In [1]:
```
import pandas as pd
import numpy as np
import requests
from urllib.request import urlopen
from bs4 import BeautifulSoup
#import get to call a get request on the site
from requests import get
import re
```

Webscraping each city

In [2]:
```
#replace each city unique link here (for example purposes this is Houston,
url = 'https://Houston.craigslist.org/d/motorcycles-scooters/search/mca'
```

Creating all the links to parse through for each city

In [5]:
```
#get the first page of motorcyle listings
response = get(url)
html_soup = BeautifulSoup(response.text, 'html.parser')
#find the total number of posts to find the limit of the pagination
results_num = html_soup.find('div', class_= 'search-legend')
results_total = int(results_num.find('span', class_='totalcount').text) #pu

#each page has 119 posts so each new page is defined as follows: s=120, s=2
pages = np.arange(0, results_total+1, 120)
```

In [6]:
```
#get the links for each pagination
links = []

for page in pages:
    #get request
    links.append(url
                 + "?s=" #the parameter for defining the page number
                 + str(page)) #the page number in the pages array from ea



links
```

Out[6]:
```
['https://Houston.craigslist.org/d/motorcycles-scooters/search/mca?s=0',
 'https://Houston.craigslist.org/d/motorcycles-scooters/search/mca?s=12
0',
 'https://Houston.craigslist.org/d/motorcycles-scooters/search/mca?s=24
0',
 'https://Houston.craigslist.org/d/motorcycles-scooters/search/mca?s=36
0',
 'https://Houston.craigslist.org/d/motorcycles-scooters/search/mca?s=48
0']
```

Scrape each page to get the individual listings on the page with their

price and unique link

In [7]:
```
price_list= []
url_list = []
title_list = []

for link in links:
    response = get(link)
    html_soup = BeautifulSoup(response.text, 'html.parser')
    results = html_soup.find(class_='rows')
    results.prettify()
    bike_elements = results.find_all('li', class_ ='result-row')
    for bike_elem in bike_elements:

        #   print(bike_elem.text)
        price_elem =bike_elem.find('span', class_='result-price')
        try:
            url_elem =bike_elem.find('a', class_='result-image gallery')['h
        except TypeError:
            pass
        title_elem =bike_elem.find('a', class_="result-title hdrlnk")
        title_list.append(title_elem.text.strip())
        url_list.append(url_elem)
        price_list.append(price_elem.text.strip())
    #   print(price_list)
        print(title_elem.text.strip())
        print(url_elem)
        print(price_elem.text.strip())
        print()
```

```
2016 Aprilia SR 50 MT 4T  -Finance it with Instant Credit Approval!
https://houston.craigslist.org/mcd/d/las-vegas-2016-aprilia-sr-50-mt-4t/7
396616386.html (https://houston.craigslist.org/mcd/d/las-vegas-2016-april
ia-sr-50-mt-4t/7396616386.html)
$1,795

Kawasaki 440 LTD
https://houston.craigslist.org/mcy/d/richmond-kawasaki-440-ltd/739660801
3.html (https://houston.craigslist.org/mcy/d/richmond-kawasaki-440-ltd/73
96608013.html)
$3,000

Rmz 450
https://houston.craigslist.org/mcy/d/spring-rmz-450/7396556117.html (http
s://houston.craigslist.org/mcy/d/spring-rmz-450/7396556117.html)
$4,500

2009 Kawasaki ZX6R
https://houston.craigslist.org/mcy/d/missouri-city-2009-kawasaki-zx6r/739
```

With the individual links that were appended to the list, scrape each link for their attributes

In [8]:
```
bike_title_2= []
attribute_list =[]

for i, url in enumerate(url_list):
    bike_url = url
    bike_page = requests.get(bike_url)
    bike_soup =BeautifulSoup(bike_page.content, 'html.parser')
    attributes = bike_soup.find_all('p', class_='attrgroup')
#     for attribute in attributes:
    if attributes:
        bike_title_2.append(attributes[0].text.strip())
        attribute_list.append((attributes[1].text.strip()))
    else:
        bike_title_2.append('none')
        attribute_list.append('none')
        print(i, url)
```

```
502 https://houston.craigslist.org/mcy/d/brazoria-buy-bikes/7383680696.ht
ml (https://houston.craigslist.org/mcy/d/brazoria-buy-bikes/7383680696.ht
ml)
```

Making sure all the lists are equal length

In [9]:
```
print(len(bike_title_2))
print(len(attribute_list))
print(len(price_list))
print(len(title_list))
print(len(url_list))
```

```
546
546
546
546
546
```

Concatanating them to a single DataFrame

In [10]:
```
df = pd.DataFrame(list(zip(price_list,url_list,title_list,bike_title_2, att
df
```

Out[10]:

	Price	URL	title	bikeTitle	attributes
0	$1,795	https://houston.craigslist.org/mcd/d/las-vegas...	2016 Aprilia SR 50 MT 4T - Finance it with Ins...	2016 Aprilia SR 50 MT 4T	fuel: gas\n\nodometer: 96\n\ntitle status: cle...
1	$3,000	https://houston.craigslist.org/mcy/d/richmond-...	Kawasaki 440 LTD	1980 KZ 440 LTD	condition: excellent\n\nengine displacement (C...
2	$4,500	https://houston.craigslist.org/mcy/d/spring-rm...	Rmz 450	suzuki rmz 450	fuel: gas\n\nodometer: 21\n\ntransmission: manua
3	$4,495	https://houston.craigslist.org/mcy/d/missouri-...	2009 Kawasaki ZX6R	2009 kawasaki ninja zx-6r	condition: good\n\ncryptocurrency: ok\n\nengine...
4	$12,000	https://houston.craigslist.org/mcy/d/humble-20...	2008 Harley Davidson custom Street Glide	2008 harley davidson	condition: excellent\n\nfuel: gas\n\nodometer:...
...	...	...	...	...	...
541	$5,500	https://houston.craigslist.org/mcy/d/houston-h...	Harley Davidson	2011 HD XL883N	condition: like new\n\nfuel: gas\n\nodometer: ...
542	$11,500	https://houston.craigslist.org/mcy/d/cleburne-...	Must see beautiful 2011 Harley Davidson custom...	2011 harley davidson street glide	condition: excellent\n\nfuel: gas\n\nodometer:...
543	$4,500	https://houston.craigslist.org/mcy/d/tomball-2...	2001 Honda VTX 1800	2001 honda vtx 1800	fuel: gas\n\nodometer: 39096\n\npaint color: b...
544	$1,900	https://houston.craigslist.org/mcy/d/sugar-lan...	Rare 1985 Honda Gyro S	honda	fuel: gas\n\nodometer: 800\n\ntransmission: au...
545	$15,000	https://houston.craigslist.org/mcy/d/richmond-...	Can am spyder	can am spyder	condition: like new\n\nfuel: gas\n\nodometer: ...

546 rows × 5 columns

Giving each attributes its own column

Attribute column cleaning to start

In [11]:
```
#Splitting attributes to make it readable
for row in range(len(df)):
    attributes = df['attributes'][row].split(', ')[1:]
    for attribute in attributes:
        sub_attributes = attribute.split(': ')
        try:
            df.at[row,sub_attributes[0]]= sub_attributes[1]
        except:
            pass
```

In [12]:
```
pd.set_option('display.max_columns',None)
```

In [14]:
```
for i in range(len(attribute_list)):
    df['attributes'][i] = df['attributes'][i].strip()
    df['attributes'][i] = df['attributes'][i].replace('\n', '')
    df['attributes'][i] = df['attributes'][i].replace('VIN', ', VIN')
    df['attributes'][i] = df['attributes'][i].replace('condition', ', condi
    df['attributes'][i] = df['attributes'][i].replace('fuel', ', fuel')
    df['attributes'][i] = df['attributes'][i].replace('paint', ', paint')
    df['attributes'][i] = df['attributes'][i].replace('title', ', title')
    df['attributes'][i] = df['attributes'][i].replace('engine', ', engine')
    df['attributes'][i] = df['attributes'][i].replace('odometer', ', odomet
    df['attributes'][i] = df['attributes'][i].replace('delivery', ', delive
    df['attributes'][i] = df['attributes'][i].replace('transmission', ', tr
    df['attributes'][i] = df['attributes'][i].replace('type', ', type')
```

Each attribute gets its own column

In [16]:
```
for row in range(len(df)):
    attributes = df['attributes'][row].split(', ')[1:]
    for attribute in attributes:
        sub_attributes = attribute.split(': ')
        try:
            df.at[row,sub_attributes[0]]= sub_attributes[1]
        except:
            pass
```

In [17]:
```
df
```

Out[17]:

	Price	URL	title	bikeTitle	attributes	fu
0	$1,795	https://houston.craigslist.org/mcd/d/las-vegas...	2016 Aprilia SR 50 MT 4T - Finance it with Ins...	2016 Aprilia SR 50 MT 4T	, fuel: gas, odometer: 96, title status: clean...	g
1	$3,000	https://houston.craigslist.org/mcy/d/richmond-...	Kawasaki 440 LTD	1980 KZ 440 LTD	, condition: excellent, engine displacement (C...	g
2	$4,500	https://houston.craigslist.org/mcy/d/spring-rm...	Rmz 450	suzuki rmz 450	, fuel: gas, odometer: 21, transmission: manual	g
3	$4,495	https://houston.craigslist.org/mcy/d/missouri-...	2009 Kawasaki ZX6R	2009 kawasaki ninja zx-6r	, condition: goodcryptocurrency ok, engine dis...	g
4	$12,000	https://houston.craigslist.org/mcy/d/humble-20...	2008 Harley Davidson custom Street Glide	2008 harley davidson	, condition: excellent, fuel: gas, odometer: 1...	g
...	...	...	...	...	...	
541	$5,500	https://houston.craigslist.org/mcy/d/houston-h...	Harley Davidson	2011 HD XL883N	, condition: like new, fuel: gas, odometer: 29...	g
542	$11,500	https://houston.craigslist.org/mcy/d/cleburne-...	Must see beautiful 2011 Harley Davidson custom...	2011 harley davidson street glide	, condition: excellent, fuel: gas, odometer: 1...	g
543	$4,500	https://houston.craigslist.org/mcy/d/tomball-2...	2001 Honda VTX 1800	2001 honda vtx 1800	, fuel: gas, odometer: 39096, paint color: bla...	g
544	$1,900	https://houston.craigslist.org/mcy/d/sugar-lan...	Rare 1985 Honda Gyro S	honda	, fuel: gas, odometer: 800, transmission: auto...	g
545	$15,000	https://houston.craigslist.org/mcy/d/richmond-...	Can am spyder	can am spyder	, condition: like new, fuel: gas, odometer: 10...	g

546 rows × 14 columns

Getting year from the bikeTitle

In [18]:
```
#getting year from BikeTitle in case it doesnt have it in the title
for row in range(len(df)):
    df.at[row, 'years'] = df['bikeTitle'][row][:4]
```

Getting year from the title to compare

In [19]:
```
import re

def regex_year(string):
    try:
        return re.match(r"[\d]+", string).group()
    except:
        return re.match(r"[\d]+", string)
```

In [20]:
```
df['year'] = list(map(regex_year, df['title']))
```

Then pickle it to use the data frame to concat with the other cities

In [142]:
```
df.to_pickle('houston_motorcycle')
```

In [143]:
```
df_pickle=pd.read_pickle('houston_motorcycle')
df_pickle
```

Out[143]:

	Price	URL	title	bikeTitle	attribut
0	$0	https://houston.craigslist.org/mcd/d/alvin-199...	1992 Harley Davidson Dyna Lowrider Dayton	1992 Harley Davidson Dyna Lowrider Dayton	, V 1HD1GAL10NY30546 fuel: gas, odometer
1	$3,700	https://houston.craigslist.org/mcy/d/houston-2...	2019 Honda Grom	2019 honda grom	, condition: goo engine displaceme (CC):
2	$23,199	https://houston.craigslist.org/mcd/d/south-hou...	2016 Harley-Davidson FLHXS - Street Glide Spec...	2016 Harley-Davidson FLHXS - Street G	, V 1HD1KRM14GB68577 fuel: , odometer: 0
3	$0	https://houston.craigslist.org/mcd/d/alvin-201...	2012 Dodge Avenger Sedan	2012 Dodge Avenger	, V 1C3CD2AB9CN11124 fuel: other, odomete
4	$0	https://houston.craigslist.org/mcd/d/alvin-200...	2007 Victory Jackpot	2007 Victory Jackpot	, V 5VPXB26D37300013 fuel: gas, odometer
...	...	...	...	...	
523	$27,560	https://houston.craigslist.org/mcd/d/south-hou...	2018 Can-Am Spyder RT Limited Chrome LTD In-li...	2018 Can-Am Spyder RT Limited Chrome	, V 2BXNBDD25JV00150 fuel: , odometer: 86
524	$2,950	https://houston.craigslist.org/mcy/d/spring-20...	2004 Honda Shadow	2007 honda shadow	, condition: excelle fuel: gas, odometer: 2
525	$15,000	https://houston.craigslist.org/mcy/d/magnolia-...	2008 Honda Goldwing 1800 -Only 1,350 miles	2008 honda goldwing gl1800	, condition: like ne fuel: gas, odomet 13
526	$3,900	https://houston.craigslist.org/mcy/d/spring-20...	2007 Honda CRF450X	2007 honda crf450x	, condition: excelle engine displaceme (C
527	$6,500	https://houston.craigslist.org/mcy/d/richmond-...	2010 HONDA STATELINE	2010 honda stateline	, fuel: gas, odomet 428, paint color: blacl

528 rows × 16 columns

In [ ]: