DATS 6313 – Time Series Analysis & Modeling

Instructor: Reza Jafari

Lab #2

Bradley Reardon

1/29/2022

1 – Abstract:

This lab pertains to developing a correlation coefficient program and applying it to some data from the tute1.csv file.

2 – Introduction:

This experiment was performed to increase understanding of what the correlation coefficient statistic represents, how to interpret it, and what it looks like visually when plotted amongst the data points.

3 – Method, Theory, and Procedures:

The correlation coefficient is a statistical measure of the strength of the relationship between the relative movements of two variables. There are two methods used in this experiment to determine the strength of the relationship between variables within the tute1.csv dataset; plotting the variables against one another to see a visual representation of the relationship and calculating the correlation coefficient of the pair of variables using the program developed.

The following figures represent the formula for calculating the correlation coefficient (r) [figure 1] and the code developed to calculate it given a proper input [figure 2]:

Text

Description automatically generated with medium confidenceText

Description automatically generated

4 – Answers to Lab Questions:

Text

Description automatically generated

1. The following images represent the code used to test the function along with the outputText

   Description automatically generatedText

   Description automatically generated
2. The calculated correlation coefficient makes sense with respect to the scatter plot since the scatter trends steeply negative and the correlation coefficient is -0.63.

Chart, scatter chart

Description automatically generated

1. The calculated correlation coefficient makes sense with respect to the scatter plot since the scatter trends steeply positive and the correlation coefficient is 0.90.Chart, scatter chart

   Description automatically generated
2. The calculated correlation coefficient makes sense with respect to the scatter plot since the scatter trends steeply negative and the correlation coefficient is -0.76.Chart, scatter chart

   Description automatically generated
3. Diagram

   Description automatically generatedChart, box and whisker chart

   Description automatically generated
4. Chart, treemap chart

   Description automatically generated

5 – Conclusion:

The correlation coefficient is used to determine the relationship between two variables. In this lab, I found whether there was a positive or negative relationship between the Sales. AdBudget, and GDP features from the tute1.csv dataset. Using the correlation coefficient (CC) program I wrote, I found the following CCs:

* Sales & AdBudget: 0.91
  + Strong positive correlation
* Sales & GDP: - 0.64
  + Slightly strong negative correlation
* AdBudget & GDP: -0.77
  + Strong negative correlation

6 – Appendix

import matplotlib.pyplot as plt  
import pandas as pd  
import numpy as np  
import seaborn as sns  
  
# Question 1  
  
def correlation\_coefficent\_cal(x,y):  
 numer = 0  
 denom\_x = 0  
 denom\_y = 0  
 for i in range(len(x)):  
 numer += ((x[i] - np.mean(x)) \* (y[i] - np.mean(y)))  
 denom\_x += (x[i] - np.mean(x))\*\*2  
 denom\_y += (y[i] - np.mean(y))\*\*2  
 r = numer / (np.sqrt(denom\_x) \* np.sqrt(denom\_y))  
 print(r)  
  
# Question 2  
x = [1,2,3,4,5]  
y = [1,2,3,4,5]  
z = [-1,-2,-3,-4,-5]  
g = [1,1,0,-1,-1,0,1]  
h = [0,1,1,1,-1,-1,-1]  
  
correlation\_coefficent\_cal(x, y)  
correlation\_coefficent\_cal(x, z)  
correlation\_coefficent\_cal(g, h)  
  
# Question 3  
df = pd.read\_csv(r'C:\Users\brear\OneDrive\Documents\GitHub\Time-Series-Analysis-and-Moldeing\Datasets\tute1.csv')  
df.rename(columns={'Unnamed: 0': 'Date'}, inplace=True)  
print(df.head())  
  
plt.scatter(df['Sales'], df['GDP'], c = 'g')  
plt.legend(loc='upper left')  
plt.xlabel('GDP')  
plt.ylabel('Sales')  
plt.title('Sales by GDP')  
plt.grid(color='gray', linestyle='-', linewidth=1)  
plt.show()  
  
correlation\_coefficent\_cal(df['Sales'], df['GDP'])  
'''  
The calculated correlation coefficient makes sense with respect  
to the scatter plot since the scatter trends negatively and the   
correlation coefficient is -0.63.  
'''  
  
# question 4  
plt.scatter(df['Sales'], df['AdBudget'], c = 'r')  
plt.legend(loc='upper left')  
plt.xlabel('AdBudget')  
plt.ylabel('Sales')  
plt.title('Sales by AdBudget')  
plt.grid(color='gray', linestyle='-', linewidth=1)  
plt.show()  
  
correlation\_coefficent\_cal(df['Sales'], df['AdBudget'])  
'''  
The calculated correlation coefficient makes sense with respect  
to the scatter plot since the scatter trends positively and the   
correlation coefficient is 0.90.  
'''  
  
# question 5  
plt.scatter(df['GDP'], df['AdBudget'], c = 'b')  
plt.legend(loc='upper left')  
plt.xlabel('AdBudget')  
plt.ylabel('GDP')  
plt.title('GDP by AdBudget')  
plt.grid(color='gray', linestyle='-', linewidth=1)  
plt.show()  
  
correlation\_coefficent\_cal(df['AdBudget'], df['GDP'])  
'''  
The calculated correlation coefficient makes sense with respect  
to the scatter plot since the scatter trends negatively and the   
correlation coefficient is -0.76.  
'''  
  
# question 6  
sns.pairplot(df, kind="kde", diag\_kind = 'hist')  
plt.show()  
sns.pairplot(df, kind="hist", diag\_kind = 'hist')  
plt.show()  
  
# question 7  
sns.heatmap(df[['Sales','AdBudget','GDP']].corr(), annot=True)  
plt.show()