Name (Last, First):

Tushar Vimalbhai Patel

NEU ID:

002080292

NEU COE INFO6105 Final Project Proposal

(due 03/25/2025)

Dataset 1: https://public.opendatasoft.com/explore/embed/dataset/us-colleges-and-universities/table/?flg=en-us&dataChart=eyJxdWVyaWVzIjpbeyJjb25maWciOnsiZGF0YXNldCI6InVzLWNvbGxlZ2VzLWFuZC11bml2ZXJzaXRpZXMiLCJvcHRpb25zIjp7ImZsZyI6ImVuLXVzIn19LCJjaGFydHMiOlt7ImFsaWduTW9udGgiOnRydWUsInR5cGUiOiJjb2x1bW4iLCJmdW5jIjoiQ09VTlQiLCJzY2llbnRpZmljRGlzcGxheSI6dHJ1ZSwiY29sb3IiOiIjRkY1MTVBIn1dLCJ4QXhpcyI6ImNpdHkiLCJtYXhwb2ludHMiOjUwLCJzb3J0IjoiIn1dLCJ0aW1lc2NhbGUiOiIiLCJkaXNwbGF5TGVnZW5kIjp0cnVlLCJhbGlnbk1vbnRoIjp0cnVlfQ%3D%3D

Dataset 2: https://collegescorecard.ed.gov/data

This Datasets contain US colleges and Universities and their data like their Location, Student Population, student aid, student outcome, enrollments, etc.

Reason: Just a year ago, while applying to universities in the US, I had no guidance and had to face a lot of issues in deciding which universities to apply to. I had to go through lots of universities to check if they fit my prefrences and this was very time consuming. Now that I have learned to handle data to a certain extent, I want to try and create a recommandation model.

Question: What are the characteristics of top-ranked colleges in terms of student population and financial aid?
Method: Identify the characteristics (such as student population size, average financial aid, and student-to-faculty ratio) of highly ranked institutions. Visualize these relationships using heatmaps and bar charts to uncover trends.

Question: How does the student-to-faculty ratio influence the student enrollment size?
Method: Investigate if smaller student-to-faculty ratios correlate with higher or lower enrollment sizes. Use scatter plots to visualize this relationship and regression analysis to quantify it.

Question: How does the region of the U.S. (e.g., East Coast vs. West Coast) affect the average enrollment size and cost of attendance?
Method: Break down the dataset into geographic regions and analyze differences in enrollment size and cost of attendance across regions. Use bar plots and geographic visualizations to compare regions.

Question: What factors contribute to the acceptance rates of top universities?
Method: Examine factors such as location, student-to-faculty ratio, and financial aid that could be contributing to higher acceptance rates at top universities. Perform regression analysis to quantify these relationships.

I am not limited to this set of questions but, this are the general questions i want to answer with this project. Any inputs and suggestions are most welcomed.

NEU COE INFO6105 Final Project Guidelines

(updated 03/18/2024)

One of the requirements for INFO6105 is the final project (***up to 20% of the grade***). **The project must be done individually.** This is an opportunity to be creative in solving a problem that interests you and demonstrate your understanding of the concepts and principles and your proficiency with R. The project should be challenging enough so that you could discuss it at future interviews with potential employers. Another benefit of this project is that it gives your professor a good topic for discussion should you ever desire a reference.

Before beginning your project of cited literature, you MUST submit via Gradescope a few sentences (250 words max, describing your project proposal (due March 25, 2023, at 5 pm ET, worth 5% of the final project). This is so your professor can assess the appropriateness of your idea.

The project is due by April 15, 2023, at 5 pm ET and includes a 3-minute presentation of your work. Your verbal and written communication and presentation skills will be judged for 20% Extra Credit points.

In addition to the presentation, you are required to submit the following through email attachment in one zip file called **<username>_final_project.zip**:

- **One PDF file of your final project report (no more than eight pages, including a URL link to a video you made that recorded the process of running your R program to complete your data analysis and visualization)**
- **One .R file (including instructions on how to run your code written as comments).**

Here are some key criteria for determining your final project and its elements.

1. It must be original work and not something that might be proprietary to your company or generated by ChatGPT or similar AI tools.
2. The presentation and well-documented code should be at the level that other students and lay people can understand your project. Do not use advanced math or industry terms that require a lot of explanation.
3. You can choose any topic that interests you and conforms to the above criteria. The most important thing to remember is that this final project is meant to demonstrate your ability to apply what was learned in the course. You will not be judged on the originality of your topic or the difficulty of implementation. That said, if your project is overly simplistic, that will be held against you. We want you to show us the concepts and principles taught in this course that are well-understood and conform to best practices. Again, think of this as something you can show an employer as an example of why they should hire you as a data engineer, a data analyst, or a data scientist.
4. Your project must include at least the minimum number of each of the following:

a. One data set from reputed sites such as (and cite the data source):
   i. UC Irvine Machine Learning Repository (https://archive.ics.uci.edu/)
   ii. CMU StatLib (https://lib.stat.cmu.edu/datasets/)

b. Two meaningful questions about the data set are asked, to be answered with two or more methods taught in this course.

c. Use R for all the data analysis and visualization procedures.

d. Write a final project report (**no more than eight pages**) with the following sections: Introduction, Methods, Results, Discussion and Conclusion, References (cite all the sources).

e. **Include a URL link to a video you made that recorded the process of running your R program to complete your data analysis and visualization.** (You can use YouTube, Vimeo, or other cloud video providers if you can provide your professor with a link. YouTube allows you to create an unlisted video that's not publicly available, but anyone with the link can view the video.)

f. It is important to remember that longer is not better, and you will lose credit for not being succinct in your presentation.

When submitting your project, include a .R file with your code plus instructions (written as comments) for running the code and installing any 3rd party modules or libraries. We cannot grade your project if it uses proprietary modules or if we cannot run it.

Finally, we emphasize that your project MUST NOT contain proprietary, nonpublic, or confidential algorithms and data from your employer or other sources and **NOT be generated by ChatGPT or similar AI tools (including your presentation slides).**

Good luck, and have fun!