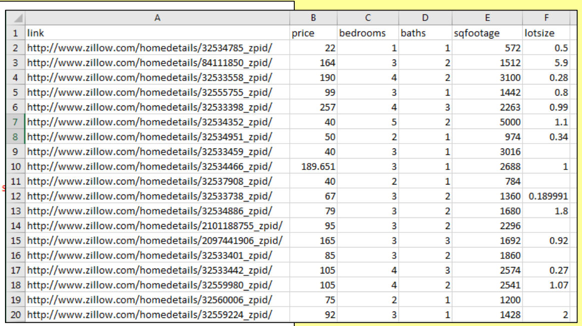
- URL to the property
- Approximate price
- # of bedrooms
- # of bathrooms
- Square footage
- Lot size

The full data option adds the following (however, this significantly increases the processing time to

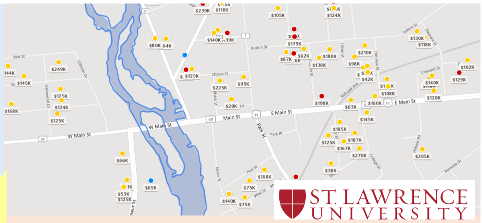
- street address
- city
- state
- zip code
- year built
- Zestimate
- Exact price

Zip Code: 13617 File Name: Canton.csv Full Data □ Max # of Records (1-300): 100 Submit

Output dataset



	A	B	C	D	E	F
1	link	price	bedrooms	baths	sqfootage	lotsize
2	http://www.zillow.com/homedetails/32534785_zpid/	22	1	1	572	0.5
3	http://www.zillow.com/homedetails/84111850_zpid/	164	3	2	1512	5.9
4	http://www.zillow.com/homedetails/32533558_zpid/	190	4	2	3100	0.28
5	http://www.zillow.com/homedetails/32555755_zpid/	99	3	1	1442	0.8
6	http://www.zillow.com/homedetails/32533398_zpid/	257	4	3	2263	0.99
7	http://www.zillow.com/homedetails/32534352_zpid/	40	5	2	5000	1.1
8	http://www.zillow.com/homedetails/32534951_zpid/	50	2	1	974	0.34
9	http://www.zillow.com/homedetails/32533459_zpid/	40	3	1	3016	
10	http://www.zillow.com/homedetails/32534466_zpid/	189.651	3	1	2688	1
11	http://www.zillow.com/homedetails/32537908_zpid/	40	2	1	784	
12	http://www.zillow.com/homedetails/32533738_zpid/	67	3	2	1360	0.189991
13	http://www.zillow.com/homedetails/32534886_zpid/	79	3	2	1680	1.8
14	http://www.zillow.com/homedetails/2101188755_zpid/	95	3	2	2296	
15	http://www.zillow.com/homedetails/2097441906_zpid/	165	3	3	1692	0.92
16	http://www.zillow.com/homedetails/32533401_zpid/	85	3	2	1860	
17	http://www.zillow.com/homedetails/32533442_zpid/	105	4	3	2574	0.27
18	http://www.zillow.com/homedetails/32559980_zpid/	105	4	2	2541	1.07
19	http://www.zillow.com/homedetails/32560006_zpid/	75	2	1	1200	
20	http://www.zillow.com/homedetails/32559224_zpid/	92	3	1	1428	2



Sample Projects/Activities

Stat 1: IMDb Ratings (in-class activity)

1. Pick a favorite TV series
2. Download IMDb ratings dataset
3. Create side-by-side boxplots to compare
   - ratings by season (as in the app)
   - ratings by episode within a season
4. Are there any outlier ratings?
5. Generate summary statistics within seasons (or episodes within seasons)
6. Compare ratings for your series to another student's choice

Stat 2: Car prices (regression project)

1. Pick a car model and zip code to sample
2. Look at a regression model to predict **price** based on **age**.
   - Is there an age at which this car is predicted to be free?
3. Look at residual plots (usually show curvature with faster early depreciation).
4. Try a quadratic regression based on age (often shows an "classic" effect for older cars)
5. Use both **age** and **mileage** as predictors of **price.** (deals with strongly related predictors).

Stat 1: House prices (homework activity)

1. Obtain housing samples from two cities
2. Construct a CI (and/or test) for the difference in mean price between the cities.

Stat 2: House prices (regression)

1. Obtain a sample of houses from a zip code and build models to predict price based on the other variables.

Caveat: You are free (and encouraged!) to use these apps with your students (or to quickly get your own data for class examples/ exams/…),

BUT be aware that there is some risk of one of the sites we are using to scrape the data objecting to many requests coming from our web server and shutting off access.