fix_datatypes_cyl

October 26, 2017

1 Fixing `cyl` Data Type

- 2008: extract int from string
- 2018: convert float to int

```
In [1]: # load datasets
        import pandas as pd
```

```
In [7]: df_08 = pd.read_csv('data_08.csv')
        df_08.head(1)
```

```
Out[7]:        model  displ      cyl    trans drive      fuel veh_class  \
        0  ACURA MDX    3.7  (6 cyl)  Auto-S5   4WD  Gasoline       SUV

          air_pollution_score city_mpg hwy_mpg cmb_mpg greenhouse_gas_score smartway
        0                   7       15      20      17                    4       no
```

```
In [8]: df_18 = pd.read_csv('data_18.csv')
        df_18.head(1)
```

```
Out[8]:        model  displ  cyl       trans drive      fuel  veh_class  \
        0  ACURA RDX    3.5  6.0  SemiAuto-6   2WD  Gasoline  small SUV

           air_pollution_score city_mpg hwy_mpg cmb_mpg  greenhouse_gas_score smartway
        0                    3       20      28      23                     5       No
```

```
In [9]: # check value counts for the 2008 cyl column
        df_08['cyl'].value_counts()
```

```
Out[9]: (6 cyl)     409
        (4 cyl)     283
        (8 cyl)     199
        (5 cyl)      48
        (12 cyl)     30
        (10 cyl)     14
        (2 cyl)       2
        (16 cyl)      1
        Name: cyl, dtype: int64
```

Read this to help you extract ints from strings in Pandas for the next step.

```
In [10]: # Extract int from strings in the 2008 cyl column
         df_08['cyl'] = df_08['cyl'].str.extract('(\d+)', expand=False).astype(int)

In [11]: # Check value counts for 2008 cyl column again to confirm the change
         df_08['cyl'].value_counts()

Out[11]: 6     409
         4     283
         8     199
         5      48
         12     30
         10     14
         2       2
         16      1
         Name: cyl, dtype: int64

In [13]: # convert 2018 cyl column to int
         df_18['cyl'] = df_18['cyl'].astype(int)

In [14]: df_18['cyl'].value_counts()

Out[14]: 4     365
         6     246
         8     153
         3      18
         12      9
         5       2
         16      1
         Name: cyl, dtype: int64

In [15]: df_08.to_csv('data_08.csv', index=False)
         df_18.to_csv('data_18.csv', index=False)

In [ ]:
```