merging_data

October 27, 2017

1 Merging Datasets

Use Pandas Merges to create a combined dataset from `clean_08.csv` and `clean_18.csv`.

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

```
In [2]: df_08 = pd.read_csv('clean_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      15.0     20.0     17.0                     4

          smartway
        0       no
```

```
In [3]: df_18 = pd.read_csv('clean_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      20.0     28.0     23.0                     5

          smartway
        0       No
```

1.0.1 Create combined dataset

```
In [4]: # rename 2008 columns
        df_08.rename(columns=lambda x: x[:10] + "_2008", inplace=True)
```

```
In [5]: # view to check names
        df_08.head()
```

Out[5]:

	model_2008	displ_2008	cyl_2008	trans_2008	drive_2008	fuel_2008	\
0	ACURA MDX	3.7	6	Auto-S5	4WD	Gasoline	
1	ACURA RDX	2.3	4	Auto-S5	4WD	Gasoline	
2	ACURA RL	3.5	6	Auto-S5	4WD	Gasoline	
3	ACURA TL	3.2	6	Auto-S5	2WD	Gasoline	
4	ACURA TL	3.5	6	Auto-S5	2WD	Gasoline	

	veh_class_2008	air_pollut_2008	city_mpg_2008	hwy_mpg_2008	cmb_mpg_2008	\
0	SUV	7.0	15.0	20.0	17.0	
1	SUV	7.0	17.0	22.0	19.0	
2	midsize car	7.0	16.0	24.0	19.0	
3	midsize car	7.0	18.0	26.0	21.0	
4	midsize car	7.0	17.0	26.0	20.0	

	greenhouse_2008	smartway_2008
0	4	no
1	5	no
2	5	no
3	6	yes
4	6	yes

In [6]:
```
# merge datasets
df_combined = pd.merge(df_08, df_18, left_on='model_2008', right_on='model')
```

In [7]:
```
# view to check merge
df_combined.head()
```

Out[7]:

	model_2008	displ_2008	cyl_2008	trans_2008	drive_2008	fuel_2008	\
0	ACURA RDX	2.3	4	Auto-S5	4WD	Gasoline	
1	ACURA RDX	2.3	4	Auto-S5	4WD	Gasoline	
2	AUDI A3	2.0	4	Man-6	2WD	Gasoline	
3	AUDI A3	2.0	4	Man-6	2WD	Gasoline	
4	AUDI A3	2.0	4	Auto-S6	2WD	Gasoline	

	veh_class_2008	air_pollut_2008	city_mpg_2008	hwy_mpg_2008	...	\
0	SUV	7.0	17.0	22.0	...	
1	SUV	7.0	17.0	22.0	...	
2	station wagon	7.0	21.0	29.0	...	
3	station wagon	7.0	21.0	29.0	...	
4	station wagon	7.0	22.0	29.0	...	

	trans	drive	fuel	veh_class	air_pollution_score	city_mpg	\
0	SemiAuto-6	2WD	Gasoline	small SUV	3.0	20.0	
1	SemiAuto-6	4WD	Gasoline	small SUV	3.0	19.0	
2	AMS-6	4WD	Gasoline	small car	7.0	24.0	
3	AMS-7	2WD	Gasoline	small car	7.0	26.0	
4	AMS-6	4WD	Gasoline	small car	7.0	24.0	

```
   hwy_mpg  cmb_mpg  greenhouse_gas_score smartway
0     28.0     23.0                     5       No
1     27.0     22.0                     4       No
2     31.0     27.0                     6       No
3     35.0     29.0                     6       No
4     31.0     27.0                     6       No

[5 rows x 26 columns]
```

```
In [9]: df_combined.shape
```

```
Out[9]: (922, 26)
```

Save the combined dataset

```
In [10]: df_combined.to_csv('combined_dataset.csv', index=False)
```

```
In [ ]:
```