Lab 2 – Time Series Data Analytics

Table of Contents

I. Forecasting

Given the following historical data of exact sales numbers:

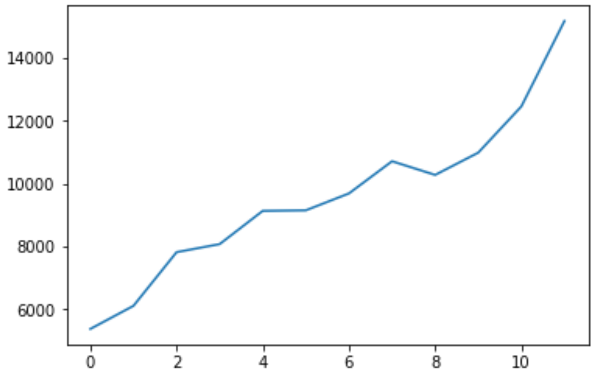
Month	Sales
1	5384
2	6118
3	7825
4	8081
5	9139
6	9156
7	9693
8	10717
9	10282
10	10990
11	12460
12	15177

1. Import necessary libraries

```
#make necesarry imports
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
from math import sqrt
import sys, os
from contextlib import contextmanager
import matplotlib as mpl
import seaborn as sns
import sklearn
```

2. Visualize and interpret the pattern of this time-series

```
df = pd.read_csv('data.csv')
df.head()
plt.plot(df['Sales'])
plt.show()
```



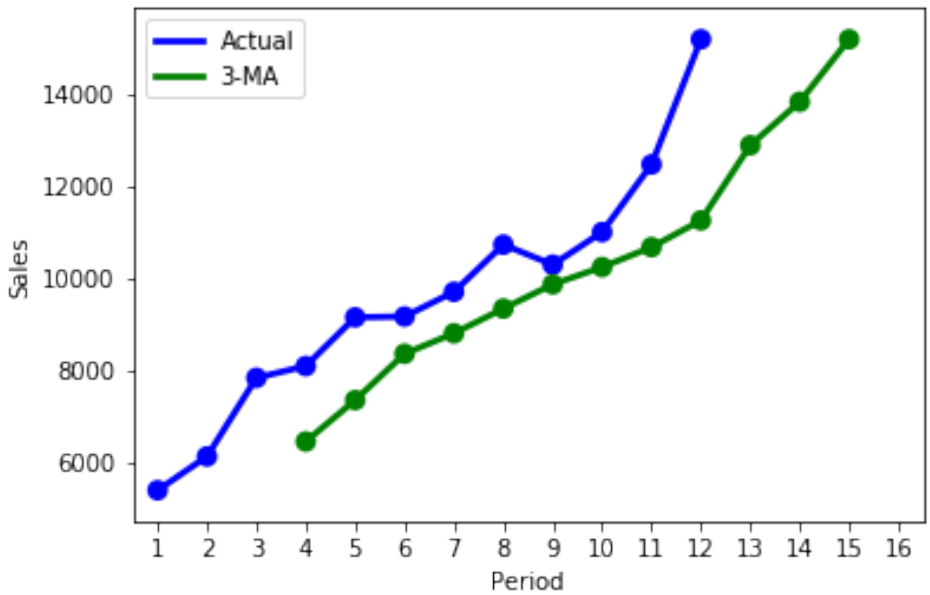
3. Predict future demand in month 13,14,15,16

a. Using moving average

```
# Using moving average
def moving_average(df, k, to_period):
    moving_average_df = pd.DataFrame(columns=['Period','Sales'])
    num_df = len(df)
    for m in range(0, to_period):
        if m < k:
            sale_predict = float('nan')
        else:
            history = df['Sales'][m-k:m]
            sale_predict = history.mean()
        moving_average_df.loc[m] = [m+1, sale_predict]

    moving_average_df['Period'] = moving_average_df['Period'].astype(int)
    return moving_average_df

f, ax = plt.subplots(1, 1)
ma_df = moving_average(df, 3, 16)
sns.pointplot(x='Period', y='Sales', data=df, color='b')
sns.pointplot(x='Period', y='Sales', data=ma_df, color='g')
ax.legend(handles=ax.lines[::len(df)+1], labels=["Actual", "3-MA"])
plt.show()
```



b. Using linear regression

More details of linear regression can be found via https://en.wikipedia.org/wiki/Linear_regression.

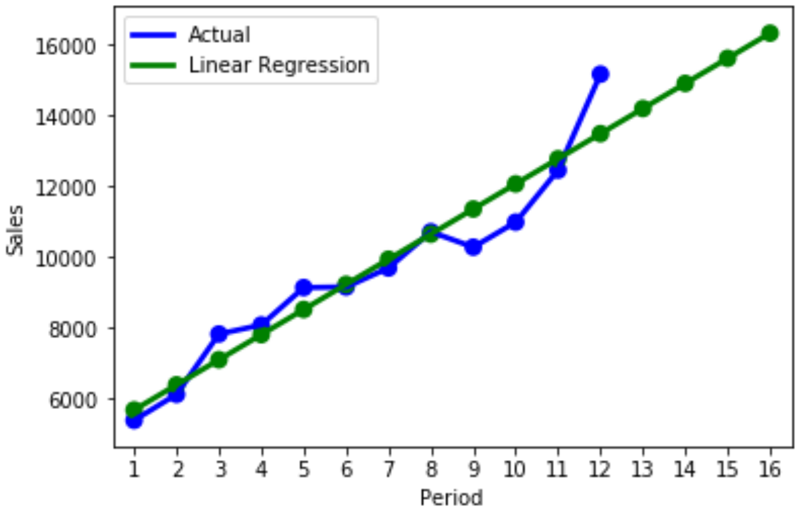
```
def linear_regression(df):
    linear_df = df.copy()
    linear_df['PeriodSales'] = linear_df['Period'] * linear_df['Sales']
    linear_df['Period_2'] = linear_df['Period'] * linear_df['Period']
    linear_df['Sales_2'] = linear_df['Sales'] * linear_df['Sales']
    linear_sum = linear_df.sum()
    linear_mean = linear_df.mean()

    b = (linear_sum['PeriodSales'] - len(df) * linear_mean['Period'] * linear_mean['Sales']) \
        / (linear_sum['Period_2']- len(df) * linear_mean['Period'] * linear_mean['Period'])
    a = linear_mean['Sales'] - b * linear_mean['Period']
    return a,b


a,b = linear_regression(df)
linear_df = pd.DataFrame(columns=['Period','Sales'])
for m in range(1, 17):
    sale = a + b * m
    linear_df.loc[m-1] = [m,sale]
linear_df['Period'] = linear_df['Period'].astype(int)

f, ax = plt.subplots(1, 1)
sns.pointplot(ax=ax, x='Period', y='Sales', data=df, color='b')
sns.pointplot(ax=ax, x='Period', y='Sales', data=linear_df, color='g')

ax.legend(handles=ax.lines[::len(df)+2], labels=["Actual", "Linear Regression"])
plt.show()
```



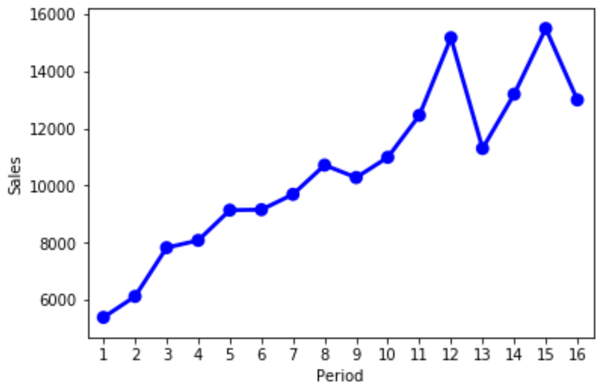
4. Evaluation: compare the above implemented methods

Give the ground truth as follow:

Month	Sales
13	11300
14	13200
15	15500
16	13000

- ✓ Visualise the ground truth

```
eval_df = pd.read_csv('actual.csv')
eval_df = pd.concat([df, eval_df], ignore_index=True)
sns.pointplot(x='Period', y='Sales', data=eval_df, color='b')
```



- ✓ Compare: we will use mean square error (MSE) to evaluate above methods. You can read https://en.wikipedia.org/wiki/Mean_squared_error for more information about this estimator.

```
MSE = (eval_df['Sales'][12:16] - moving_average(eval_df, 3, 16)['Sales'][12:16])**2
MSE = MSE.mean()
print("MSE of {0}: {1}".format("Exponential smoothing with alpha=0.8", MSE))
MSE = (eval_df['Sales'][12:16] - linear_df['Sales'][12:16])**2
MSE = MSE.mean()
print("MSE of {0}: {1}".format("Linear Regression", MSE))
```

- ✓ Which one produced the best prediction?

II. Exercises

Given the climate data (climate_data.csv), you need to predict the "LandAverageTemperature" of year 2015 using the following method:

- ✓ Using moving average with suitable k.
- ✓ Using linear regression.

The ground-truth of the climate data can be found in actual_climate.csv. Using this ground-truth and finish the following tasks:

- ✓ Visualize the ground-truth.
- ✓ Compare the accuracy of above methods.