Case Study Rubric

DS 4002 - Spring 2025 - Elizabeth Hunter
Due: April 29th

Submission format: Upload link to GitHub repository on UVA Canvas

Individual Assignment

Why am I doing this?
This case study allows you to apply your data science and statistical modeling skills to a real-world issue facing the Charlottesville community. Investigating the relationship between UVA's enrollment growth and rising housing prices will give you experience in data cleaning, feature selection, and predictive modeling. As you work through this assignment, you will see how data analysis can inform public understanding of economic trends and help address critical challenges.

What am I going to do?
You will independently analyze Charlottesville's housing market, using public real estate data, UVA enrollment data, and CPI inflation data. You'll begin by merging and cleaning datasets, adjusting home sale prices to 2024 dollars using CPI values, and removing outliers. Then, you'll explore how sale prices vary based on features like bedrooms, grade, year built, and square footage. You'll also examine how changes in UVA's enrollment may relate to local housing prices through visualizations and regression modeling. To further explore prediction, you will implement machine learning models including decision trees, bootstrapped ensembles, and random forests.

Final deliverables should include:
- A data dictionary
- Data files (raw and processed)
- Well-documented, commented code
- A GitHub repository containing all materials used
- A written PDF with images of graphs, analysis, and conclusions

Tips for Success:
- Ask for help. Don't hesitate to reach out to instructors or TAs. This case study involves multiple data sources and advanced techniques, which can be challenging if you're working on a project like this for the first time. Clarifying questions early can save you time later.

- Stay organized. Clearly label your data frames and variables. With multiple datasets (housing, CPI, enrollment) and several rounds of cleaning and merging, organization is key to keeping your workflow clear and avoiding confusion.
- Work Iteratively. Test your code step-by-step rather than trying to do everything at once. Break the project into logical chunks: cleaning, merging, adjusting, analyzing, and modeling.

How will I know I have succeeded?
You will meet expectations on this case study when you follow the criteria in the rubric below.

Spec Category	**Spec Details**
Formatting	● One GitHub repository (submitted via a link on Canvas) ○ Create a new GitHub repository for this assignment titled "CS2_Charlottesville_Housing" that contains ■ README.md ■ SCRIPTS folder ■ DATA folder ■ Written PDF ■ REFERENCES.md
README.md	● Section 1: Software and platform section ○ Include type of software, add-on packages, and platform (Windows, Mac, Linux) ● Section 2: A Map of your documentation ○ Outline the structure of your files and folders ● Section 3: Instructions for reproducing your results. ○ Provide clear, step-by-step instructions to recreate the case study results
SCRIPTS folder	● Include all scripts used in your project ● Add comments for others to easily follow

DATA folder	● Include the initial raw data files ● Include the final data sets used for analysis ● Include a data appendix file as a PDF explaining the contents of each dataset
Written PDF	● Include a paragraph with your overall findings, include quantitative statistics ● Include a paragraph with analysis and supporting graphs to back up your overall findings ● Include a paragraph to explain your experience with the case study, include positive and negative
REFRENCES.md	● Include a list of references used during this project ● Use IEEE citation style ● Add a brief annotation under each citation explaining how the reference informed your work