รู้จักกับ pandas

author: Wes McKinney http://wesmckinney.com/

Pandas คืออะไร เป็นไลบราลี่ในภาษา Python สำหรับจัดการและวิเคราะห์ข้อมูลที่เป็นแบบโครงสร้างทั้งรูปแบบมิติเดียวและหลายมิติ

pandas ย่อมาจาก "panel data" longitudinal data เป็นการเก็บข้อมูลที่เราสนใจต่อเนื่องหลาย ๆ ช่วงเวลา เช่น

- ยอดขายรถแยกตามยี่ห้อในแต่ละเดือน
- ราคาหุ้นปิดรายวันของหุ้นแต่ละตัว

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

ยอดจดทะเบียนรถยนต์นั่งส่วนบุคคลไม่เกิน 7 คน ปี 2559

http://apps.dlt.go.th/statistics_web/statistics.html

In [2]:
```
df = pd.read_csv('https://github.com/prasertcbs/basic-dataset/raw/master/par
df.head()
```

Out[2]:

	month	brand	1300cc	1600cc	1800cc	2000cc	gt2000cc	elec	total
0	1	ALFA ROMEO	0	0	0	0	1	0	1
1	1	AUDI	0	2	0	14	0	0	16
2	1	BENTLEY	0	0	0	0	2	0	2
3	1	BMW	1	45	0	586	64	0	696
4	1	CADILLAC	0	0	0	0	1	0	1

In [3]:
```
df[df.brand.str.contains('AUDI|BMW|BENZ')]
```

Out[3]:

	month	brand	1300cc	1600cc	1800cc	2000cc	gt2000cc	elec	total
1	1	AUDI	0	2	0	14	0	0	16
3	1	BMW	1	45	0	586	64	0	696
20	1	MERCEDES BENZ	0	161	46	372	605	0	1184
38	2	AUDI	0	0	0	10	0	0	10
39	2	BMW	0	60	0	689	52	0	801
55	2	MERCEDES BENZ	0	193	42	475	672	0	1382
77	3	AUDI	0	0	0	15	1	0	16
79	3	BMW	0	55	0	658	55	0	768
92	3	MERCEDES BENZ	0	174	41	554	650	0	1419
110	4	AUDI	0	0	0	7	0	0	7
112	4	BMW	1	39	0	443	44	0	527
127	4	MERCEDES BENZ	0	133	29	311	427	0	900
144	5	AUDI	0	1	0	13	0	0	14
146	5	BMW	0	42	0	522	35	0	599
162	5	MERCEDES BENZ	0	156	24	482	516	0	1178
181	6	AUDI	0	0	0	11	0	0	11
183	6	BMW	0	62	0	568	56	0	686
200	6	MERCEDES BENZ	0	151	36	509	503	0	1199
220	7	AUDI	0	1	0	5	1	0	7
222	7	BMW	0	32	0	483	48	0	563
239	7	MERCEDES BENZ	0	128	28	460	353	0	969
259	8	AUDI	0	0	0	9	0	0	9
260	8	BMW	0	52	0	543	71	0	666
277	8	MERCEDES BENZ	0	140	27	543	429	0	1139
295	9	AUDI	0	0	0	2	2	0	4
296	9	BMW	0	55	0	486	55	0	596
310	9	MERCEDES BENZ	0	108	19	578	404	0	1109

	month	brand	1300cc	1600cc	1800cc	2000cc	gt2000cc	elec	total
329	10	AUDI	0	0	0	9	0	0	9
330	10	BMW	0	50	0	451	28	0	529
346	10	MERCEDES BENZ	0	96	19	457	280	0	852
362	11	AUDI	0	0	0	9	3	0	12
364	11	BMW	0	64	0	461	35	0	560
378	11	MERCEDES BENZ	1	94	19	449	276	0	839
396	12	AUDI	0	0	0	7	2	0	9
398	12	BMW	0	40	0	269	19	0	328
413	12	MERCEDES BENZ	0	49	5	251	156	0	461

In [4]:
```
# df.sample(n=10)
```

In [5]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 430 entries, 0 to 429
Data columns (total 9 columns):
 #   Column    Non-Null Count  Dtype 
---  ------    --------------  ----- 
 0   month     430 non-null    int64 
 1   brand     430 non-null    object
 2   1300cc    430 non-null    int64 
 3   1600cc    430 non-null    int64 
 4   1800cc    430 non-null    int64 
 5   2000cc    430 non-null    int64 
 6   gt2000cc  430 non-null    int64 
 7   elec      430 non-null    int64 
 8   total     430 non-null    int64 
dtypes: int64(8), object(1)
memory usage: 30.4+ KB
```

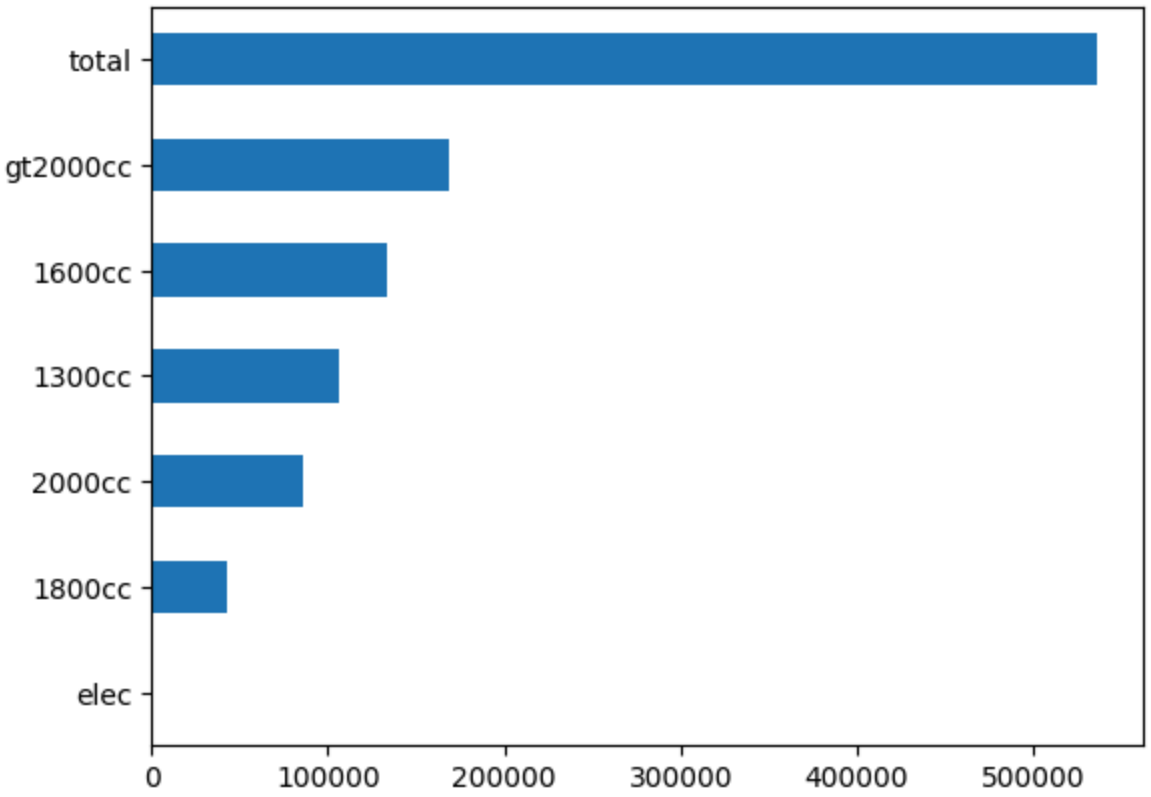
In [6]:
```
t = df.sum(numeric_only=True) # tip: set numeric_only=True
t
```

Out[6]:
```
month         2761
1300cc      106020
1600cc      133028
1800cc       42628
2000cc       85363
gt2000cc    168878
elec             2
total       535919
dtype: int64
```

In [7]:
```
t = df.drop(['month'], axis=1).sum(numeric_only=True) # tip: set numeric_onl
t
```

Out[7]:
```
1300cc      106020
1600cc      133028
1800cc       42628
2000cc       85363
gt2000cc    168878
elec             2
total       535919
dtype: int64
```

In [8]:
```
t.sort_values().plot(kind='barh');
```

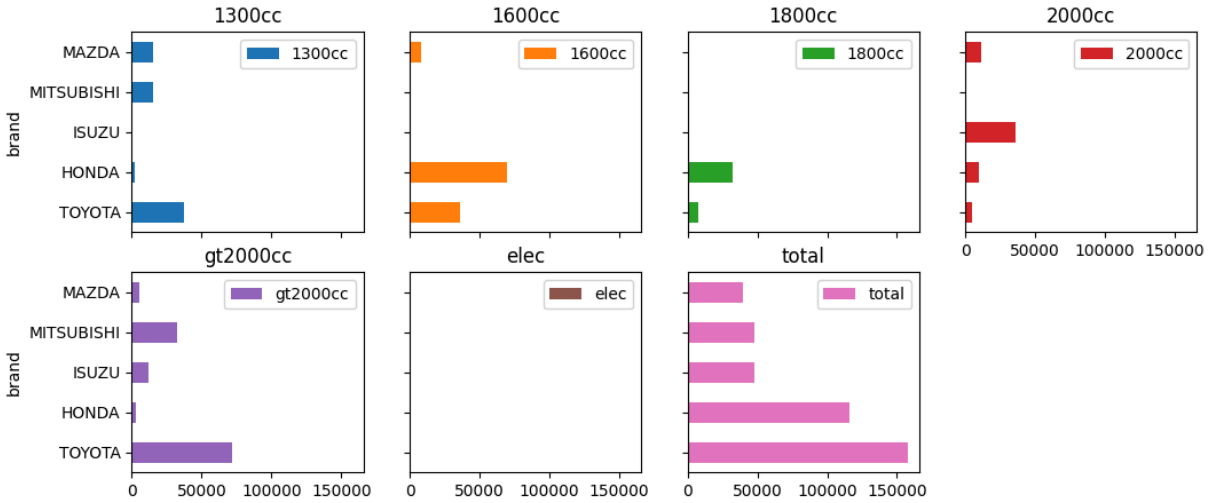


In [9]:
```
df.drop(['month'], axis=1).groupby('brand').sum().nlargest(5, 'total')
```

Out[9]:

brand	1300cc	1600cc	1800cc	2000cc	gt2000cc	elec	total
TOYOTA	37713	36105	7715	4359	72058	0	157950
HONDA	1890	69590	32547	9524	2575	0	116126
ISUZU	0	0	0	35943	12144	0	48087
MITSUBISHI	15342	3	64	173	32286	0	47868
MAZDA	14962	8423	0	11384	5137	0	39906

In [10]:
```
df.drop(['month'], axis=1).groupby('brand').sum().nlargest(5, 'total') \
    .plot(kind='barh', subplots=True, layout=(2, 4), figsize=(12, 5), sharey
```



In [11]:
```
df.groupby('brand').sum().nsmallest(10, 'total')
```

Out[11]:

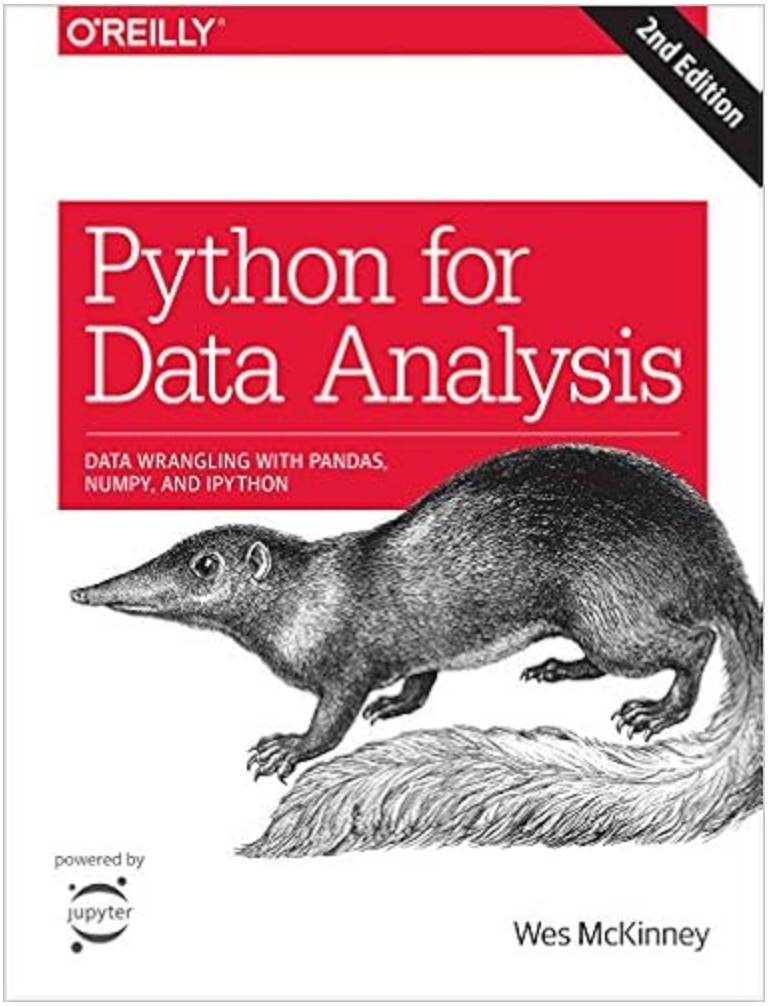
brand	month	1300cc	1600cc	1800cc	2000cc	gt2000cc	elec	total
CHRYSLER	10	0	0	0	1	0	0	1
HUMMER	10	0	0	0	0	1	0	1
LOTUS	12	0	0	0	0	1	0	1
NAZA	21	1	0	0	0	1	0	2
CHERY	12	0	0	0	3	0	0	3
DAIHATSU	7	4	0	0	0	0	0	4
MITSUOKA	6	0	0	0	3	1	0	4
PERODUA	8	0	4	0	0	0	0	4
SKODA	30	0	1	2	1	0	0	4
CADILLAC	32	0	0	0	0	5	0	5

In [12]:
```
# df.to_excel('data/panel_data.xlsx', index=False)
```

In [13]:
```
import requests
from PIL import Image # pillow package (Python Image Library)
import io

r = requests.get('https://images-na.ssl-images-amazon.com/images/I/51cUNf8zu
img = Image.open(io.BytesIO(r.content))
img
```

Out[13]:


O'REILLY
2nd Edition
Python for Data Analysis
DATA WRANGLING WITH PANDAS, NUMPY, AND IPYTHON
powered by
jupyter
Wes McKinney

In [ ]: