**Data Science Program Final Project**

**Executive Summary**

The Data Science program at Bethel School of Technology concludes with a capstone project to showcase the skills each student has learned and apply them to a question and dataset of their choosing. They have six weeks to complete and present the project from start to finish.

This project proposal document will outline the final project proposed by Erin Gancer, an exploration of women in engineering degree seeking programs in the United States from 2000-2020.

**Business Objectives**

The objective of this project is to showcase the data science skills Erin has acquired during her time in the Bethel Tech program, and also demonstrate her curiosity and continued commitment to learning by incorporating additional skills gained during the final project phase which were not covered in the course materials.

The visualization and data analysis done will provide insight into the enrollment trends of women in an engineering discipline over time. This will help universities understand these trends as they seek to attract diverse talent into these courses of study. It will also give women seeking entry into these fields information on the climate they are entering.

At the end of this project Erin will present her findings and additional questions to explore in a final presentation designed to be understood by experts and laypeople alike.

**Background**

The final project capstone project allows students to showcase their skills and marry those skills to an area of interest that they would naturally like to explore.

Erin has chosen to explore the changing ratios of women to men enrolled in engineering disciplines in United States universities. Erin was enrolled as an engineering undergraduate at Michigan Technological University from 2002-2006 and experienced the very male dominated atmosphere first hand. At that time the school’s ratio of men to women was 2:1. Although she did not personally experience discrimination based on her sex in the university setting first hand, she did find it challenging in many ways to be a part of a very male dominated field. She believes this makeup is changing, and anecdotally she has heard that this is so, however, she is curious what the data may say about this.

**Scope**

Erin will primarily be using Python and data science packages as taught in the course, as well as GeoPandas and Folium python packages that were not covered. Tableau, Excel, Powerpoint and other software tools may be used as necessary.

**Functional requirements**

Data Wrangling: Datasets from the National Center for Educational Statistics (NCES) containing enrollment information for each US university will be downloaded and wrangled for analyzing. The university information along with program enrollment information for each year will be isolated, and datasets with different years will be concatenated into usable forms for analysis. The use of another dataset with university geo location information will also be used for visualization.

Data Analysis: Dictionaries of terms used in the NCES data will be studied and understood so that the necessary columns in each dataset are preserved. Requirements for ANOVA analysis will be met, and the relationship between university, time, and enrollment ratios will be studied.

Data Visualization: Using GeoPandas and Folium packages, Erin will create visualizations of the distribution of engineering universities and the ratios of men and women in their program over time. Tableau will also be incorporated as an extra exercise demonstrating competence if time allows.

Presentation: In conjunction with her instructor, Erin will produce a 20 minute power point presentation that will outline the project and insights gained. This presentation will be accessible to all interested parties via zoom, and will not require data science understanding to understand.

**Personnel requirements**

Erin is a Data Scientist will be the sole person working on this project. She will be responsible for all stages of the project from data acquisition to presentation.

She will meet at least weekly with her instructor for guidance and accountability. She will also provide updates and ask questions via slack as needed.

**Delivery schedule**

Week 1: Explore possible topics and datasets that could answer interesting questions. Settle on a topic and begin exploration.

Weeks 2-3: Study the datasets and narrow down topic. Begin wrangling data into usable forms for needed analysis. Become familiar with geo and location dependent data analysis. Learn to use GeoPandas and Folium.

Week 4: Produce visualizations with GeoPandas and Folium. Begin ANOVA analysis based on what the visuals show is interesting in the data.

Week 5: Begin Final Presentation slides and wrap up analysis.

Week 6: Finish final presentation and practice with Instructor. Present!

**Other requirements**

All data must be free, and tutorials for the new skills needed will be free as well.

**Assumptions**

Datasets will be available, computer software and hardware will be in good working order.

**Limitations**

If roadblocks in the project arise the project may be delayed. Unavoidable scheduling issues for Erin or her instructor may cause the project to be delayed.

**Risks**

Risks that may arise include hardware malfunctions causing delays or loss of saved work. These risks are being minimized by keeping backups in cloud sources. Delay due to family emergencies are also possible, however Erin is motivated to finish the project so that risks is also minimal and could be managed. Risk of project delay and incompletion are both minimal.