网销案件EDA

目的：对网销案件从不同维度，运用制表、作图、特征分析等方法进行探索性数据分析，以发现赔案的规律和特点，对业务的发展提供决策建议。

分析思路：

1. 数据导入与概览
2. 案件整体情况
    - 赔款金额描述统计与分布情况
    - 投保人与被保险人关系分布情况
    - 出险原因分布情况
3. 时间维度
    - 每日赔款金额趋势和赔案数量分布
4. 机构维度
    - 各机构案件数量与赔款金额对比
    - 各机构赔案业务来源占比
    - 各机构赔案险种占比
5. 险种维度
    - 各险种赔案数量对比
    - 各险种案均赔款对比
    - 各险种出险案件数量随时间变化趋势
    - 各出险原因出险案件数量时间趋势
6. 出险省份维度
    - 各省份案件数量分布
    - 各省份案均赔款分布
    - 各省份险种构成
7. 业务来源维度
    - 各业务来源赔案数量对比
    - 各业务来源案均赔款对比
    - 各业务来源赔款总额对比
8. 终端维度
    - 各终端来源赔案数量对比
    - 各终端来源案均赔款对比
9. 年龄维度
    - 各年龄段案件数量对比
    - 各年龄段案均赔款对比
    - 各年龄段赔案出险原因分布
10. 相关性探索
    - 各类日期差值分布情况
    - 各日期差值与赔案金额相关性分析

1. 数据导入与概览

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
plt.rcParams['font.sans-serif']=['SimHei']
plt.rcParams['axes.unicode_minus'] = False
plt.style.use('ggplot')
from  pyecharts.charts import Bar
import pandas_profiling
from eplot import eplot
import warnings
warnings.filterwarnings('ignore')
```

```
data = pd.read_csv('process_data.csv')
```

```
data.info()
```

```
RangeIndex: 1426 entries, 0 to 1425
Data columns (total 26 columns):
company                 1426 non-null object
organization            1426 non-null object
id                      1426 non-null object
birthday                1426 non-null object
relationship_insured    1426 non-null object
insurance_code          1426 non-null int64
```

```
date_case               1426 non-null object
start_date_insurance    1426 non-null object
report_date             1426 non-null object
decision_time           1426 non-null object
loss_type               1426 non-null int64
pay_money               1426 non-null float64
pay_counts              1426 non-null int64
business_source         1426 non-null object
terminal_source         1426 non-null object
channel                 1426 non-null object
insurance_type          1426 non-null object
case_reason             1426 non-null object
case_province           1426 non-null object
case_city               1426 non-null object
case_town               1382 non-null object
case_road               1426 non-null object
age                     881 non-null float64
diff_case_date          1426 non-null float64
diff_report_date        1426 non-null float64
diff_dic_data           1426 non-null float64
dtypes: float64(5), int64(3), object(18)
```

```
data.describe()
```

	insurance_code	loss_type	pay_money	pay_counts	age	diff_case_date	diff_report_date	diff_d
count	1426.000000	1426.000000	1426.000000	1426.000000	881.000000	1426.000000	1426.000000	1426.0
mean	619.865358	0.019635	2068.582574	0.009818	45.253121	139.692146	25.390603	21.624
std	69.447404	0.138792	4968.942240	0.105506	11.680333	97.872727	35.454893	28.578
min	604.000000	0.000000	11.200000	0.000000	19.000000	0.000000	0.000000	0.0000
25%	604.000000	0.000000	349.385000	0.000000	37.000000	57.000000	2.000000	3.0000
50%	606.000000	0.000000	863.220000	0.000000	46.000000	127.000000	14.000000	11.000
75%	606.000000	0.000000	2434.730000	0.000000	54.000000	210.000000	33.000000	29.000
max	1229.000000	1.000000	109100.000000	2.000000	67.000000	449.000000	310.000000	208.00

```
data.shape
```

```
(1426, 26)
```

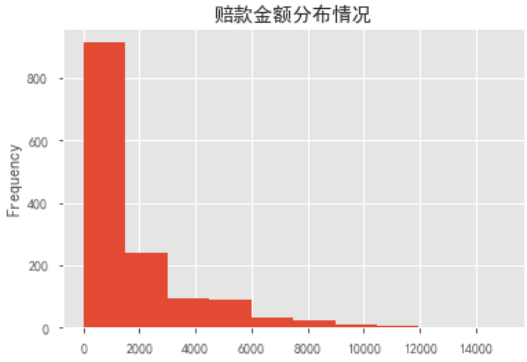
```
# 根据先验经验，填充学平险年龄为均值8，标准差1的正态分布随机数
age = np.random.normal(loc=8, scale=1.0, size=484)
age = age.astype('int64')
data['age'][data['insurance_type']== '学平险'] = age
```

2. 案件整体情况

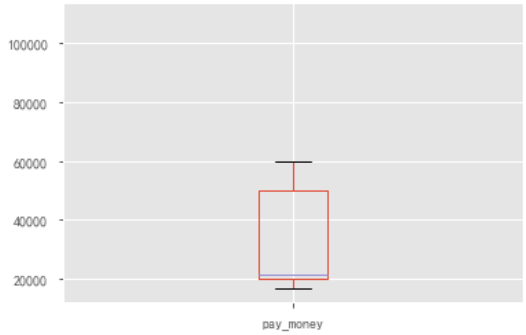
2.1 赔款金额描述统计与分布情况

```
print(data.pay_money.describe())
print('总赔款金额:', data.pay_money.sum())
# 注意应该保留报案号，可以去重复案件，describe已经给出
print('案均金额:', data.pay_money.sum()/ data.id.count())

# 存在极值，用切比雪夫
plt.title('赔款金额分布情况')
data.pay_money[data.pay_money<15000].plot.hist(bins=10)
print('赔款大于>15000的案件数量:', data.pay_money[data.pay_money>15000].count())
```

```
count      1426.000000
mean       2068.582574
std        4968.942240
min          11.200000
25%         349.385000
50%         863.220000
75%        2434.730000
max      109100.000000

总赔款金额: 2949798.75
案均金额: 2068.5825736325387
赔款大于>15000的案件数量: 11
```

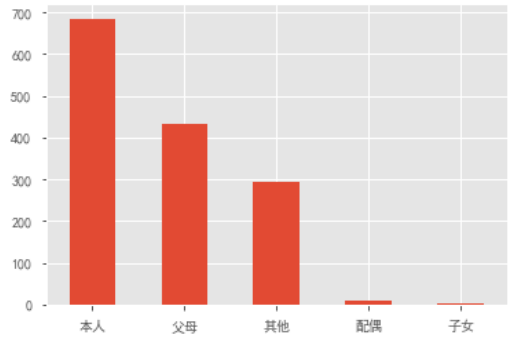



```
# 赔款金额箱线图
data.pay_money[data.pay_money>15000].plot.box()
plt.show()
```



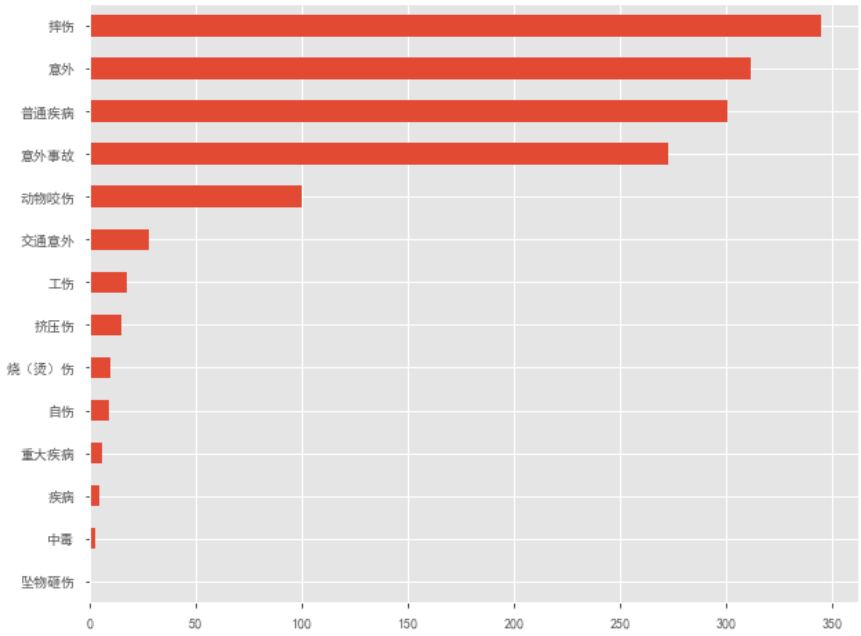
2.2 投保人与被保险人人关系分布情况

```
# 投保人与被保险人关系分布情况
data.relationship_insured.value_counts().plot.bar()
plt.xticks(rotation=0)
plt.show()
```



2.3 出险原因分布情况

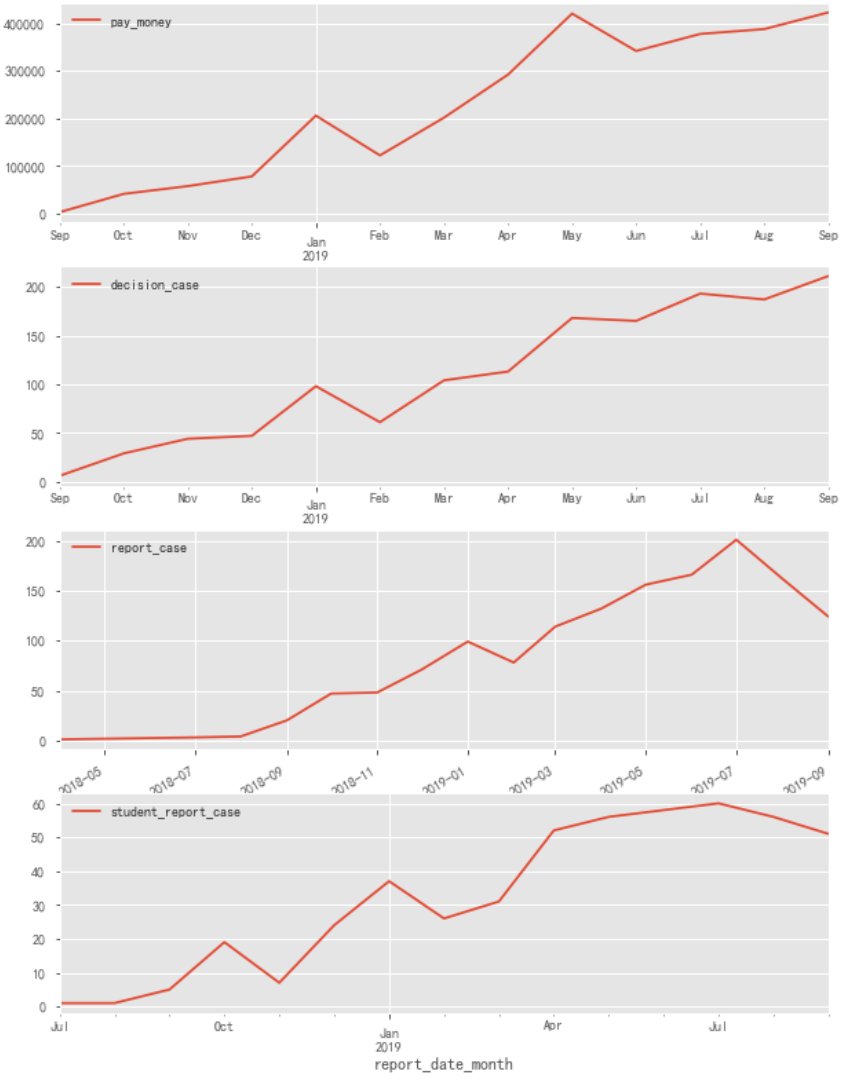
```
plt.figure(figsize=(10,8))
data.case_reason.value_counts().sort_values().plot.barh()
plt.show()
```



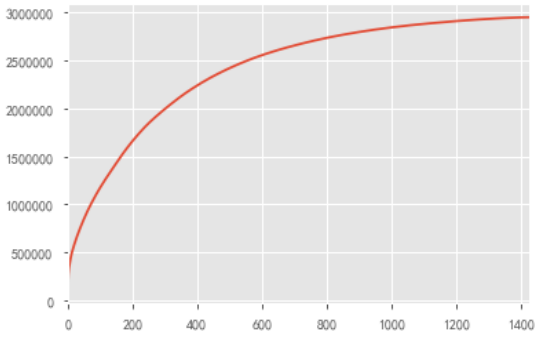
3. 时间维度

3.1 每日赔款金额趋势和赔案数量分布

```
plt.figure(figsize=(10,15))
plt.subplot(411)
data.groupby('decision_time_month').pay_money.sum().plot(label='pay_money')
plt.legend()
plt.subplot(412)
data.groupby('decision_time_month').id.count().plot(label='decision_case')
plt.legend()
plt.subplot(413)
# 2019-7到2019-9有下降，猜测与放暑假有关，进一步钻取学平险报案情况
data.groupby('report_date_month').id.count().plot(label='report_case')
plt.legend()
plt.subplot(414)
data[data.insurance_type=='学平险'].groupby('report_date_month').id.count().plot(label='student_report_case')
plt.legend()
plt.show()
```



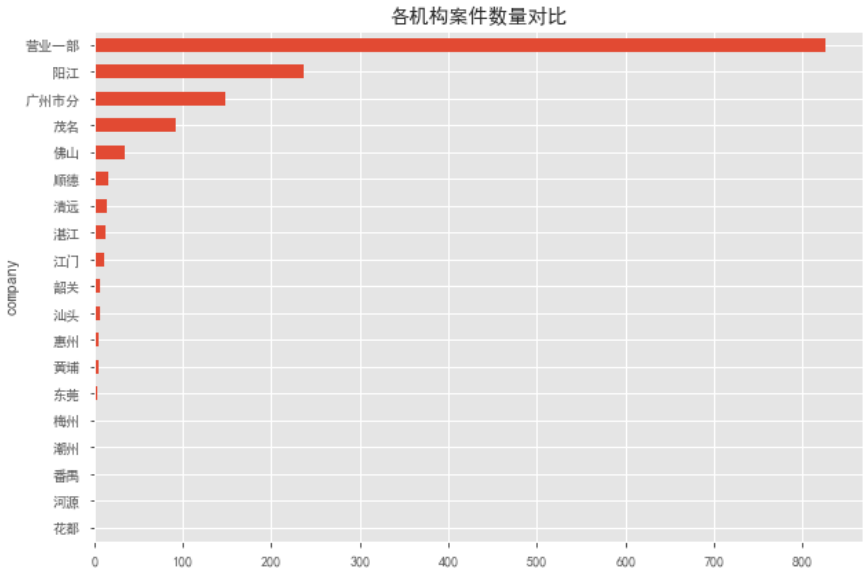

```
# 赔款金额累加折线图 接近二八法则，3:7
data.pay_money.sort_values(ascending=False).cumsum().reset_index(drop=True).plot()
plt.show()
```

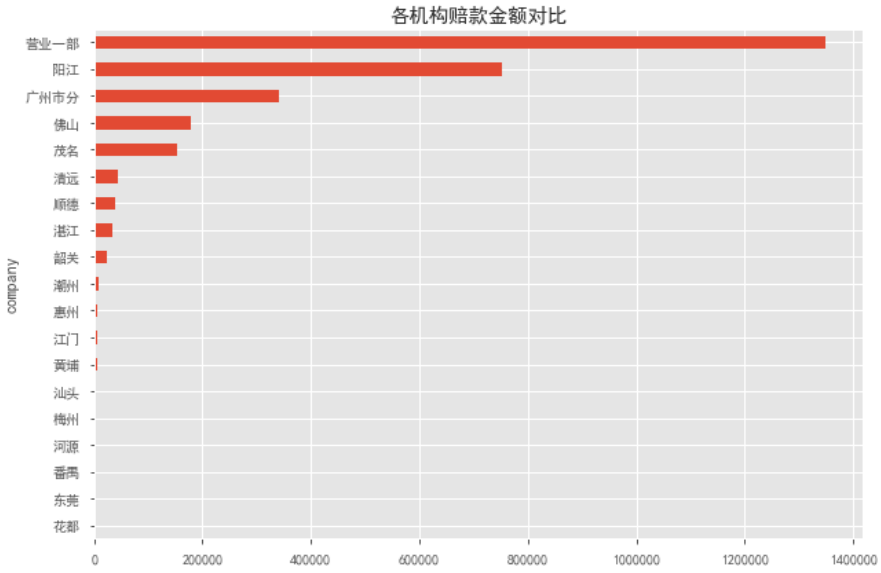


4. 机构维度

4.1 各机构案件数量与赔款金额对比

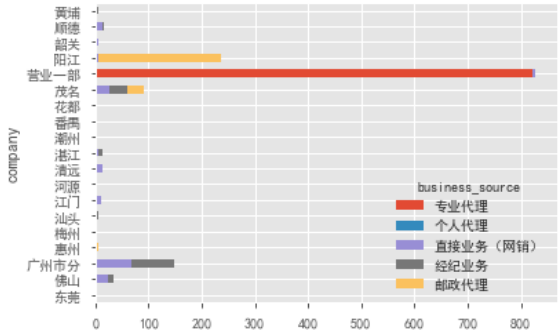
```
plt.figure(figsize=(10,15))
plt.subplot(211)
data.groupby('company').id.count().sort_values().plot.barh()
plt.title('各机构案件数量对比')
plt.subplot(212)
data.groupby('company').pay_money.sum().sort_values().plot.barh()
plt.title('各机构赔款金额对比')
plt.show()
```





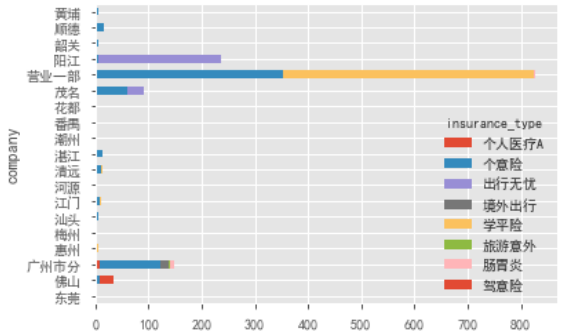
4.2 各机构赔案业务来源占比

```
df = data.pivot_table(index='company', columns='business_source',\
    values='id', aggfunc='count')
plt.figure(figsize=(10,15))
df.plot.barh(stacked=True)
plt.show()
```



4.3 各机构赔案险种占比

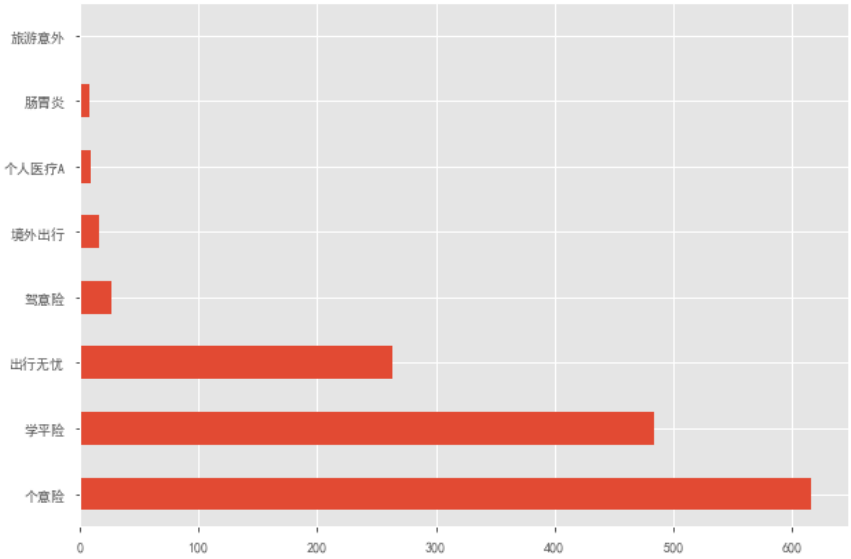
```
plt.figure(figsize=(10,15))
df = data.pivot_table(index='company', columns='insurance_type',\
    values='id', aggfunc='count')
df.plot.barh(stacked=True)
plt.show()
```



5. 险种维度

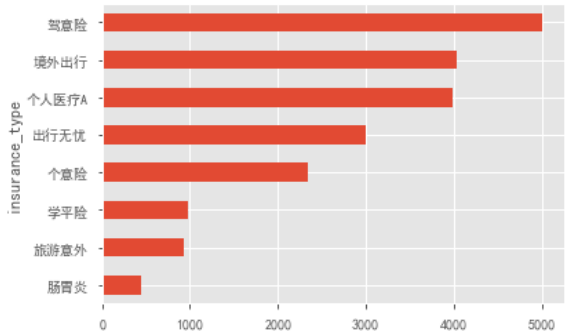
5.1 各险种赔案数量对比

```
plt.figure(figsize=(10,7))
data.insurance_type.value_counts().plot.barh()
plt.show()
```



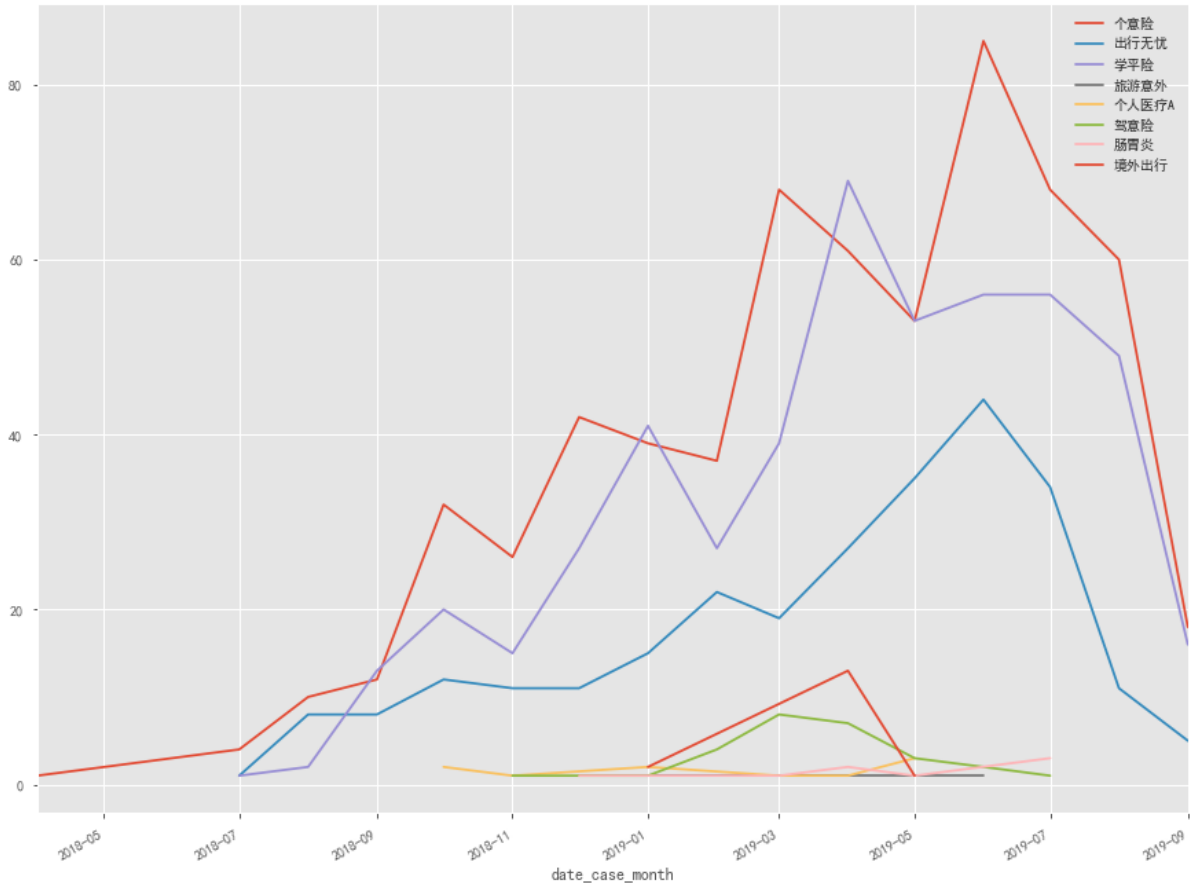
5.2 各险种案均赔款对比

```
df = data.groupby('insurance_type').apply(lambda x : x.pay_money.sum()/ x.pay_money.count())
df.sort_values().plot.barh()
plt.show()
```



5.3 各险种出险案件数量随时间变化趋势

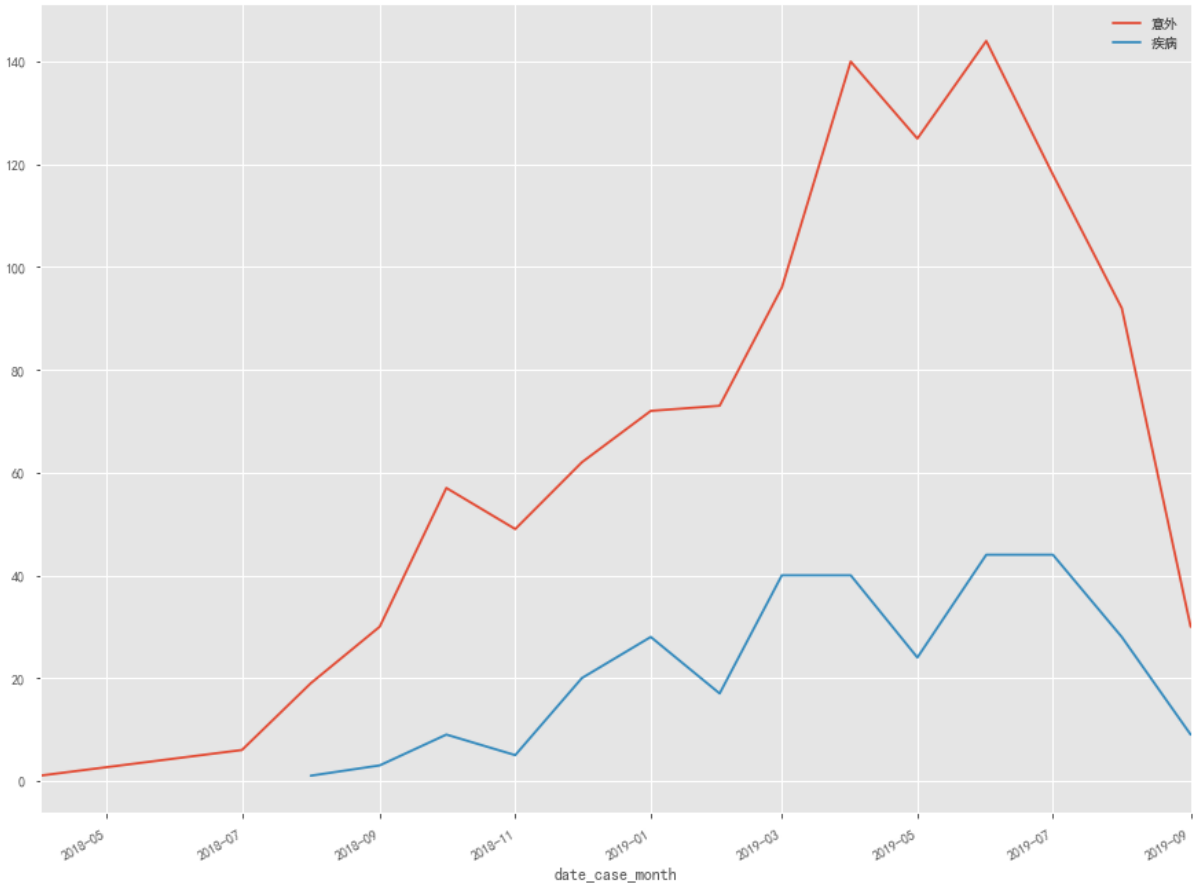
```
insurance_type = ['个意险', '出行无忧', '学平险', '旅游意外', '个人医疗A', '驾意险', '肠胃炎', '境外出行']
num = data.insurance_type.nunique()
df = data.groupby(['insurance_type', 'date_case_month']).count().id

plt.figure(figsize=(15,12))
for i in range(num):
    df[insurance_type[i]].plot(label=insurance_type[i])
plt.legend()
plt.show()
```



5.4 各出险原因出险案件数量时间趋势

```
# 将出险原因映射为 疾病、意外、其他
dict = {
    '摔伤'        : '意外',
    '其他'        : '其他',
    '意外事故'    : '意外',
    '普通疾病'    : '疾病',
    '动物咬伤'    : '意外',
    '意外'        : '意外',
    '交通意外'    : '意外',
    '工伤'        : '意外',
    '挤压伤'      : '意外',
    '烧（烫）伤' : '意外',
    '自伤'        : '意外',
    '重大疾病'    : '疾病',
    '疾病'        : '疾病',
    '中毒'        : '意外',
    '坠物砸伤'    : '意外',
}
data['case_reason_type'] = data.case_reason.apply(lambda x : dict[x])
```

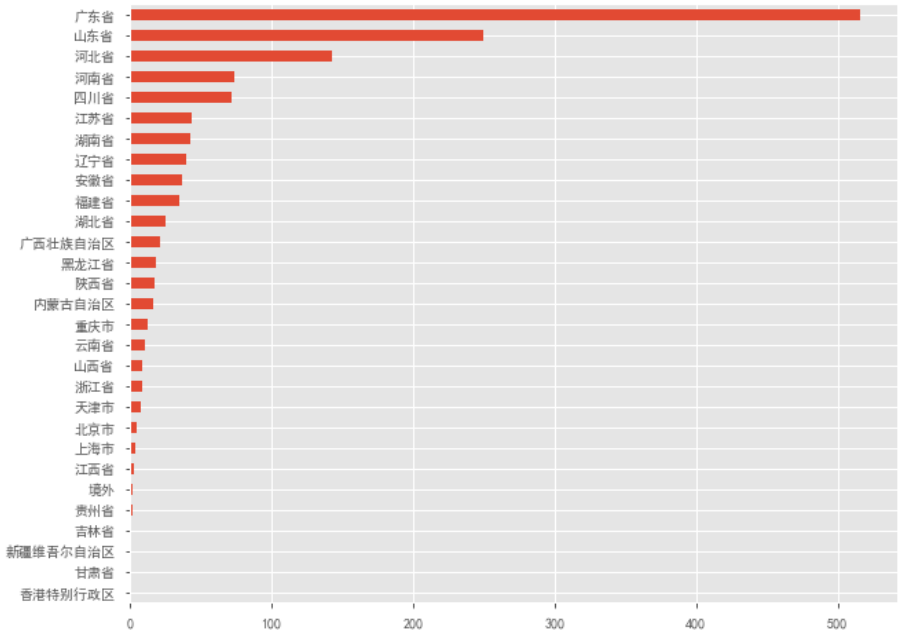
```
case_reason_type = ['意外', '疾病', '其他']
num = data.case_reason_type.nunique()
df = data.groupby(['case_reason_type', 'date_case_month']).count().id

plt.figure(figsize=(15,12))
for i in range(num):
    df[case_reason_type[i]].plot(label=case_reason_type[i])
plt.legend()
plt.show()
```



6. 出险省份维度

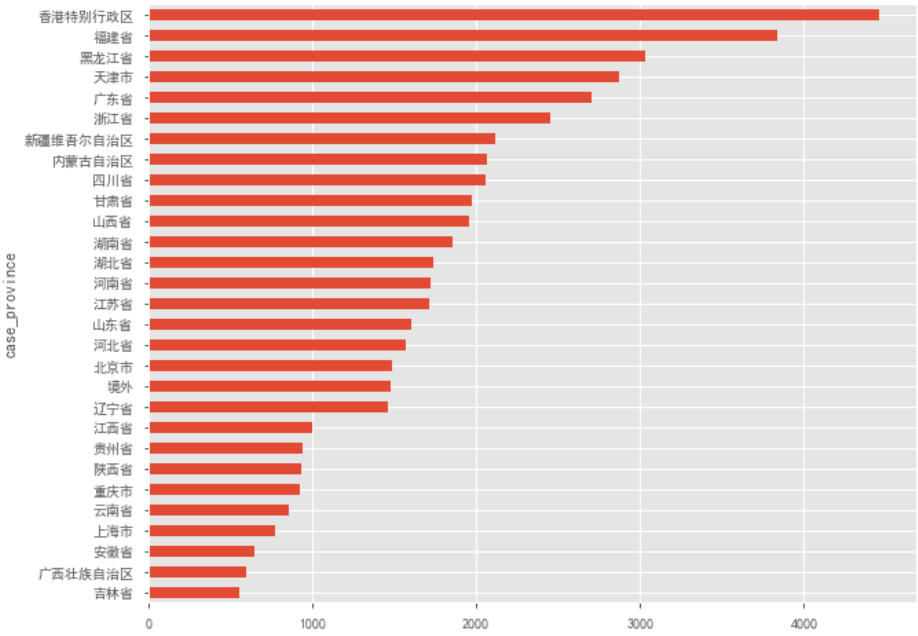
6.1 各省份案件数量分布

```
plt.figure(figsize=(10,8))
data.case_province.value_counts().sort_values().plot.barh()
plt.show()
```



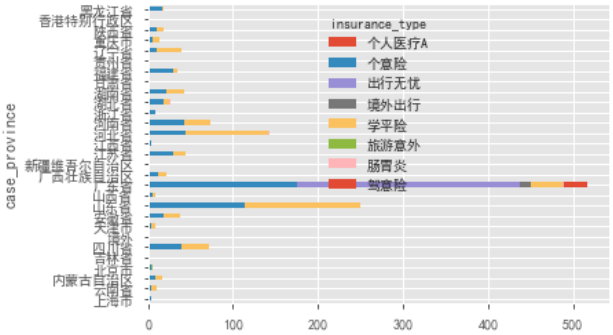
6.2 各省份案均赔款分布

```
plt.figure(figsize=(10,8))
data.groupby('case_province').apply(lambda x : x.pay_money.sum()/x.id.count()).sort_values().plot.barh()
plt.show()
```



6.3 各省份险种构成

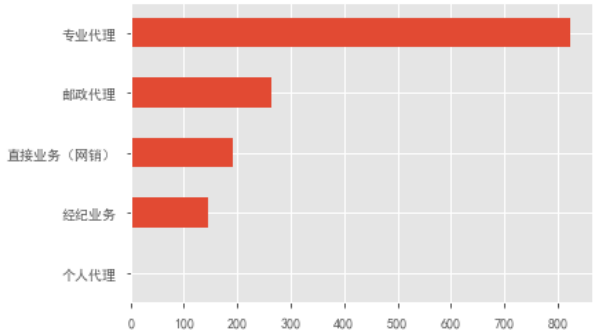
```
plt.figure(figsize=(10,50))
data.pivot_table(index='case_province', columns='insurance_type', values='id', aggfunc='count').plot.barh(stacked=True)
plt.show()
```



7. 业务来源维度

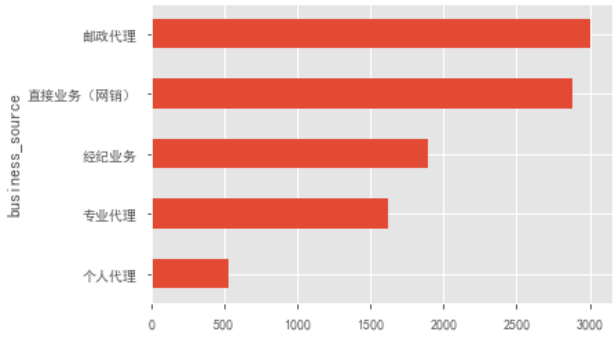
7.1 各业务来源赔案数量对比

```
data.business_source.value_counts().sort_values().plot.barh()
plt.show()
```



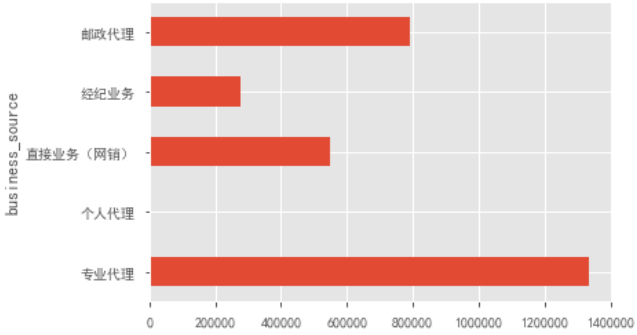
7.2 各业务来源案均赔款对比

```
data.groupby('business_source').apply(lambda x : x.pay_money.sum()/x.id.count()).sort_values().plot.barh()
plt.show()
```



7.3 各业务来源赔款总额对比

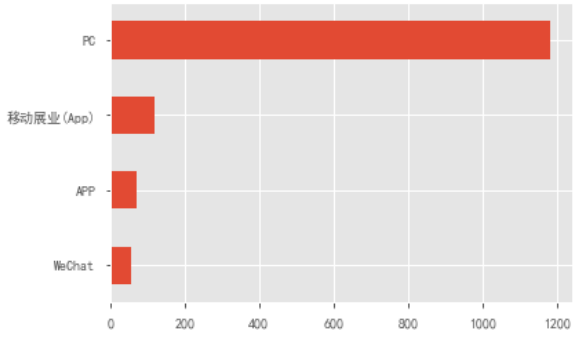
```
data.groupby('business_source').sum().pay_money.plot.barh()
plt.show()
```



8. 终端维度

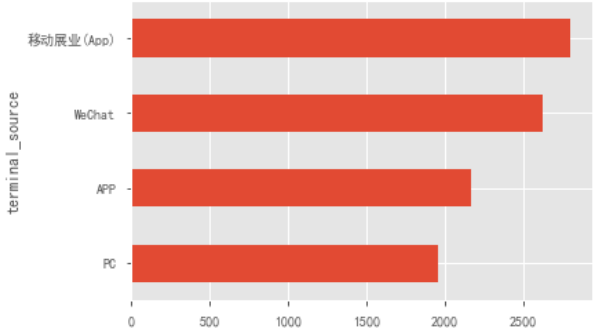
8.1 各终端来源赔案数量对比

```
data.terminal_source.value_counts().sort_values().plot.barh()
plt.show()
```



8.2 各终端来源案均赔款对比

```
data.groupby('terminal_source').apply(lambda x : x.pay_money.sum()/x.id.count()).sort_values().plot.barh()
plt.show()
```



9. 年龄维度

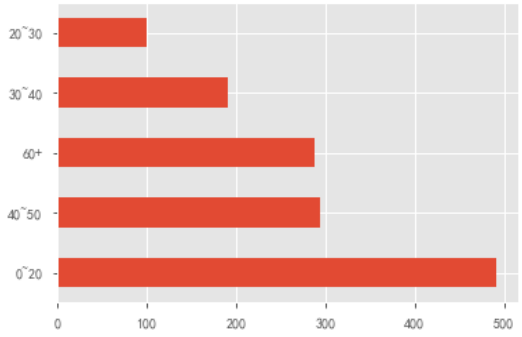
```
# 对年龄分析，去掉年龄为空值的数据
data_age = data.dropna(axis=0,subset=["age"]) #指定列，按行删除
data_age.isnull().sum()
print(data_age.shape)
print(data_age.age.describe())
```

```
count    1364.000000
mean       31.875367
std        20.365170
min         5.000000
25%         8.000000
50%        35.000000
75%        49.000000
max        67.000000
```

```
# 划分年龄段
data_age['age_label'] = pd.cut(data_age.age, bins=[0,20,30,40,50,150], labels=['0~20', '20~30',\
                                                    '30~40','40~50','60+'])
```

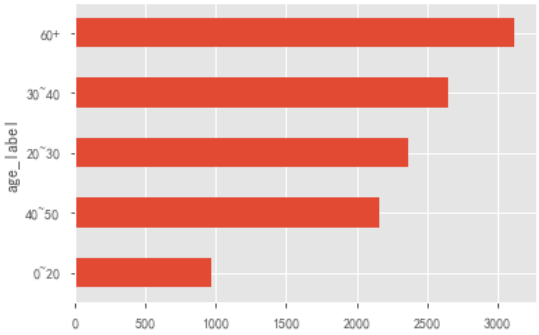
9.1 各年龄段案件数量对比

```
data_age.age_label.value_counts().plot.barh()
plt.show()
```

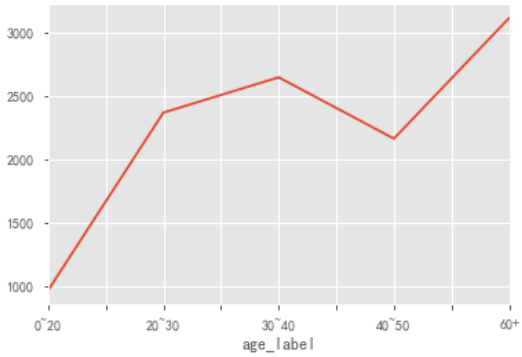


9.2 各年龄段案均赔款对比

```
# 柱形图
data_age.groupby('age_label').apply(lambda x : x.pay_money.sum()/x.id.count()).sort_values().plot.barh()
plt.show()
```

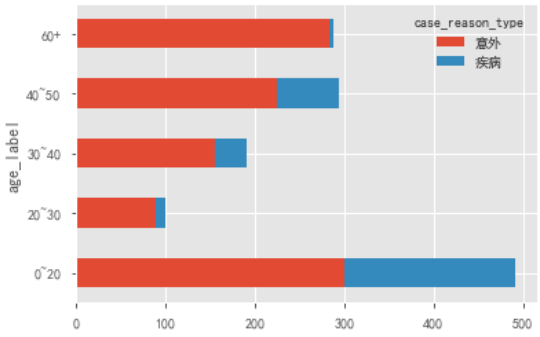



```
# 条形图
data_age.groupby('age_label').apply(lambda x : x.pay_money.sum()/x.id.count()).plot()
plt.show()
```



9.3 各年龄段赔案出险原因分布

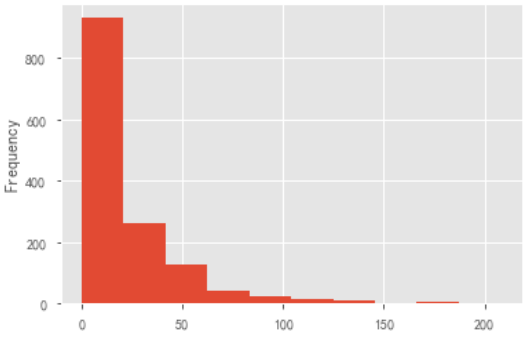
```
plt.figure(figsize=(10,50))
data_age.pivot_table(index='age_label', columns='case_reason_type', values='id', aggfunc='count').plot.barh(stacked=True)
plt.show()
```



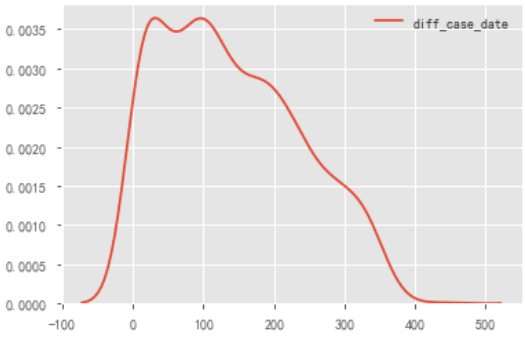
10. 相关性探索

10.1 各类日期差值分布情况

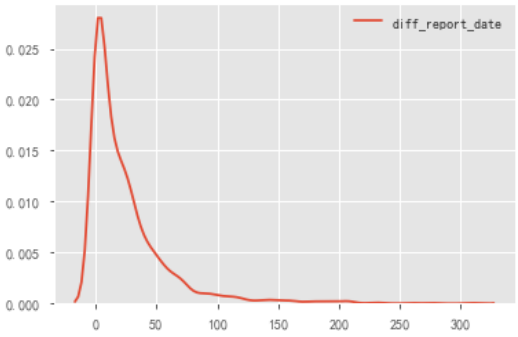
```
# 赔款日期与报案日期之差的分布
data.diff_dic_data.plot.hist()
plt.show()
```




```
# 出险日期与保单起期之差的分布
sns.kdeplot(data.diff_case_date)
plt.show()
```

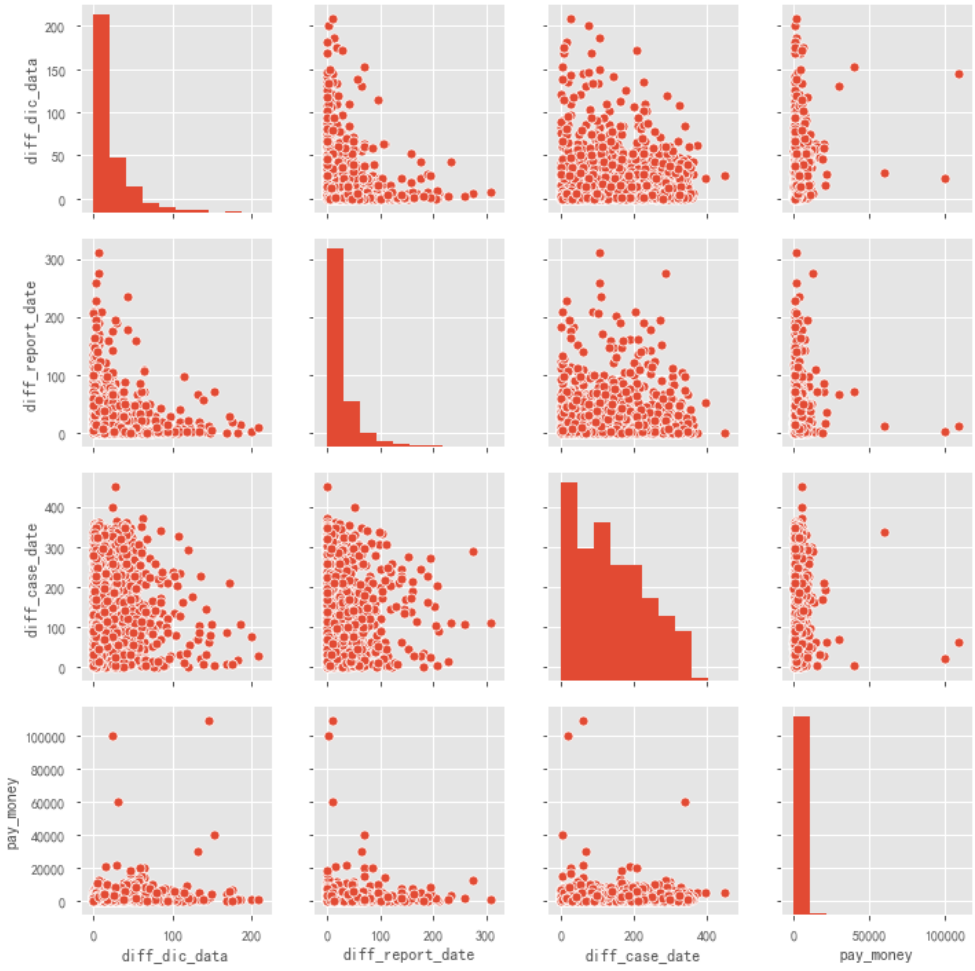


```
# 报案日期与出险日期之差的分布
sns.kdeplot(data.diff_report_date)
plt.show()
```



10.2 各日期差值与赔案金额相关性分析

```
sns.pairplot(data[['diff_dic_data', 'diff_report_date', 'diff_case_date', 'pay_money']])
plt.show()
```




```
# 导出数据
data_age.to_csv('data_age.csv',encoding='utf_8_sig', index=0)
data_age.to_excel('data_age.xlsx', index=0)
```