CSE 5520 Fall 2021
Homework 3 (Due 11:59 pm September 23, Thu, at HuskyCT)
Linear Regression, Pearson Correlation and Heat map

This homework is to help you practice with data visualization fundamentals related to various correlation analysis methods. You are expected to use these visualization techniques and others in doing your final project. You are required to do this exercise in Python. All plots/graphs must have titles and x-y coordinate tick labels and any necessary legend if desired.

Part 1: Linear Regression

Consider a simple set of data points {<2, 2 >, <3, 3 >, {<4, 5>, <6, 4>}.

Step 1: Show the given <x, y> data points in a 2D plot.

Step 2: Draw the regression line over the plot shown in Step 1. Steps 1 and 2 should be done in one cell of Jupyter notebook.

Step 3: Create a separate markdown cell and include your manual calculation of the components of the equation designed to produce the linear regression line. Hint: You can do calculation using MS Excel and import the snipped image in the markdown cell. There are other ways also. This hand calculation should show how intercept and slope are calculated.

Step 4: Show how covariance matrix is calculated using python. Print the value.

Step 5: Create a separate markdown cell and include your manual calculation of the components of the equation designed to produce covariance matrix.

Step 6: Show how Pearson's correlation coefficient is computed using python. Print the value.

Step 7: Create a separate markdown cell and include your manual calculation of the components of the equation designed to produce Pearson's correlation coefficient.

Step 8. This time, you use the nba.csv from HuskyCT. Produce a regression line for this data set using weight as dependent variable and height as independent variable. The plot should include data points and the regression line. The intercept and slope values should be included in the title of the plot where should appear at the top (centered in bold) above the plot.

Step 9. Using the regression line in Step 8, compute and print the "predicted" weight for a rookie player whose height is known 91.

Part II: Scatter Plot with Pearson Correlation Coefficient

Consider the article "Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease" Cell. 2013 Apr 25;153(3):707-20. PMID: 23622250; PMCID: PMC3677161, by Zhang et al. The gene expression dataset published for this article GSE44768 is available at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE44768. This dataset has been preprocessed and some portion is available in HuskyCT's Data folder. The article

"GSE44768_article" is also available in HW3 folder. Do Part II steps using only the data from the Dementia group (GSE44768_CR_alz_female_reduced.csv).

Step 1. Create and show two scatter plots, one comparing gene expression values between TYROBP and DOCK2 and the other comparing gene expression values between TYROBP and GSTA4. Each plot should have the appropriate title including r (Pearson correlation coefficient value).

Step 2. Create and show two scatter plots, one comparing gene expression values between TYROBP and FCER1G and the other comparing gene expression values between ACBD5 and LMAN1. This time, both plots should appear side by side. Each plot should have the appropriate title including r.

Part III: Correlation Matrix and Heatmap

Now you are analyzing multiple pairs of genes at the same time, i.e., you like to know which pairs would exhibit good positive or negative correlations? Instead of all possible pairs, you will examine pairs from only 10 genes.

Step 1. Create and show a 10 x 10 correlation matrix in which TYROBP, DOCK2 and GSTA4 are included. For the other 7 genes, you can choose any from the list for the analysis. Do this step using only the data from the Dementia group (GSE44768_CR_alz_female_reduced.csv). Retain your gene order for the rest of Part III analysis.

Step 2. Convert correlation matrix you generated in Step 2 into a correlation coefficient heatmap and show. You are required to add title and axis labels appropriately. Do this step using only the data from the Dementia group (GSE44768_CR_alz_female_reduced.csv).

Step 3. This time repeat Steps 1 and 2 for the data from the Non-Dementia group (GSE44768_CR_nd_female_reduced.csv).

Step 4. Compare the heatmaps (not correlation matrix) you generated from Steps 2 and 3 side by side.

Step 5. Create a markdown cell and discuss your comparison between the two heatmaps you are showing side by side in Step 4. Do you see any noticeable difference?

Part IV: Review of Histogram/Boxplot/Violinplot

Consider again the gene expression data set introduced in Part II. You would like to compare histograms for TYROBP's gene expression levels between the two cohorts, the Dementia group (GSE44768_CR_alz_female_reduced.csv) and the Non-Dementia group (GSE44768_CR_nd_female_reduced.csv). Label each plot appropriately including color legend.

Step 1. Create and show two histograms for TYROBP's gene expression level for the Dementia group and the Non-Dementia group, individually, side by side.

Step 2. This time, merge both histograms with different colors, blue for the Dementia group and red for the Non-Dementia group, into one plot.

Step 3. This time, create and show boxplots for TYROBP's gene expression level both for the Dementia group and the Non-Dementia group in one plot, side by side.

Step 4. Repeat Step 3 using violinplot again in one plot, side by side.

Step 5. Create a markdown cell to include your interpretation for comparing TYROBP gene expression levels from the two cohorts. We are not interested in checking if your answer is right or not from the perspective of Alzheimer's Disease but we will check if your reasoning based on your plots is sound or not.

You upload your Jupyter notebook in HuskyCT. The file name should be of the following format: HWn_Doe where n is the homework number and Doe denotes last name.

HWs and Projects, 5% penalty for one day late submission. No acceptance after 5 days late. Extension is allowed only with the supporting medical record.