Python - Analiza danych z modułem PANDAS

www.udemy.com (http://www.udemy.com) (R)

LAB - S03-L008-LAB Kontrola i optymalizacja typów danych

1. Zaimportuj moduł pandas oraz numpy i nadaj im standardowe aliasy. Do zmiennej **fuel** wczytaj zawartość pliku **fuel.csv**. Podczas wczytywania skorzystaj z dodatkowego argumentu **low_memory=False**, pobierz tylko następujące kolumny: 'Vehicle ID','Year','Make','Model','Class','Fuel Type','Combined MPG (FT1)'. Wyświetl nagłówek tak utworzonego Data Frame
2. Wyświetl informację o kolumnach (wykorzystywane typy, ilość obiektów nienullowych, **dokładna** informacja o pamięci)
3. Skonwertuj kolumnę **Year** do typu int. Ponownie wyświetl informację o typach kolumn i użyciu pamięci
4. Sprawdź ile razy powtarzają się wartości w kolumnie **Make**
5. Zmień typ kolumny **Make** na **category** i ponownie wyświetl informację o pamięci konsumowanej przez ten typ
6. Powtórz kroki 4-5 dla kolumn: **Model, Class, 'Fuel Type', 'Combined MPG (FT1)'**
7. Porównaj rozmiar obiektu **fuel** na początku (krok 2) i na końcu (krok 6) - czy różnica jest wyraźna?

Rozwiązania:

Poniżej znajdują się propozycje rozwiązań zadań. Prawdopodobnie istnieje wiele dobrych rozwiązań, dlatego jeżeli rozwiązujesz zadania samodzielnie, to najprawdopodobniej zrobisz to inaczej, może nawet lepiej :) Możesz pochwalić się swoimi rozwiązaniami w sekcji Q&A

In [1]:
```
import pandas as pd
import numpy as np
fuel = pd.read_csv("fuel.csv",
                   usecols=['Vehicle ID','Year','Make',
                   'Model','Class','Fuel Type',
                   'Combined MPG (FT1)'],
                   index_col = 'Vehicle ID')
fuel.head()
```

Out[1]:

Vehicle ID	Year	Make	Model	Class	Fuel Type	Combined MPG (FT1)
26587	1984	Alfa Romeo	GT V6 2.5	Minicompact Cars	Regular	20.0
27705	1984	Alfa Romeo	GT V6 2.5	Minicompact Cars	Regular	20.0
26561	1984	Alfa Romeo	Spider Veloce 2000	Two Seaters	Regular	21.0
27681	1984	Alfa Romeo	Spider Veloce 2000	Two Seaters	Regular	21.0
27550	1984	AM General	DJ Po Vehicle 2WD	Special Purpose Vehicle 2WD	Regular	17.0

In [2]:
```
fuel.info(memory_usage='deep')
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 31684 entries, 26587 to 32106
Data columns (total 6 columns):
Year                  31684 non-null int64
Make                  31684 non-null object
Model                 31683 non-null object
Class                 31683 non-null object
Fuel Type             31683 non-null object
Combined MPG (FT1)    31683 non-null float64
dtypes: float64(1), int64(1), object(4)
memory usage: 8.9 MB
```

In [3]:
```
fuel['Year'] = fuel['Year'].astype('int')
fuel.info(memory_usage='deep')
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 31684 entries, 26587 to 32106
Data columns (total 6 columns):
Year                  31684 non-null int32
Make                  31684 non-null object
Model                 31683 non-null object
Class                 31683 non-null object
Fuel Type             31683 non-null object
Combined MPG (FT1)    31683 non-null float64
dtypes: float64(1), int32(1), object(4)
memory usage: 8.8 MB
```

In [4]:
```
fuel['Make'].value_counts().head()
```

Out[4]:
```
Chevrolet    3389
Ford         2721
Dodge        2361
GMC          2174
Toyota       1599
Name: Make, dtype: int64
```

In [5]:
```
fuel['Make'] = fuel['Make'].astype('category')
fuel.info(memory_usage='deep')
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 31684 entries, 26587 to 32106
Data columns (total 6 columns):
Year                  31684 non-null int32
Make                  31684 non-null category
Model                 31683 non-null object
Class                 31683 non-null object
Fuel Type             31683 non-null object
Combined MPG (FT1)    31683 non-null float64
dtypes: category(1), float64(1), int32(1), object(3)
memory usage: 6.9 MB
```

In [6]:
```
fuel['Model'].value_counts().head()
```

Out[6]:
```
F150 Pickup 2WD      194
Truck 2WD            187
F150 Pickup 4WD      172
Ranger Pickup 2WD    169
Mustang              156
Name: Model, dtype: int64
```

In [7]:
```
fuel['Model'] = fuel['Model'].astype('category')
fuel.info(memory_usage='deep')
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 31684 entries, 26587 to 32106
Data columns (total 6 columns):
Year                  31684 non-null int32
Make                  31684 non-null category
Model                 31683 non-null category
Class                 31683 non-null object
Fuel Type             31683 non-null object
Combined MPG (FT1)    31683 non-null float64
dtypes: category(2), float64(1), int32(1), object(2)
memory usage: 5.2 MB
```

In [8]:
```
fuel['Class'].value_counts().head()
```

Out[8]:
```
Compact Cars                   4489
Subcompact Cars                4299
Midsize Cars                   3393
Standard Pickup Trucks         2354
Sport Utility Vehicle - 4WD    2034
Name: Class, dtype: int64
```

In [9]:
```
fuel['Class'] = fuel['Class'].astype('category')
fuel.info(memory_usage='deep')
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 31684 entries, 26587 to 32106
Data columns (total 6 columns):
Year                  31684 non-null int32
Make                  31684 non-null category
Model                 31683 non-null category
Class                 31683 non-null category
Fuel Type             31683 non-null object
Combined MPG (FT1)    31683 non-null float64
dtypes: category(3), float64(1), int32(1), object(1)
memory usage: 3.0 MB
```

In [10]:
```
fuel['Fuel Type'].value_counts().head()
```

Out[10]:
```
Regular           22439
Premium            7375
Diesel              936
Gasoline or E85     767
CNG                  55
Name: Fuel Type, dtype: int64
```

In [11]:
```
fuel['Fuel Type'] = fuel['Fuel Type'].astype('category')
fuel.info(memory_usage='deep')
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 31684 entries, 26587 to 32106
Data columns (total 6 columns):
Year                  31684 non-null int32
Make                  31684 non-null category
Model                 31683 non-null category
Class                 31683 non-null category
Fuel Type             31683 non-null category
Combined MPG (FT1)    31683 non-null float64
dtypes: category(4), float64(1), int32(1)
memory usage: 1.1 MB
```

Initial size 8.9 MB and after optimization 1.1 (!!!)