Decoding University Admissions Trends Case Study Rubric

DS 4002 – Fall 2024 – Bianca Linares
Due: TBD
Submission format: GitHub repository link

Individual Assignment

General Descriptoin: Submit to Canvas a link to the GitHub repository for this project,

Preparatory Assignments: Class discussions and assignments prior to this case study.

Why am I doing this?
This case study allows you to apply your data science knowledge to analyze trends in higher education over the past 20 years. By examining admissions and tuition rates across the North and South regions of the United States, you will explore how data analysis can uncover meaningful insights with potential implications for education policy and equity. This assignment exposes you to real-world challenges in data cleaning, modeling, and visualization, providing valuable experience in tackling complex problems that impact accessibility and affordability in education.

What am I going to do?
The GitHub repository for this case study can be found at https://github.com/blinares-cs/DS4002_CS3. You will analyze university admissions and tuition data from the US Department of Education, focusing on trends over the past 20 years in the North and South regions of the United States. Your work will be documented in the GitHub repository, which will act as a comprehensive report containing your outputs, data, and code scripts. After inspecting and cleaning the dataset to address missing values or outliers, you will use Python to perform your analysis. Start by creating visualizations, such as line charts or heatmaps, to highlight trends and disparities between regions. Next, you may apply techniques such as time series modeling (e.g., ARIMA or Seasonal ETS) to analyze patterns and fluctuations in admissions and tuition rates, using stationarity tests like the ADF test when necessary. You will also conduct statistical tests, such as paired t-tests, to compare rates between the North and South regions. Your GitHub repository should include clear explanations of your methodology, annotated code scripts, and organized outputs to ensure reproducibility. By the end of the project, you will have a detailed analysis that provides insights into regional differences and trends in higher education, supporting informed policy decisions.

Tips for success:
- Start with a Clear Plan: Before diving into the analysis, outline your steps for cleaning the data, exploring trends, and performing comparisons. A clear roadmap will help you stay organized and efficient.
- Choose Descriptive Names: Use clear and descriptive names for your files, variables, and outputs. For example, "north_admissions_cleaned.csv" is much easier to interpret than "data1.csv."
- Familiarize Yourself with Time Series Analysis: If you plan to use methods such as ARIMA or Seasonal ETS, take some time to review their requirements, such as stationarity testing, and experiment with tools such as ADF tests.
- Focus on Readability: Ensure your code is well-documented with comments explaining each step. This not only helps others understand your work but also makes it easier for you to debug and refine your analysis.
- Leverage Visualizations Effectively: Use charts and graphs to convey your findings. Make sure each visualization is clearly labeled and includes a brief explanation of what it represents.
- Iterate and Experiment: Test different approaches, models, and visualizations to find the most insightful results. Don’t hesitate to revisit earlier steps if needed to improve your analysis.
- Keep Your Repository Organized: Ensure your GitHub repository is structured logically, with separate folders for data, code, and outputs. A well-organized repository makes your work accessible and easy to navigate.

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

Spec Category	Spec Details
Formatting	● One GitHub repository (submitted via link on Canvas). ● Create a new GitHub repository for this assignment titled ."CS3_AdmissionTrends" that contains: ○ README.md file ○ LICENSE.md (use MIT as default) ○ CODE Folder ○ DATA Folder. ○ FINDINGS Folder. ○ REFERENCES Folder
README.md	Purpose: This file serves as an place for anyone accessing your repository, enabling them to quickly understand its purpose and structure. Guidelines: ● Section 1: Overall Summary ○ Provide a brief overview of the project, including its goals and objectives. ○ Summarize the scope of the analysis and the key findings or outcomes. ● Section 2: Reproducibility ○ Describe all software and tools used for the project (e.g.,pandas, statsmodels, matplotlib) and include installation commands if applicable (e.g., pip install pandas). ○ Explain how to reproduce the analysis, including steps to download and prepare the dataset, run scripts for preprocessing and analysis, and generate visualizations and outputs. ○ Mention the folders and subfolders in the repository and describe their purpose and how they are used in the workflow. ● Section 3: Challenges and Reflections ○ Highlight any challenges encountered during the project, such as issues with the dataset, modeling, or visualizations. ○ Explain how these challenges were addressed and what was learned from the process. ● Section 4: References ○ List all references used in the project, including data sources, technical documentation, and relevant research articles. ○ Use consistent citation format (e.g., IEEE).
LICENSE.md	Purpose: This file explains to a visitor the terms under which they may use and cite your repository. ● Select appropriate license (MIT is appropriate) from the GitHub options list on repository creation.
CODE Folder	Purpose: This folder contains all the source code necessary for your project. It should allow anyone to easily execute your analysis to reproduce the results. Guidelines: ● Include all scripts used in the project, naming them in the order they should be executed. This ensures clarity for users following the workflow. ● Each script should begin with a header comment providing essential information, including: the script's purpose, any prerequisites or dependencies required to run the script (e.g., packages, datasets), and expected outputs. ● Throughout your scripts, include clear and detailed comments to explain

	what each command or sequence of commands accomplishes.
DATA Folder	Purpose: This folder contains all the data for the project, including both raw and processed datasets, as well as a detailed paper describing the data and the initial analysis performed before building any models. This ensures transparency and provides context for the analysis. Guidelines: ● Include Raw and Processed Data: ○ The raw data should be provided in its original format (e.g., raw_data.csv) to ensure reproducibility. ○ Include the processed data used for analysis (e.g., cleaned_data.csv or processed_data.csv). ○ If the raw and processed data are the same, only include that single dataset. ● Data Storage: ○ If the dataset is small enough, include it directly in this folder. ○ If the dataset is too large for GitHub, provide a file (e.g., data_readme.txt) explaining how to obtain the dataset, including relevant links or instructions. ● Data Appendix File: ○ Include a PDF that describes the data in detail, providing context and an initial analysis. ■ Begin each section with a description of the dataset, including what each row represents. ■ Create a subsection for each variable, describing its meaning, units, and any notable characteristics (e.g., missing values or outliers). ● Initial Analysis: ○ Provide tables, figures, and descriptive statistics that summarize the data. This includes counts, means, standard deviations, or visualizations to highlight important features.
OUTPUTS Folder	Purpose: This folder contains all the output generated by your project, including figures, tables, and other visualizations. It also includes a detailed paper describing the meaning of the outputs and how they contributed to reaching your conclusions. Guidelines: ● Figures and Tables: ○ Include all figures, tables, and visualizations generated during the project, especially those referenced in your presentation. ○ Ensure filenames are informative and descriptive (e.g., admissions_trends_north_vs_south.png or tuition_rate_analysis_table.csv). ● Paper on Outputs and Conclusions: ○ Provide a PDF document that explains the outputs in detail. ○ For each figure or table: ■ Describe what it represents and its relevance. ■ Discuss how the visualization or table contributed to understanding the data and reaching conclusions. ● Summarize how the outputs collectively support the findings and final conclusions of the project.
REFERENCES Folder	Purpose: This folder contains all the references used in the project, providing proper attribution and ensuring transparency. Guidelines: ● Include a comprehensive list of all references used, such as data sources, research articles, technical documentation, and relevant online materials. ● Use IEEE consistent citation format.