cleaning_column_labels-zh

August 15, 2018

1

all_alpha_08.csv all_alpha_18.csv

In [1]:
```
#
import pandas as pd
df_08 = pd.read_csv('all_alpha_08.csv')
df_18 = pd.read_csv('all_alpha_18.csv')
```

In [2]:
```
#  2008
df_08.head(1)
```

Out[2]:

	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

In [3]:
```
#  2018
df_18.head(1)
```

Out[3]:

	Model	Displ	Cyl	Trans	Drive	Fuel	Cert Region	Stnd	Stnd Description	Underhood ID	Veh Class	Air Pollution Score	City MPG	Hwy MPG	Cmb MPG	Greenhouse Gas Score	SmartWay	Comb CO2
0	ACURA RDX	3.5	6.0	SemiAuto-6	2WD	Gasoline	FA	T3B125	Federal Tier 3 Bin 125	JHNXT03.5GV3	small SUV	3	20	28	23	5	No	386

1.0.1

Pandas drop

2008 : 'Stnd''Underhood ID''FE Calc Appr''Unadj Cmb MPG' 2018 : 'Stnd''Stnd Description''Underhood ID''Comb CO2'

In [4]:
```
#  2008
df_08.drop(['Stnd', 'Underhood ID', 'FE Calc Appr', 'Unadj Cmb MPG'], axis=1, inplace=Tr

#
df_08.head(1)
```

Out[4]:

	Model	Displ	Cyl	Trans	Drive	Fuel	Sales Area	Veh Class	Air Pollution Score	City MPG	Hwy MPG	Cmb MPG	Greenhouse Gas Score	SmartWay
0	ACURA MDX	3.7	(6 cyl)	Auto-S5	4WD	Gasoline	CA	SUV	7	15	20	17	4	no

In [5]:
```
#  2018
df_18.drop(['Stnd','Stnd Description','Underhood ID','Comb CO2'],axis=1,inplace=True)

#
df_18.head()
```

Out[5]:

	Model	Displ	Cyl	Trans	Drive	Fuel	Cert Region	Veh Class	Air Pollution Score	City MPG	Hwy MPG	Cmb MPG	Greenhouse Gas Score	SmartWay
0	ACURA RDX	3.5	6.0	SemiAuto-6	2WD	Gasoline	FA	small SUV	3	20	28	23	5	No
1	ACURA RDX	3.5	6.0	SemiAuto-6	2WD	Gasoline	CA	small SUV	3	20	28	23	5	No
2	ACURA RDX	3.5	6.0	SemiAuto-6	4WD	Gasoline	FA	small SUV	3	19	27	22	4	No
3	ACURA RDX	3.5	6.0	SemiAuto-6	4WD	Gasoline	CA	small SUV	3	19	27	22	4	No
4	ACURA TLX	2.4	4.0	AMS-8	2WD	Gasoline	CA	small car	3	23	33	27	6	No

1.0.2

1. 2008 “Sales Area”””Cert Region“ Python df.column_name df[‘column_name’] query()

In [6]:
```
#
df_08.rename(columns={'Sales Area':'Cert Region'},inplace=True)

#
df_08.head(1)
```

Out[6]:

	Model	Displ	Cyl	Trans	Drive	Fuel	Cert Region	Veh Class	Air Pollution Score	City MPG	Hwy MPG	Cmb MPG	Greenhouse Gas Score	SmartWay
0	ACURA MDX	3.7	(6 cyl)	Auto-S5	4WD	Gasoline	CA	SUV	7	15	20	17	4	no

In [7]:
```
#  2008
df_08.rename(columns=lambda x: x.strip().lower().replace(" ", "_"), inplace=True)
```

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

Out[7]:

	model	displ	cyl	trans	drive	fuel	cert_region	veh_class	air_pollution_score	city_mpg	hwy_mpg	cmb_mpg	greenhouse_gas_score	smartway
0	ACURA MDX	3.7	(6 cyl)	Auto-S5	4WD	Gasoline	CA	SUV	7	15	20	17	4	no

```
In [8]: #  2018
        df_18.rename(columns=lambda x: x.strip().lower().replace(" ", "_"),inplace=True)

        #
        df_18.head(1)
```

Out[8]:

	model	displ	cyl	trans	drive	fuel	cert_region	veh_class	air_pollution_score	city_mpg	hwy_mpg	cmb_mpg	greenhouse_gas_score	smartway
0	ACURA RDX	3.5	6.0	SemiAuto-6	2WD	Gasoline	FA	small SUV	3	20	28	23	5	No

```
In [9]: #  2008  2018
        df_08.columns == df_18.columns
```

```
Out[9]: array([ True,  True,  True,  True,  True,  True,  True,  True,  True,
                True,  True,  True,  True,  True], dtype=bool)
```

```
In [10]: #
         (df_08.columns == df_18.columns).all()
```

```
Out[10]: True
```

```
In [11]: #
         df_08.to_csv('data_08.csv', index=True)
         df_18.to_csv('data_18.csv', index=False)
```

```
In [18]: df_08.head()
```

Out[18]:

	model	displ	cyl	trans	drive	fuel	cert_region	veh_class	air_pollution_score	city_mpg	hwy_mpg	cmb_mpg
0	ACURA MDX	3.7	(6 cyl)	Auto-S5	4WD	Gasoline	CA	SUV	7	15	20	17
1	ACURA MDX	3.7	(6 cyl)	Auto-S5	4WD	Gasoline	FA	SUV	6	15	20	17
2	ACURA RDX	2.3	(4 cyl)	Auto-S5	4WD	Gasoline	CA					
3	ACURA RDX	2.3	(4 cyl)	Auto-S5	4WD	Gasoline	FA					
4	ACURA RL	3.5	(6 cyl)	Auto-S5	4WD	Gasoline	CA					

```
2          SUV                  7       17      22      19
3          SUV                  6       17      22      19
4  midsize car                  7       16      24      19

  greenhouse_gas_score smartway
0                    4       no
1                    4       no
2                    5       no
3                    5       no
4                    5       no
```