A group of people cheering

Description automatically generated with low confidence

## Scenario

You are in charge of marketing for a well-known prestigious university catering to international students specializing in international business, and analytics. As part of the marketing effort, you have an outreach program to entice prospective students to learn more about the program and eventually enroll. These ambassador students are showcased on the university website with bios and interests so they are relatable and prospective students can learn from their experiences.

## Contextual Information:

As a welcoming and diverse university, you, as the marketing leader, need to ensure the biographies of the student ambassadors do not show bias towards campuses, topics or other observed information.

## Technical Considerations:

Review the most frequent terms, interesting associations, or phrases and compare the language across programs, campuses or other groups to ensure the showcased ambassadors present a welcoming diverse and inclusive learning atmosphere without leaning towards one or another aspects of the data too heavily.

A third-party service called namsor was used to infer the gender, countries and regions of the student ambassadors. As an inference there is likely some incorrect assumptions from the service, but these *may* used in identifying any marketing tendencies. Use of the namsor columns is optional and meant to add more context to the data set.

## Non-Technical:

This data was web-scraped directly from the university site. It is publicly available although the unique identifiers were changed, and first/last names removed. While it is not impossible to reconstruct the identities, these steps were taken so that it would be laborious and not worth the effort. The point of this exercise is not to reconstruct identities but to understand the typical bios and interests of showcased students along with identifying any opportunities for improving the public face of the university.

Ideally, the language used will be consistent showing no meaningfully discriminatory biases in selecting the identified ambassadors. However, differences may be observed as unintentional consequences. In either case, your presentation should support your conclusion that the language is consistent, or if not, how the language differs so that the marketing team can make adjustments.

## Project Deliverables include

1. R scripts for data processing & exploration “lastName\_TM\_ambassador\_case.R”
   1. Your script(s) must account for all aspects of the material in your presentation to ensure the presentation is data driven (no cheating with Excel or other tools!)
2. PowerPoint of any visualizations, findings and descriptions of non-technical results as if presented to the thinktank’s leader. “lastName\_TM\_ambassador\_case\_.pptx”
   1. The PowerPoint must be accompanied by a voice over embedded in the file, or screenshare video uploaded

## Example Data

|  |  |
| --- | --- |
| Variable | Description |
| "cid" | Unique Identifier |
| "programTitle" | The type of program the student has matriculated in |
| "campus" | The campus location |
| "interests" | A text field with interest of the student |
| "bio" | A text field with biographical information of the student |
| namSorGender.likelyGender | Using the namsor service, gender is inferred from name |
| namSorCountry.country | Using the namsor service, primary country is inferred from name |
| namSorCountry.countryAlt | Using the namsor service, alternative country is inferred from name |
| namSorCountry.region | Using the namsor service, world region is inferred from name |
| namSorCountry.subRegion | Using the namsor service, world subregion is inferred from name |
| isoCode.Country | Using the ISO 2 letter codes, the full name of the country is presented, corresponds with namSorCountry.country |
| isoCode.countryAlt | Using the ISO 2 letter codes, the full name of the alternative country is presented, corresponds with namSorCountry.countryAlt |

## Criteria for Success

## **Organization of content**– Logical ordering of ideas, modeling artifacts, applicable visualizations in slides

## **Organization of code**- R Code is well organized, concise, and free from error

## **Text mining process** – Recognize the type of data mining problem, adherence to established main data and text mining steps.

## **Completeness** – Understood impact, and mined the data for relevant insights/recommendations