The Engineering World #DataScience 16 & 17

May 31, 2018

AKKAL BAHADUR BIST
DATA SCIENTIST AT
KATHMANDU INSTITUTE OF APPLIED SCIENCES (KIAS)
Center for Conservation Biology (CCB)

1 SCALING AND DISTRIBUTION OF DATA

1.0.1 Transforming dataset distributions

In [1]:
```
import numpy as np
import pandas as pd
from pandas import Series, DataFrame
import matplotlib.pyplot as plt

from pylab import rcParams
import seaborn as sb

import scipy

import sklearn
from sklearn import preprocessing
from sklearn.preprocessing import scale
from scipy.stats.stats import pearsonr
```

In [2]:
```
%matplotlib inline
rcParams ['figure.figsize'] = 5,4
sb.set_style ('whitegrid')
```

1.0.2 Normalizing an transform features with MinMaxScalar() and fit_transform

In [3]:
```
address = 'mtcars.csv'
cars = pd.read_csv(address)
cars.columns = ['car_names', 'mpg', 'cyl', 'disp', 'hp', 'drat', 'wt', 'qsec', 'vs', 'am
cars.head()
```

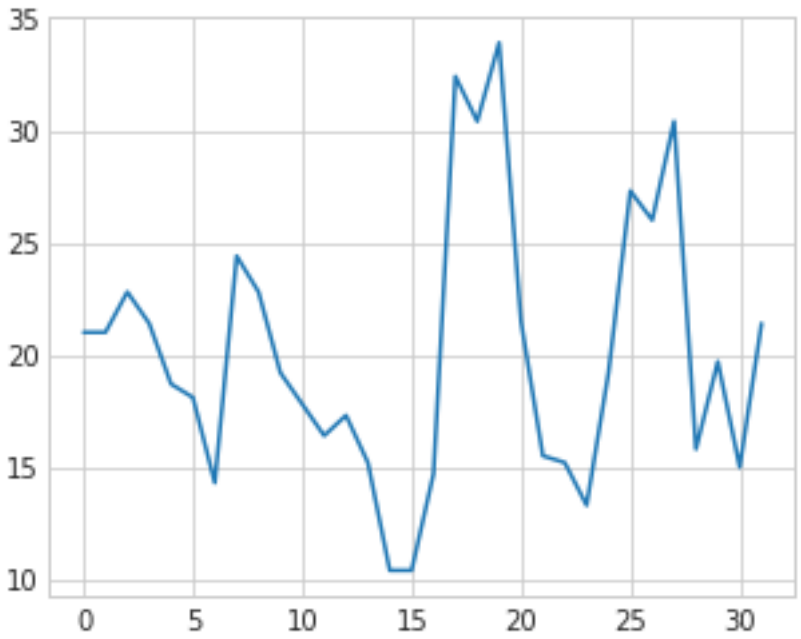
Out[3]:

	car_names	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	\
0	Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	

```
1      Mazda RX4 Wag  21.0    6  160.0  110  3.90  2.875  17.02   0   1      4
2         Datsun 710  22.8    4  108.0   93  3.85  2.320  18.61   1   1      4
3     Hornet 4 Drive  21.4    6  258.0  110  3.08  3.215  19.44   1   0      3
4  Hornet Sportabout  18.7    8  360.0  175  3.15  3.440  17.02   0   0      3

   carb
0     4
1     4
2     1
3     1
4     2
```

In [4]:
```
mpg = cars.mpg
plt.plot(mpg)
```

Out[4]:
```
[<matplotlib.lines.Line2D at 0x7f6bbd0b6f60>]
```

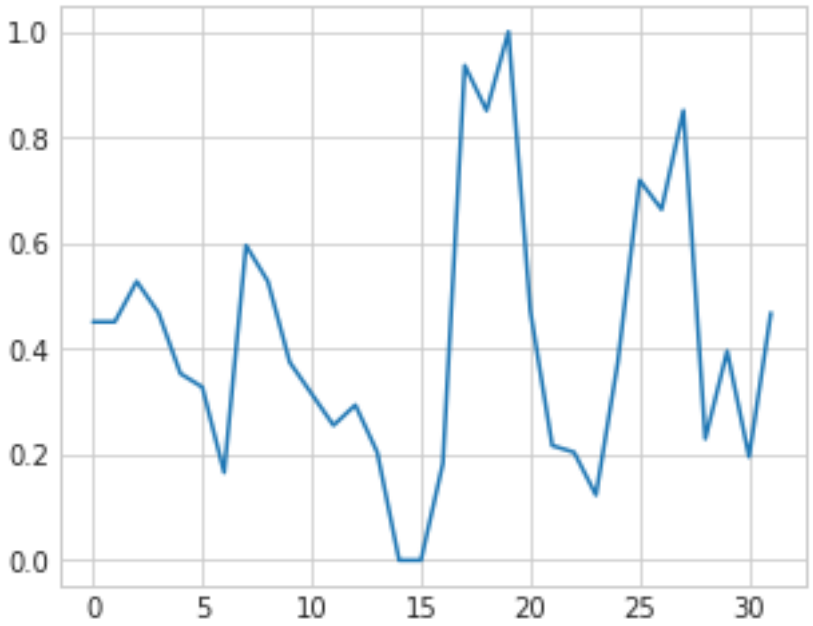


In [5]:
```
cars[['mpg']].describe()
```

Out[5]:
```
             mpg
count  32.000000
mean   20.090625
std     6.026948
min    10.400000
25%    15.425000
50%    19.200000
75%    22.800000
max    33.900000
```

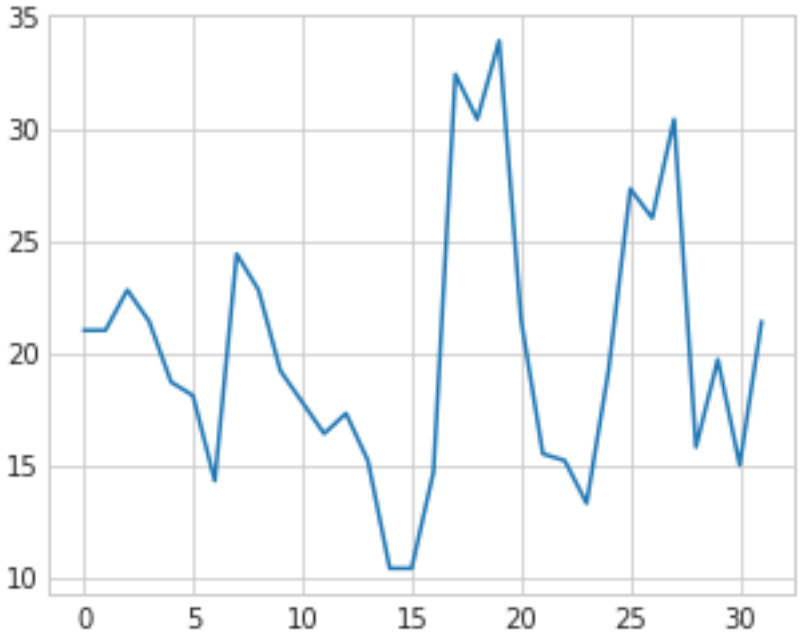
In [6]:
```
mpg_matrix = mpg.values.reshape(-1,1) #scaled
scaled = preprocessing.MinMaxScaler()

scaled_mpg = scaled.fit_transform(mpg_matrix)
plt.plot(scaled_mpg)
```

Out[6]: [<matplotlib.lines.Line2D at 0x7f6bbcfed978>]



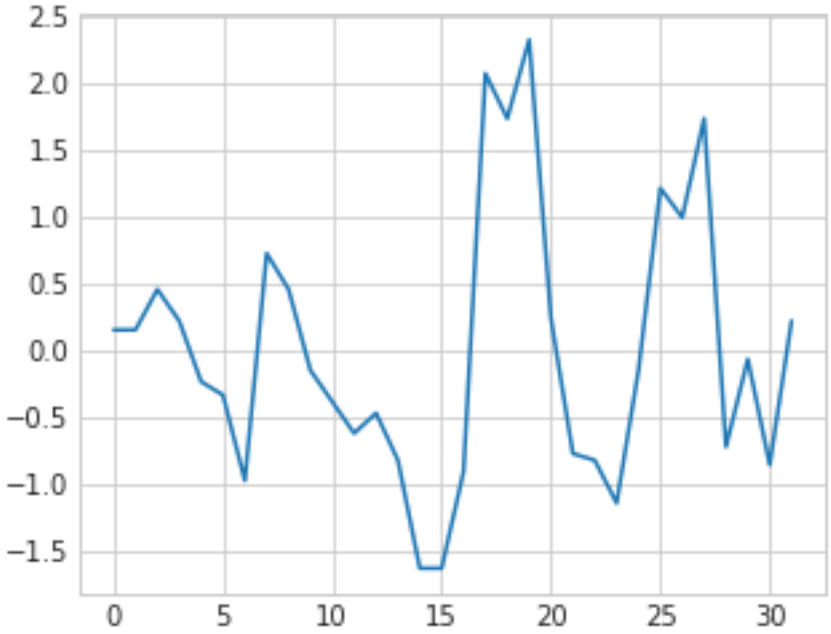
In [7]:
```
standardized_mpg = scale(mpg, axis = 0, with_mean = False, with_std = False)
plt.plot(standardized_mpg)
```

Out[7]: [<matplotlib.lines.Line2D at 0x7f6bbcfd9940>]



In [8]:
```
standardized_mpg = scale(mpg)
plt.plot(standardized_mpg)
```

Out[8]:
```
[<matplotlib.lines.Line2D at 0x7f6bbcf41518>]
```



1.0.3 Using scale() to scale your features

2 INTRODUCTION TO MACHINE LEARNING