Final Project

Data Cleaning for Informed Craigslist Car Purchases

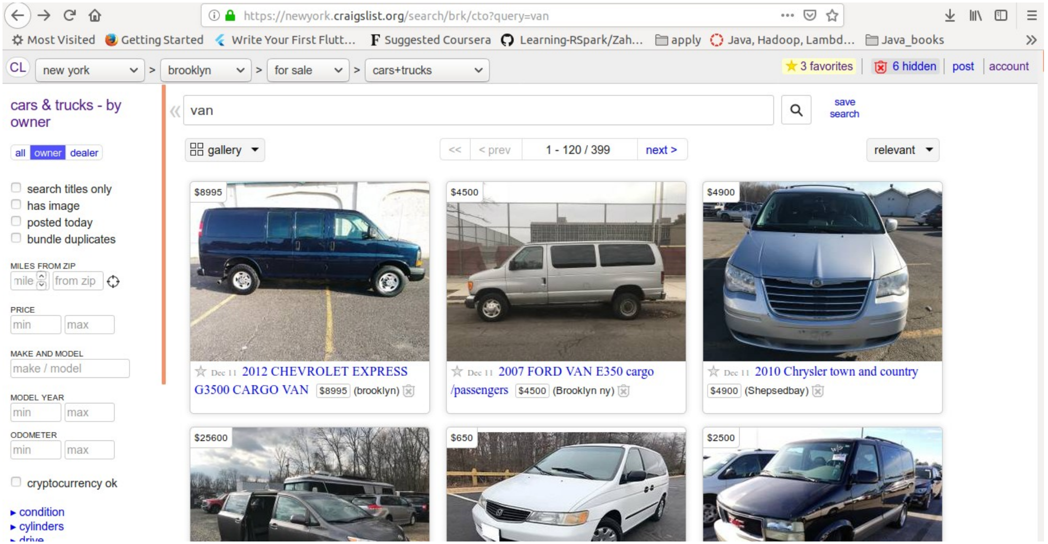
Carolyn Mason

CSCI E-63 Big Data Analytics
Harvard University Extension School
Prof. Zoran B. Djordjević

Problem Statement

Problem: People spend a large amount of time searching Craigslist for deals that best suite them. There are many options that can be difficult to sort through while making timely informed decisions. Sometimes the item will sell and the opportunity will be missed or users pass up a ‘good-deal’ without realizing it. There are tools that can aggregate Craigslist data which can help users sort through the listings and have a better understanding of the best listings out there. This project will take this set of data and use big data tools to make useful insights for user feedback.

Background:
Site: https://www.craigslist.org
Launch date: 1995
HQ: San Francsiso, CA
Founder: Craig Newmark
What: a place to buy, sell, rent, etc.
Users: 60 million users
Monthly page views: 50 billion
Monthly ads poster: 80 million
Vehicle listings: 3000+ in Hartford area



https://newyork.craigslist.org/search/brk/cto?query=van

new york > brooklyn > for sale > cars+trucks

cars & trucks - by owner

van

1 - 120 / 399

$8995 ☆ Dec 11 2012 CHEVROLET EXPRESS G3500 CARGO VAN $8995 (brooklyn)

$4500 ☆ Dec 11 2007 FORD VAN E350 cargo /passengers $4500 (Brooklyn ny)

$4900 ☆ Dec 11 2010 Chrysler town and country $4900 (Shepsedbay)

$25600

$650

$2500

Software and Technology

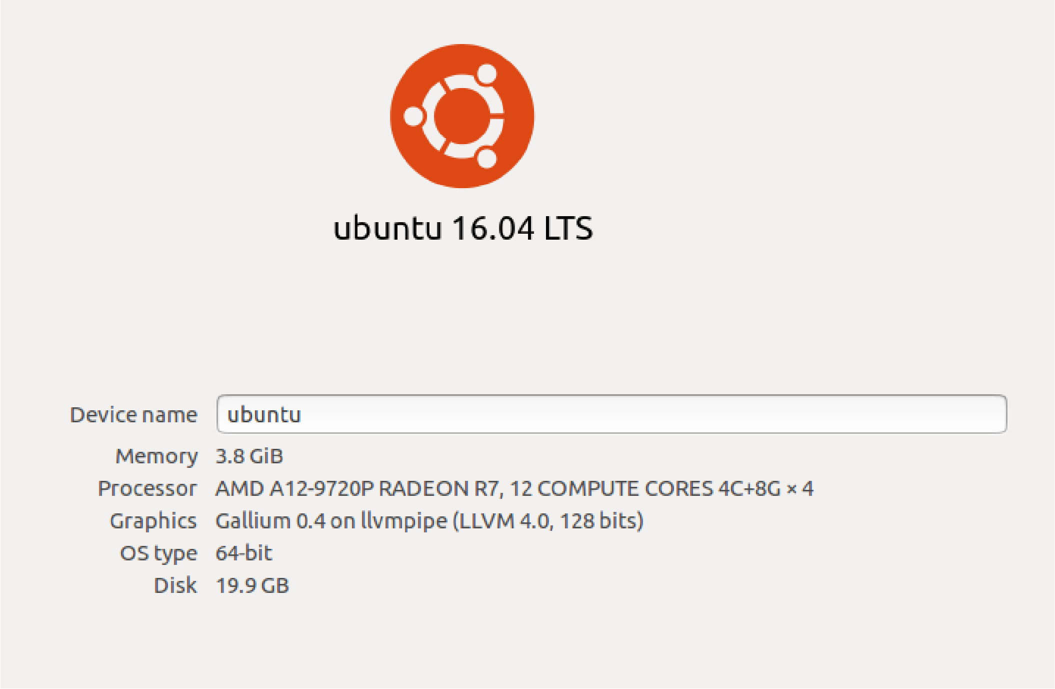
- This project demonstrates all examples using:
    - Python
    - Pyspark

The packages used include:

- Sql
- Scrapy webcrawler
- Matplotlib

My Hardware Environment

- Operating system Ubuntu
- Run on a Linux 16.04 Virtual Machine



ubuntu 16.04 LTS

Device name	ubuntu
Memory	3.8 GiB
Processor	AMD A12-9720P RADEON R7, 12 COMPUTE CORES 4C+8G × 4
Graphics	Gallium 0.4 on llvmpipe (LLVM 4.0, 128 bits)
OS type	64-bit
Disk	19.9 GB

Data Sets

This project takes on two large sets of data. These include fuel economy listings from the US Department of Engery and aggregated Craigslist data. The fuel economy data is joined with the Craigslist data to provide more information for the user. The websites I used are listed below:

Fuel economy data (Datasets for All Model Years (1984-2019)):

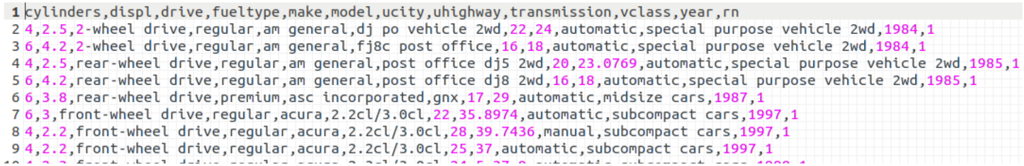
- Website: https://www.fueleconomy.gov/feg/download.shtml
- Raw data set contains a list of 40,692 vehicles with 83 columns
- Cleaned and condensed for: cylinders, displacement, drive, fuel type, make, model, mpg city, mpg highway, transmission, vehicle class, year

Craigslist:

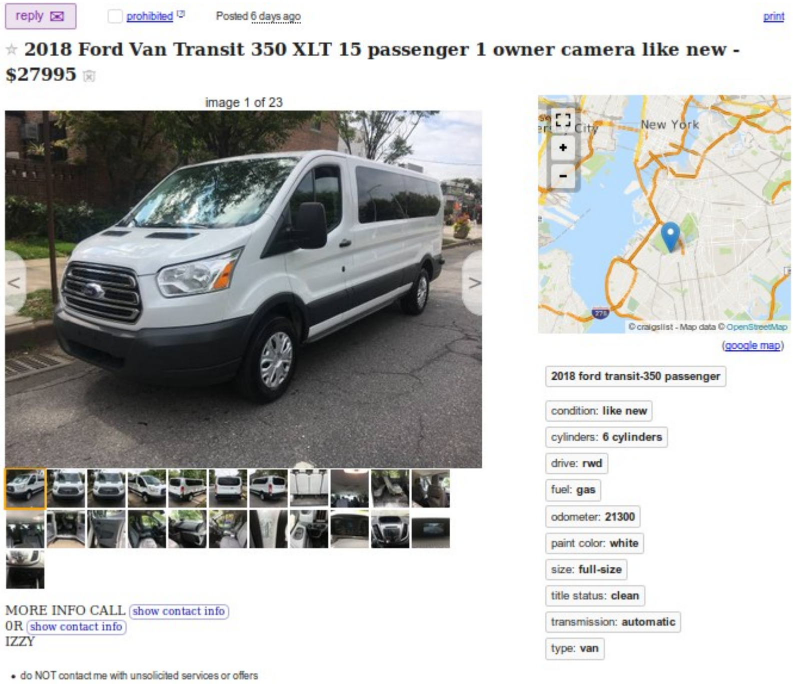
- Website: https://newyork.craigslist.org/search/brk/cto
- Raw data is dependent on the search, generally < 3,000 vehicles
- Column names: title, url, price, address, vin, odometer, condition, cylinders, drive, fuel, paint color, size, title status, transmission, type, year, make, model, description

Overview of Steps

There are three major parts to this tool:
1) Prepare the fuel economy data set
2) Gather and clean Craigslist data
3) Join the data sets for informed learning



```
cylinders,displ,drive,fueltype,make,model,ucity,uhighway,transmission,vclass,year,rn
4,2.5,2-wheel drive,regular,am general,dj po vehicle 2wd,22,24,automatic,special purpose vehicle 2wd,1984,1
6,4.2,2-wheel drive,regular,am general,fj8c post office,16,18,automatic,special purpose vehicle 2wd,1984,1
4,2.5,rear-wheel drive,regular,am general,post office dj5 2wd,20,23.0769,automatic,special purpose vehicle 2wd,1985,1
6,4.2,rear-wheel drive,regular,am general,post office dj8 2wd,16,18,automatic,special purpose vehicle 2wd,1985,1
6,3.8,rear-wheel drive,premium,asc incorporated,gnx,17,29,automatic,midsize cars,1987,1
6,3,front-wheel drive,regular,acura,2.2cl/3.0cl,22,35.8974,automatic,subcompact cars,1997,1
4,2.2,front-wheel drive,regular,acura,2.2cl/3.0cl,28,39.7436,manual,subcompact cars,1997,1
4,2.2,front-wheel drive,regular,acura,2.2cl/3.0cl,25,37,automatic,subcompact cars,1997,1
```



☆ **2018 Ford Van Transit 350 XLT 15 passenger 1 owner camera like new - $27995**

2018 ford transit-350 passenger

condition: **like new**

cylinders: **6 cylinders**

drive: **rwd**

fuel: **gas**

odometer: **21300**

paint color: **white**

size: **full-size**

title status: **clean**

transmission: **automatic**

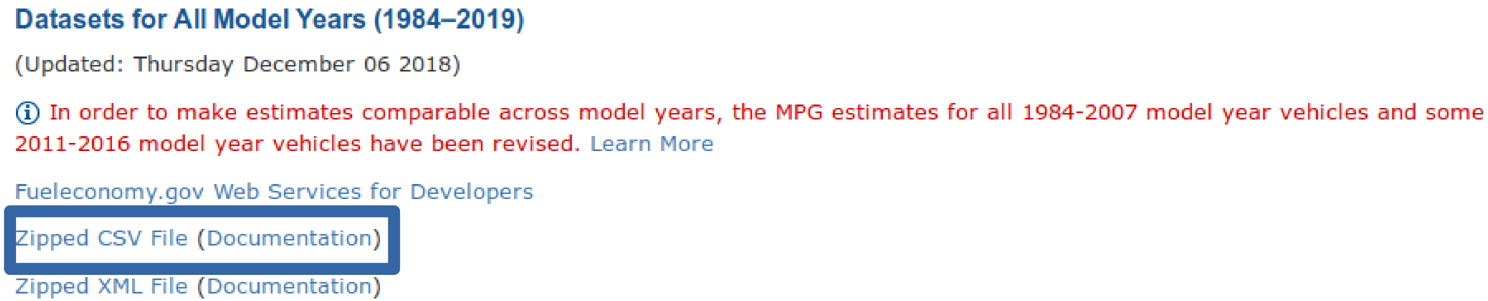
type: **van**

MORE INFO CALL show contact info
0R show contact info
IZZY

- do NOT contact me with unsolicited services or offers

Overview: Fuel Economy Data Set

1) Download data from: https://www.fueleconomy.gov/feg/download.shtml



Datasets for All Model Years (1984–2019)

(Updated: Thursday December 06 2018)

ⓘ In order to make estimates comparable across model years, the MPG estimates for all 1984-2007 model year vehicles and some 2011-2016 model year vehicles have been revised. Learn More

Fueleconomy.gov Web Services for Developers

Zipped CSV File (Documentation)

Zipped XML File (Documentation)

2) Import necessary functions and read in csv

3) Clean data:

a) In order to match how data is stored in the Craigslist data file, I updated the transmission information

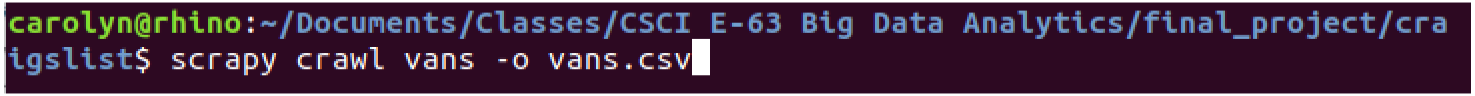
b) Removed duplicates. There were several occurrences of the same make, model, year, transmission, and number of cylinders. These data entries only differed by displacement and/or fuel economy and were not matchable with Craigslist

4) Only keep useful metics: Cleaned and condensed for: cylinders, displacement, drive, fuel type, make, model, mpg city, mpg highway, transmission, vehicle class, year

5) Save csv. By hand change all text to lower case. This helps with joining the datasets.

Overview: Craigslist Data Set

1) Open scraper called vans.py. It has a JobsSpider class and two functions called: parse and parse_page
2) In the JobsSpider class setup the scraper file with a starter url:
start_urls = ["https://newyork.craigslist.org/search/cto?min_price=1000"]
3) Run ‘scrapy crawl vans -o vans.csv’ in the command line.



```
carolyn@rhino:~/Documents/Classes/CSCI E-63 Big Data Analytics/final_project/cra
igslist$ scrapy crawl vans -o vans.csv
```

4) The scraper only keeps good output. The code is setup to find data in the description if the user does not supply it as a tag. Every listing should provide the minimum of make, model, year, transmission, and number of cylinders, else it will be thrown out.
5) The output columns are setup in settings.py. They are set to: title, url, price, address, vin, odometer, condition, cylinders, drive, fuel, paint color, size, title status, transmission, type, year, make, model, description
6) Aggregated data is saved as vans.csv

Joining the Data

1) After the first two data sets (‘fuel_simple.csv’ and ‘vans.csv’) are created, the files can be joined.
2) Import functions, set schemas, and load the two csvs into data frames.
3) Create tables and join on make, model, year, transmission, and cylinders. This join should be good, since there was initial cleaning

```
df_combined = spark.sql("SELECT data.title,data.make,data.model,data.year,data.transmission,fuel.ucity,fuel.uhighway,fuel.vclass,data.size,fuel.drive,fuel.fueltype,data.cylinders,data.fuel,fuel.displ,data.url,data.price,data.odometer,data.condition,data.address,data.vin,data.paint_color,data.title_status,data.type,data.description FROM data LEFT OUTER JOIN fuel ON fuel.make=data.make AND fuel.model=data.model AND fuel.transmission=data.transmission AND fuel.year=data.year AND fuel.cylinders=data.cylinders")
```

4) Run new queries for useful information!

```
df_ans = spark.sql("SELECT make,model,year,price,ucity,uhighway,vclass,condition FROM query WHERE uhighway IS NOT NULL ORDER BY uhighway DESC, price")
```

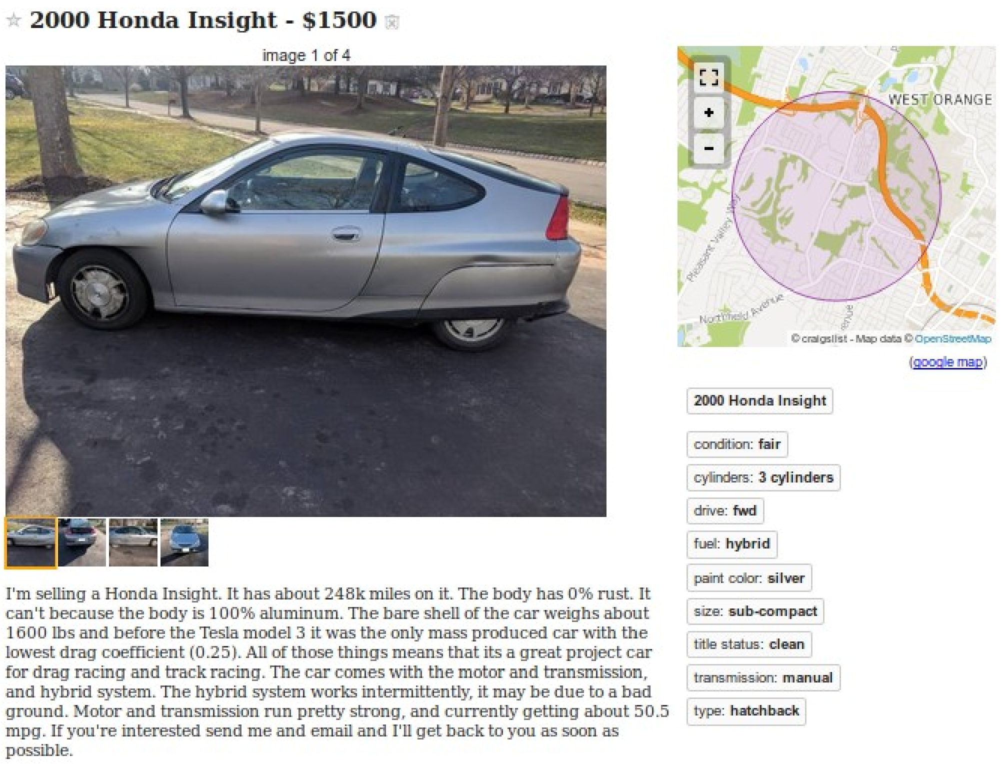
Preliminary Results

```
>>> df_ans.show(20,False)
+----------+-------+----+-------+-------+--------+------------+---------+
|make      |model  |year|price  |ucity  |uhighway|vclass      |condition|
+----------+-------+----+-------+-------+--------+------------+---------+
|honda     |insight|2000|1500.0 |68.1881|89.2029 |two seaters |fair     |
|honda     |insight|2000|2400.0 |68.1881|89.2029 |two seaters |excellent|
|toyota    |prius  |2016|21900.0|76.0467|71.5838 |midsize cars|like new |
|chevrolet |cruze  |2018|9890.0 |40.5   |70.2    |compact cars|like new |
|chevrolet |cruze  |2017|12690.0|40.5   |70.2    |compact cars|excellent|
|lexus     |ct 200h|2015|17900.0|72.0295|69.6895 |compact cars|like new |
|toyota    |prius  |2011|4900.0 |71.8162|69.5514 |midsize cars|good     |
|toyota    |prius  |2012|6499.0 |71.7588|69.5142 |midsize cars|like new |
|toyota    |prius  |2012|8500.0 |71.7588|69.5142 |midsize cars|good     |
|toyota    |prius  |2014|13900.0|71.651 |69.4488 |midsize cars|like new |
|chevrolet |cruze  |2014|6000.0 |34.8   |66.2994 |midsize cars|excellent|
|chevrolet |cruze  |2014|10000.0|34.8   |66.2994 |midsize cars|excellent|
|chevrolet |cruze  |2014|10000.0|34.8   |66.2994 |midsize cars|excellent|
|toyota    |prius  |2006|2200.0 |66.6   |64.8    |midsize cars|good     |
|toyota    |prius  |2005|3990.0 |66.6   |64.8    |midsize cars|like new |
|toyota    |prius  |2009|4200.0 |66.6   |64.8    |midsize cars|good     |
|toyota    |prius  |2008|4800.0 |66.6   |64.8    |midsize cars|like new |
|toyota    |prius  |2009|5500.0 |66.6   |64.8    |midsize cars|good     |
|bmw       |328d   |2014|13999.0|41.5772|64.7919 |compact cars|like new |
|mitsubishi|mirage |2015|4995.0 |49.4465|63.3897 |compact cars|excellent|
+----------+-------+----+-------+-------+--------+------------+---------+
```

Preliminary Results Cont.

```
>>> df_ans = spark.sql("SELECT url FROM query WHERE uhighway IS NOT NULL ORDER BY uhighway DESC, price")>>>
df_ans.show(20,False)
+--------------------------------------------------------------------------------------+
|url                                                                                   |
+--------------------------------------------------------------------------------------+
|https://newyork.craigslist.org/jsy/cto/d/2000-honda-insight/6771193156.html           |
|https://newyork.craigslist.org/lgi/cto/d/2000-honda-insight/6759610035.html           |
|https://phoenix.craigslist.org/evl/cto/d/2016-toyota-prius-touring/6754419444.html    |
|https://phoenix.craigslist.org/nph/cto/d/2018-chevy-cruze-ls/6756759665.html          |
|https://newyork.craigslist.org/stn/cto/d/2017-chevrolet-cruze-lt-great/6761191984.html|
|https://phoenix.craigslist.org/evl/cto/d/2015-lexus-ct200h-warranty/6753137983.html   |
|https://gulfport.craigslist.org/cto/d/2011-toyota-prius/6742823416.html               |
|https://newyork.craigslist.org/brk/cto/d/2012-toyota-prius-iii-leather/6768312410.html|
|https://newyork.craigslist.org/brk/cto/d/2012-prius-for-sale/6752384858.html          |
|https://phoenix.craigslist.org/nph/cto/d/2014-toyota-prius-wagon/6762160316.html      |
|https://newyork.craigslist.org/jsy/cto/d/2014-chevy-cruze68kfree-temp/6770824008.html |
|https://cosprings.craigslist.org/cto/d/2014-chevy-cruze/6769888826.html               |
|https://denver.craigslist.org/cto/d/2014-chevy-cruze/6769889821.html                  |
|https://newyork.craigslist.org/lgi/cto/d/2006-toyota-prius-hybrid-179k/6768891950.html|
|https://newyork.craigslist.org/que/cto/d/2005-toyota-prius-hybrid/6751978060.html     |
|https://newyork.craigslist.org/lgi/cto/d/2009-toyota-prius/6763990165.html            |
|https://newyork.craigslist.org/brk/cto/d/toyota-prius-2008/6770886287.html            |
|https://newyork.craigslist.org/fct/cto/d/2009-toyota-prius/6768673255.html            |
|https://newyork.craigslist.org/brk/cto/d/2014-bmw-328d-xdrive/6765070181.html         |
|https://phoenix.craigslist.org/nph/cto/d/2015-mitsubishi-mirage/6770293393.html       |
+--------------------------------------------------------------------------------------+
```

Preliminary Results Cont.



☆ **2000 Honda Insight - $1500**

image 1 of 4

2000 Honda Insight

condition: **fair**

cylinders: **3 cylinders**

drive: **fwd**

fuel: **hybrid**

paint color: **silver**

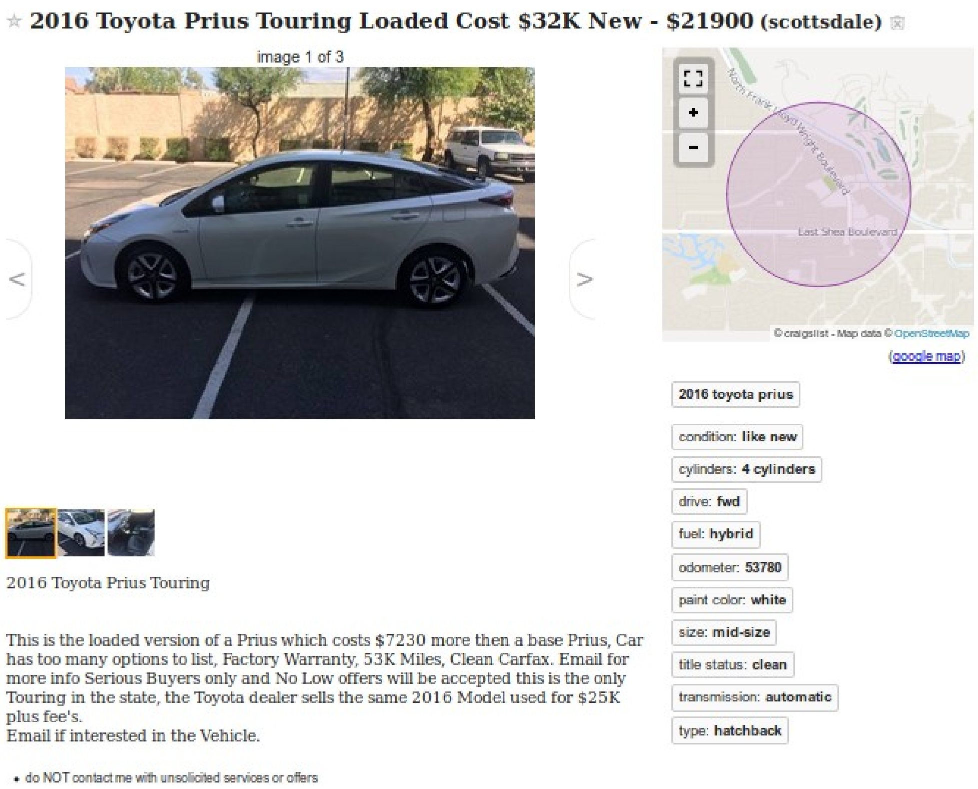
size: **sub-compact**

title status: **clean**

transmission: **manual**

type: **hatchback**

I'm selling a Honda Insight. It has about 248k miles on it. The body has 0% rust. It can't because the body is 100% aluminum. The bare shell of the car weighs about 1600 lbs and before the Tesla model 3 it was the only mass produced car with the lowest drag coefficient (0.25). All of those things means that its a great project car for drag racing and track racing. The car comes with the motor and transmission, and hybrid system. The hybrid system works intermittently, it may be due to a bad ground. Motor and transmission run pretty strong, and currently getting about 50.5 mpg. If you're interested send me and email and I'll get back to you as soon as possible.



☆ **2016 Toyota Prius Touring Loaded Cost $32K New - $21900** (scottsdale)

image 1 of 3

<
>
North Frank Lloyd Wright Boulevard
East Shea Boulevard
© craigslist - Map data © OpenStreetMap
(google map)

2016 toyota prius

condition: **like new**

cylinders: **4 cylinders**

drive: **fwd**

fuel: **hybrid**

odometer: **53780**

paint color: **white**

size: **mid-size**

title status: **clean**

transmission: **automatic**

type: **hatchback**

2016 Toyota Prius Touring

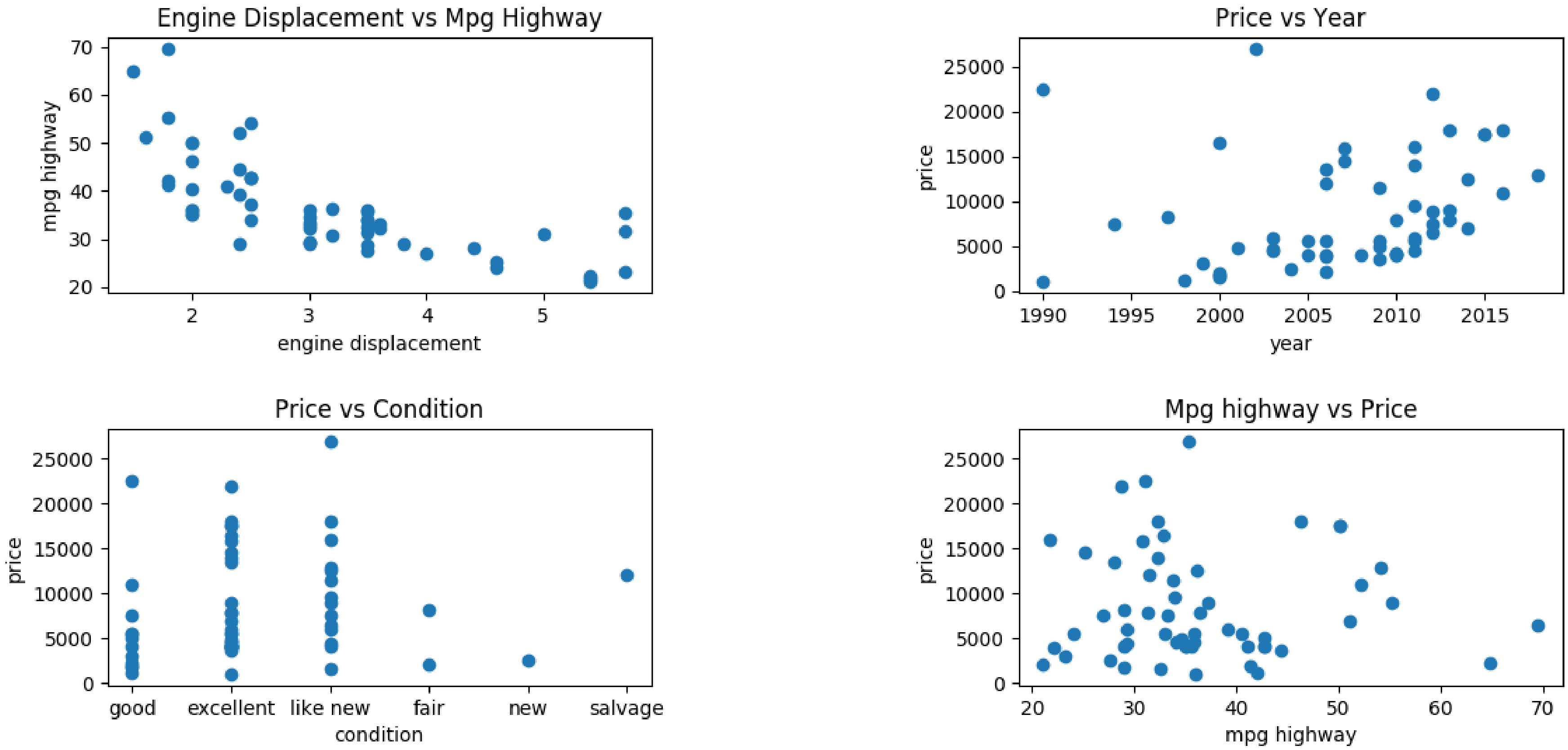
This is the loaded version of a Prius which costs $7230 more then a base Prius, Car has too many options to list, Factory Warranty, 53K Miles, Clean Carfax. Email for more info Serious Buyers only and No Low offers will be accepted this is the only Touring in the state, the Toyota dealer sells the same 2016 Model used for $25K plus fee's.
Email if interested in the Vehicle.

- do NOT contact me with unsolicited services or offers

https://newyork.craigslist.org/jsy/cto/d/2000-honda-insight/6771193156.html

https://phoenix.craigslist.org/evl/cto/d/2016-toyota-prius-touring/6754419444.html

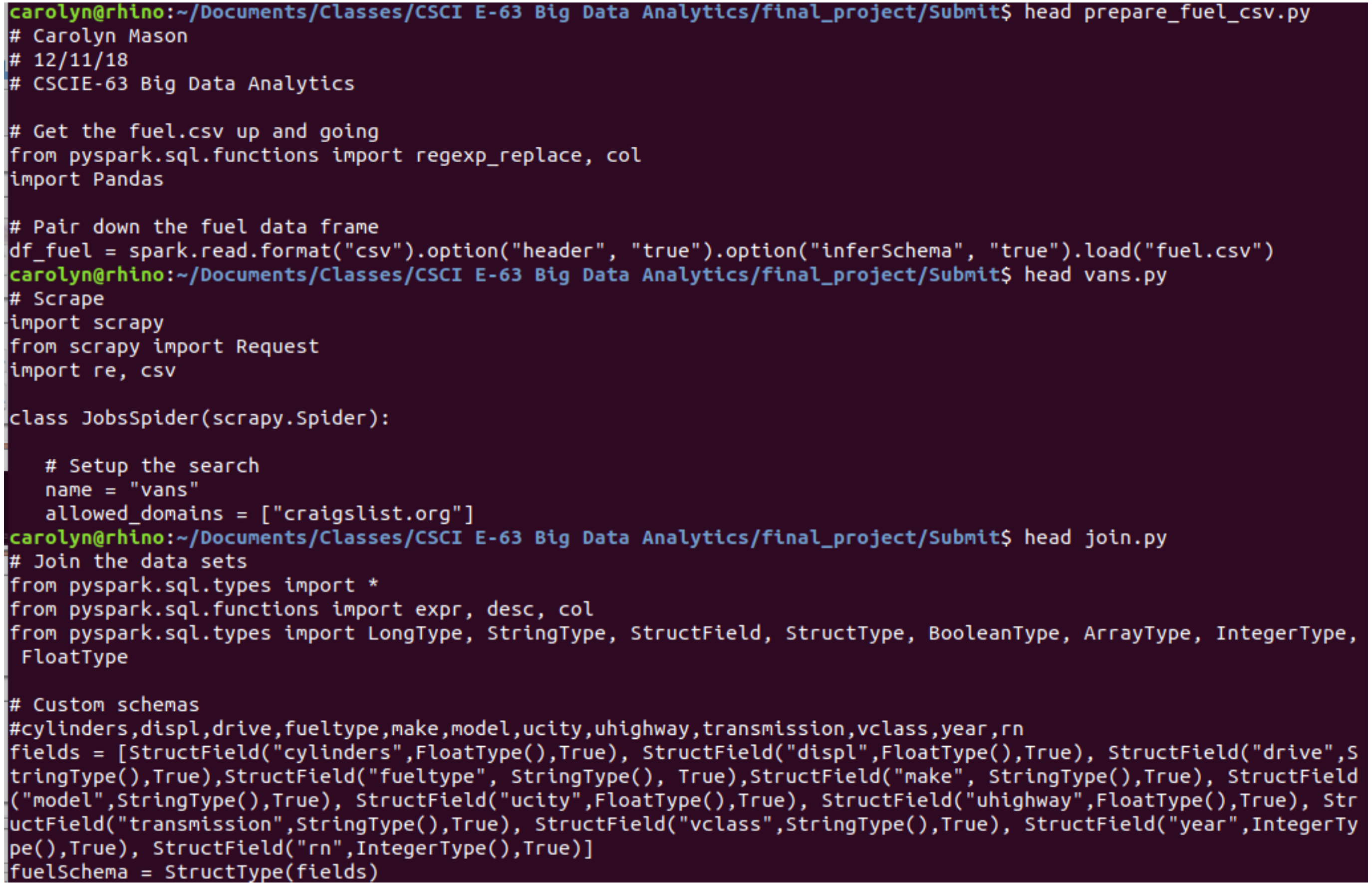
Data Analysis & Visualization



Code Overview

There are three parts to my code as described in the steps above. The files are:

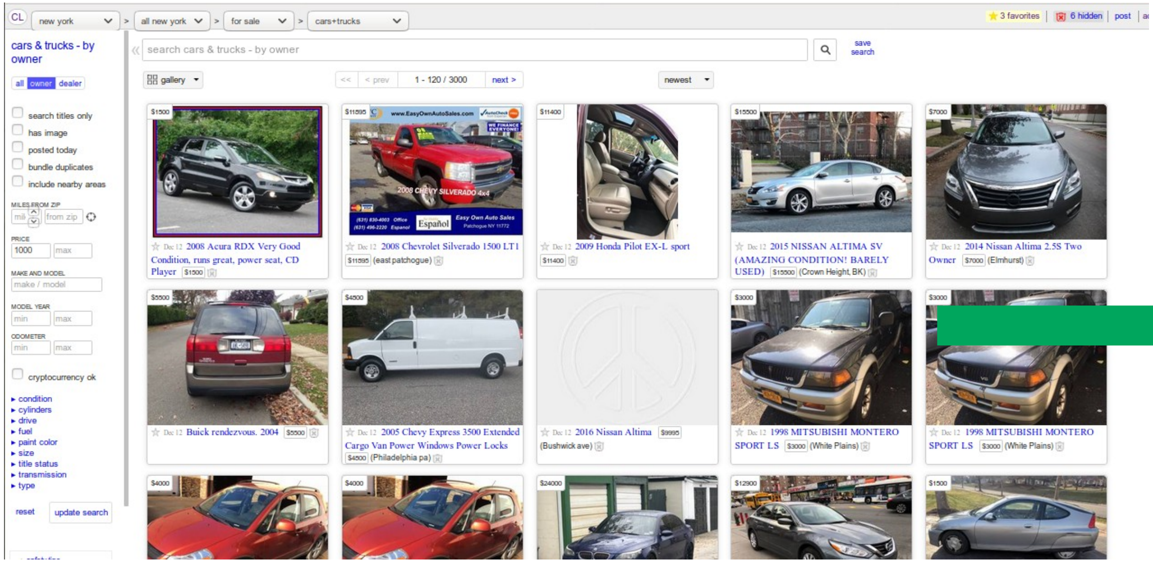
- prepare_fuel_csv.py
- vans.py
- join.py



```
carolyn@rhino:~/Documents/Classes/CSCI E-63 Big Data Analytics/final_project/Submit$ head prepare_fuel_csv.py
# Carolyn Mason
# 12/11/18
# CSCIE-63 Big Data Analytics

# Get the fuel.csv up and going
from pyspark.sql.functions import regexp_replace, col
import Pandas

# Pair down the fuel data frame
df_fuel = spark.read.format("csv").option("header", "true").option("inferSchema", "true").load("fuel.csv")
carolyn@rhino:~/Documents/Classes/CSCI E-63 Big Data Analytics/final_project/Submit$ head vans.py
# Scrape
import scrapy
from scrapy import Request
import re, csv

class JobsSpider(scrapy.Spider):

   # Setup the search
   name = "vans"
   allowed_domains = ["craigslist.org"]
carolyn@rhino:~/Documents/Classes/CSCI E-63 Big Data Analytics/final_project/Submit$ head join.py
# Join the data sets
from pyspark.sql.types import *
from pyspark.sql.functions import expr, desc, col
from pyspark.sql.types import LongType, StringType, StructField, StructType, BooleanType, ArrayType, IntegerType,
 FloatType

# Custom schemas
#cylinders,displ,drive,fueltype,make,model,ucity,uhighway,transmission,vclass,year,rn
fields = [StructField("cylinders",FloatType(),True), StructField("displ",FloatType(),True), StructField("drive",S
tringType(),True),StructField("fueltype", StringType(), True),StructField("make", StringType(),True), StructField
("model",StringType(),True), StructField("ucity",FloatType(),True), StructField("uhighway",FloatType(),True), Str
uctField("transmission",StringType(),True), StructField("vclass",StringType(),True), StructField("year",IntegerTy
pe(),True), StructField("rn",IntegerType(),True)]
fuelSchema = StructType(fields)
```

Final Results

The tool that I created can successfully search Craigslist for the vehicle that best matches their needs. This should decrease search time and make users feel more comfortable with their final decision It is also easy to see two like cars and compare the year, condition, and price. Since the data is user input and the code is not infallible it is not unlikely that errors will occur. More niche searches- like for vans may produce worse matches. Overall the code will produce useful user feedback and help view data trends.



☆ Dec 12 2008 Acura RDX Very Good Condition, runs great, power seat, CD Player $1500

☆ Dec 12 2008 Chevrolet Silverado 1500 LT1 $11595 (east patchogue)

☆ Dec 12 2009 Honda Pilot EX-L sport $11400

☆ Dec 12 2015 NISSAN ALTIMA SV (AMAZING CONDITION! BARELY USED) $15500 (Crown Height, BK)

☆ Dec 12 2014 Nissan Altima 2.5S Two Owner $7000 (Elmhurst)

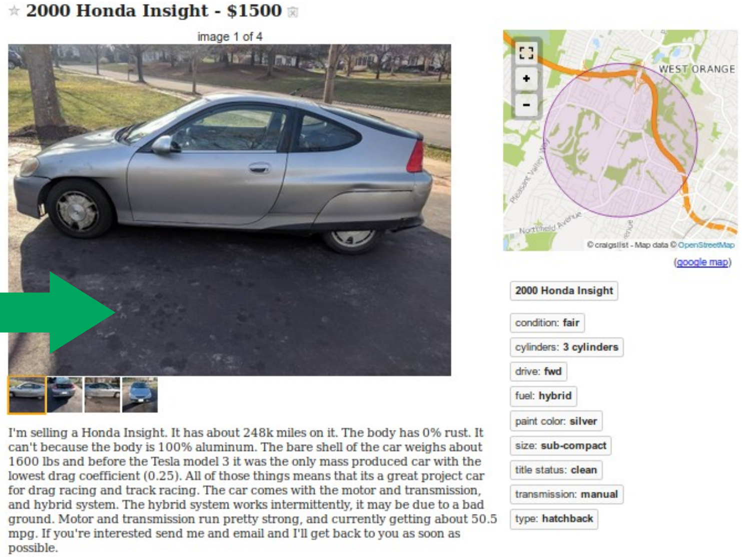
☆ Dec 12 Buick rendezvous. 2004 $5500

☆ Dec 12 2005 Chevy Express 3500 Extended Cargo Van Power Windows Power Locks $4500 (Philadelphia pa)

☆ Dec 12 2016 Nissan Altima $9995 (Bushwick ave)

☆ Dec 12 1998 MITSUBISHI MONTERO SPORT LS $3000 (White Plains)

☆ Dec 12 1998 MITSUBISHI MONTERO SPORT LS $3000 (White Plains)



☆ 2000 Honda Insight - $1500

I'm selling a Honda Insight. It has about 248k miles on it. The body has 0% rust. It can't because the body is 100% aluminum. The bare shell of the car weighs about 1600 lbs and before the Tesla model 3 it was the only mass produced car with the lowest drag coefficient (0.25). All of those things means that its a great project car for drag racing and track racing. The car comes with the motor and transmission, and hybrid system. The hybrid system works intermittently, it may be due to a bad ground. Motor and transmission run pretty strong, and currently getting about 50.5 mpg. If you're interested send me and email and I'll get back to you as soon as possible.

2000 Honda Insight

condition: **fair**

cylinders: **3 cylinders**

drive: **fwd**

fuel: **hybrid**

paint color: **silver**

size: **sub-compact**

title status: **clean**

transmission: **manual**

type: **hatchback**

Lessons Learned

- From this project, I have a greater appreciation of what it takes to have good data. Data that is sourced from user input is rife- with irregularities. If data is missing points or contains bad information it can be difficult to root out and rooting out bad data can come at a cost.

- Running a more irregular search- like for vans can be more difficult. About half the data needed to be thrown out and many matches were wrong. Vans are more unique and may require better code insights and more fuel economy information.

- This project is a success if the user wants Craigslist data matched with fuel economy. The user just has to realize that not all data is good and it is difficult to draw the intended conclusions from the data. I originally hoped to use machine learning to make guesses on un-matched Craigslist data, but feel there is not enough (or consistent) enough information supplied from Craigslist to be useful.

Future Work

- Get more niche searches working! I am interested in using my own code to search for a van. I would like the code to be able to produce good/ consistent results. There are other considerations like van-size that might be useful too.

- It would be interesting to use Craigslist location search and make the same comparisons with vehicles around the United States. Some people might be willing to fly out to a location to see a vehicle they have been looking for.

- Fuel economy data can be downloaded from the website- but it can also be imported to Python. I discovered this a little late in my project and think that would be a good addition. That would keep the fuel economy data up-to-date.

YouTube URLs

2 minute: https://youtu.be/XTy6c-Z5twU
15 minute: https://youtu.be/CLf7mLvR4Hc

References

Zoran’s class notes

Fuel economy data: https://www.fueleconomy.gov/feg/download.shtml

Craigslist: https://newyork.craigslist.org/search/brk/cto

Craigslist Facts: https://expandedramblings.com/index.php/craigslist-statistics/

Initial scrapy setup:
https://python.gotrained.com/scrapy-tutorial-web-scraping-craigslist/