Project 3. Data Exploration

In this project, you shall explore the data downloaded from StackExchange sites.

You will be given two CSV files containing questions and answer. Write a program that shall:

1. Read the tables of questions and answers using the Pandas CSV reader. Some numeric columns may use thousands' separators (e.g., 25,355), and the post times must be represented as DateTime objects. Pandas can handle all these things, but you must ask for them.
2. Merge the tables based on the question IDs.
3. Report each boolean and numeric variable's mean, standard deviation, skew, and kurtosis. Arrange the results in a dataframe with five columns (variable name and its four moments) and as many rows as the number of variables. Save the table as a CSV file.
4. Report the ten most productive question authors (by the number of answers) and their reputation scores. Save the results as a CSV file.
5. Report the ten most commonly used tags (regardless of position) and their counts. Save the results as a CSV file.
6. Repeat the procedure for the questions that are closed vs. not closed, edited vs. not edited, and have accepted answer vs. do not have accepted answer. You will have a total of FOUR CSV files.
7. Using method .resample(), calculate and plot the number of daily posts. You will need to set the DataFrame index to the question post times. Save the plot as a PNG file.
8. Draw the histogram of the answers and questions' scores (in the same chart - make one of the histograms semi-transparent), the number of comments to the question, and the number of comments to the answers. Save the histograms as PNG files.

9. Scatter plot question scores vs. answer scores. Calculate their correlation and its p-value. Is the correlation significant? Save the plot as a PNG file.

10. Using method .corr(), find the two most negatively correlated and two most positively correlated variables and report their correlations (no p-values). Do not include user and question IDs (they are not genuinely numerical variables).

11. Create a report in Word, Latex, LibreOffice, or Google Docs that includes all the results that you obtained (all tables, images, and scalar data). Write a short interpretation of each result. (E.g., “One can see weekly oscillations in the daily plot chart, and there is a sharp decline around Christmas.”)

Deliverable: All charts, CSV files, and the report as one zipped/gzipped file. If the report is on Google Docs, submit a link to it.