**Data Cleaning for Informed Purchases**

Carolyn Mason

**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.

**Data Sets:**

This project takes on two large sets of data:

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

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

**Featured** **Technology:**

This project demonstrates all examples using Python and Pyspark. The packages include sql, a scrapy webcrawler, and matplotlib.

**High Level Overview of Steps:**

1) Prepare the fuel economy data set

2) Gather and clean Craigslist data

3) Join the data sets for informed learning

**Challenges:**

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 an 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.

**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.

**URLs for YouTube recordings:**

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/>