Data Mining Applications – Frederik Darwin

M10601836

Homework 1 – Linear Regression

Problem:

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Revenue(m)(X)	7	2	6	4	14		16	12	14	20	15	7
Profit(m)(Y)	0.15	0.10	0.13	0.15	0.25	0.27	0.24	0.20	0.27	0.44	0.34	0.17

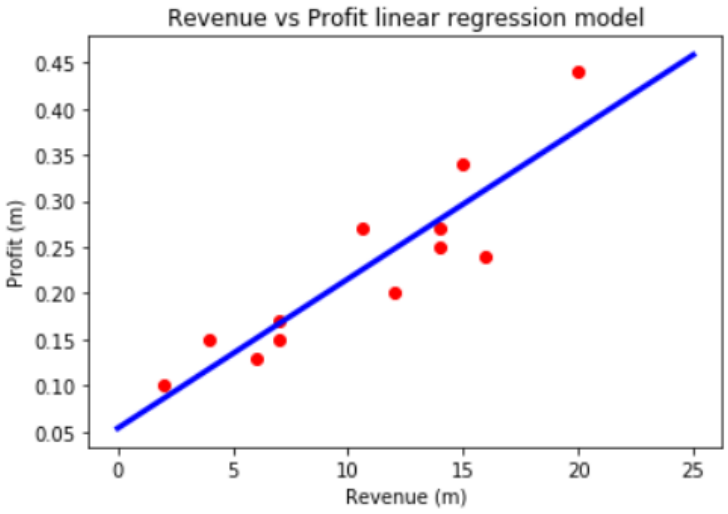
The following table shows the monthly revenues and the corresponding profits for a franchise company in 2017. Please write a computer program to find the linear regression model and predict the profit for January, 2018 if its revenue is 10 million dollars. There is a missing value in the data. Try to solve this problem yourself. Any kind of computer language is allowed for this homework.

Language Used: Python 3.6 in Jupyter Notebook

1. The missing value
   Can be easily replaced by calculating the total average of all revenue from that year, then the resulting average is used to fill the empty value.

```
dataset_X_train.replace(np.nan, np.nanmean(dataset_X_train), inplace=True)
```

2. Regression Model
   Is using the linear regression model with Revenue as x-axis and Profit as y-axis. The model will be trained using 12 data from the table (including already replaced mean values), resulting in this plot:



3. Prediction of January 2018
   With Revenue of 10 million, which mean x = 10, then the predicted value of the profit is around 0.21552963

```
In [11]:  1  dataset_Y_pred
Out[11]: array([0.21552963])
```

4. Source Code

```
# coding: utf-8
# In[1]:

import pandas as pd
import numpy as np
import math
from sklearn import datasets, linear_model
from sklearn.metrics import mean_squared_error, r2_score
import matplotlib.pyplot as plt

# In[2]:

dataset = pd.read_csv('hw1data.csv')

# In[3]:

dataset_X_train = dataset.iloc[:-1,1]
dataset_X_test = np.array(dataset.iloc[-1:,1]).reshape(-1,1)
dataset_Y_train = np.array(dataset.iloc[:-1,2])
dataset_Y_test = np.array(dataset.iloc[-1:,2])

# In[4]:

dataset_X_train.replace(np.nan, np.nanmean(dataset_X_train), inplace=True)

# In[5]:

dataset_X_train = np.array(dataset.iloc[:-1,1])
dataset_X_train = dataset_X_train.reshape(-1,1)

# In[6]:

linreg = linear_model.LinearRegression()

# In[7]:

linreg.fit(dataset_X_train, dataset_Y_train)

# In[8]:

dataset_Y_pred = linreg.predict(dataset_X_test)

# In[9]:
```

```
#Creating linear regression line
xfit = np.linspace(0, 25, 50)
yfit = linreg.predict(xfit[:, np.newaxis])

# In[14]:

#plotting the graph
plt.scatter(dataset_X_train, dataset_Y_train, color='red')
plt.plot(xfit, yfit, color='blue', linewidth=3)
plt.xlabel('Revenue (m)')
plt.ylabel('Profit (m)')
plt.title('Revenue vs Profit linear regression model')

# In[11]:

dataset_Y_pred
```