BUSN:5760 Applied Business Statistics

Assignment 1: Data Wrangling With Pandas

**NOT FOR FALL 1 2024**

Perform all of the steps below in a Jupyter Notebook. For each step on the list, use a new notebook cell. Place comments at the top of each cell with a brief explanation of what you’re doing. When you’ve finished all steps, put your Jupyter Notebook, both input files, and both output files in a folder, zip the folder, and submit it to Canvas.

The input files needed for this assignment can be found here or in the course GitHub repository: [Assignment 1 Files](https://drive.google.com/drive/folders/1dg-L5Qlw90T0b0IYGWd6ttXaKKU3jwCX?usp=drive_link)

WARNING!

I do not mind students working together. However, I despise plagiarism and cheating. Do not attempt to submit someone else’s work and do not copy someone else’s work word for word. Further, do not use Chat GPT or any other AI tools. You might be able to get away with it, but I’m pretty tricky and manage to catch a dozen or so students cheating every semester. If I think you used someone else’s work or used AI, I will give you an F on this assignment and ban you from extra credit. This means that no matter what, you’ll get no higher than a C in the class.

1. Import Pandas and assign it the alias pd
2. Load “grades.csv” into a Pandas dataframe called “df1”
3. Display the top 10 rows of data
4. Display the bottom 10 rows of the data
5. Display a table of the variable names, as well as their associated counts and data types
6. Display basic summary statistics for the dataset’s numerical variables
7. Check for duplicate rows and display them if they exist
8. Drop any duplicate rows from the dataset and redisplay duplicate rows to check that they’ve been removed
9. Check for rows with missing values and display them if they exist
10. Drop any rows with missing values from the dataset and redisplay the rows with missing values to check that they’ve been removed
11. Create a column called “Honor Student” and put a “Yes” if their GPA exceeds 3.5 and a “No” if it doesn’t. Use a custom function and the map function
12. Load “tuition.xlsx” into a dataframe called “df2”
13. Check df2 for duplicates and display them if they exist
14. Drop any duplicate rows from the df2 and redisplay duplicate rows to check that they’ve been removed
15. Check df2 for rows with missing values and display them if they exist
16. Drop any rows with missing values from the dataset and redisplay the rows with missing values to check that they’ve been removed
17. Create a column called “Cost” that is the result of subtracting financial aid from tuition.
18. Merge the two dataframes into a new dataframe called “df3” using an inner join on the “ID” column
19. Rename the “Name\_x” column “Full Name” using a dictionary
20. Drop / remove the redundant “Name\_y” column from df3
21. Create a “First Name” and “Last Name” column by splitting the “Full Name” column
22. Group the students by major and display the mean GPA and mean net cost by major
23. Filter out everyone with a GPA below a 2.0 from df3
24. Sort df3 by ID in descending order
25. Export df3 as a CSV file and as an Excel file. Do not export with an index column.

Congratulations, you’re now able to automate 80% of what people do with spreadsheets! If there are any other operations / functions / techniques you’d like to learn, let me know and I’ll add them to the demonstration notebook.