A logo for college computing

Description automatically generated

**Assessment Cover Page**

|  |  |
| --- | --- |
| *Student Full Name* | Yumiko Maria Bejarano Azogue |
| *Student Number* | 2024144 |
| *Module Title* | Strategic Thinking |
| *Assessment Title* | Capstone Project Proposal |
| *Assessment Due Date* | 19th May 2024 23:59 |
| *Date of Submission* | 19.04.2024 |

**Declaration**

By submitting this assessment, I confirm that I have read the CCT policy on academic misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source.

I declare it to be my own work and that all material from third parties has been appropriately referenced.

I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution.

Contents

[Introduction 1](#_Toc162638295)

[Scope 1](#_Toc162638296)

[Problem definition 1](#_Toc162638297)

[Objectives 2](#_Toc162638298)

[General Objective: 2](#_Toc162638299)

[Specific Objectives: 2](#_Toc162638300)

[Data Sources (Technologies used) 2](#_Toc162638301)

[Libraries 2](#_Toc162638302)

[Accomplishment Data 2](#_Toc162638303)

[Attributes 2](#_Toc162638304)

[Multiple\_choice\_responses.csv 2](#_Toc162638305)

[Other\_text\_responses.csv 2](#_Toc162638306)

[Questions\_only.csv 3](#_Toc162638307)

[Survey\_schema.csv 3](#_Toc162638308)

[Data Analysis and Insights 3](#_Toc162638309)

[Gender Distribution Among Participants 3](#_Toc162638310)

[Age Distribution 3](#_Toc162638311)

[General Data Analysis 5](#_Toc162638312)

[Career and Education 5](#_Toc162638313)

[Salary 6](#_Toc162638314)

[Conclusion 7](#_Toc162638315)

[References 8](#_Toc162638316)

**Women in Technology Industry**

## Project proposal

### Introduction

The present report focuses on analysing women's participation in the technology industry, highlighting the challenges and opportunities we constantly face. The low representation of women in technical and leadership roles in the technology industry is a problem that affects gender equity and limits the potential for innovation and growth in that sector.

Women face persistent gender barriers in the technology industry, the lack of role models and stereotypes that perpetuate gender inequality in this field, to quote the civil rights activist Marian Wright Edelman, “You can’t be what you can’t see”.

### Objectives

#### General Objective:

To investigate and understand the gender gap in the technology industry, focusing on women's participation in technical, leadership, and entrepreneurial roles.

#### Specific Objectives:

Analyze gender distribution in different contexts, including by country and age group, using data from the Kaggle ML & DS Survey.

Identify the most common job titles among surveyed women and their educational levels to better understand trends in women's participation in technology.

Explore potential areas for improvement to promote gender equity in the technology industry by analyzing the challenges and barriers faced by women in technical and leadership roles.

### Problem definition

The gender gap in different roles and leadership positions in the technology industry limits professional advancement opportunities and decision-making for women. Gender stereotypes deeply ingrained in society perpetuate the perception that women are not as competent in technological fields as men, affecting their confidence and recognition at work.

The male-dominated work culture in technology companies creates an environment that is not inclusive for women, making it difficult for them to integrate and progress within the organization. Furthermore, gender bias in evaluation and promotion processes result in lower salaries, limited promotion opportunities, and less recognition for women.

## Scope and Methodology

### Scope

This project will focus on investigating the gender gap in the technology industry, exploring women's participation in the technical, leadership, and entrepreneurial roles. Gender distribution will be analyzed across different contexts, including representation by country and age group, and will suggest potential areas for improvement to promote gender equity in technology will be identified.

The project scope will include analyzing gender distribution by country and age group, as well as exploring the most common job titles among surveyed women and their educational levels. Advanced data analysis methods such as machine learning and text mining will be used to identify patterns and trends in women's participation in technology. Any analysis not directly related to the gender gap in the technology industry will be excluded.

### Methodology

For this project, we will use the CRISP-DM methodology (Cross-Industry Standard Process for Data Mining):

Business Understanding:

* Identify business objectives.
* Determine data mining goals aligned with business objectives.
* Define success criteria.
* Develop a preliminary plan to achieve objectives.

Data Understanding:

* Gather relevant data sources.
* Explore data to understand its quality, content, and structure.
* Identify data issues, anomalies, and potential biases.
* Determine data requirements for modeling.

Data Preparation:

* Clean and preprocess data to handle missing values, outliers, and inconsistencies.
* Select relevant features for analysis.
* Transform and engineer features as needed for modeling.
* Split data into training and testing sets.

Modeling:

* Select appropriate modeling techniques based on data characteristics and objectives.
* Build initial models using selected techniques.
* Evaluate model performance using relevant metrics.
* Fine-tune models by adjusting parameters and algorithms.

Evaluation:

* Assess model performance against success criteria and business objectives.
* Validate models using cross-validation or holdout datasets.
* Interpret model results and identify areas for improvement.
* Document findings and recommendations for stakeholders.

### Timeline

Semester 1: February 2024 - May 2024

* Initial Research (Week 1 - Week 4):
  + Literature review on the project topic.
  + Identification and selection of possible methods, techniques, and approaches to be used.
  + Detailed definition of project objectives.
* Data Acquisition (Week 5 - Week 10):
  + Dataset search and selection.
  + Identification of necessary permissions to access and use selected data.
  + Evaluation of integrity and quality of acquired data.
  + Data preparation for subsequent analysis, including standardization and normalization if necessary.
* Exploratory Data Analysis (Week 11 - Week 16):
  + Initial data exploration to understand its structure and characteristics.
  + Identification of possible patterns, trends, or anomalies in the data.
  + Preparation of preliminary visualizations and descriptive analysis.
* Model Development and Evaluation (Week 17 - Week 22):
  + Implementation of data analysis models or machine learning algorithms.
  + Evaluation of model effectiveness in relation to project objectives.
  + Adjustment and refinement of models as needed.

Semester 2: August 2024 - December 2024

* Optimization and Validation (Week 23 - Week 28):
  + Further optimization of selected models.
  + Cross-validation and robustness testing of models.
  + Evaluation of accuracy and reliability of obtained results.
* Conclusion Development (Week 29 - Week 32):
  + Interpretation of results obtained from data analysis.
  + Formulation of conclusions and recommendations based on project findings.
  + Preparation of preliminary reports and presentation of results.
* Refinement and Final Report Elaboration (Week 33 - Week 36):
  + Review and refinement of the final project report.
  + Incorporation of feedback and suggestions received during preliminary presentation.
  + Preparation of final presentations and additional materials for delivery.
* Final Delivery and Presentation (Week 37 - Week 40):
  + Final project presentation to evaluation committee.
  + Delivery of final project report and other related materials.
  + Closure and completion of all project-related activities.

## Data and Ethical Considerations

### Data Source Overview

The dataset was obtained from Kaggle and focuses on incomes for various job titles by gender. Below is the link to access the dataset.

<https://www.kaggle.com/competitions/kaggle-survey-2019>.

### Ethical considerations

dfgdfgfddg.

## Attributes

### Multiple\_choice\_responses.csv

This file contains responses to single-choice questions in separate columns. For questions with multiple responses, each option was split into its own column. Text responses were encoded to safeguard user privacy, and countries with fewer than 50 respondents were grouped as the “other.”

* The dataset comprises 19,718 rows and 246 columns.
* Each column represents a different question or provides additional metadata related to the survey.

### Other\_text\_responses.csv

If "Other" was selected, respondents had the option to provide a text response. These responses were separated and shuffled to protect privacy.

* This DataFrame includes responses to open-ended survey questions.
* It consists of 19,718 rows and 28 columns.
* Each column represents an open-ended question, with responses stored as text.

### Questions\_only.csv

This file lists the questions from the 2019 Kaggle Data Science and Machine Learning Survey.

* All columns are of the 'object' data type.
* The DataFrame has dimensions of 1 row and 35 columns.

### Survey\_schema.csv

This dataset describes which questions were presented to which respondents in the survey. Generally, respondents with more experience were asked more questions.

* It contains 10 rows and 35 columns.
* All columns have the data type 'object'.
* The survey scheme dataset serves as a key reference for understanding the structure and content of the data.

### Questions\_only.csv

This file lists the questions from the 2019 Kaggle Data Science and Machine Learning Survey.

* All columns are of the 'object' data type.
* The DataFrame has dimensions of 1 row and 35 columns.

### Survey\_schema.csv

This dataset describes which questions were presented to which respondents in the survey. Generally, respondents with more experience were asked more questions.

* It contains 10 rows and 35 columns.
* All columns have the data type 'object'.
* The survey scheme dataset serves as a key reference for understanding the structure and content of the data.

### Questions\_only.csv

This file lists the questions from the 2019 Kaggle Data Science and Machine Learning Survey.

* All columns are of the 'object' data type.
* The DataFrame has dimensions of 1 row and 35 columns.

### Survey\_schema.csv

This dataset describes which questions were presented to which respondents in the survey. Generally, respondents with more experience were asked more questions.

* It contains 10 rows and 35 columns.
* All columns have the data type 'object'.
* The survey scheme dataset serves as a key reference for understanding the structure and content of the data.

# References

Mooney, P. (2019). 2019 Kaggle Machine Learning & Data Science Survey. Kaggle. URL: <https://www.kaggle.com/competitions/kaggle-survey-2019>.

Engelbert, C. 2016, 'Starting A New Conversation for Women: [Deloitte CEO Catherine Engelbert on Gender Equality', TIME](https://time.com/4587281/catherine-engelbert-gender-equality/).

OpenAI. (2024). ChatGPT [Artificial Intelligence model]. Conversation with ChatGPT on March 29, 2024, at 10:00 a.m. <https://chat.openai.com/>

Douglas M. Branson (2018) The Future of Tech Is Female : How to Achieve Gender Diversity. New York: NYU Press. Available at: https://research.ebsco.com/linkprocessor/plink?id=e3b09355-de83-37e4-b1bf-678f33012b68 (Accessed: 27 March 2024).