BUSN:5760 Applied Business Statistics

Assignment 1: Data Wrangling With Pandas

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.

If you need help, use the official Pandas documentation and the Jupyter Notebook from class:

Pandas Documentation: <https://pandas.pydata.org/>

Pandas Notebook: <https://github.com/rjwrobel86/Python4Statistics/blob/main/Notebooks/Pandas101.ipynb>

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 basic summary statistics for the dataset
6. Check for duplicate rows and display them if they exist
7. Drop any duplicate rows from the dataset
8. Check rows with missing values and drop them if they exist
9. Create a column called “Honor Student” and put a “Yes” if their GPA exceeds 3.5 and a “No” if it doesn’t
10. Load “tuition.xlsx” into a dataframe called “df2”
11. Check df2 for duplicates and display them if they exist
12. Drop any duplicate rows
13. Check df2 for rows with missing values and drop them if they exist
14. Create a column called “Cost” that is the result of subtracting financial aid from tuition.
15. Merge the two dataframes into a new dataframe called “df3” using a full join on the “ID” column
16. Check df3 for missing values and drop the rows if they exist
17. Group the students by major and display the mean GPA and mean cost by major
18. Filter out everyone with a GPA below a 2.0
19. Sort by ID in descending order
20. Export df3 as a CSV file and as an Excel file