DA0101EN-Review-Introduction

August 13, 2019

```
< a\ href="https://cocl.us/DA0101EN_edx_link_Notebook_link_top"> \\ < img\ src="https://s3-api.us-geo.objectstorage.softlayer.net/cf-courses-data/CognitiveClass/DA0101EN/Images </a> <math display="block">< /a>
```

Data Analysis with Python

Introduction

Welcome!

In this section, you will learn how to approach data acquisition in various ways, and obtain necessary insights from a dataset. By the end of this lab, you will successfully load the data into Jupyter Notebook, and gain some fundamental insights via Pandas Library.

Table of Contents

Data Acquisition

Basic Insight of Dataset

Estimated Time Needed: 10 min

Data Acquisition

There are various formats for a dataset, .csv, .json, .xlsx etc. The dataset can be stored in different places, on your local machine or sometimes online. In this section, you will learn how to load a dataset into our Jupyter Notebook. In our case, the Automobile Dataset is an online source, and it is in CSV (comma separated value) format. Let's use this dataset as an example to practice data reading.

data source: https://archive.ics.uci.edu/ml/machine-learning-databases/autos/imports-85.data

data type: csv

The Pandas Library is a useful tool that enables us to read various datasets into a data frame; our Jupyter notebook platforms have a built-in Pandas Library so that all we need to do is import Pandas without installing.

```
[]: # import pandas library import pandas as pd
```

Read Data

We use pandas.read_csv() function to read the csv file. In the bracket, we put the file path along with a quotation mark, so that pandas will read the file into a data frame from that address. The file path can be either an URL or your local file address. Because the data does not include headers, we can add an argument headers = None inside the read_csv() method, so that pandas will not automatically set the first row as a header. You can also assign the dataset to any variable you create.

This dataset was hosted on IBM Cloud object click HERE for free storage.

```
[3]: # Import pandas library
    import pandas as pd
    # Read the online file by the URL provides above, and assign it to variable "df"
    other path = "https://s3-api.us-geo.objectstorage.softlayer.net/cf-courses-data/CognitiveClass/
     →DA0101EN/auto.csv"
    df = pd.read csv(other path, header=None)
```

After reading the dataset, we can use the dataframe.head(n) method to check the top n rows of the dataframe; where n is an integer. Contrary to dataframe.head(n), dataframe.tail(n) will show you the bottom n rows of the dataframe.

```
[4]: # show the first 5 rows using dataframe.head() method
    print("The first 5 rows of the dataframe")
    df.head(5)
```

The first 5 rows of the dataframe

```
[4]:
           ? alfa-romero gas std
                                  two convertible rwd front 88.6 ...
             alfa-romero gas std
                                  two convertible rwd front 88.6 ...
    2
           ? alfa-romero gas std
                                        hatchback rwd front 94.5 ...
                                  two
    3
       2 164
                                          sedan fwd front 99.8 ...
                   audi gas std four
       2 164
                                          sedan 4wd front 99.4 ...
                   audi gas std four
       16
           17
                18
                     19
                          20 21
                                   22 23 24
                                                25
    0 130 mpfi 3.47 2.68
                           9.0 111 5000 21 27 13495
    1 130 mpfi 3.47 2.68
                           9.0 111 5000 21 27 16500
    2 152 mpfi 2.68 3.47
                           9.0 154 5000 19 26 16500
    3 109 mpfi 3.19 3.40 10.0 102 5500 24 30 13950
    4 136 mpfi 3.19 3.40 8.0 115 5500 18 22 17450
    [5 \text{ rows x } 26 \text{ columns}]
```

Question #1:

check the bottom 10 rows of data frame "df".

```
[5]: # Write your code below and press Shift+Enter to execute
    print("The last 10 rows of the dataframe")
    df.tail(10)
```

The last 10 rows of the dataframe

```
[5]:
       0
           1
                                        7
                                             8
                                                  9
                                                    ... 16 \
    195 - 1
           74 volvo
                       gas
                             std four wagon rwd front 104.3 ... 141
    196 -2 103 volvo
                             std four sedan rwd front 104.3 ... 141
                        gas
    197 -1 74 volvo
                             std four wagon rwd front 104.3 ... 141
    198 -2 103 volvo
                        gas turbo four sedan rwd front 104.3 ... 130
    199 -1 74 volvo
                       gas turbo four wagon rwd front 104.3 ... 130
```

```
200 -1 95 volvo
                         std four sedan rwd front 109.1 ... 141
201 - 1
        95 volvo
                    gas turbo four sedan rwd front 109.1 ... 141
202 - 1
        95 volvo
                    gas
                         std four sedan rwd front 109.1 ... 173
203 - 1
        95 volvo diesel turbo four sedan rwd front 109.1 ... 145
204 - 1
        95 volvo
                    gas turbo four sedan rwd front 109.1 ... 141
    17
              19
         18
                   20
                       21
                            22 \ 23 \ 24
                                         25
195 mpfi 3.78 3.15
                    9.5 114 5400 23 28 13415
196 mpfi 3.78 3.15
                    9.5 114 5400 24 28 15985
                    9.5 114 5400 24 28 16515
197 mpfi 3.78 3.15
198 mpfi 3.62 3.15
                    7.5 \ 162 \ 5100 \ 17 \ 22 \ 18420
199 mpfi 3.62 3.15
                    7.5 \ 162 \ 5100 \ 17 \ 22 \ 18950
200 mpfi 3.78 3.15
                    9.5 114 5400 23 28 16845
201 mpfi 3.78 3.15
                    8.7 160 5300 19 25 19045
202 mpfi 3.58 2.87 8.8 134 5500 18 23 21485
203 idi 3.01 3.40 23.0 106 4800 26 27 22470
204 mpfi 3.78 3.15 9.5 114 5400 19 25 22625
```

[10 rows x 26 columns]

Question #1 Answer:

Run the code below for the solution!

Double-click here for the solution.

Add Headers

Take a look at our dataset; pandas automatically set the header by an integer from 0.

To better describe our data we can introduce a header, this information is available at: https://archive.ics.uci.edu/ml/datasets/Automobile

Thus, we have to add headers manually.

Firstly, we create a list "headers" that include all column names in order. Then, we use dataframe.columns = headers to replace the headers by the list we created.

```
[6]: # create headers list
headers = ["symboling","normalized-losses","make","fuel-type","aspiration",

→"num-of-doors","body-style",

"drive-wheels","engine-location","wheel-base",

→"length","width","height","curb-weight","engine-type",

"num-of-cylinders",

→"engine-size","fuel-system","bore","stroke","compression-ratio","horsepower",

"peak-rpm","city-mpg","highway-mpg","price"]
print("headers\n", headers)
```

headers

['symboling', 'normalized-losses', 'make', 'fuel-type', 'aspiration', 'num-of-doors', 'body-style', 'drive-wheels', 'engine-location', 'wheel-base', 'length', 'width', 'height', 'curb-weight', 'engine-type', 'num-of-cylinders', 'engine-size', 'fuel-system', 'bore', 'stroke', 'compression-ratio', 'horsepower', 'peak-rpm', 'city-mpg', 'highway-mpg', 'price']

We replace headers and recheck our data frame

```
[7]: df.columns = headers
     df.head(10)
[7]:
       symboling normalized-losses
                                               make fuel-type aspiration num-of-doors \
               3
                               ? alfa-romero
                                                               \operatorname{std}
                                                                           two
                                                     gas
               3
                               ? alfa-romero
     1
                                                               \operatorname{std}
                                                    gas
                                                                           two
     2
               1
                               ? alfa-romero
                                                               \operatorname{std}
                                                    gas
                                                                           two
     3
               2
                             164
                                        audi
                                                              \operatorname{std}
                                                                         four
                                                    gas
               2
     4
                             164
                                        audi
                                                                         four
                                                              \operatorname{std}
                                                    gas
     5
               2
                               ?
                                        audi
                                                   gas
                                                             \operatorname{std}
                                                                         two
     6
               1
                             158
                                        audi
                                                              \operatorname{std}
                                                                         four
                                                   gas
     7
               1
                               ?
                                        audi
                                                                        four
                                                             \operatorname{std}
                                                   gas
     8
               1
                             158
                                        audi
                                                             turbo
                                                                          four
                                                   gas
               0
     9
                               ?
                                        audi
                                                           turbo
                                                                          two
                                                   gas
        body-style drive-wheels engine-location wheel-base ...
                                                                         engine-size \
     0 convertible
                             rwd
                                          front
                                                       88.6 ...
                                                                          130
     1
        convertible
                             rwd
                                          front
                                                       88.6 ...
                                                                         130
                                                       94.5 \dots
     2
          hatchback
                             rwd
                                           front
                                                                          152
     3
                                                      99.8 ...
             sedan
                            fwd
                                         front
                                                                        109
     4
             sedan
                            4 wd
                                          front
                                                      99.4 ...
                                                                         136
                            \operatorname{fwd}
                                                      99.8 ...
     5
             sedan
                                         front
                                                                        136
     6
                            fwd
                                         front
                                                     105.8 ...
             sedan
                                                                         136
     7
                                                      105.8 ...
             wagon
                            fwd
                                          front
                                                                         136
     8
             sedan
                            fwd
                                         front
                                                     105.8 \dots
                                                                         131
          hatchback
                                                        99.5 ...
     9
                              4 wd
                                            front
                                                                          131
       fuel-system bore stroke compression-ratio horsepower peak-rpm city-mpg \
     0
              mpfi 3.47
                                                                 5000
                                                                             21
                             2.68
                                              9.0
                                                         111
                                                                             21
     1
              mpfi 3.47
                             2.68
                                              9.0
                                                         111
                                                                 5000
     2
              mpfi 2.68
                             3.47
                                              9.0
                                                         154
                                                                 5000
                                                                             19
     3
              mpfi 3.19
                             3.40
                                             10.0
                                                         102
                                                                  5500
                                                                             24
     4
              mpfi 3.19
                                              8.0
                                                                             18
                             3.40
                                                         115
                                                                 5500
     5
              mpfi 3.19
                             3.40
                                              8.5
                                                        110
                                                                 5500
                                                                             19
     6
              mpfi 3.19
                             3.40
                                              8.5
                                                        110
                                                                 5500
                                                                             19
     7
              mpfi 3.19
                             3.40
                                              8.5
                                                                             19
                                                         110
                                                                 5500
     8
                                              8.3
                                                                             17
              mpfi 3.13
                             3.40
                                                        140
                                                                 5500
     9
                                              7.0
              mpfi 3.13
                             3.40
                                                         160
                                                                 5500
                                                                             16
      highway-mpg price
     0
               27 13495
               27 16500
     1
     2
               26 16500
     3
               30 13950
     4
               22 17450
     5
               25 15250
```

6

25 17710

```
7 25 18920
8 20 23875
9 22 ?
```

204

141

mpfi 3.78

3.15

[10 rows x 26 columns]

we can drop missing values along the column "price" as follows

[8]: df.dropna(subset=["price"], axis=0) [8]: symboling normalized-losses make fuel-type aspiration \ ? alfa-romero 0 3 gas std 1 3 alfa-romero std gas 2 1 alfa-romero gas std 3 2 164 audi gas std 4 2 164 audi std gas . . . 200 95 -1 volvo gas std 201 -1 95 turbo volvo gas 95 -1 202 volvo gas std 203 -1 95 volvo turbo diesel 204 -1 95 volvo turbo gas num-of-doors body-style drive-wheels engine-location wheel-base ... \ 0 88.6 ... two convertible rwd front front 88.6 ... 1 convertible rwd two 2 hatchback front $94.5 \dots$ two rwd 3 four sedanfwd front 99.8 ... 4 four sedan front $99.4 \dots$ 4wd . . . 200 four 109.1 ... sedan rwd front 201 109.1 ... four sedan rwd front 202 109.1 ... four sedan rwd front sedan 203 four front 109.1 ... rwd 204 four sedan rwd front109.1 ... engine-size fuel-system bore stroke compression-ratio horsepower \ 0 130 mpfi 3.47 2.689.0 111 1 130 mpfi 3.47 2.68 9.0 111 2 152 mpfi 2.68 3.47 9.0 154 3 109 mpfi 3.19 3.40 10.0 102 mpfi 3.19 4 136 8.0 3.40 115 . . . 200 mpfi 3.78 9.5141 3.15 114 201 141 mpfi 3.78 3.15 8.7 160 202 173 mpfi 3.58 2.87 8.8 134 203 3.40145 idi 3.0123.0106

9.5

114