My project uses craigslist data, and the best part is it works for any category! for my analysis, I will be using cars in 4 various areas of new jersey (north, south, central, and jersey shore). The links are below:

<https://newjersey.craigslist.org/search/cta#search=1~gallery~0~0> north

<https://southjersey.craigslist.org/search/cta#search=1~gallery~0~0> south

<https://cnj.craigslist.org/search/cta#search=1~gallery~0~0> central

<https://jerseyshore.craigslist.org/search/cta#search=1~gallery~0~0> jersey shore

The questions I'll be asking of the car dataset are:

1. What is the distribution of car colors? which are most/least prevalent
2. Is the relationship between price and miles very strong?
3. how many different brands were captured? which was the most/least common?
4. which car brand has the most and least expensive cars on average?
5. what’s the relationship between price and color? which colors are most and least expensive?

To scrape from 5 pages, I wrote a function that given the URL, will scrape the data from each posting on the page. It starts by scraping the given URL for the links to all of the listings on the page. After that, a for loop happens that runs for each URL that we pulled and scrapes the data from each page. The reason we must do this is because we are scraping from the search results page initially, and there isn't sufficient data on that webpage to answer these questions. The program has issues sometimes, because of the large number of requests put into their servers, but if you run the function on one link at a time with a strong vpn on, it does iterate all the way through with no data loss. The function also has built in cleaning to get rid off excessive All NA columns. while the function hasn’t been tested on other links, it should theoretically work on any craigslist search results page

robots.txt:

User-agent: \*

Disallow: /reply

Disallow: /fb/

Disallow: /suggest

Disallow: /flag

Disallow: /mf

Disallow: /mailflag

Disallow: /eaf

Sitemap: https://newjersey.craigslist.org/sitemap/index.xml

Sitemap: https://newjersey.craigslist.org/sitemap/index-by-date-cat.xml