cleaning_column_labels

October 25, 2017

1 Cleaning Column Labels

Use `all_alpha_08.csv` and `all_alpha_18.csv`

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

```
In [5]: # load datasets

        df_08 = pd.read_csv('all_alpha_08.csv')
```

```
In [6]: df_18 = pd.read_csv('all_alpha_18.csv')
```

```
In [7]: # view 2008 dataset
        df_08.head(1)
```

```
Out[7]:        Model  Displ      Cyl     Trans Drive      Fuel Sales Area Stnd  \
        0  ACURA MDX    3.7  (6 cyl)  Auto-S5   4WD  Gasoline         CA   U2

          Underhood ID Veh Class Air Pollution Score FE Calc Appr City MPG Hwy MPG  \
        0  8HNXT03.7PKR       SUV                   7          Drv       15      20

          Cmb MPG  Unadj Cmb MPG Greenhouse Gas Score SmartWay
        0      17        22.0527                    4       no
```

```
In [8]: # view 2018 dataset
        df_18.head(1)
```

```
Out[8]:        Model  Displ  Cyl       Trans Drive      Fuel Cert Region    Stnd  \
        0  ACURA RDX    3.5  6.0  SemiAuto-6   2WD  Gasoline          FA  T3B125

                Stnd Description  Underhood ID  Veh Class  Air Pollution Score  \
        0  Federal Tier 3 Bin 125  JHNXT03.5GV3  small SUV                    3

          City MPG Hwy MPG Cmb MPG  Greenhouse Gas Score SmartWay Comb CO2
        0       20      28      23                     5       No      386
```

1.0.1 Drop Extraneous Columns

```
In [9]: # drop columns from 2008 dataset
        df_08.drop(['Stnd', 'Underhood ID', 'FE Calc Appr', 'Unadj Cmb MPG'], axis=1, inplace=Tr

        # confirm changes
        df_08.head(1)
```

```
Out[9]:        Model  Displ      Cyl    Trans Drive      Fuel Sales Area Veh Class  \
        0  ACURA MDX    3.7  (6 cyl)  Auto-S5   4WD  Gasoline         CA       SUV

          Air Pollution Score City MPG Hwy MPG Cmb MPG Greenhouse Gas Score SmartWay
        0                   7       15      20      17                    4       no
```

```
In [14]: # drop columns from 2018 dataset
         #df_18.drop(['Stnd', 'Underhood ID', 'Stnd Description'], axis=1, inplace=True)
         #df_18.drop(['Stnd Description'], axis=1, inplace=True)
         df_18.drop(['Comb CO2'], axis=1, inplace=True)

         # confirm changes
         df_18.head(1)
```

```
Out[14]:        Model  Displ  Cyl       Trans Drive      Fuel Cert Region   Veh Class  \
         0  ACURA RDX    3.5  6.0  SemiAuto-6   2WD  Gasoline          FA   small SUV

            Air Pollution Score City MPG Hwy MPG Cmb MPG  Greenhouse Gas Score SmartWay
         0                    3       20      28      23                     5       No
```

1.0.2 Rename Columns

```
In [11]: # rename Sales Area to Cert Region
         df_08.rename(columns={'Sales Area': 'Cert Region'}, inplace=True)

         # confirm changes
         df_08.head(1)
```

```
Out[11]:        Model  Displ      Cyl    Trans Drive      Fuel Cert Region Veh Class  \
         0  ACURA MDX    3.7  (6 cyl)  Auto-S5   4WD  Gasoline          CA       SUV

           Air Pollution Score City MPG Hwy MPG Cmb MPG Greenhouse Gas Score SmartWay
         0                   7       15      20      17                    4       no
```

```
In [15]: # replace spaces with underscores and lowercase labels for 2008 dataset
         df_08.rename(columns=lambda x: x.strip().lower().replace(" ", "_"), inplace=True)

         # confirm changes
         df_08.head(1)
```

```
Out[15]:        model  displ      cyl    trans drive      fuel cert_region veh_class  \
         0  ACURA MDX    3.7  (6 cyl)  Auto-S5   4WD  Gasoline          CA       SUV
```

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

```
In [16]: # replace spaces with underscores and lowercase labels for 2018 dataset
         df_18.rename(columns=lambda x: x.strip().lower().replace(" ", "_"), inplace=True)

         # confirm changes
         df_18.head(1)
```

```
Out[16]:       model  displ  cyl        trans drive      fuel cert_region  veh_class  \
         0  ACURA RDX    3.5  6.0  SemiAuto-6   2WD  Gasoline          FA  small SUV

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

```
In [17]: # confirm column labels for 2008 and 2018 datasets are identical
         df_08.columns == df_18.columns
```

```
Out[17]: array([ True,  True,  True,  True,  True,  True,  True,  True,  True,
                 True,  True,  True,  True,  True], dtype=bool)
```

```
In [18]: # make sure they're all identical like this
         (df_08.columns == df_18.columns).all()
```

```
Out[18]: True
```

```
In [19]: # save new datasets for next section
         df_08.to_csv('data_08.csv', index=False)
         df_18.to_csv('data_18.csv', index=False)
```

```
In [ ]:
```

```
In [ ]:
```