Assignment 2

Jiarong Ye

October 15, 2018

Note: I use python to scrape car info from the dealer site and connect to Redis with its python wrapper for this assignment*

Import packages

```
In [192]: import redis
          from urllib.request import urlopen
          import pandas as pd
          import re
          from bs4 import BeautifulSoup
          import json
```

Scrape car retail information from a chosen car dealer website

Titles

```
In [3]: url = 'http://www.statecollege.com/auto/dealers/stocker-chevrolet,9803/'
        page = urlopen(url=url)
        soup = BeautifulSoup(page, 'html.parser')
        titles = soup.findAll('strong', {"class": "title"})
        titles_df = pd.DataFrame(list(map(lambda x: x.getText(), titles)), columns=['Title'])
```

Embedded urls

```
In [4]: urls_df = pd.DataFrame(['http://www.statecollege.com/'+ a['href']
                               for i in soup.findAll('td', {'align': 'left'})
                               for a in i.findAll('a', href=True)], columns=['urls'])
```

Addresses, phones of sellers & descriptions of cars

```
In [5]: addresses = []
        phones = []
        descriptions = []
        for url in urls_df.values:
            page = urlopen(url=url[0])
            soup = BeautifulSoup(page, 'html.parser')
            for address in soup.findAll('p', {'class': 'business_address_address'}):
                addresses.append(re.sub('\s+', ' ', address.getText().replace('\n', '')))
```

```
            for phone in soup.find('p', {'class': 'business_address_phone'}):
                phones.append(phone)
            for i, c in enumerate(soup.findAll('p')):
                if i==5:
                    descriptions.append(c.getText())
        address_df = pd.DataFrame(addresses, columns=['address'])
        phones_df = pd.DataFrame(phones, columns=['phone'])
        descriptions_df = pd.DataFrame(descriptions, columns=['description'])
```

Other attributes

```
In [6]: tables = []
        for url in urls_df.values:
            page = urlopen(url=url[0])
            soup = BeautifulSoup(page, 'html.parser')
            for table in soup.findAll('table', {'class': 'auto_detail_data_table'}):
                table = pd.read_html(str(table))
                tables.append(table)
```

```
In [29]: attributes = ['Year', 'Mileage', 'Make', 'Model', 'Trim', 'Style',
                       'Engine', 'Exterior Color', 'Interior Color', 'VIN', 'Stock #']
         attributes_dict = {}
         for i in attributes:
             attributes_dict[i] = []
         for table in tables:
             table = table[0]
             for attribute in attributes:
                 attributes_dict[attribute].append(table[table.iloc[:, 0] ==
                                                        attribute+':'].iloc[:, 1].values[0])
         cars = pd.concat([titles_df, pd.DataFrame(attributes_dict), descriptions_df], axis=1)
```

Combine all attributes

```
In [30]: cars
```

```
Out[30]:                          Title  Year Mileage       Make           Model  \
         0            2014 Subaru Outback  2014   75655     Subaru         Outback
         1       2013 Subaru XV Crosstrek  2013   69331     Subaru    XV Crosstrek
         2           2014 Chevrolet Cruze  2014   61147  Chevrolet           Cruze
         3          2015 Chevrolet Malibu  2015   25757  Chevrolet          Malibu
         4  2015 Chevrolet Silverado 1500  2015   41869  Chevrolet  Silverado 1500
         5  2016 Chevrolet Malibu Limited  2016   52582  Chevrolet  Malibu Limited
         6           2016 Subaru Forester  2016   16994     Subaru        Forester
         7  2015 Chevrolet Silverado 1500  2015   49503  Chevrolet  Silverado 1500
         8             2017 Subaru Legacy  2017       2     Subaru          Legacy
         9            2017 Subaru Outback  2017       3     Subaru         Outback

                   Trim            Style Engine        Exterior Color  \
         0  2.5i Premium    Sport Utility    4 -  Twilight Blue Metall
```

```
1       Premium    Station Wagon    4 -  Tangerine Orange Pea
2            LS          4dr Car    4 -  Atlantis Blue Metall
3            LS          4dr Car    4 -   Ashen Gray Metallic
4  High Country  Crew Cab Pickup    8 -    Deep Ruby Metallic
5            LS          4dr Car    4 -  Champagne Silver Met
6         2.5i    Sport Utility    4 -   Ice Silver Metallic
7            LT  Crew Cab Pickup    8 -          Summit White
8       Limited          4dr Car    6 -     Tungsten Metallic
9       Limited    Sport Utility    4 -  Twilight Blue Metall

          Interior Color                VIN  Stock #  \
0                  Black  4S4BRBCC4E3219562  606614A
1                  Black  JF2GPACCXD1843307  204842B
2  Jet Black/Medium Tit  1G1PA5SG4E7343131  204275A
3    Jet Black/Titanium  1G11B5SL2FF313357  204597A
4                Saddle  3GCUKTEC9FG401065  204778A
5    Jet Black/Titanium  1G11B5SAXGF135777   15498A
6                  Black  JF2SJAAC3GG451169  606499A
7              Jet Black  3GCUKREC9FG459437  204882A
8            Warm Ivory  4S3BNEN6XH3029939   604992
9           Slate Black  4S4BSANC0H3391975   605565

                                          description
0  CARFAX One-Owner. Clean CARFAX. Blue 2014 Suba...
1  Clean CARFAX. Orange 2013 Subaru XV Crosstrek ...
2  CARFAX One-Owner. Clean CARFAX. Blue 2014 Chev...
3  CARFAX One-Owner. Clean CARFAX. Grey 2015 Chev...
4  CARFAX One-Owner. Clean CARFAX. Burgundy 2015 ...
5  CARFAX One-Owner. Gold 2016 Chevrolet Malibu L...
6  CARFAX One-Owner. Clean CARFAX. Ice Silver Met...
7  CARFAX One-Owner. Clean CARFAX. White 2015 Che...
8  $750 off MSRP!We at Stocker Chevrolet apprecia...
9  We at Stocker Chevrolet appreciate your time a...
```

In [32]:
```
sellers = pd.concat([address_df, phones_df], axis=1)
sellers
```

Out[32]:
```
                                    address           phone
0   701 Benner Pike State College PA, 16801   (866) 235-0270
1   701 Benner Pike State College PA, 16801   (866) 235-0270
2   701 Benner Pike State College PA, 16801   (866) 235-0270
3   701 Benner Pike State College PA, 16801   (866) 235-0270
4   701 Benner Pike State College PA, 16801   (866) 235-0270
5   701 Benner Pike State College PA, 16801   (866) 235-0270
6   701 Benner Pike State College PA, 16801   (866) 235-0270
7   701 Benner Pike State College PA, 16801   (866) 235-0270
8   701 Benner Pike State College PA, 16801   (866) 235-0270
9   701 Benner Pike State College PA, 16801   (866) 235-0270
```

Insert data into Redis

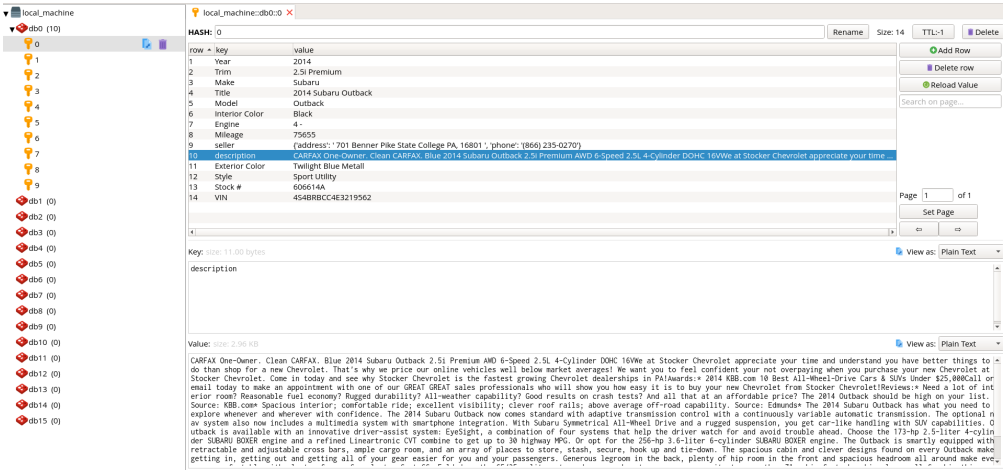
```
In [191]: r = redis.StrictRedis(host='localhost', port=6379, db=0)
          for idx in range(len(cars)):
              v = dict(zip(cars.columns, cars.values[idx]))
              serialized = json.dumps(dict(zip(sellers.columns,sellers.values[idx])))
              v['seller'] = serialized
              r.hmset(idx, v)
```

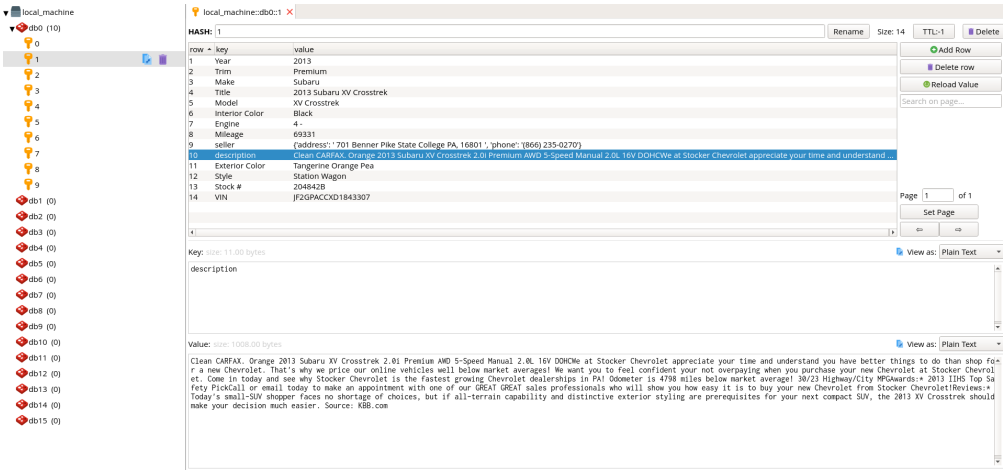
Data model design

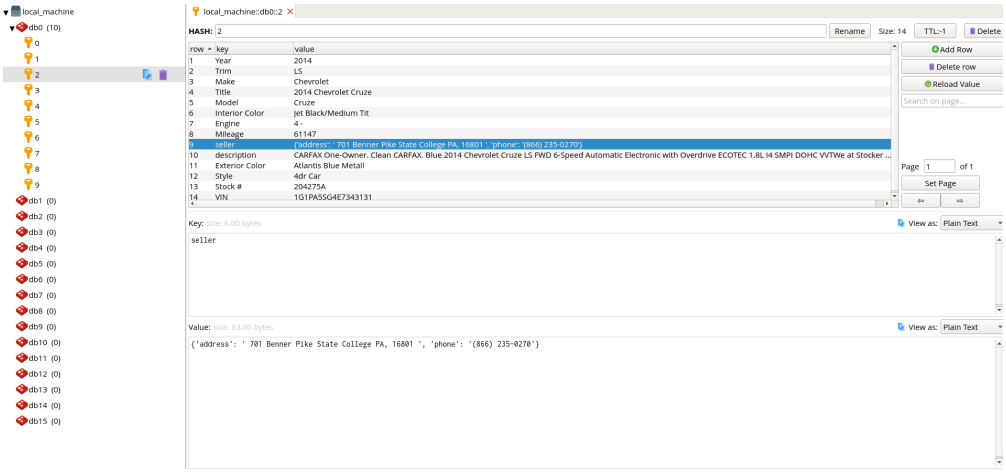
- Attributes
  - Title
  - Year
  - Engine
  - Exterior Color
  - Interior Color
  - Make
  - Mileage
  - Model
  - Stock
  - Style
  - Trim
  - VIN
  - Description
  - Seller
    - address
    - phone

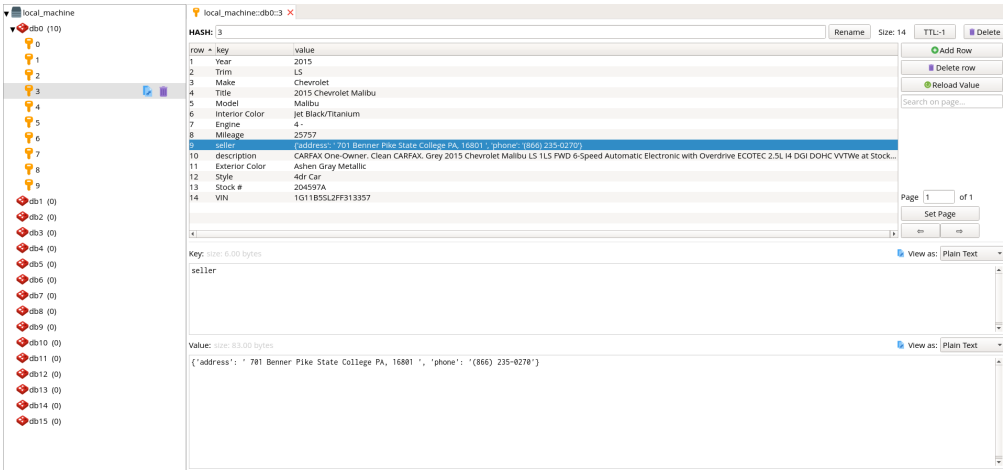
The data model covers all basic attributes of a car that could be found on the retail website and that buyers should be aware of.

Display the result









Filter contents examples

```
In [214]: for i in range(10):
              # find 'Subaru' cars
              if 'Subaru' in r.hget(name=i, key=b'Title').decode('utf-8'):
                  print(r.hget(name=i, key=b'Title'))
```

```
b'2014 Subaru Outback'
b'2013 Subaru XV Crosstrek'
b'2016 Subaru Forester'
b'2017 Subaru Legacy'
b'2017 Subaru Outback'
```

```
In [197]: for i in range(10):
              # find cars with Malibu model:
```

```
            if 'Malibu' in r.hget(name=i, key=b'Model').decode('utf-8'):
                print(r.hmget(name=i, keys=[b'Title', b'Model']))
```

```
[b'2015 Chevrolet Malibu', b'Malibu']
[b'2016 Chevrolet Malibu Limited', b'Malibu Limited']
```

```
In [195]: for i in range(10):
            # find VIN numbers ends with 7:
            if r.hget(name=i, key=b'VIN').decode('utf-8').endswith('7'):
                print(r.hmget(name=i, keys=[b'Title', b'VIN']))
```

```
[b'2013 Subaru XV Crosstrek', b'JF2GPACCXD1843307']
[b'2015 Chevrolet Malibu', b'1G11B5SL2FF313357']
[b'2016 Chevrolet Malibu Limited', b'1G11B5SAXGF135777']
[b'2015 Chevrolet Silverado 1500', b'3GCUKREC9FG459437']
```

```
In [212]: for i in range(3):
            # get seller info
            seller = json.loads(r.hget(name=i, key=b'seller').decode('utf-8'))
            seller_address = seller['address']
            seller_phone = seller['phone']
            print('The seller lives in {}, phone number is {}'
                  .format(seller_address, seller_phone))
```

```
The seller lives in  701 Benner Pike State College PA, 16801 , phone number is (866) 235-0270
The seller lives in  701 Benner Pike State College PA, 16801 , phone number is (866) 235-0270
The seller lives in  701 Benner Pike State College PA, 16801 , phone number is (866) 235-0270
```