In [1]:
```python
import scipy
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import statsmodels.api as sm

import warnings
warnings.filterwarnings("ignore")
```

Data Outline

In [2]:
```python
data=pd.read_excel("variable in TAZ.xlsx",index_col=[0])
data.head()
```

Out[2]:

FID	Open Area	Traffic Area	Residential Area	Commercial Area	POI Entropy	POI Richness	POI Simpson	POI Gini Coefficient	Road Density	Population Density	Distan to Covi Breako Locati
0	71.645258	4.294096	13.019560	11.041090	5.091138	9	4.198463	0.639767	9.735901	5.580306	14.6319
1	69.738119	6.621425	13.815199	9.825255	6.259978	11	5.198540	0.560863	10.755058	8.039338	13.6522
2	52.448830	10.773829	5.802397	30.974951	5.895796	8	4.880866	0.583916	25.677644	24.438971	12.5832
3	81.918734	1.370870	4.296954	12.413444	4.434488	7	3.417664	0.687313	2.469991	7.604859	13.0013
4	75.649980	10.455473	4.733141	9.161406	3.371602	5	2.578947	0.753247	11.680421	10.101772	12.4896

In [3]:
```python
data.describe()
```

Out[3]:

	Open Area	Traffic Area	Residential Area	Commercial Area	POI Entropy	POI Richness	POI Simpson	POI Gini Coefficient	Road Density	Populati Dens
count	619.000000	619.000000	619.000000	619.000000	619.000000	619.000000	619.000000	619.000000	619.000000	619.0000
mean	30.343190	10.494497	42.866562	16.294614	6.706165	10.119548	5.458242	0.521041	33.438890	136.0596
std	26.084727	6.332472	25.562116	11.299997	1.491362	1.560616	1.522791	0.107144	25.350221	97.1283
min	0.000000	0.000000	0.000000	0.000000	1.000000	1.000000	1.000000	0.248521	0.000000	1.5572
25%	8.036251	6.683807	20.063942	8.064498	5.888762	10.000000	4.428854	0.449962	16.392934	55.1907
50%	23.696392	9.509252	42.831070	14.091025	6.995776	11.000000	5.594995	0.508385	27.730754	127.3770
75%	50.620470	12.901970	64.753794	21.873750	7.731973	11.000000	6.528810	0.583542	41.158415	187.9509
max	100.000000	73.583656	100.000000	77.899569	9.917969	11.000000	9.198390	0.909091	198.452300	537.0502

In [4]:
```python
data.columns
```

Out[4]:
```
Index(['Open Area', 'Traffic Area', 'Residential Area', 'Commercial Area',
       'POI Entropy', 'POI Richness', 'POI Simpson', 'POI Gini Coefficient',
       'Road Density', 'Population Density',
       'Distance to Covid-19 Breakout Location', 'Risk Level'],
      dtype='object')
```

In [5]:
```python
cols=data.columns.tolist()

fig,ax=plt.subplots(3,4,figsize=(15,8))
plt.subplots_adjust(wspace =0.2, hspace =0.3)
axs=ax.flatten()
for col in cols:
    _=axs[cols.index(col)].hist(data[col],rwidth=0.5)
    axs[cols.index(col)].set_xlabel(col,fontsize=11)

plt.show()
```

In [6]:
```python
corr_list=[]
for i in range(len(cols)):
    for j in range(len(cols)):
        x=data[cols[i]]
        y=data[cols[j]]
        correlation_pearson, pvalue_pearson = scipy.stats.pearsonr(x, y)
        corr_list.append({"x":cols[i],"y":cols[j],"corr":correlation_pearson,"p-value":pvalu
e_pearson})

corr_df=pd.DataFrame(corr_list)

print(" r：")
corr_df1=corr_df.pivot_table(index="x",columns="y",values="corr")
corr_df1=corr_df1.reindex(index=cols,columns=cols)
corr_df1.to_csv("pearson_corr.csv")
display(corr_df1)

print(" p：")
corr_df2=corr_df.pivot_table(index="x",columns="y",values="p-value")
corr_df2=corr_df2.reindex(index=cols,columns=cols)
corr_df2.to_csv("pearson_corr_pvalue.csv")
display(corr_df2)
```

r：

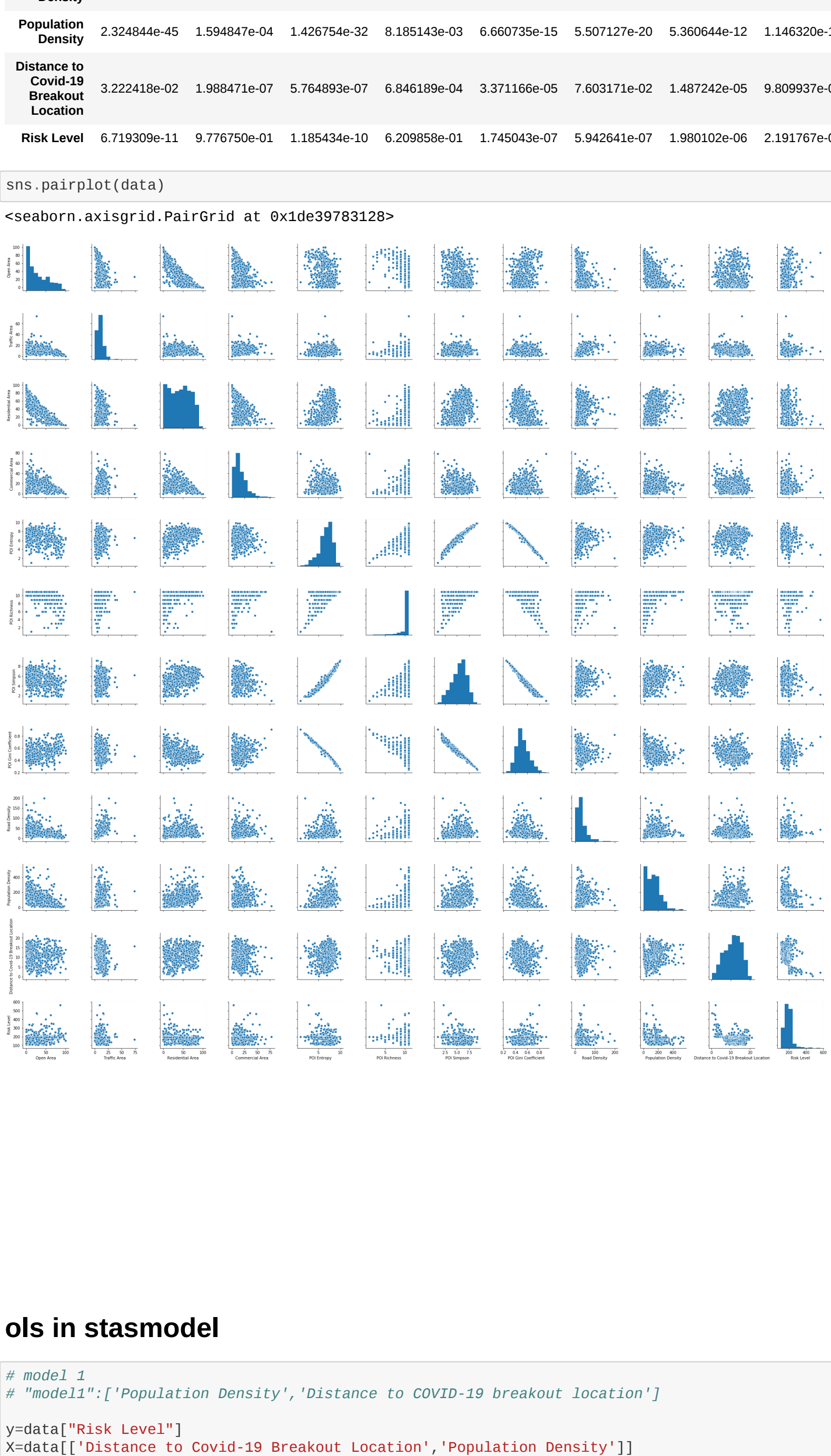
y / x	Open Area	Traffic Area	Residential Area	Commercial Area	POI Entropy	POI Richness	POI Simpson	POI Gini Coefficient	Road Density	Population Density
Open Area	1.000000	-0.282023	-0.844146	-0.240805	-0.390280	-0.506744	-0.328219	0.370190	-0.300703	-0.526110
Traffic Area	-0.282023	1.000000	-0.084538	0.281876	0.008710	0.097651	0.009007	-0.011634	0.501339	0.151195
Residential Area	-0.844146	-0.084538	1.000000	-0.266137	0.471444	0.439070	0.423015	-0.456373	0.107101	0.452489
Commercial Area	-0.240805	0.281876	-0.266137	1.000000	-0.170362	0.121880	-0.204244	0.184286	0.170950	0.106199
POI Entropy	-0.390280	0.008710	0.471444	-0.170362	1.000000	0.619359	0.973512	-0.992663	0.054265	0.306204
POI Richness	-0.506744	0.097651	0.439070	0.121880	0.619359	1.000000	0.482504	-0.579733	0.033140	0.356479
POI Simpson	-0.328219	0.009007	0.423015	-0.204244	0.973512	0.482504	1.000000	-0.980355	0.049994	0.272451
POI Gini Coefficient	0.370190	-0.011634	-0.456373	0.184286	-0.992663	-0.579733	-0.980355	1.000000	-0.048731	-0.303625
Road Density	-0.300703	0.501339	0.107101	0.170950	0.054265	0.033140	0.049994	-0.048731	1.000000	0.211977
Population Density	-0.526110	0.151195	0.452489	0.106199	0.306204	0.356479	0.272451	-0.303625	0.211977	1.000000
Distance to Covid-19										

p：

y / x	Open Area	Traffic Area	Residential Area	Commercial Area	POI Entropy	POI Richness	POI Simpson	POI Gir Coefficien
Open Area	0.000000e+00	8.800190e-13	3.399519e-169	1.290493e-09	5.904662e-24	1.090873e-41	5.167163e-17	1.537782e-2
Traffic Area	8.800190e-13	0.000000e+00	3.548307e-02	9.053225e-13	8.287764e-01	1.508127e-02	8.230330e-01	7.726746e-0
Residential Area	3.399519e-169	3.548307e-02	0.000000e+00	1.699390e-11	1.423510e-35	1.471727e-30	2.893681e-28	3.585718e-3
Commercial Area	1.290493e-09	9.053225e-13	1.699390e-11	0.000000e+00	2.033605e-05	2.385125e-03	2.969371e-07	3.925058e-0
POI Entropy	5.904662e-24	8.287764e-01	1.423510e-35	2.033605e-05	0.000000e+00	7.671093e-67	0.000000e+00	0.000000e+0
POI Richness	1.090873e-41	1.508127e-02	1.471727e-30	2.385125e-03	7.671093e-67	0.000000e+00	2.057379e-37	7.289094e-5
POI Simpson	5.167163e-17	8.230330e-01	2.893681e-28	2.969371e-07	0.000000e+00	2.057379e-37	0.000000e+00	0.000000e+0
POI Gini Coefficient	1.537782e-21	7.726746e-01	3.585718e-33	3.925058e-06	0.000000e+00	7.289094e-57	0.000000e+00	0.000000e+0
Road Density	2.107377e-14	1.049905e-40	7.654585e-03	1.901973e-05	1.775443e-01	4.104729e-01	2.141989e-01	2.260170e-0
Population Density	2.324844e-45	1.594847e-04	1.426754e-32	8.185143e-03	6.660735e-15	5.507127e-20	5.360644e-12	1.146320e-1
Distance to Covid-19 Breakout Location	3.222418e-02	1.988471e-07	5.764893e-07	6.846189e-04	3.371166e-05	7.603171e-02	1.487242e-05	9.809937e-0
Risk Level	6.719309e-11	9.776750e-01	1.185434e-10	6.209858e-01	1.745043e-07	5.942641e-07	1.980102e-06	2.191767e-0

In [7]:
```python
sns.pairplot(data)
```

Out[7]: <seaborn.axisgrid.PairGrid at 0x1de39783128>



ols in stasmodel

In [8]:
```python
# model 1
# "model1":['Population Density','Distance to COVID-19 breakout location']

y=data["Risk Level"]
X=data[['Distance to Covid-19 Breakout Location','Population Density']]
X = sm.add_constant(X) # 添加常数项
model = sm.OLS(y,X)
results = model.fit()
results.summary()

# ols 回归模型一般看这么几个指标
# Adj. R-squared            -调整后的拟合优度 因为 r2 会随着 自变量个数增加而增大 有缺陷，所以有这个指标
# Adj. coef                 -各个自变量在回归模型中的系数
# Adj. std err              -系数的标准误
# t                         -系数的t检验统计量
```

Out[8]:

OLS Regression Results

Dep. Variable:	Risk Level	R-squared:	0.167
Model:	OLS	Adj. R-squared:	0.165
Method:	Least Squares	F-statistic:	61.84
Date:	Thu, 27 Aug 2020	Prob (F-statistic):	3.35e-25
Time:	18:15:57	Log-Likelihood:	-3212.7
No. Observations:	619	AIC:	6431.
Df Residuals:	616	BIC:	6445.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	239.6807	5.047	47.491	0.000	229.769	249.592
Distance to Covid-19 Breakout Location	-3.4914	0.420	-8.316	0.000	-4.316	-2.667
Population Density	-0.0962	0.019	-5.180	0.000	-0.133	-0.060

Omnibus:	332.587	Durbin-Watson:	1.750
Prob(Omnibus):	0.000	Jarque-Bera (JB):	3561.330
Skew:	2.156	Prob(JB):	0.00
Kurtosis:	13.931	Cond. No.	484.

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

In [9]:
```python
# model 2
# "model1":['Population Density','Distance to COVID-19 breakout location', 'Road density']

y=data["Risk Level"]
X=data[['Distance to Covid-19 Breakout Location','Population Density','Road Density']]
X = sm.add_constant(X) # 添加常数项
model = sm.OLS(y,X)
results = model.fit()
results.summary()
```

Out[9]:

OLS Regression Results

Dep. Variable:	Risk Level	R-squared:	0.167
Model:	OLS	Adj. R-squared:	0.163
Method:	Least Squares	F-statistic:	41.22
Date:	Thu, 27 Aug 2020	Prob (F-statistic):	2.80e-24
Time:	18:19:16	Log-Likelihood:	-3212.6
No. Observations:	619	AIC:	6433.
Df Residuals:	615	BIC:	6451.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	238.9630	5.407	44.196	0.000	228.345	249.581
Distance to Covid-19 Breakout Location	-3.4876	0.420	-8.298	0.000	-4.313	-2.662
Population Density	-0.0977	0.019	-5.138	0.000	-0.135	-0.060
Road Density	0.0263	0.071	0.372	0.710	-0.113	0.165

Omnibus:	332.230	Durbin-Watson:	1.750
Prob(Omnibus):	0.000	Jarque-Bera (JB):	3552.148
Skew:	2.153	Prob(JB):	0.00
Kurtosis:	13.917	Cond. No.	526.

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

In [10]:
```python
# model 3
# "model1":['Population Density','Distance to COVID-19 breakout location', 'POI Entropy','PO
I Richness','POI Simpson']

y=data["Risk Level"]
X=data[['Distance to Covid-19 Breakout Location','Population Density','POI Entropy','POI Ric
hness','POI Simpson']]
X = sm.add_constant(X) # 添加常数项
model = sm.OLS(y,X)
results = model.fit()
results.summary()
```

Out[10]:

OLS Regression Results

Dep. Variable:	Risk Level	R-squared:	0.182
Model:	OLS	Adj. R-squared:	0.176
Method:	Least Squares	F-statistic:	27.32
Date:	Thu, 27 Aug 2020	Prob (F-statistic):	5.38e-25
Time:	18:22:35	Log-Likelihood:	-3207.1
No. Observations:	619	AIC:	6426.
Df Residuals:	613	BIC:	6453.
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	275.7921	12.321	22.383	0.000	251.595	299.989
Distance to Covid-19 Breakout Location	-3.4484	0.421	-8.191	0.000	-4.275	-2.622
Population Density	-0.0729	0.020	-3.684	0.000	-0.112	-0.034
POI Entropy	-4.0281	7.679	-0.525	0.600	-19.108	11.052
POI Richness	-2.3464	1.955	-1.200	0.231	-6.187	1.494
POI Simpson	2.0149	6.748	0.299	0.765	-11.237	15.267

Omnibus:	311.290	Durbin-Watson:	1.752
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2854.821
Skew:	2.033	Prob(JB):	0.00
Kurtosis:	12.703	Cond. No.	1.20e+03

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.2e+03. This might indicate that there are
strong multicollinearity or other numerical problems.

In [11]:
```python
# model 4
# "model1":['Population Density','Distance to COVID-19 breakout location', 'POI Entropy','PO
I Richness','POI Simpson','POI Gini Coefficient']

y=data["Risk Level"]
X=data[['Distance to Covid-19 Breakout Location','Population Density','POI Entropy','POI Ric
hness','POI Simpson','POI Gini Coefficient']]
X = sm.add_constant(X) # 添加常数项
model = sm.OLS(y,X)
results = model.fit()
results.summary()
```

Out[11]:

OLS Regression Results

Dep. Variable:	Risk Level	R-squared:	0.182
Model:	OLS	Adj. R-squared:	0.174
Method:	Least Squares	F-statistic:	22.73
Date:	Thu, 27 Aug 2020	Prob (F-statistic):	2.98e-24
Time:	18:24:27	Log-Likelihood:	-3207.1
No. Observations:	619	AIC:	6428.
Df Residuals:	612	BIC:	6459.
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	284.5578	153.565	1.853	0.064	-17.020	586.136
Distance to Covid-19 Breakout Location	-3.4502	0.422	-8.168	0.000	-4.280	-2.621
Population Density	-0.0729	0.020	-3.678	0.000	-0.112	-0.034
POI Entropy	-4.5168	11.485	-0.393	0.694	-27.071	18.037
POI Richness	-2.3399	1.960	-1.194	0.233	-6.190	1.510
POI Simpson	1.8600	7.275	0.256	0.798	-12.426	16.146
POI Gini Coefficient	-8.9826	156.857	-0.057	0.954	-317.025	299.060

Omnibus:	311.325	Durbin-Watson:	1.752
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2855.822
Skew:	2.033	Prob(JB):	0.00
Kurtosis:	12.705	Cond. No.	2.12e+04

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.12e+04. This might indicate that there are
strong multicollinearity or other numerical problems.

In [13]:
```python
# model 5
# "model1":['Population Density','Distance to COVID-19 breakout location', 'POI Entropy','PO
I Richness','POI Simpson','Open Area','Traffic Area','Residential Area','Commercial Area']

y=data["Risk Level"]
X=data[['Distance to Covid-19 Breakout Location','Population Density','POI Entropy','POI Ric
hness','POI Simpson','Open Area','Traffic Area','Residential Area','Commercial Area']]
X = sm.add_constant(X) # 添加常数项
model = sm.OLS(y,X)
results = model.fit()
results.summary()
```

Out[13]:

OLS Regression Results

Dep. Variable:	Risk Level	R-squared:	0.203
Model:	OLS	Adj. R-squared:	0.192
Method:	Least Squares	F-statistic:	17.28
Date:	Thu, 27 Aug 2020	Prob (F-statistic):	1.46e-25
Time:	18:29:17	Log-Likelihood:	-3199.0
No. Observations:	619	AIC:	6418.
Df Residuals:	609	BIC:	6462.
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-5.102e+04	2.05e+04	-2.489	0.013	-9.13e+04	-1.08e+04
Distance to Covid-19 Breakout Location	-3.3988	0.437	-7.785	0.000	-4.256	-2.541
Population Density	-0.0482	0.022	-2.217	0.027	-0.091	-0.005
POI Entropy	-3.8674	7.707	-0.502	0.616	-19.003	11.268
POI Richness	-0.9501	2.016	-0.471	0.638	-4.908	3.008
POI Simpson	1.6045	6.737	0.238	0.812	-11.625	14.834
Open Area	513.0184	204.975	2.503	0.013	110.474	915.563
Traffic Area	512.7808	204.989	2.501	0.013	110.209	915.353
Residential Area	512.7794	204.979	2.502	0.013	110.228	915.330
Commercial Area	512.6415	204.967	2.501	0.013	110.113	915.170

Omnibus:	320.985	Durbin-Watson:	1.759
Prob(Omnibus):	0.000	Jarque-Bera (JB):	3044.999
Skew:	2.103	Prob(JB):	0.00
Kurtosis:	13.019	Cond. No.	2.08e+06

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.08e+06. This might indicate that there are
strong multicollinearity or other numerical problems.

In [ ]: