Python - Analiza danych z modułem PANDAS

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

LAB - S03-L005-LAB Oczyszczanie danych - fillna()

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. Wykonaj poniższe polecenia, żeby wprowadzić nowe wartości NaN do pierwszych wierszy zmiennej **fuel**

fuel.loc[27705, 'Class'] = np.NaN

fuel.loc[26561, 'Class'] = np.NaN

fuel.loc[27550, 'Fuel Type'] = np.NaN

fuel.loc[27705, 'Combined MPG (FT1)'] = np.NaN

fuel.loc[27681, 'Combined MPG (FT1)'] = np.NaN

3. Wyświetl nagłówek **fuel** żeby upewnić się, że wiersze rzeczywiście zawierają wartości **NaN**
4. Zamień wszystkie występujące w obiekcie **fuel** wartości na -1. Zmodyfikowane dane mają być wyświetlone na ekranie (wystarczy 5 pierwszych wierszy)
5. Utwórz słownik o nazwie **replaceRules**, który uda się wykorzystać do zamiany wartości wg następującej reguły:

w kolumnie 'Class' brakujące wartości należy uzupełnić przez '---',

w kolumnie 'Fuel Type' brakujące wartości należy uzupełnić przez '---',

w kolumnie 'Combined MPG (FT1)' brakujące wartości należy uzupełnić przez -1

6. Zamień brakujące wartości w obiekcie **fuel** korzystając z reguł zapisanych w zmiennej **replaceRules**. Zmodyfikowane dane mają być wyświetlone na ekranie (wystarczy 5 pierwszych wierszy)
7. Oblicz średnią wartość z kolumny **'Combined MPG (FT1)'** i zapisz ją w zmiennej **avgMPG**. Wyświetl wartość tej zmiennej
8. Korzystając z metody pozwalającej na usuwanie wartości **NaN** z data frame, kolumna po kolumnie zamień NaN wg zasad opisanych poniżej, a potem wyświetl obiekt **fuel** (wystarczy nagłówek):

NaN w kolumnie 'Class' na '?'

NaN w kolumnie 'Fuel Type na '?'

NaN w kolumnie 'Combined MPG (FT1)' na wartość wyzanaczoną w avgMPG

9. Uruchom ponownie kod z pkt. 2, żeby na nowo wprowadzić wartości **NaN**
10. Korzystając z jednej z automatycznych metod uzupełniania wartości, uzupełnij kolumnę **'Combined MPG (FT1)'** poprzednią nie-nullową wartością z tej kolumny. Wyświetl nagłówek oczyszczonych danych

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.loc[27705, 'Class'] = np.NaN
fuel.loc[26561, 'Class'] = np.NaN
fuel.loc[27550, 'Fuel Type'] = np.NaN
fuel.loc[27705, 'Combined MPG (FT1)'] = np.NaN
fuel.loc[27681, 'Combined MPG (FT1)'] = np.NaN
```

In [3]:
```
fuel.head()
```

Out[3]:

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	NaN	Regular	NaN
26561	1984	Alfa Romeo	Spider Veloce 2000	NaN	Regular	21.0
27681	1984	Alfa Romeo	Spider Veloce 2000	Two Seaters	Regular	NaN
27550	1984	AM General	DJ Po Vehicle 2WD	Special Purpose Vehicle 2WD	NaN	17.0

In [4]:
```
fuel.fillna(-1).head()
```

Out[4]:

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	-1	Regular	-1.0
26561	1984	Alfa Romeo	Spider Veloce 2000	-1	Regular	21.0
27681	1984	Alfa Romeo	Spider Veloce 2000	Two Seaters	Regular	-1.0
27550	1984	AM General	DJ Po Vehicle 2WD	Special Purpose Vehicle 2WD	-1	17.0

In [5]:
```
replaceRules = {'Class':'---',
                'Fuel Type':'---',
                'Combined MPG (FT1)':-1}
```

In [6]:
```
fuel.fillna(replaceRules).head()
```

Out[6]:

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	---	Regular	-1.0
26561	1984	Alfa Romeo	Spider Veloce 2000	---	Regular	21.0
27681	1984	Alfa Romeo	Spider Veloce 2000	Two Seaters	Regular	-1.0
27550	1984	AM General	DJ Po Vehicle 2WD	Special Purpose Vehicle 2WD	---	17.0

In [7]:
```
avgMPG = fuel['Combined MPG (FT1)'].mean()
avgMPG
```

Out[7]:
```
19.444177898424922
```

In [8]:
```
fuel['Class'].fillna('?',inplace=True)
fuel['Fuel Type'].fillna('?',inplace=True)
fuel['Combined MPG (FT1)'].fillna(avgMPG, inplace=True)
fuel.head()
```

Out[8]:

Vehicle ID	Year	Make	Model	Class	Fuel Type	Combined MPG (FT1)
26587	1984	Alfa Romeo	GT V6 2.5	Minicompact Cars	Regular	20.000000
27705	1984	Alfa Romeo	GT V6 2.5	?	Regular	19.444178
26561	1984	Alfa Romeo	Spider Veloce 2000	?	Regular	21.000000
27681	1984	Alfa Romeo	Spider Veloce 2000	Two Seaters	Regular	19.444178
27550	1984	AM General	DJ Po Vehicle 2WD	Special Purpose Vehicle 2WD	?	17.000000

In [9]:
```
fuel.loc[27705, 'Class'] = np.NaN
fuel.loc[26561, 'Class'] = np.NaN
fuel.loc[27550, 'Fuel Type'] = np.NaN
fuel.loc[27705, 'Combined MPG (FT1)'] = np.NaN
fuel.loc[27681, 'Combined MPG (FT1)'] = np.NaN
fuel.head()
```

Out[9]:

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	NaN	Regular	NaN
26561	1984	Alfa Romeo	Spider Veloce 2000	NaN	Regular	21.0
27681	1984	Alfa Romeo	Spider Veloce 2000	Two Seaters	Regular	NaN
27550	1984	AM General	DJ Po Vehicle 2WD	Special Purpose Vehicle 2WD	NaN	17.0

In [10]:
```
fuel['Combined MPG (FT1)'].fillna(method='ffill', inplace=True)
fuel.head()
```

Out[10]:

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	NaN	Regular	20.0
26561	1984	Alfa Romeo	Spider Veloce 2000	NaN	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	NaN	17.0