SARIMA Time Series Forecast Crime

April 15, 2017

In [2]:
```
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.ticker as tkr
from vertica_util import *
from scipy import stats
import math
import statsmodels
import statsmodels.api as sm
from dateutil.relativedelta import relativedelta
import datetime
%matplotlib inline
```

In [3]:
```
df = pd.read_csv('monthly_crime.csv', sep=',', header=0)
```

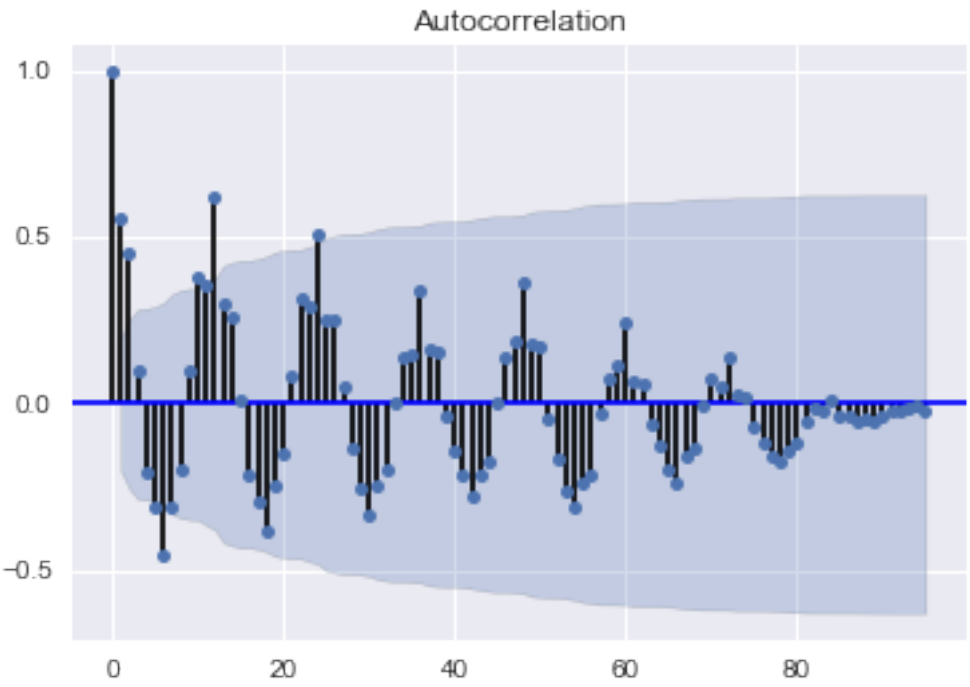
In [4]:
```
df.head()
```

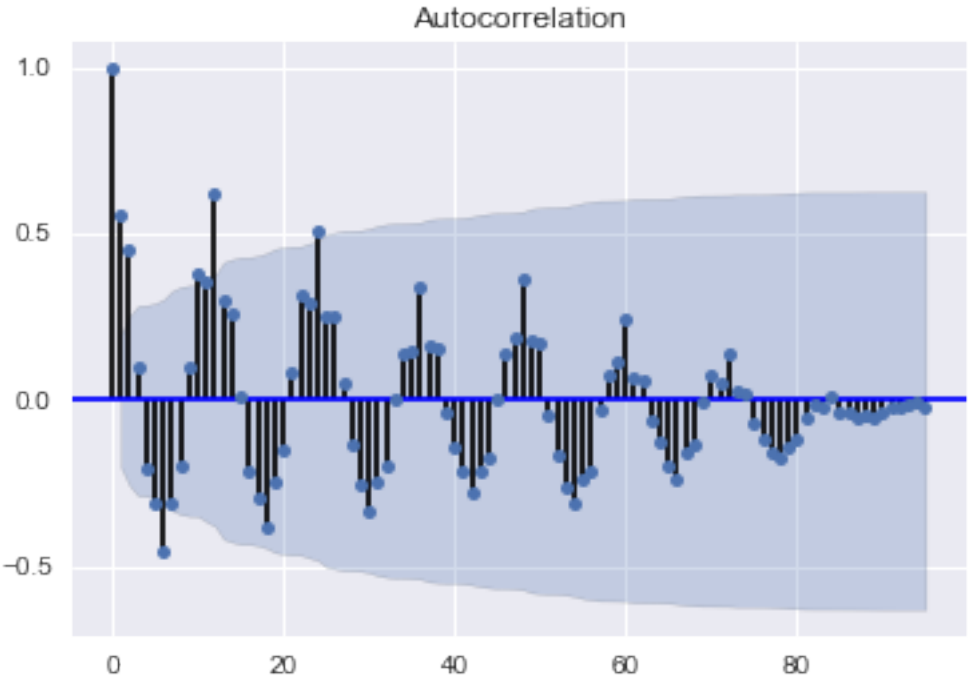
Out[4]:

	Month	Count
0	Jan-08	45391
1	Feb-08	39741
2	Mar-08	44187
3	Apr-08	43086
4	May-08	45261

In [5]:
```
sm.graphics.tsa.plot_acf(df.Count)
```

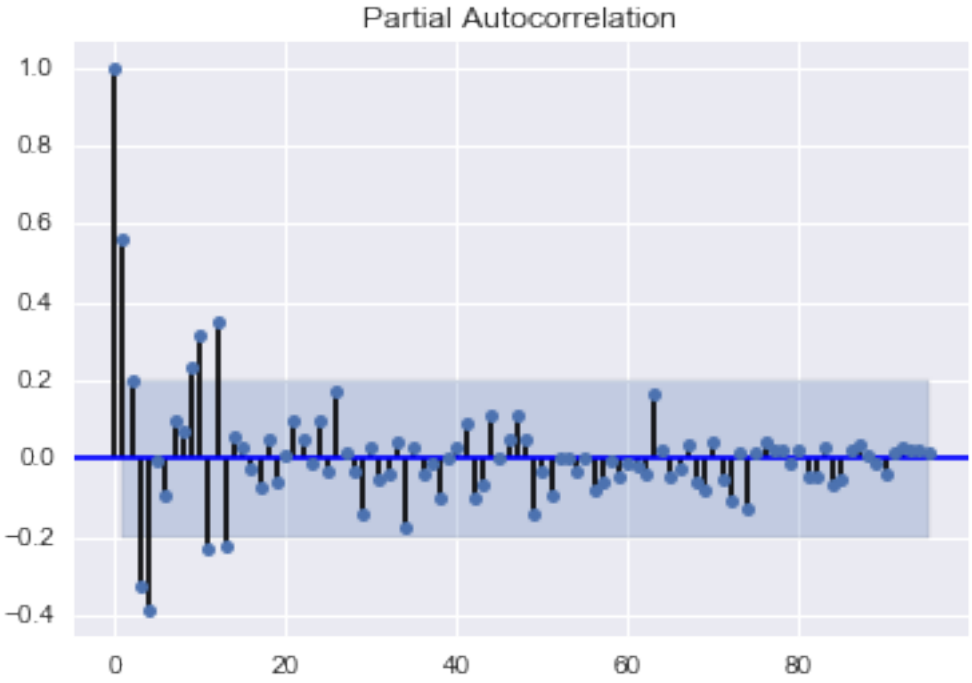
Out[5]:

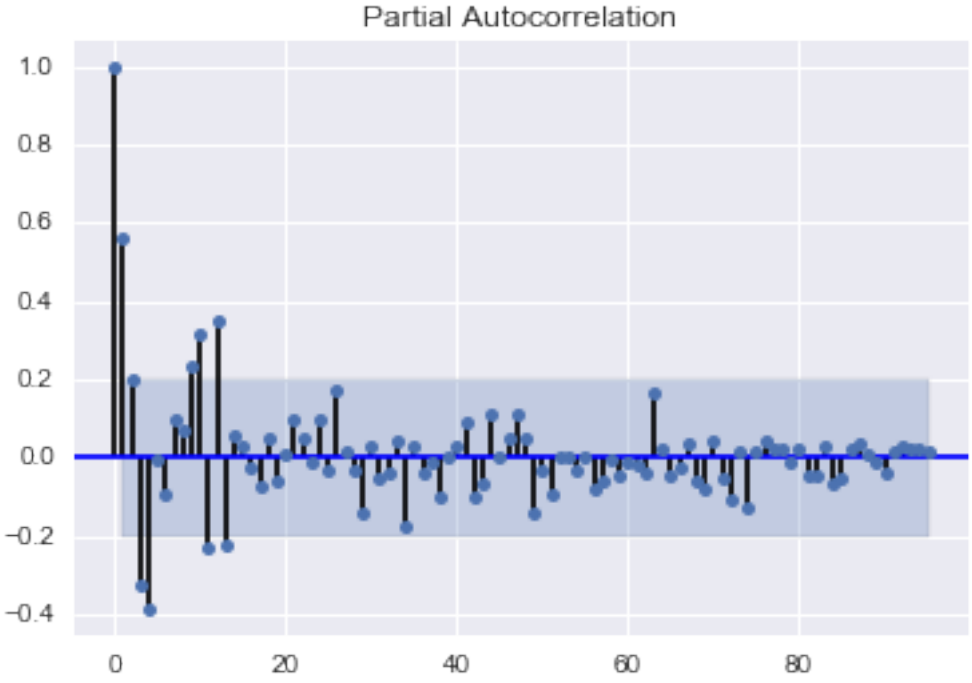




In [6]: `sm.graphics.tsa.plot_pacf(df.Count)`

Out[6]:


Partial Autocorrelation
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
0
20
40
60
80


Partial Autocorrelation
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
0
20
40
60
80

```
In [7]: model = sm.tsa.statespace.SARIMAX(df['Count'], type='ct',
                                          seasonal_order=(4, 1, 1, 12))
        results = model.fit()
        print results.summary()

/Users/asimonoff/anaconda/lib/python2.7/site-packages/statsmodels/base/model.py:496: Convergence
  "Check mle_retvals", ConvergenceWarning)
```

Statespace Model Results

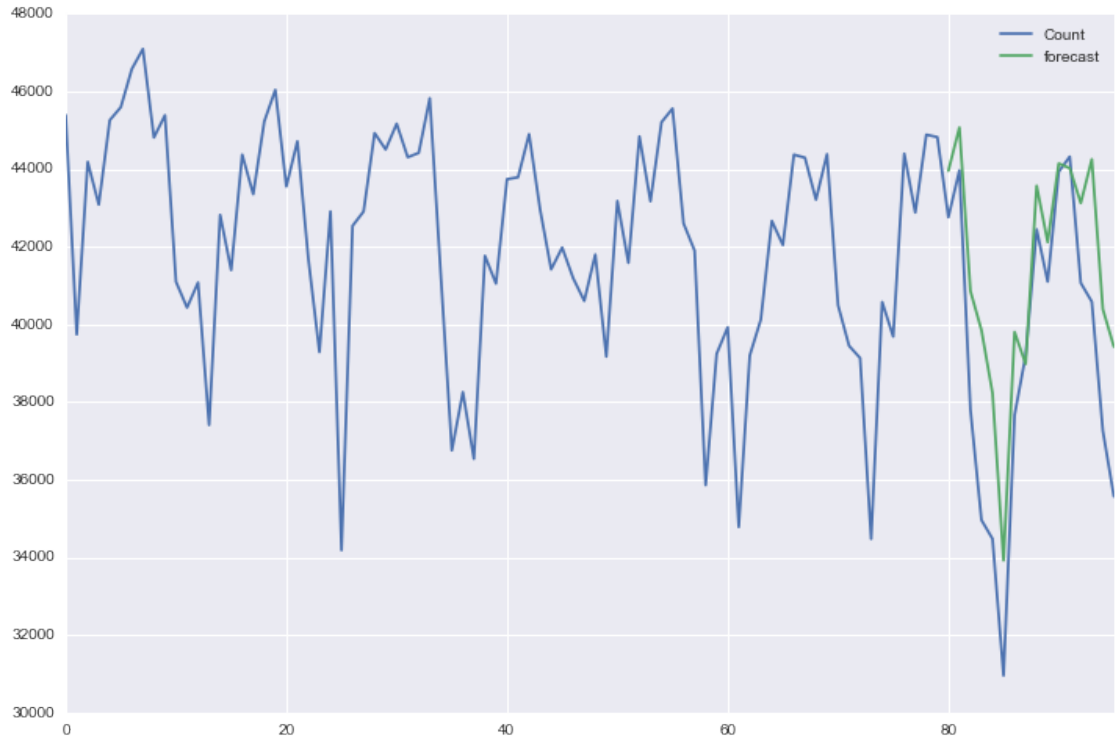
Dep. Variable:	Count	No. Observations:	96
Model:	SARIMAX(1, 0, 0)x(4, 1, 1, 12)	Log Likelihood	-753.427
Date:	Sat, 15 Apr 2017	AIC	1520.854
Time:	12:51:24	BIC	1538.805
Sample:	0	HQIC	1528.110
	- 96		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.0638	0.034	1.895	0.058	-0.002	0.130
ar.S.L12	-1.0340	0.116	-8.881	0.000	-1.262	-0.806
ar.S.L24	-0.0626	0.034	-1.856	0.063	-0.129	0.004
ar.S.L36	-0.0372	0.024	-1.553	0.120	-0.084	0.010
ar.S.L48	-0.0120	0.016	-0.735	0.462	-0.044	0.020
ma.S.L12	0.9871	0.228	4.330	0.000	0.540	1.434
sigma2	3.029e+06	1.62e-07	1.87e+13	0.000	3.03e+06	3.03e+06

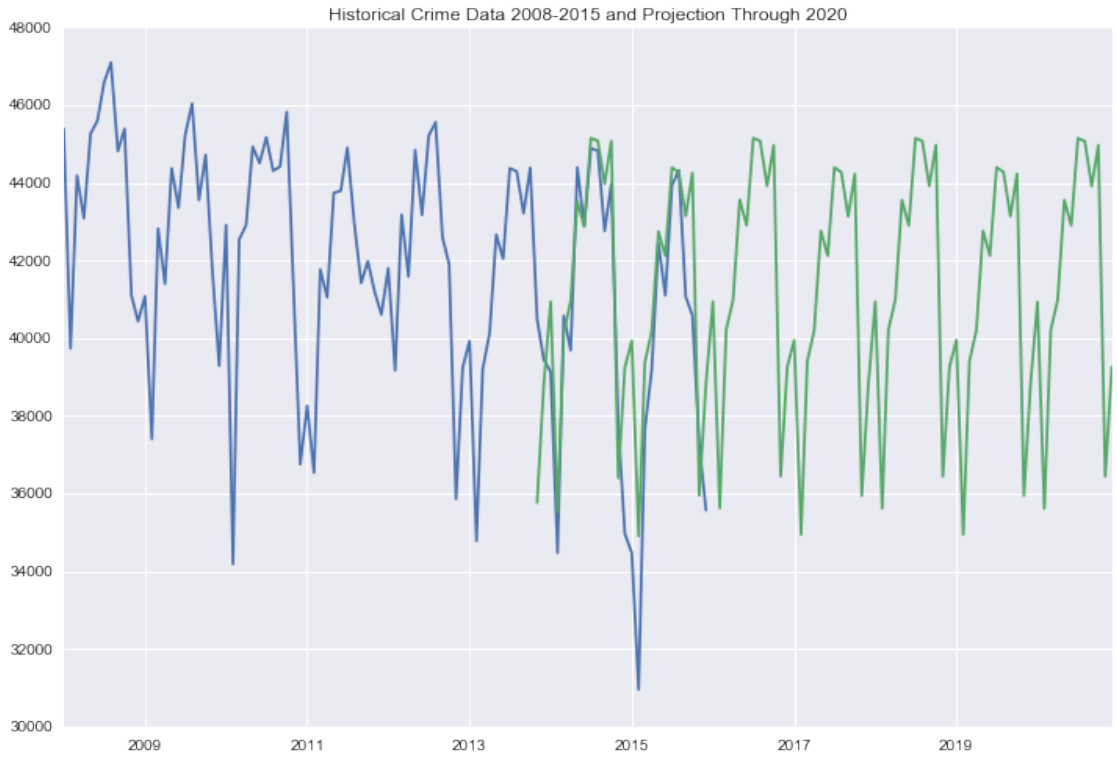
Ljung-Box (Q):	89.54	Jarque-Bera (JB):	3.41
Prob(Q):	0.00	Prob(JB):	0.18
Heteroskedasticity (H):	1.33	Skew:	-0.21
Prob(H) (two-sided):	0.45	Kurtosis:	3.89

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
[2] Covariance matrix is singular or near-singular, with condition number 8.05e+28. Standard err

```
In [8]: df['forecast'] = results.predict(start = 80, end= 98, dynamic= True)
        df[['Count', 'forecast']].plot(figsize=(12, 8))

Out[8]: <matplotlib.axes._subplots.AxesSubplot at 0x1192029d0>
```




```
In [9]: start = datetime.datetime.strptime("2016-01-01", "%Y-%m-%d")
        date_list = [start + relativedelta(months=x) for x in range(0,60)]

In [10]: dates_formatted=[]
         for i in range(60):
             dates_formatted.append(str(date_list[i].strftime('%b'))
                                    + "-" + str(date_list[i].year))

In [11]: future = pd.DataFrame(index=dates_formatted, columns= df.columns).reset_index()
         future['Month'] = future['index']
         del future['index']
         df2 = pd.concat([df, future]).reset_index(drop=True)

In [12]: start = datetime.datetime.strptime("2008-01-01", "%Y-%m-%d")
         date_list = [start + relativedelta(months=x) for x in range(0,156)]
         df2['date'] = date_list

In [13]: df2['forecast'] = results.predict(start = 70, end = 156, dynamic= True)
         df2=df2.set_index('date')
         plt.figure(figsize=(12,8))
         plt.plot(df2[['Count', 'forecast']])
         plt.title('Historical Crime Data 2008-2015 and Projection Through 2020')

Out[13]: <matplotlib.text.Text at 0x11bbd4390>
```


Historical Crime Data 2008-2015 and Projection Through 2020
48000
46000
44000
42000
40000
38000
36000
34000
32000
30000
2009
2011
2013
2015
2017
2019