Advanced Programming in Python Workshop – MiniProject #1

Due till 20/12/24

Part 1 – NumPy, Matplotlib/Seaborn:

Q1 (35pts): Convert covariance matrix into correlation matrix using numpy

A. Briefly explain in 1-2 sentences the concepts of Covariance and Correlation, and describe their relationship. You may use equations for clarification

B. Load the Iris dataset, using any method of your choice (e.g.: pd.read_csv("iris.csv") in Pandas).
Use visualization to explore the relationship between the different features.

C. Implement the following functions:
   a. A function to calculate the covariance between two variables:
   def calcCov(x, y): <your implementation here>,
   b. A function to compute the Covariance matrix:
   def covMat(data): <your implementation here>
   covMat(data) should return an n by n covariance matrix, where n is the number of features (in case of the iris dataset n=4)

D. Test1: compare the results of your function with NumPy's np.cov(data,rowvar=False) using the iris dataset.

E. Using your covariance function, implement a function to calculate the correlation matrix:
def corrMat(data):
    <should use covMat(data) and return the correlation matrix>

F. Test2: validate your correlation matrix implementation by comparing it with the results of NumPy's np.corrcoef(data,rowvar=False), using the iris dataset.

G. Use visualizations to communicate the tests results. Include appropriate titles, axis labels, and colorbars where relevant.

Part 2 – NumPy, Pandas, MatPlotLib/Seaborn/Plotly:

The purpose of this part is to practice using libraries that were introduced in lecture. These libraries include pandas, numpy and matplotlib/seaborn.

Please use the git commands that you were taught while completing this project and upload this project to your github account.

When submitting the assignment, please include the code file as well as the URL to your git account to show us you understand navigating projects using git.

Q2 (50pts): For this question you will use the dataset titled “laptop-price – dataset.csv”.

Import the libraries mentioned above and import the dataset from your filesystem into the code.

Please write code to complete the following tasks with this dataset:

- Plot the price of all the laptops
- Which company has on average the most expensive laptop? What is the average laptop price for each company?
- Find the different types of Operating systems present in the data - under the column name "OpSys".
    - Please note - there are operating systems that are the same systems and just written differently in the column - please fix them to be uniform.
- Plot for each of the operating system types the distribution of the prices, so that the number of plots equals to the number of unique operating systems.
- What is the relationship between RAM and computer price? add an adequate plot to support your findings.
- Create a new column for the dataframe called "Storage type" that extracts the storage type from the column "Memory".
    - For example, in the first row in the column "Memory" it states "128GB SSD", the new column will have just "SSD" in its first row.

All plots must be plotted with axes titles and units as well as plot titles.

Q3 (15pts): Think of additional questions related to this data. What types of analyses and visualizations would you use to address them? Select two questions from your list and implement. Submit your list of questions, suggested analyses and visualizations and the implementation.

Submission: your submission will include one jupyter notebook file for the code (including comments in the code), a diagram of the class hierarchy, and the URL to your github page (where we will find your uploaded project).

Grading criteria:

Correctness: 60%

Structure and readability: 15%

Explanation and documentation: 15%

Adherence to guidelines and submission requirements: 10%

Innovation and creativity (bonus): 5pts