fix_datatypes_mpg_greenhouse

October 26, 2017

0.1 Fix `city_mpg`, `hwy_mpg`, `cmb_mpg` datatypes

2008 and 2018: convert string to float

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

```
In [2]: df_08 = pd.read_csv('data_08.csv')
        df_08.head(1)
```

```
Out[2]:        model  displ  cyl    trans drive      fuel veh_class  \
        0  ACURA MDX    3.7    6  Auto-S5   4WD  Gasoline       SUV

           air_pollution_score  city_mpg  hwy_mpg  cmb_mpg  greenhouse_gas_score  \
        0                  7.0        15       20       17                     4

          smartway
        0       no
```

```
In [3]: df_18 = pd.read_csv('data_18.csv')
        df_18.head(1)
```

```
Out[3]:        model  displ  cyl       trans drive      fuel  veh_class  \
        0  ACURA RDX    3.5    6  SemiAuto-6   2WD  Gasoline  small SUV

           air_pollution_score  city_mpg  hwy_mpg  cmb_mpg  greenhouse_gas_score  \
        0                  3.0        20       28       23                     5

          smartway
        0       No
```

```
In [4]: # convert mpg columns to floats
        mpg_columns = ['city_mpg', 'hwy_mpg', 'cmb_mpg']
        for c in mpg_columns:
            df_18[c] = df_18[c].astype(float)
            df_08[c] = df_08[c].astype(float)
```

```
In [12]: print('s/b float: {}'.format(type(df_08['city_mpg'][0])))
         print('s/b float: {}'.format(type(df_18['city_mpg'][0])))
         print('s/b int: {}'.format(type(df_08['cyl'][0])))
         print('s/b int: {}'.format(type(df_18['cyl'][0])))
         print('s/b float: {}'.format(type(df_08['air_pollution_score'][0])))
         print('s/b float: {}'.format(type(df_18['air_pollution_score'][0])))
         print('s/b int: {}'.format(type(df_08['greenhouse_gas_score'][0])))
         print('s/b int: {}'.format(type(df_18['greenhouse_gas_score'][0])))
```

```
s/b float: <class 'numpy.float64'>
s/b float: <class 'numpy.float64'>
s/b int: <class 'numpy.int64'>
s/b int: <class 'numpy.int64'>
s/b float: <class 'numpy.float64'>
s/b float: <class 'numpy.float64'>
s/b int: <class 'numpy.int64'>
s/b int: <class 'numpy.int64'>
```

0.2 Fix `greenhouse_gas_score` datatype

`2008: convert from float to int`

```
In [13]: # convert from float to int
         df_08['greenhouse_gas_score'] = df_08['greenhouse_gas_score'].astype(int)
```

0.3 All the dataypes are now fixed! Take one last check to confirm all the changes.

```
In [14]: df_08.dtypes
```

```
Out[14]: model                    object
         displ                   float64
         cyl                       int64
         trans                    object
         drive                    object
         fuel                     object
         veh_class                object
         air_pollution_score     float64
         city_mpg                float64
         hwy_mpg                 float64
         cmb_mpg                 float64
         greenhouse_gas_score      int64
         smartway                 object
         dtype: object
```

```
In [15]: df_18.dtypes
```

```
Out[15]: model                    object
         displ                   float64
```

```
cyl                       int64
trans                    object
drive                    object
fuel                     object
veh_class                object
air_pollution_score     float64
city_mpg                float64
hwy_mpg                 float64
cmb_mpg                 float64
greenhouse_gas_score      int64
smartway                 object
dtype: object
```

In [16]:
```
df_08.dtypes == df_18.dtypes
```

Out[16]:
```
model                    True
displ                    True
cyl                      True
trans                    True
drive                    True
fuel                     True
veh_class                True
air_pollution_score      True
city_mpg                 True
hwy_mpg                  True
cmb_mpg                  True
greenhouse_gas_score     True
smartway                 True
dtype: bool
```

In [17]:
```
# Save your new CLEAN datasets as new files!
df_08.to_csv('clean_08.csv', index=False)
df_18.to_csv('clean_18.csv', index=False)
```

In [ ]: