assessing

October 25, 2017

1 Assessing

Use the space below to explore `all_alpha_08.csv` and `all_alpha_18.csv` to answer the quiz questions below.

In [1]:
```
import pandas as pd
import numpy as np

df08 = pd.read_csv('all_alpha_08.csv')
df08.head()
```

Out[1]:

	Model	Displ	Cyl	Trans	Drive	Fuel	Sales Area	Stnd	Underhood ID	Veh Class	Air Pollution Score	FE Calc Appr	City MPG	Hwy MPG	Cmb MPG	Unadj Cmb MPG	Greenhouse Gas Score	SmartWay
0	ACURA MDX	3.7	(6 cyl)	Auto-S5	4WD	Gasoline	CA	U2	8HNXT03.7PKR	SUV	7	Drv	15	20	17	22.0527	4	no
1	ACURA MDX	3.7	(6 cyl)	Auto-S5	4WD	Gasoline	FA	B5	8HNXT03.7PKR	SUV	6	Drv	15	20	17	22.0527	4	no
2	ACURA RDX	2.3	(4 cyl)	Auto-S5	4WD	Gasoline	CA	U2	8HNXT02.3DKR	SUV	7	Drv	17	22	19	24.1745	5	no
3	ACURA RDX	2.3	(4 cyl)	Auto-S5	4WD	Gasoline	FA	B5	8HNXT02.3DKR	SUV	6	Drv	17	22	19	24.1745	5	no
4	ACURA RL	3.5	(6 cyl)	Auto-S5	4WD	Gasoline	CA	U2	8HNXV03.5HKR	midsize car	7	Drv	16	24	19	24.5629	5	no

In [2]:
```
df18 = pd.read_csv('all_alpha_18.csv')
df18.head()
```

Out[2]:

	Model	Displ	Cyl	Trans	Drive	Fuel	Cert Region	Stnd
0	ACURA RDX	3.5	6.0	SemiAuto-6	2WD	Gasoline	FA	T3B125

```
1  ACURA RDX    3.5  6.0  SemiAuto-6   2WD  Gasoline          CA         U2
2  ACURA RDX    3.5  6.0  SemiAuto-6   4WD  Gasoline          FA     T3B125
3  ACURA RDX    3.5  6.0  SemiAuto-6   4WD  Gasoline          CA         U2
4  ACURA TLX    2.4  4.0       AMS-8   2WD  Gasoline          CA  L3ULEV125

                Stnd Description  Underhood ID  Veh Class  Air Pollution Score  \
0      Federal Tier 3 Bin 125  JHNXT03.5GV3  small SUV                    3
1      California LEV-II ULEV  JHNXT03.5GV3  small SUV                    3
2      Federal Tier 3 Bin 125  JHNXT03.5GV3  small SUV                    3
3      California LEV-II ULEV  JHNXT03.5GV3  small SUV                    3
4  California LEV-III ULEV125  JHNXV02.4WH3  small car                    3

  City MPG Hwy MPG Cmb MPG  Greenhouse Gas Score SmartWay Comb CO2
0       20      28      23                     5       No      386
1       20      28      23                     5       No      386
2       19      27      22                     4       No      402
3       19      27      22                     4       No      402
4       23      33      27                     6       No      330
```

In [3]:
```
df08.shape
```

Out[3]:
```
(2404, 18)
```

In [4]:
```
sum(df08.duplicated())
```

Out[4]:
```
25
```

In [5]:
```
# trying to count rows with missing data. these produce the same result.
# np.count_nonzero(df08.isnull())
# sum(df08.isnull().sum())

# but taking the max of this is apparently correct
df08.isnull().sum()
```

Out[5]:
```
Model                    0
Displ                    0
Cyl                    199
Trans                  199
Drive                   93
Fuel                     0
Sales Area               0
Stnd                     0
Underhood ID             0
Veh Class                0
Air Pollution Score      0
FE Calc Appr           199
City MPG               199
Hwy MPG                199
Cmb MPG                199
```

```
        Unadj Cmb MPG           199
        Greenhouse Gas Score    199
        SmartWay                  0
        dtype: int64
```

```
In [6]: df08.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2404 entries, 0 to 2403
Data columns (total 18 columns):
Model                   2404 non-null object
Displ                   2404 non-null float64
Cyl                     2205 non-null object
Trans                   2205 non-null object
Drive                   2311 non-null object
Fuel                    2404 non-null object
Sales Area              2404 non-null object
Stnd                    2404 non-null object
Underhood ID            2404 non-null object
Veh Class               2404 non-null object
Air Pollution Score     2404 non-null object
FE Calc Appr            2205 non-null object
City MPG                2205 non-null object
Hwy MPG                 2205 non-null object
Cmb MPG                 2205 non-null object
Unadj Cmb MPG           2205 non-null float64
Greenhouse Gas Score    2205 non-null object
SmartWay                2404 non-null object
dtypes: float64(2), object(16)
memory usage: 338.1+ KB
```

```
In [7]: df08.nunique()
```

```
Out[7]: Model                   436
        Displ                    47
        Cyl                       8
        Trans                    14
        Drive                     2
        Fuel                      5
        Sales Area                3
        Stnd                     12
        Underhood ID            343
        Veh Class                 9
        Air Pollution Score      13
        FE Calc Appr              2
        City MPG                 39
        Hwy MPG                  43
        Cmb MPG                  38
```

```
        Unadj Cmb MPG           721
        Greenhouse Gas Score     20
        SmartWay                  2
        dtype: int64
```

```
In [8]: df18.shape
```

```
Out[8]: (1611, 18)
```

```
In [9]: # taking the max of this is apparently correct for rows w/ missing data
        df18.isnull().sum()
```

```
Out[9]: Model                   0
        Displ                   2
        Cyl                     2
        Trans                   0
        Drive                   0
        Fuel                    0
        Cert Region             0
        Stnd                    0
        Stnd Description        0
        Underhood ID            0
        Veh Class               0
        Air Pollution Score     0
        City MPG                0
        Hwy MPG                 0
        Cmb MPG                 0
        Greenhouse Gas Score    0
        SmartWay                0
        Comb CO2                0
        dtype: int64
```

```
In [10]: df18.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1611 entries, 0 to 1610
Data columns (total 18 columns):
Model                   1611 non-null object
Displ                   1609 non-null float64
Cyl                     1609 non-null float64
Trans                   1611 non-null object
Drive                   1611 non-null object
Fuel                    1611 non-null object
Cert Region             1611 non-null object
Stnd                    1611 non-null object
Stnd Description        1611 non-null object
Underhood ID            1611 non-null object
Veh Class               1611 non-null object
Air Pollution Score     1611 non-null int64
```

```
City MPG                1611 non-null object
Hwy MPG                 1611 non-null object
Cmb MPG                 1611 non-null object
Greenhouse Gas Score    1611 non-null int64
SmartWay                1611 non-null object
Comb CO2                1611 non-null object
dtypes: float64(2), int64(2), object(14)
memory usage: 226.6+ KB
```

```
In [11]: df18.nunique()
```

```
Out[11]: Model                   367
         Displ                    36
         Cyl                       7
         Trans                    26
         Drive                     2
         Fuel                      5
         Cert Region               2
         Stnd                     19
         Stnd Description         19
         Underhood ID            230
         Veh Class                 9
         Air Pollution Score       6
         City MPG                 58
         Hwy MPG                  62
         Cmb MPG                  57
         Greenhouse Gas Score     10
         SmartWay                  3
         Comb CO2                299
         dtype: int64
```

```
In [12]: sum(df18.duplicated())
```

```
Out[12]: 0
```

```
In [13]: type(df08['Cyl'][0])
```

```
Out[13]: str
```

```
In [14]: type(df18['Cyl'][0])
```

```
Out[14]: numpy.float64
```

```
In [15]: type(df18['City MPG'][0])
```

```
Out[15]: str
```

```
In [16]: type(df08['Greenhouse Gas Score'][0])
```

```
Out[16]: str

In [17]: type(df18['Greenhouse Gas Score'][0])

Out[17]: numpy.int64

In [4]: df08.Fuel.unique()

Out[4]: array(['Gasoline', 'ethanol/gas', 'ethanol', 'diesel', 'CNG'], dtype=object)

In [5]: df18.Fuel.unique()

Out[5]: array(['Gasoline', 'Gasoline/Electricity', 'Diesel', 'Ethanol/Gas',
               'Electricity'], dtype=object)

In [ ]:
```