PyCon 2015 大会 上海 - Python大数据分析与可视化

丁来强

- Email: wjo1212 at 163.com
- Wechat: LaiQiangDing

In [36]:

```
%matplotlib inline
%run env.py
```

```
<matplotlib.figure.Figure at 0x10b11d6d0>
```

2011-2013中国轿车销量分析

加载, 整理

In [37]:

```
df = pd.read_excel(u"./汽车销量.xls")
```

查看前几行

In [38]:

```
df[:10]
```

Out[38]:

	乘用车新注册量	乘用车新注册量:分品牌:讴歌	乘用车新注册量:分品牌:安驰	乘用车新注册量:分品牌:阿斯顿马丁	乘用车新注册量:分品牌:奥迪	乘用车新注册量:分品牌:北汽
国家/地区	中国	中国	中国	中国	中国	中国
频率	月, 月	月, 月	月, 月	月, 月	月, 月	月, 月
成功	辆	辆	辆	辆	辆	辆
数据来源	POLK	POLK	POLK	POLK	POLK	POLK
状况	继续	继续	继续	继续	继续	继续
数列码	314368301 (CRATAAA)	314368401 (CRATAAB)	314368501 (CRATAAC)	314368601 (CRATAAD)	314368701 (CRATAAE)	314368801 (CRATAAF)
开始时间段	2011-01-01 00:00:00	2011-01-01 00:00:00	2011-01-01 00:00:00	2011-01-01 00:00:00	2011-01-01 00:00:00	2011-01-01 00:00:00
最后时间段	2013-12-01 00:00:00	2013-12-01 00:00:00	2013-12-01 00:00:00	2013-12-01 00:00:00	2013-12-01 00:00:00	2013-12-01 00:00:00
最新更新时间	2014-01-28 00:00:00	2014-01-28 00:00:00	2014-01-28 00:00:00	2014-01-28 00:00:00	2014-01-28 00:00:00	2014-01-28 00:00:00
2011-01-01 00:00:00	1305810	740	56	15	22292	481

10 rows × 112 columns

删除开头几行

In [39]:

```
df = df.drop(df.index[:9])
```

In [40]:

```
df.ix[:10, :10]
```

Out[40]:

	乘用车新注册量	乘用车新注册量:分品牌:讴歌	乘用车新注册量:分品牌:安驰	乘用车新注册量:分品牌:阿斯顿马丁	乘用车新注册量:分品牌:奥迪	乘用车新注册量:分品牌:北汽	乘用车新注册量:分品牌:宾利	乘用车新注册量:分品牌:奔腾	乘用车新注册量:分品牌:宝马	乘用车新注册量:分品牌:别克
2011-01-01	1305810	740	56	15	22292	481	85	13283	17918	55275
2011-02-01	631088	219	2	6	10533	148	50	5519	7863	25459
2011-03-01	1124005	378	3	22	20400	391	81	8639	19108	54636
2011-04-01	977721	325	2	13	19688	377	92	8415	18099	48827
2011-05-01	1025588	369	4	22	22880	484	95	8886	19602	49790
2011-06-01	964851	327	2	11	22396	641	90	7584	20624	46542
2011-07-01	912849	308	7	8	22854	541	110	7118	20250	46541
2011-08-01	1087484	307	4	8	26742	588	100	8290	22078	53124
2011-09-01	1371862	304	10	13	30972	585	124	10736	22761	65955
2011-10-01	936413	250	1	12	22442	553	101	7148	16451	42448

In [41]:

```
df.index
```

Out[41]:

```
DatetimeIndex(['2011-01-01', '2011-02-01', '2011-03-01', '2011-0
4-01',
               '2011-05-01', '2011-06-01', '2011-07-01', '2011-0
8-01',
               '2011-09-01', '2011-10-01', '2011-11-01', '2011-1
2-01',
               '2012-01-01', '2012-02-01', '2012-03-01', '2012-0
4-01',
               '2012-05-01', '2012-06-01', '2012-07-01', '2012-0
8-01',
               '2012-09-01', '2012-10-01', '2012-11-01', '2012-1
2-01',
               '2013-01-01', '2013-02-01', '2013-03-01', '2013-0
4-01',
               '2013-05-01', '2013-06-01', '2013-07-01', '2013-0
8-01',
               '2013-09-01', '2013-10-01', '2013-11-01', '2013-1
2-01'],
              dtype='datetime64[ns]', freq=None, tz=None)
```

In [42]:

```
len(df.columns) # 111个品牌
```

Out[42]:

```
112
```

In [43]:

```
df.columns = [x.split(':')[2] if ':' in x else x for x in df.columns]
```

In [44]:

```
df.ix[:10, :10]
```

Out[44]:

	乘用车新注册量	讴歌	安驰	阿斯顿马丁	奥迪	北汽	宾利	奔腾	宝马	别克
2011-01-01	1305810	740	56	15	22292	481	85	13283	17918	55275
2011-02-01	631088	219	2	6	10533	148	50	5519	7863	25459
2011-03-01	1124005	378	3	22	20400	391	81	8639	19108	54636
2011-04-01	977721	325	2	13	19688	377	92	8415	18099	48827
2011-05-01	1025588	369	4	22	22880	484	95	8886	19602	49790
2011-06-01	964851	327	2	11	22396	641	90	7584	20624	46542
2011-07-01	912849	308	7	8	22854	541	110	7118	20250	46541
2011-08-01	1087484	307	4	8	26742	588	100	8290	22078	53124
2011-09-01	1371862	304	10	13	30972	585	124	10736	22761	65955
2011-10-01	936413	250	1	12	22442	553	101	7148	16451	42448

In [45]:

```
df2 = df[:10].copy()
```

In [46]:

```
df2["Total"] = df2.sum(axis=1)
```

In [47]:

```
df2[u'乘用车新注册量'] / df2.Total
```

Out[47]:

```
2011-01-01          0.5
2011-02-01          0.5
2011-03-01    0.5000002
2011-04-01    0.5000141
2011-05-01     0.500268
2011-06-01    0.5003205
2011-07-01    0.5003445
2011-08-01    0.5003709
2011-09-01    0.5004845
2011-10-01    0.5007961
dtype: object
```

In [48]:

```
df2["Other"] = df2[u'乘用车新注册量'] - ( df2.Total - df2[u'乘用车新注册量'])
```

In [49]:

```
df2.Other
```

Out[49]:

```
2011-01-01       0
2011-02-01       0
2011-03-01       1
2011-04-01      55
2011-05-01    1099
2011-06-01    1236
2011-07-01    1257
2011-08-01    1612
2011-09-01    2656
2011-10-01    2977
Name: Other, dtype: object
```

In [50]:

```
df["Other"] = df[u'乘用车新注册量'] - ( df.sum(axis=1) - df[u'乘用车新注册量'])
```

In [51]:

```
df = df.drop(u'乘用车新注册量', axis=1)
```

In [52]:

```
df2 = df.T
```

In [53]:

```
df.ix[:5,:5]
```

Out[53]:

	讴歌	安驰	阿斯顿马丁	奥迪	北汽
2011-01-01	740	56	15	22292	481
2011-02-01	219	2	6	10533	148
2011-03-01	378	3	22	20400	391
2011-04-01	325	2	13	19688	377
2011-05-01	369	4	22	22880	484

In [54]:

```
df2.ix[:5,:5]
```

Out[54]:

	2011-01-01 00:00:00	2011-02-01 00:00:00	2011-03-01 00:00:00	2011-04-01 00:00:00	2011-05-01 00:00:00
讴歌	740	219	378	325	369
安驰	56	2	3	2	4
阿斯顿马丁	15	6	22	13	22
奥迪	22292	10533	20400	19688	22880
北汽	481	148	391	377	484

2011, 2012, 2013 汽车总体销量情况

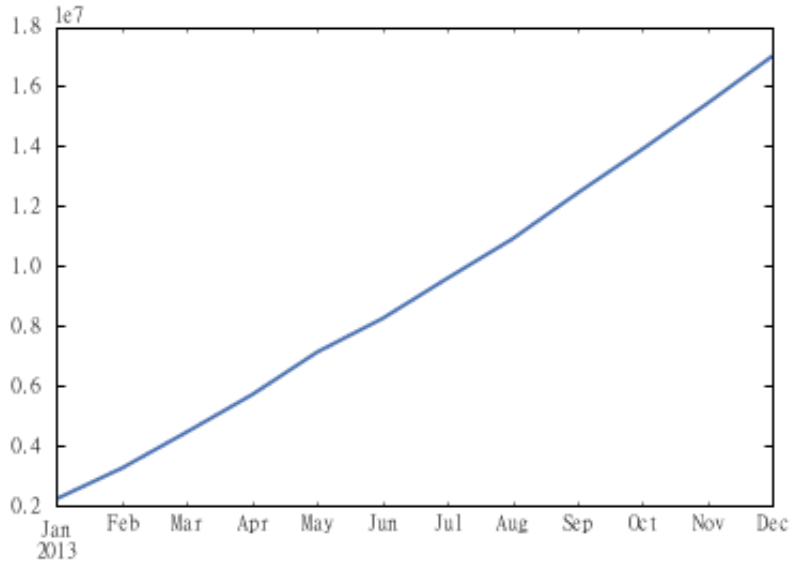
2013年累计销量

In [55]:

```
df['2013'].sum(axis=1).cumsum().plot()
```

Out[55]:

<matplotlib.axes._subplots.AxesSubplot at 0x10b1bba90>



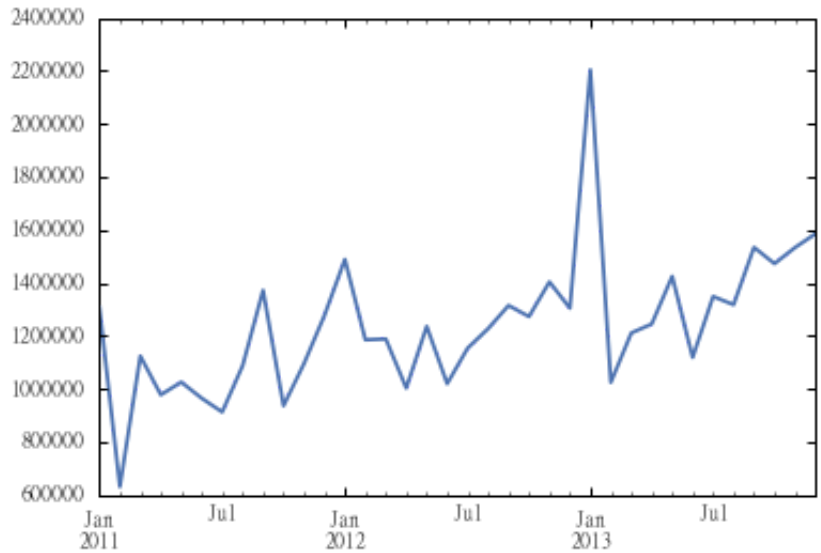
月度销量情况

In [56]:

```
df.sum(axis=1).plot()
```

Out[56]:

<matplotlib.axes._subplots.AxesSubplot at 0x10b1bb8d0>



月度销售热力图

In [88]:

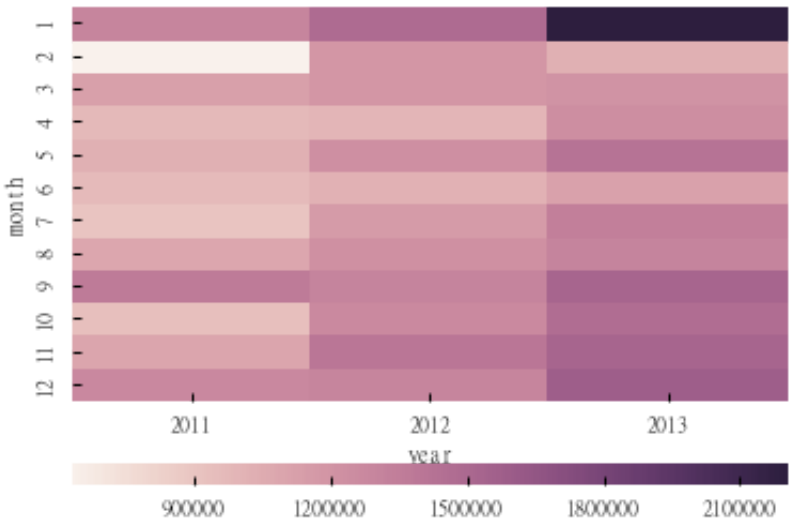
```
df_heat = df.sum(axis=1).to_frame()
df_heat = df_heat.assign(month=df_heat.index.month, year=df_heat.index.year)
df_heat.columns = ['sell', 'month', 'year']
df_heat = df_heat.pivot("month", "year", "sell")
df_heat.head()
```

Out[88]:

year	2011	2012	2013
month			
1	1305810	1488666	2203799
2	631088	1185260	1024571
3	1124005	1188454	1210883
4	977721	1003202	1244892
5	1025588	1235918	1424747

In [57]:

```
grid_kws = {"height_ratios": (.9, .05), "hspace": .3}
f, (ax, cbar_ax) = plt.subplots(2, gridspec_kw=grid_kws)
ax = sns.heatmap(df_heat, ax=ax,
                 cbar_ax=cbar_ax,
                 cbar_kws={"orientation": "horizontal"})
```



月度销量回归

In [58]:

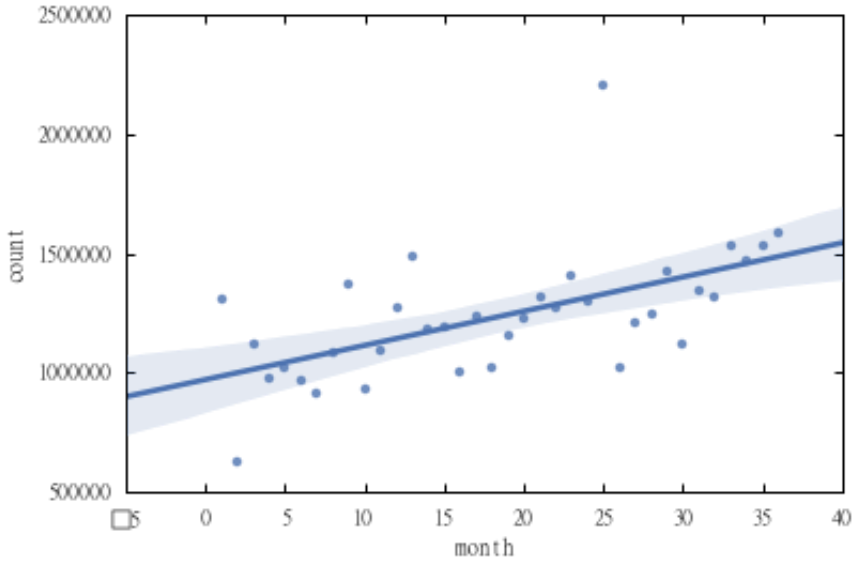
```
df_13_monthly = df.sum(axis=1)
df_13_monthly = df_13_monthly.to_frame().reset_index()
df_13_monthly["index"] = range(1, len(df_13_monthly)+1)
df_13_monthly.columns = ["month", "count"]
```

In [59]:

```
sns.regplot(x="month", y="count", data=df_13_monthly)
```

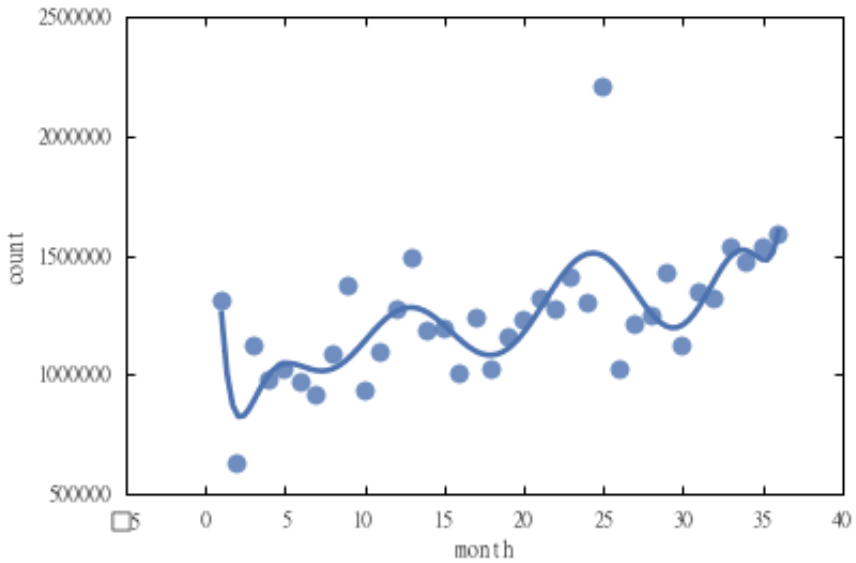
Out[59]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x10b33b910>
```



In [60]:

```
ax = sns.regplot(x="month", y="count",
                 data=df_13_monthly,
                 scatter_kws={"s": 80},
                 order=10, ci=None, truncate=True)
```



2011, 2012, 2013 Top10 的销售趋势

In [61]:

```
df_year = df.resample('A-DEC', how='sum').ix[:,1:]
df_year.ix[:,:10]
```

Out[61]:

	安驰	阿斯顿马丁	奥迪	北汽	宾利	奔腾	宝马	别克	比亚迪	凯迪拉克
2011-12-31	91	150	275927	6048	1185	103164	225445	592420	365883	24948
2012-12-31	1	101	381464	4563	1906	79868	269717	660970	379176	31877
2013-12-31	477	285	478230	5277	1898	109911	374052	792227	401892	43107

In [62]:

```
def get_year(y, count=5):
    sell = df_year.ix[0].copy().drop([u'五菱', u'长安'])
    sell.sort(ascending=False)
    return sell[:count]
sell_2011_top = get_year(0)
sell_2012_top = get_year(1)
sell_2013_top = get_year(2)
```

In [63]:

```
df[sell_2013_top.index.values][:10]
```

Out[63]:

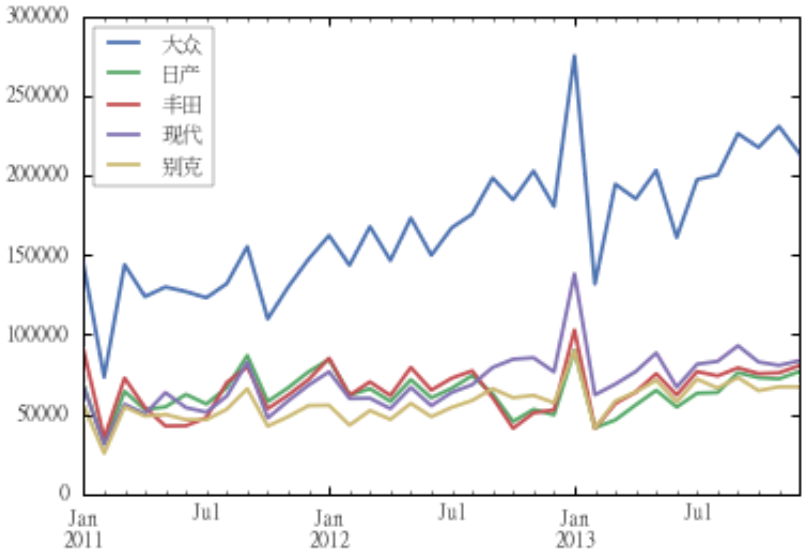
	大众	日产	丰田	现代	别克
2011-01-01	143518	67732	89901	67103	55275
2011-02-01	73330	31380	34892	31030	25459
2011-03-01	143837	64336	72653	56218	54636
2011-04-01	123768	52804	54051	50095	48827
2011-05-01	129834	54547	42549	63504	49790
2011-06-01	126891	62403	42739	54111	46542
2011-07-01	123063	56358	47850	51210	46541
2011-08-01	131945	66739	69872	61384	53124
2011-09-01	155141	86796	79724	82512	65955
2011-10-01	109785	57916	53401	47613	42448

In [64]:

```
df[sell_2013_top.index.values].plot()
```

Out[64]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x109322cd0>
```



季度销售热力图

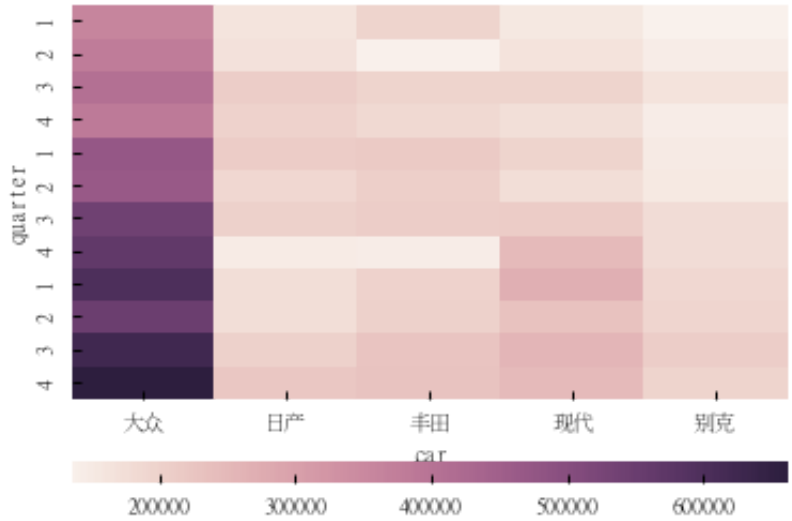
In [65]:

```
data = df[sell_2013_top.index.values].resample("Q-DEC", how="sum")
data.index=[i.quarter for i in data.index]
data.index.name = "quarter"
data.columns.name = "car"
data = data.astype(int)
sns.heatmap(data)
```

...

In [66]:

```
grid_kws = {"height_ratios": (.9, .05), "hspace": .3}
f, (ax, cbar_ax) = plt.subplots(2, gridspec_kw=grid_kws)
ax = sns.heatmap(data, ax=ax,
                 cbar_ax=cbar_ax,
                 cbar_kws={"orientation": "horizontal"})
```



In [67]:

```
sell_2013_top = df[sell_2013_top.index.values].unstack().reset_index()
sell_2013_top.columns = [u'品牌', u'月份', u'销量']
sell_2013_top[u'销量'] = sell_2013_top[u'销量'].astype('int64')
```

In [68]:

```
sell_2013_top[:5]
```

Out[68]:

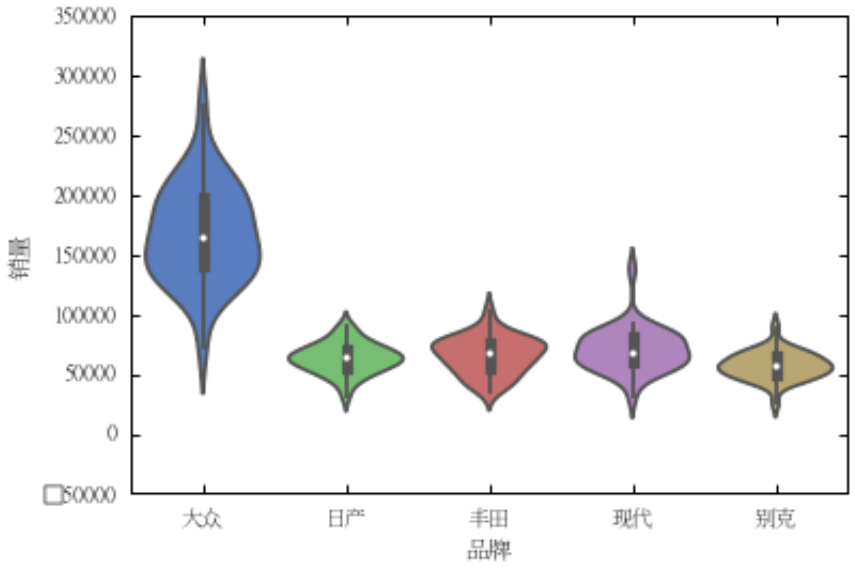
	品牌	月份	销量
0	大众	2011-01-01	143518
1	大众	2011-02-01	73330
2	大众	2011-03-01	143837
3	大众	2011-04-01	123768
4	大众	2011-05-01	129834

In [69]:

```
sns.violinplot(x=u'品牌', y=u'销量', #hue="smoker",
               data=sell_2013_top, palette="muted",
               scale="count") #scale="area/count"
```

Out[69]:

<matplotlib.axes._subplots.AxesSubplot at 0x10c5a1250>



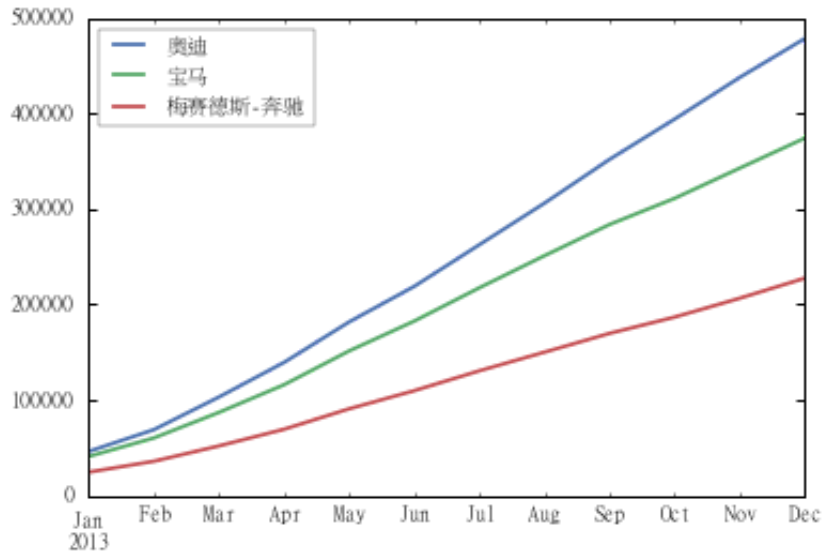
ABB之间销售情况

In [71]:

```
df_abb = df[[u'奥迪', u'宝马', u'梅赛德斯-奔驰']].copy()
df_abb['2013'].cumsum().plot()
```

Out[71]:

<matplotlib.axes._subplots.AxesSubplot at 0x1097932d0>



月度销量情况

In [72]:

```
get_year(2, 30)
df_abb.plot()
```

...

季度销售情况

In [73]:

```
df_abb.resample("Q-DEC", how='sum').plot()
```

...

In [74]:

```
#df_abb["季度"] = df_abb.index.quarter
df_abb = df_abb.unstack().reset_index()
```

In [75]:

```
df_abb.columns = [u'brand', u'month', u'sell']
```

In [76]:

```
df_abb.dtypes
```

Out[76]:

```
brand            object
month    datetime64[ns]
sell             object
dtype: object
```

In [77]:

```
df_abb[u"quarter"] = df_abb[u"month"].dt.quarter
df_abb[u"year"] = df_abb[u"month"].dt.year
df_abb[u"month"] = df_abb[u"month"].dt.month
```

In [78]:

```
df_abb[u"brand"] = df_abb[u"brand"].map({u"奥迪": "Audi",
                                         u"宝马": "BMW",
                                        u"梅赛德斯-奔驰": "Benz"})
```

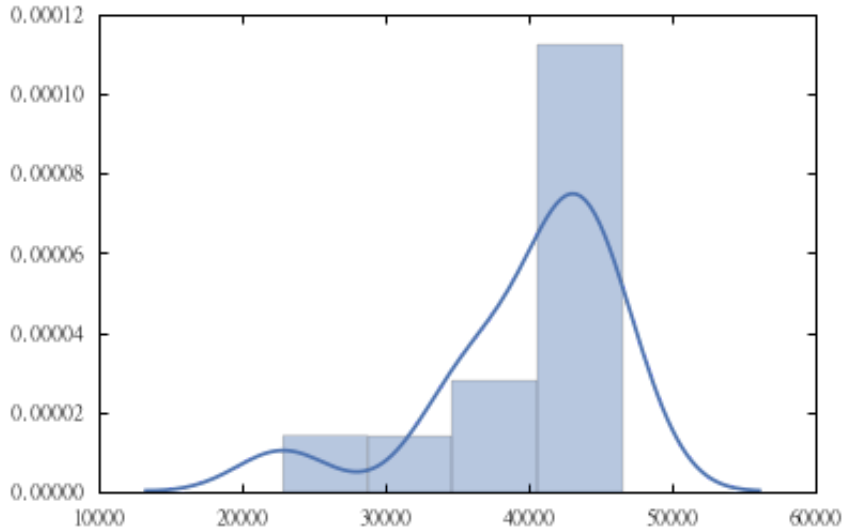
月销量分布 (2013)

In [79]:

```
sns.distplot(df_abb.query('brand=="Audi" and year==2013').sell.values, bins=4)
```

Out[79]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x10d0dd210>
```

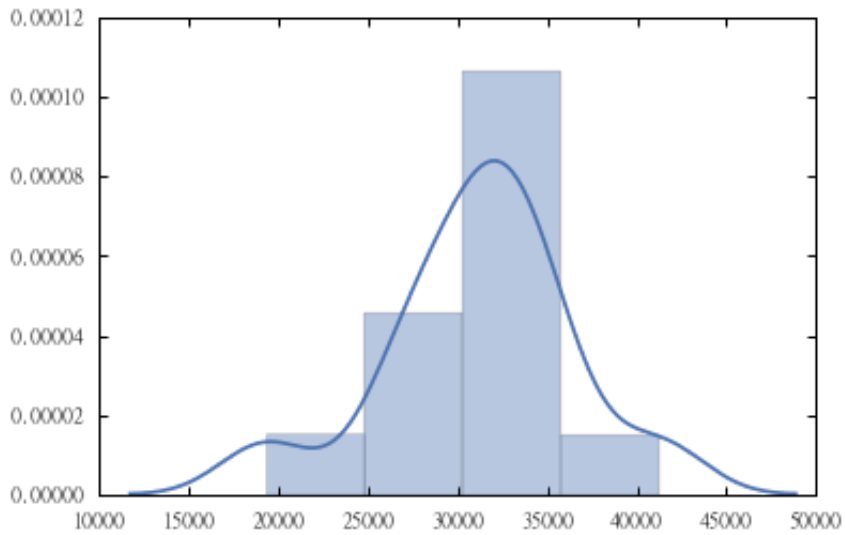


In [80]:

```
sns.distplot(df_abb.query('brand=="BMW" and year==2013').sell.values, bins=4)
```

Out[80]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x10d158210>
```

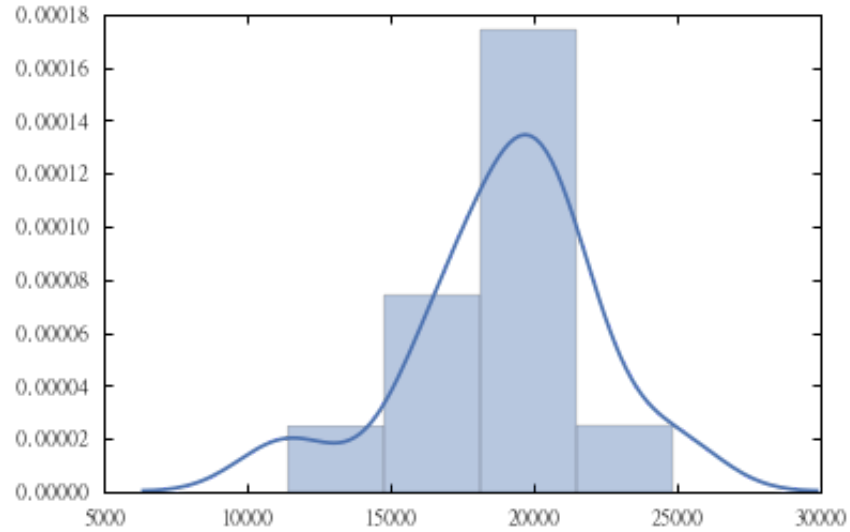


In [81]:

```
sns.distplot(df_abb.query('brand=="Benz" and year==2013').sell.values, bins=4)
```

Out[81]:

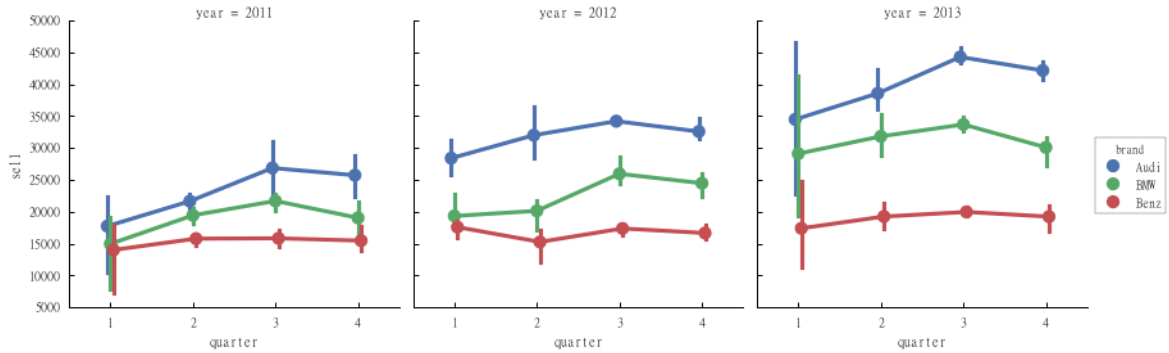
```
<matplotlib.axes._subplots.AxesSubplot at 0x10d358150>
```



因素图

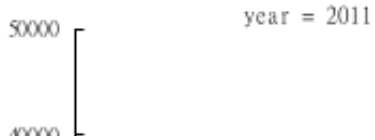
In [82]:

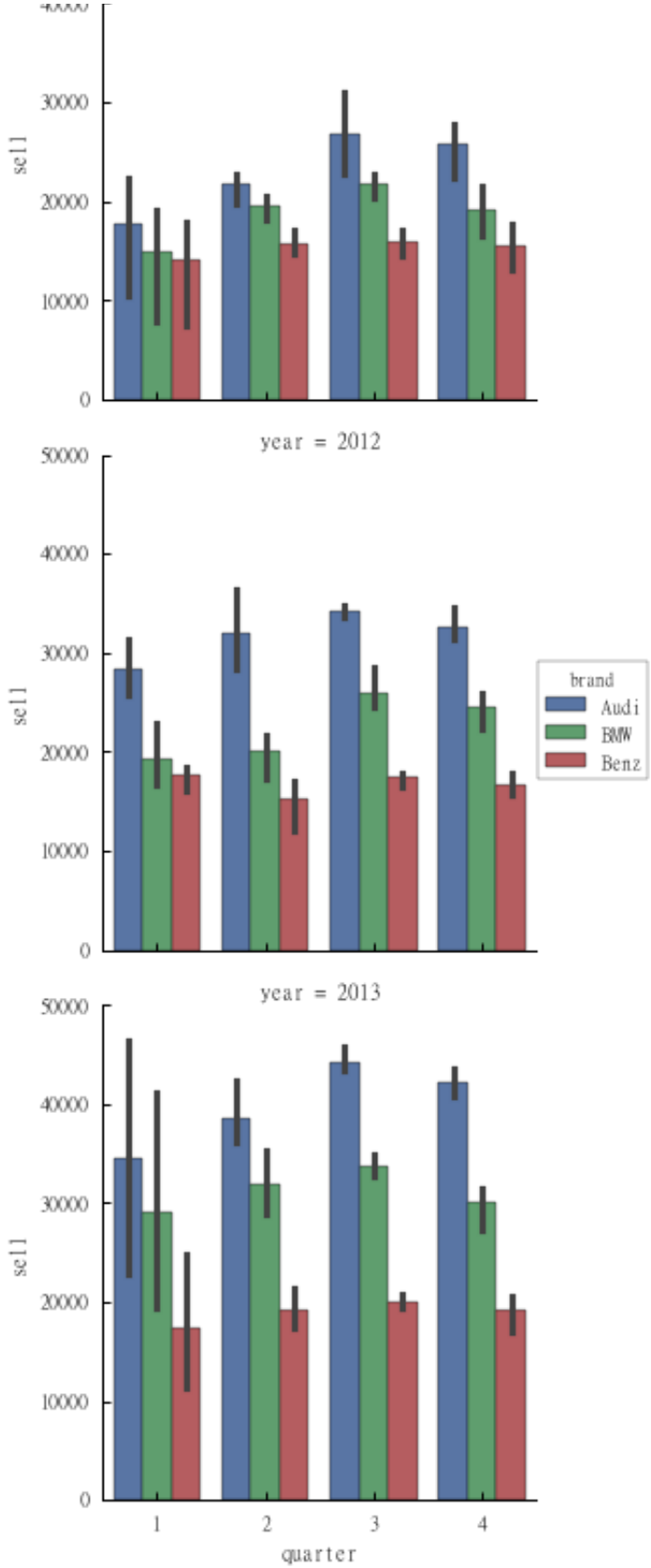
```
g = sns.factorplot(x=u"quarter", y=u"sell",
                   hue=u"brand",
                   col=u"year",    # row="year"
                   data=df_abb,
                   dodge=True)
```



In [83]:

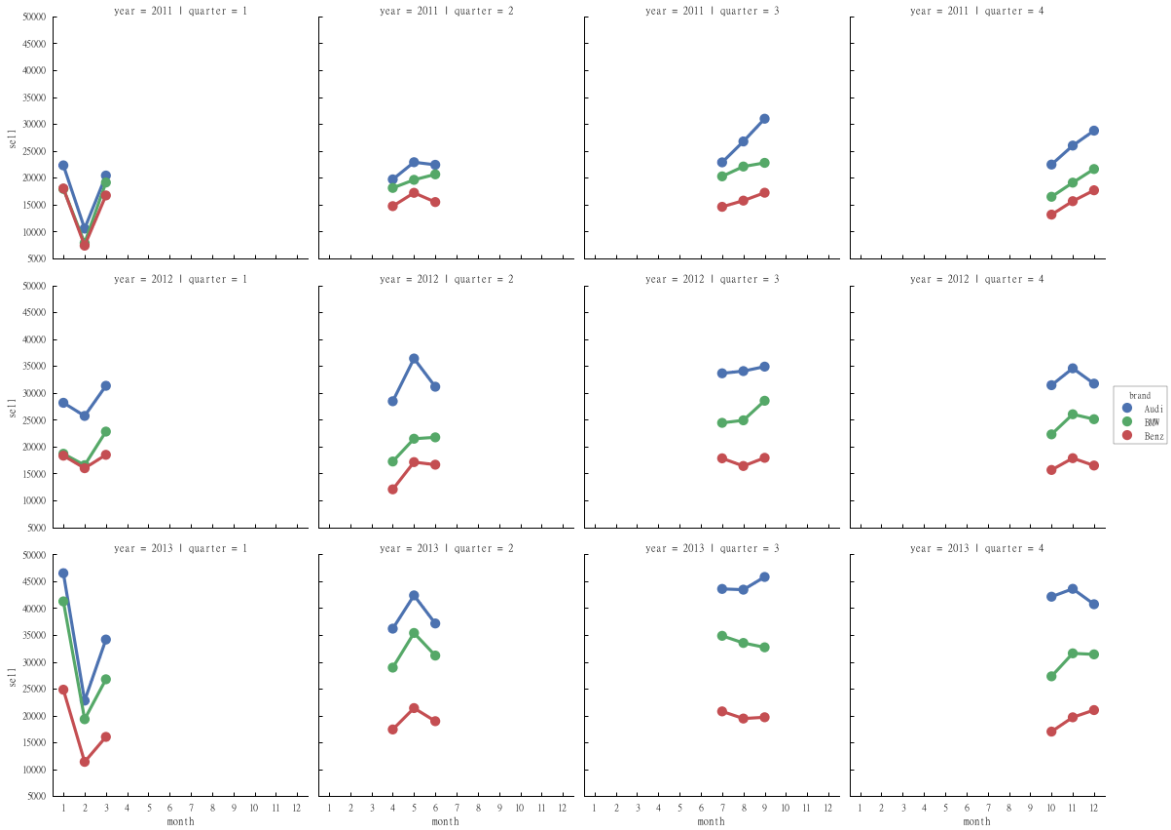
```
g = sns.factorplot(x=u"quarter", y=u"sell",
                   hue=u"brand",
                   row=u"year",
                   data=df_abb,
                   kind="bar")
```




sell
year = 2012
year = 2013
brand
Audi
BMW
Benz
quarter

In [84]:

```
g = sns.factorplot(x=u"month", y=u"sell",
                   hue=u"brand", row='year',
                   col=u"quarter", data=df_abb)
```



关联关系

ABB 与总体月度销量的关系

In [85]:

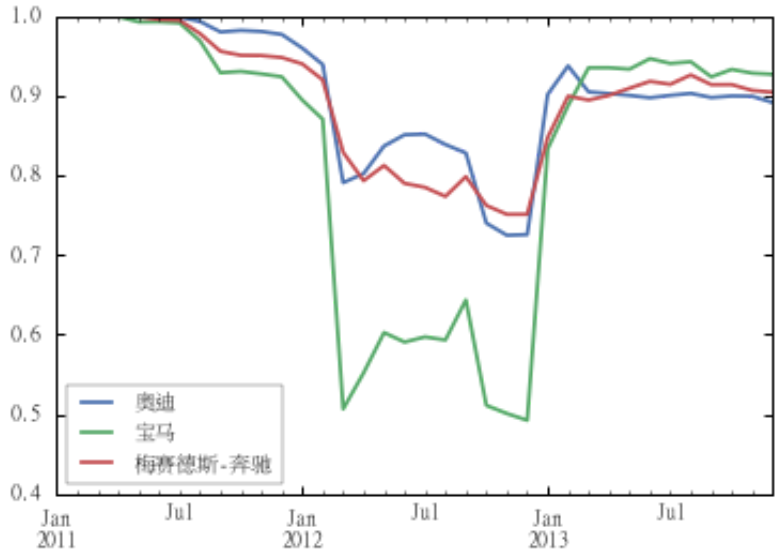
```
df_all = df.sum(axis=1)
df_all_rets = df_all/df_all.shift(1) - 1
returns = df.pct_change()
corr = pd.rolling_corr(returns[[u"奥迪", u"宝马", u"梅赛德斯-奔驰"]],
                       df_all_rets, 12, min_periods=2)
```

In [86]:

```
corr.plot()
```

Out[86]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x1106fec50>
```



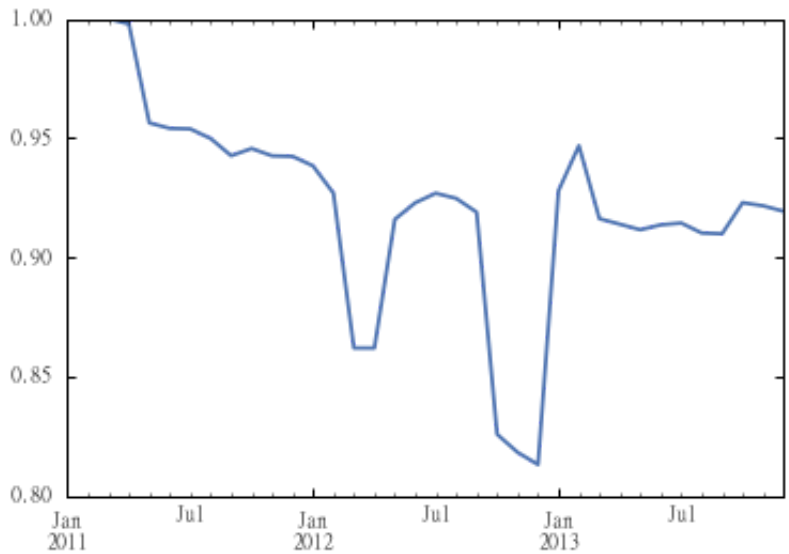
现代 别克月度销量关系

In [87]:

```
hynda_rets = df[u'现代'].pct_change()
bieke_rets = df[u"别克"].pct_change()
pd.rolling_corr(hynda_rets, bieke_rets, 12, min_periods=2).plot()
```

Out[87]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x111023790>
```



In [ ]:

```
```