Assignment 32 - ARIMA - shampoo dataset¶

In [1]:

```
# Import Packages

from pandas import read_csv
from pandas import datetime
from matplotlib import pyplot
%matplotlib inline
import warnings
warnings.filterwarnings("ignore")
from matplotlib.pylab import rcParams
rcParams['figure.figsize'] = 15,6
```

In [2]:

```
# Import dataset
series = read_csv('shampoo-sales.csv',header=0, parse_dates=[0], index_col=0, squeeze=True)
```

In [3]:

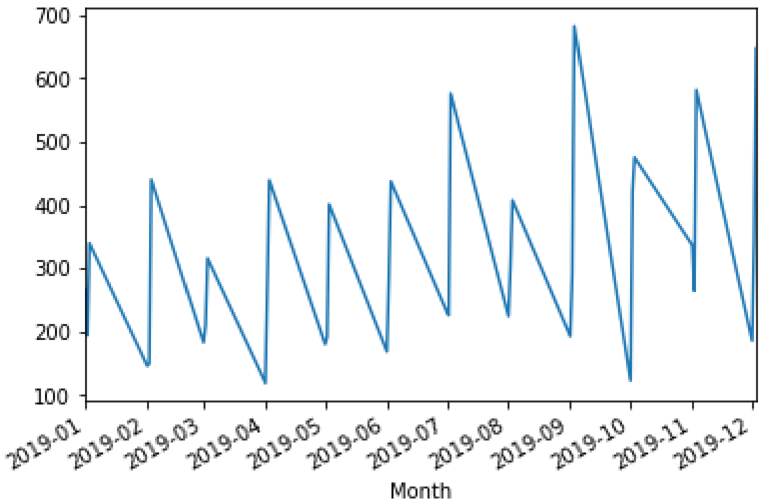
```
series.head(5)
```

Out[3]:

```
Month
2019-01-01    266.0
2019-02-01    145.9
2019-03-01    183.1
2019-04-01    119.3
2019-05-01    180.3
Name: Sales of shampoo over a three year period, dtype: float64
```

In [4]:

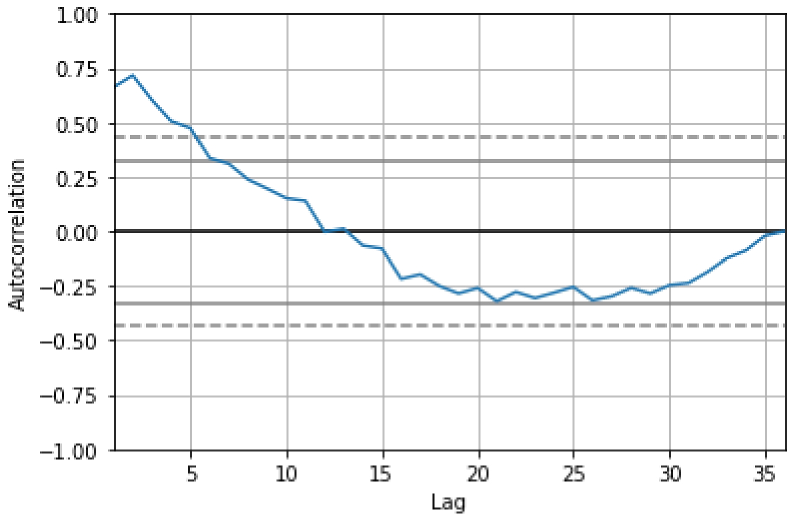
```
#Plot the dataset
series.plot()
pyplot.show()
```



The above plot shows that there is a trend in data. So we can apply differincing by 1 in ARIMA model to make it as stationary

In [5]:

```
#autocorrelation plot of the time series
from pandas.plotting import autocorrelation_plot
autocorrelation_plot(series)
pyplot.show()
```



Autoregressive Integrated Moving Average (ARIMA)

In this model in addition to AR , MA model it also has I i.e integration of differenciation which helps in converting the non- stationary (trend & seasionality) to stationary.

The method is suitable for univariate time series with trend and without seasonal components

AR: Autoregression. A model that uses the dependent relationship between an observation and some number of lagged observations.

I: Integrated. The use of differencing of raw observations (e.g. subtracting an observation from an observation at the previous time step) in order to make the time series stationary.

MA: Moving Average. A model that uses the dependency between an observation and its residual error fand applied to lagged observations.

The parameters of the ARIMA model are defined as follows:

p (Lag Order): The number of lag observations included in the model, also called the lag order.This would be similar to stating that it is likely to be warm tomorrow if it has been warm the past 3 days.

d (Differencing degree): The number of times that the raw observations are differenced, also called the degree of differencing. This would be similar to stating that it is likely to be same temperature tomorrow if the difference in temperature in the last three days has been very small.

q (MA order): The size of the moving average window, also called the order of moving average. This allows us to set the error of our model as a linear combination of the error values observed at previous time points in the past. \

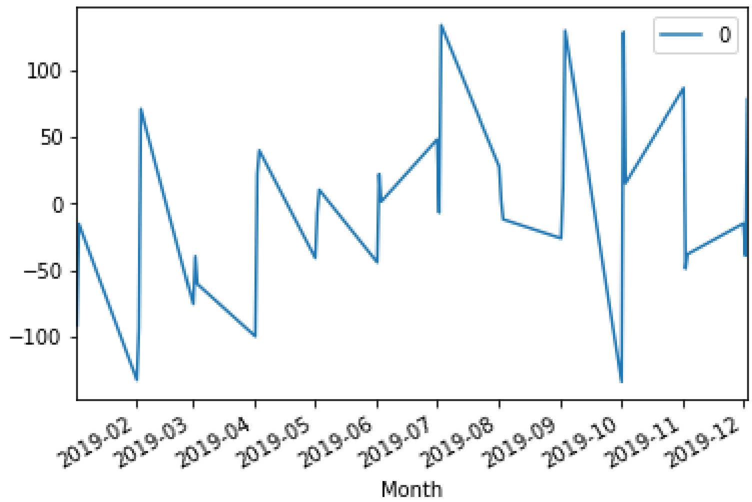
First, we fit an ARIMA(5,1,0) model which sets the lag value to 5 for AR, uses a difference order of 1 as I to make the time series stationary, and uses a moving average (MA) model of 0.

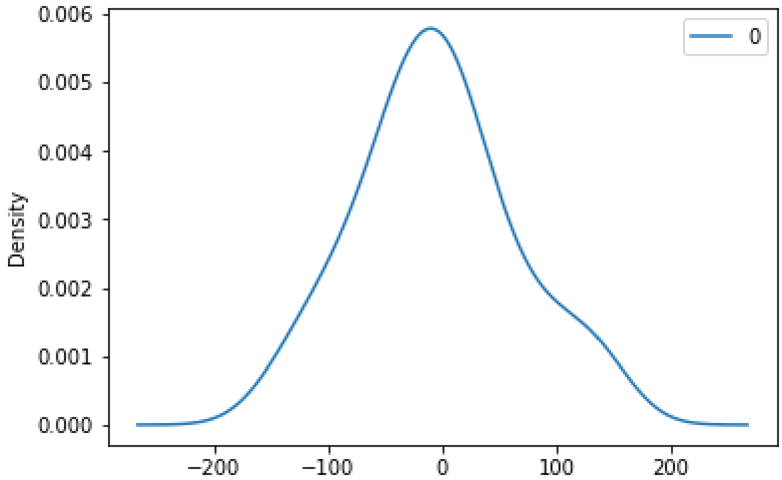
In [6]:

```
from statsmodels.tsa.arima_model import ARIMA
from pandas import DataFrame

# fit model
model = ARIMA(series, order=(5,1,0))
model_fit = model.fit(disp=0)
print(model_fit.summary())

# plot residual errors
residuals = DataFrame(model_fit.resid)

residuals.plot()
pyplot.show()
residuals.plot(kind='kde')
pyplot.show()

print(residuals.describe())
```

```
C:\Users\mallikarjuna.m\AppData\Local\Continuum\anaconda\lib\site-packages\s
tatsmodels\tsa\base\tsa_model.py:225: ValueWarning: A date index has been pr
ovided, but it has no associated frequency information and so will be ignore
d when e.g. forecasting.
  ' ignored when e.g. forecasting.', ValueWarning)
C:\Users\mallikarjuna.m\AppData\Local\Continuum\anaconda\lib\site-packages\s
tatsmodels\tsa\base\tsa_model.py:225: ValueWarning: A date index has been pr
ovided, but it has no associated frequency information and so will be ignore
d when e.g. forecasting.
  ' ignored when e.g. forecasting.', ValueWarning)

                                                    ARIMA Model Results
============================================================================
===========================
Dep. Variable:     D.Sales of shampoo over a three year period   No. Observa
tions:                   35
Model:                                           ARIMA(5, 1, 0)   Log Likelih
ood                -196.170
Method:                                                 css-mle   S.D. of inn
ovations             64.241
Date:                                          Mon, 18 Mar 2019   AIC
406.340
Time:                                                  19:24:51   BIC
417.227
Sample:                                                       1   HQIC
410.098

============================================================================
=========================================
                                                              coef    std err
z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------------
-----------------------------------------
const                                                      12.0649      3.652
3.304      0.003       4.908      19.222
ar.L1.D.Sales of shampoo over a three year period    -1.1082      0.183
-6.063      0.000      -1.466      -0.750
ar.L2.D.Sales of shampoo over a three year period    -0.6203      0.282
-2.203      0.036      -1.172      -0.068
ar.L3.D.Sales of shampoo over a three year period    -0.3606      0.295
-1.222      0.231      -0.939       0.218
```

```
ar.L4.D.Sales of shampoo over a three year period    -0.1252      0.280
-0.447       0.658       -0.674       0.424
ar.L5.D.Sales of shampoo over a three year period     0.1289      0.191
0.673       0.506       -0.246       0.504
                                    Roots
============================================================================
=
                  Real           Imaginary           Modulus         Frequenc
y
----------------------------------------------------------------------------
-
AR.1           -1.0617           -0.5064j            1.1763           -0.429
2
AR.2           -1.0617           +0.5064j            1.1763            0.429
2
AR.3            0.0816           -1.3804j            1.3828           -0.240
6
AR.4            0.0816           +1.3804j            1.3828            0.240
6
AR.5            2.9315           -0.0000j            2.9315           -0.000
0
----------------------------------------------------------------------------
-
```






```
                0
count   35.000000
mean    -5.495218
std     68.132882
min   -133.296637
```

```
25%    -42.477890
50%     -7.186512
75%     24.748330
max    133.237936
```

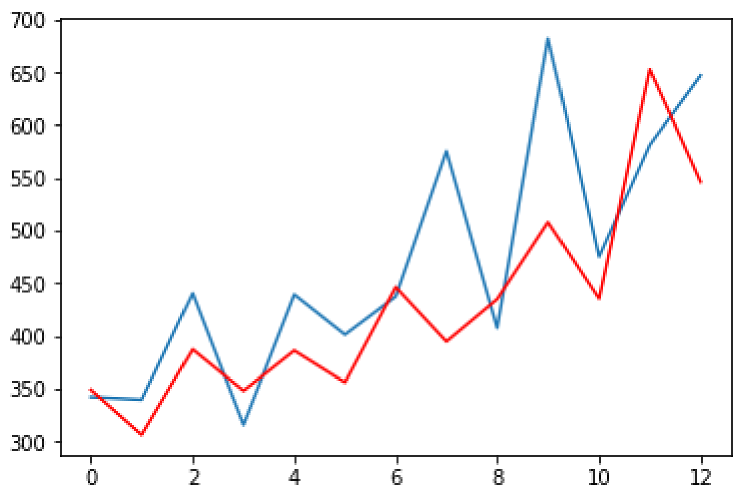
The distribution of the residual errors is displayed. The results show that indeed there is a bias in the prediction (a non-zero mean in the residuals).

Rolling Forecast ARIMA Model

In [7]:

```
from sklearn.metrics import mean_squared_error

X = series.values
size = int(len(X) * 0.66)
train, test = X[0:size], X[size:len(X)]
history = [x for x in train]
predictions = list()

for t in range(len(test)):
    model = ARIMA(history, order=(5,1,0))
    model_fit = model.fit(disp=0)
    output = model_fit.forecast()
    yhat = output[0]
    predictions.append(yhat)
    obs = test[t]
    history.append(obs)
    print('predicted=%f, expected=%f' % (yhat, obs))
error = mean_squared_error(test, predictions)
print('Test MSE: %.3f' % error)

# plot
pyplot.plot(test)
pyplot.plot(predictions, color='red')
pyplot.show()
```

```
predicted=349.117712, expected=342.300000
predicted=306.512952, expected=339.700000
predicted=387.376449, expected=440.400000
predicted=348.154255, expected=315.900000
predicted=386.308818, expected=439.300000
predicted=356.082087, expected=401.300000
predicted=446.379462, expected=437.400000
predicted=394.737224, expected=575.500000
predicted=434.915402, expected=407.600000
predicted=507.923547, expected=682.000000
predicted=435.482779, expected=475.300000
predicted=652.743826, expected=581.300000
predicted=546.343519, expected=646.900000
Test MSE: 6958.324
```



The MSE of the model is 6958.324

In [ ]:

```

```