CS 457 - Homework Assignment 2: Data Wrangling

Due Date: Monday, September 11 at 11:59 pm

Purpose:
Demonstrate Pandas library in Python to clean up messy dataset and gain useful insights from the data. This skill set will be used frequently in future data analytics.

Part 1 (40 points)

Question

Format the 580SurveyCleanup.csv dataset using Python. See the Week2-Class-DataWrangling.ipynb file for starter code.

Once the data is formatted, export it to output file in csv format with header. Use prefix such as "formatted_" in the output filename.

Part 2 (30 points)

Use the formatted data from Part 1, identify all the columns which have missing values. List those columns.

Investigate each column and fill in the missing values. Discuss why did you choose specific method of filling in missing values for each column.

Clean the values for each column if needed. Cleaning task can include handling whitespaces, upper/lower/title case, multiple names of the same entity etc.

Once the data is clean, export it to output file in csv format with header. Use prefix such as "cleaned_" in the output filename.

Part 3 (30 points)

Use the cleaned data from Part 2, prepare statistical summaries (counts, mean, unique etc.) for Questions Q1-Q12.

Tell the useful story/insights for each question based on statistical summary from this survey data. Use the 580SurveyQuestion.pdf file to review the actual question of the survey and use it in your insights to make it more engaging.

Deliverable: Submit a ipynb file containing your code for all Parts 1, 2 and 3. Submit output csv files for Part 1 and Part 2. For Part 3 make sure all the outputs are visible in ipynb along with all your insights and stories with respect to original survey questions given in pdf file. Include homework title, your name and your email on top of your ipynb code file.