MET CS 555 | TERM PROJECT

RESEARCH SCENARIO AND QUESTIONS

The United States has been a popular destination for international students seeking higher education opportunities. Understanding the trends, patterns, and factors influencing international student enrolment in the USA is crucial for educational institutions, policymakers, and stakeholders to make informed decisions and adapt to the evolving landscape of international education.

This research aims to investigate the following questions:

a) What are the key factors that significantly influence the number of international students coming to the USA?

b) Which countries contribute the most and least to the international student population in the USA in 2022? Also, which countries have shown a significant increase or decrease?

c) Which fields of study and educational levels are most popular among international students in the USA?

d) Which institute is most popular among international students?

e) What is the predicted trend of international students in the USA for the next 5-10 years?

By addressing these questions, this research will provide valuable insights into the dynamics of international student enrollment in the USA, help identify potential opportunities and challenges, and inform strategies for attracting and supporting international students.

DATA DESCRIPTION

Initial data cleaning and pre-processing steps were performed using Jupyter Notebook to combine the 6 CSV files into the "CombinedData" dataset, while keeping the country file separate "CountryDistribution".
The analysis utilizes two datasets.

1. Combined Data dataset:

- Obtained by combining 6 CSV files from various sources such as IEE, Kaggle, and IEE Open Doors (https://opendoorsdata.org/data/international-students/). The dataset contains 56 columns, with parameter up to 31.
- Columns 32 to 56 relates to international students enrolled in specific US universities. These columns will be used as EDA to understand leading institutions.
- Key variables: relates to international student enrolment, demographics, funding sources, and university.

2. Country Distribution dataset:

- Obtained from a separate CSV file.
- Provides information on the distribution of international students in the USA by country of origin.
- Includes variables such as country name and the number of students from each country from 2002 to 2022.
- This dataset will be used for exploratory data analysis (EDA) to understand the country-wise distribution of international students.

MET CS 555 | TERM PROJECT

Data Cleaning and Pre-processing

The data cleaning and pre-processing steps were performed using R programming language in RStudio.

1. Combined Data dataset:

   - Subset the data to include only records from 2012 onwards.
   - Removed commas and converted all columns to numeric except Year, Degree, and Field of Study.
   - Standardized the numeric columns.
   - One-hot encoded the categorical variables "Degree" and "Field of Study".
   - Cleaned column names by removing periods and renaming specific columns.
   - Saved the cleaned dataset as "Cleaned_CombinedData.csv" for further analysis and modelling.

2. Country Distribution dataset:

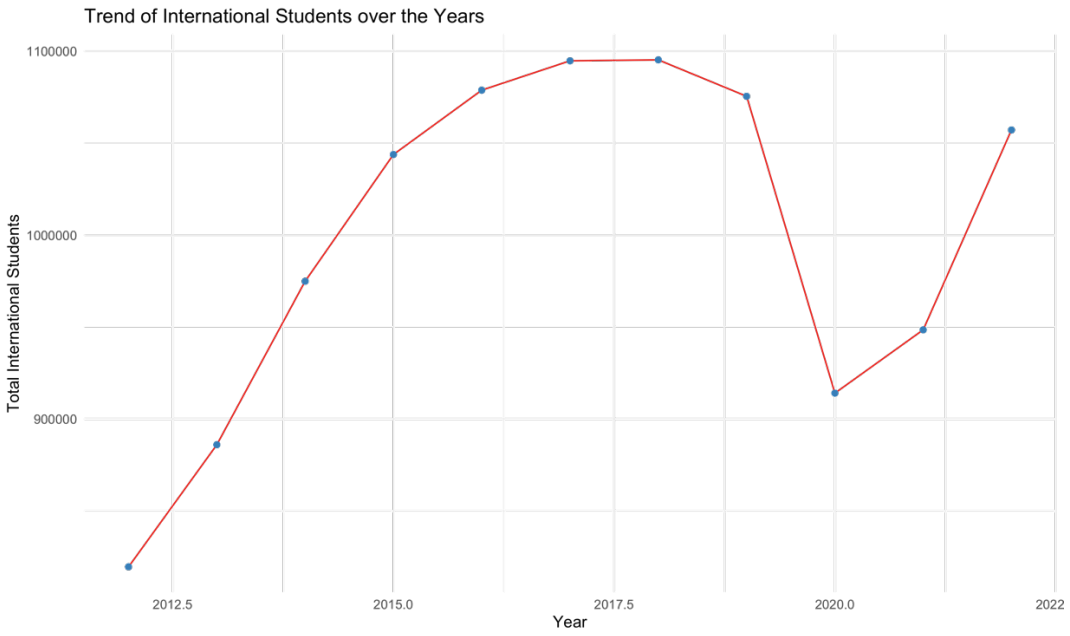
   - Removed columns from 2002 to 2011 to focus on 10 years trend.
   - Converted all columns except “Place of Origin” (containing country names) to numeric and removed commas and other signs.
   - Replaced missing values with row means.
   - Removed rows where all values were zero.
   - Saved the cleaned dataset as "Cleaned_CountryDistribution.csv" for further analysis.

Exploratory Data Analysis (EDA)

After cleaning and pre-processing the data, exploratory data analysis was conducted to gain insights into the trends, patterns, and relationships within the **Combined Data dataset**.

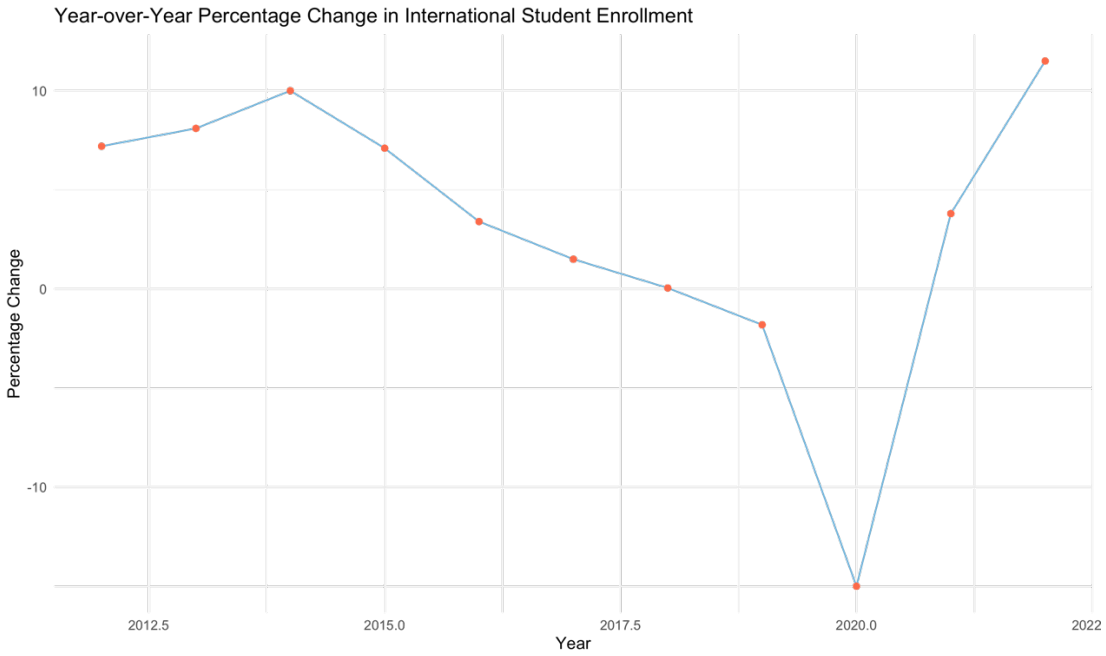
The following analyses were performed:

1. Trend Analysis for Total International Students:



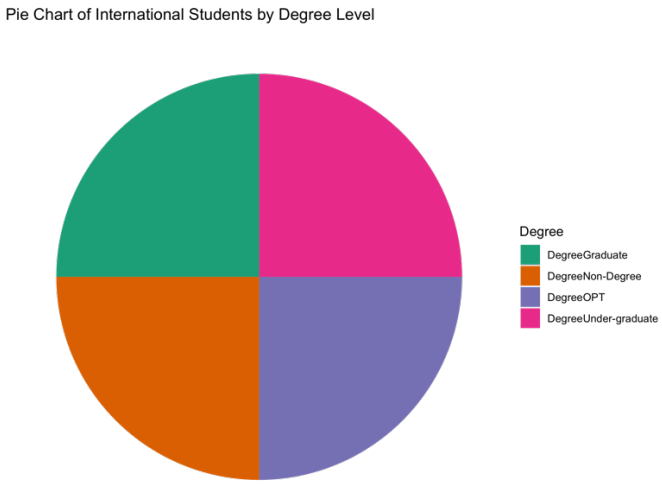
- The maximum number of international students is 1,095,299 in 2018.
- The minimum number of international students is 819,644 in 2012.
- This analysis provides an overview of the overall trend in international student enrolment in the United States.

2. Percentage Change Analysis:



- The maximum increase in international students is 11.5% in 2022.
- The maximum decrease in international students is -15% in 2020.
- This analysis helps identify significant changes in enrolment patterns and potential factors influencing these changes.

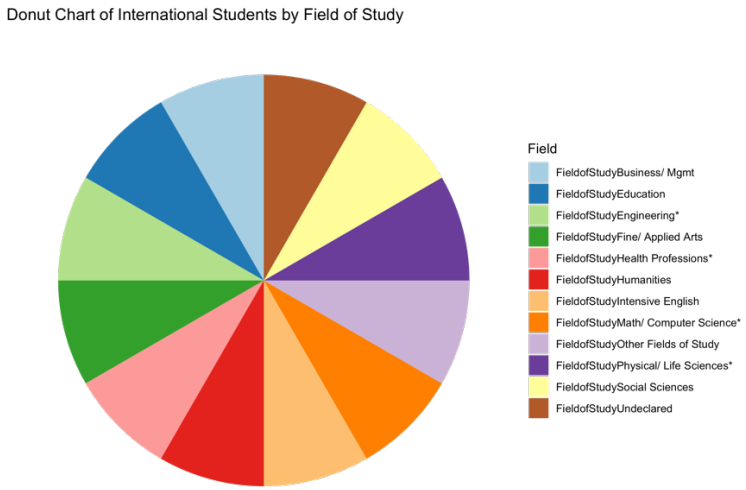
3. Degree Level Analysis:



- Shockingly, the print statements for the maximum and minimum international students by degree both show 132 for Degree (Graduate) while the graph suggest it's 132 across all degree level.
- This analysis provides insights into the popularity of different degree levels among international students.

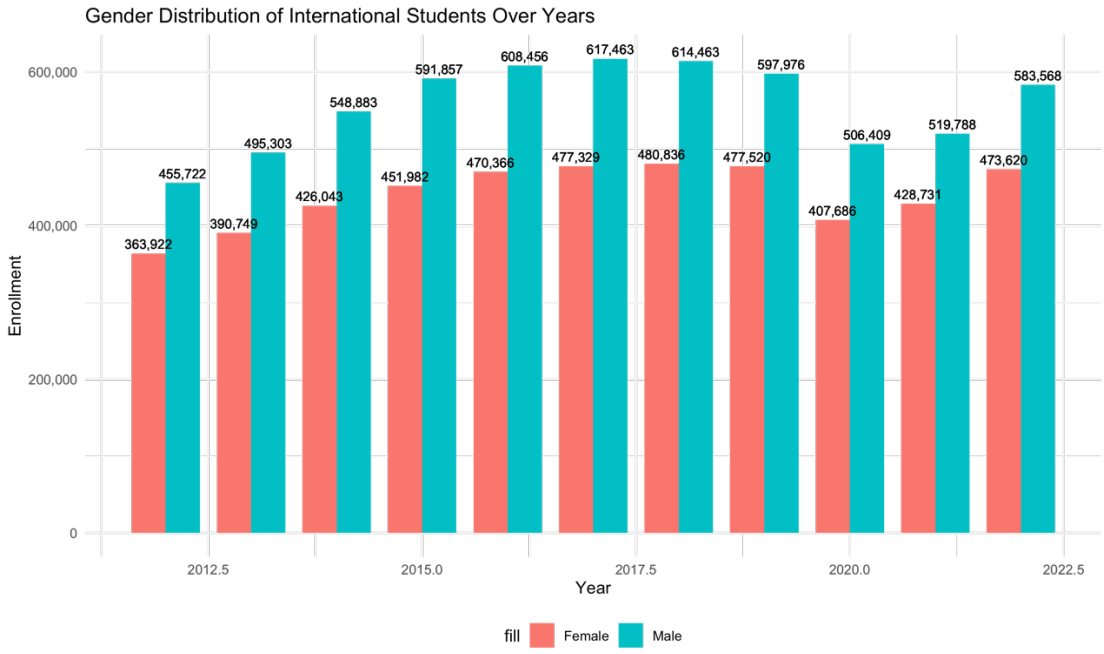
MET CS 555 | TERM PROJECT

4. Field of Study Analysis:



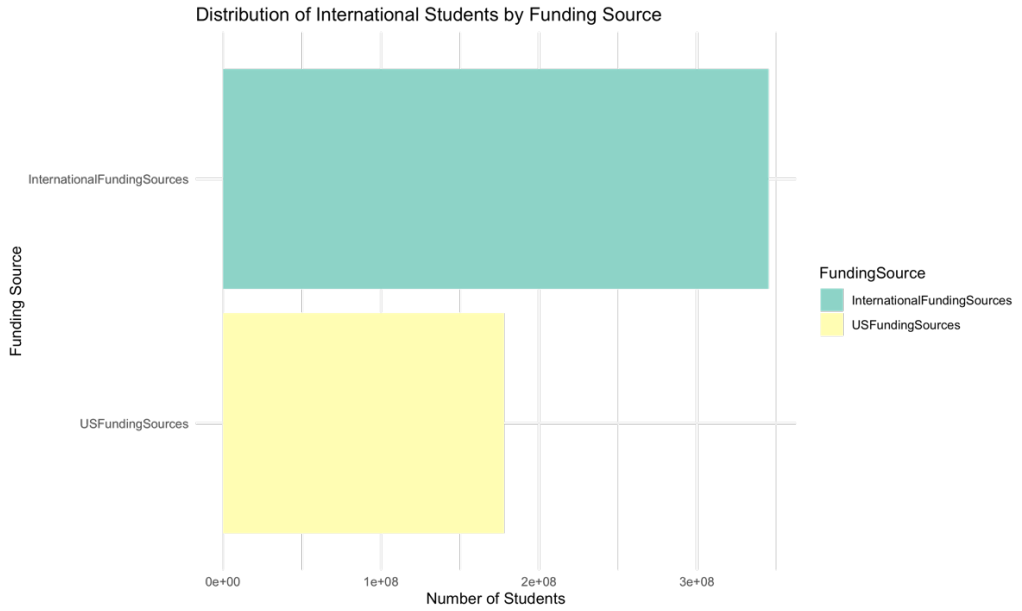
- Similar to the Degree Level Analysis, the print statements for the maximum and minimum international students by field of study both show 44 for Field of Study (Business/ Management) while the graph suggest it's 44 across all field of study.
- This analysis helps identify the most popular fields of study among international students.

5. Demographic Analysis - Gender Distribution by Year:



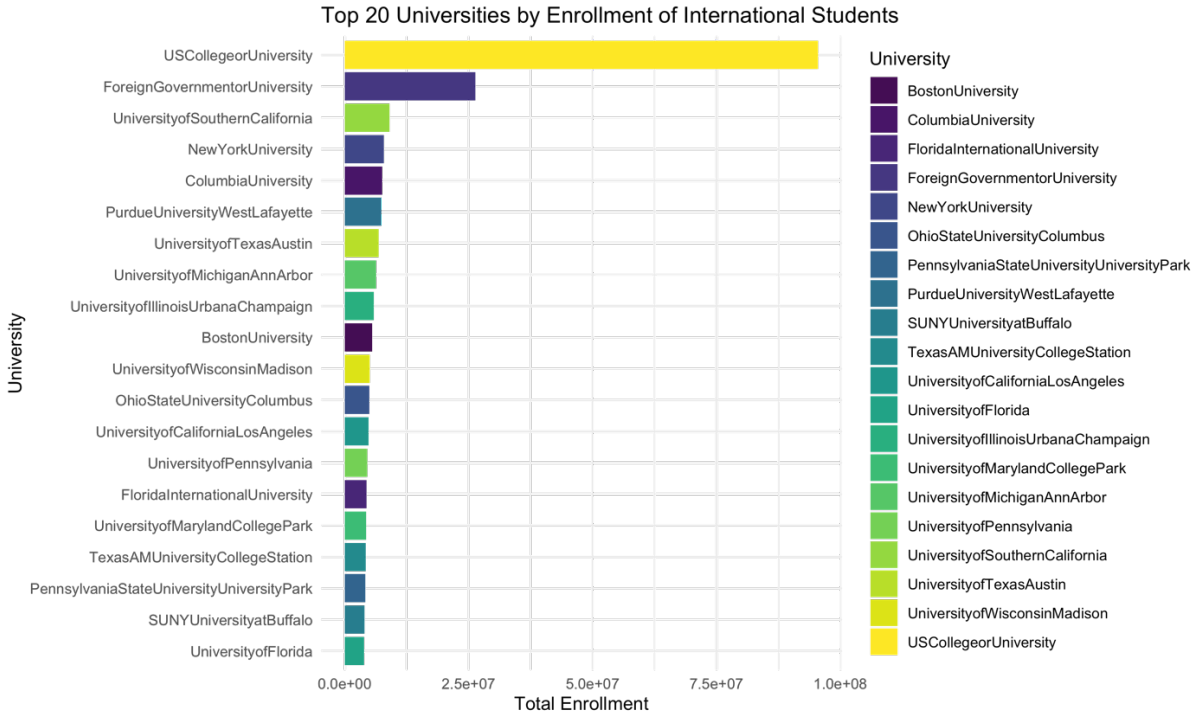
- The maximum number of female international students is 480,836 in 2018.
- The minimum number of female international students is 363,922 in 2012.
- The maximum number of male international students is 617,463 in 2017.
- The minimum number of male international students is 455,722 in 2012.
- This analysis provides insights into gender trends among international students.

6. Funding Source Analysis:



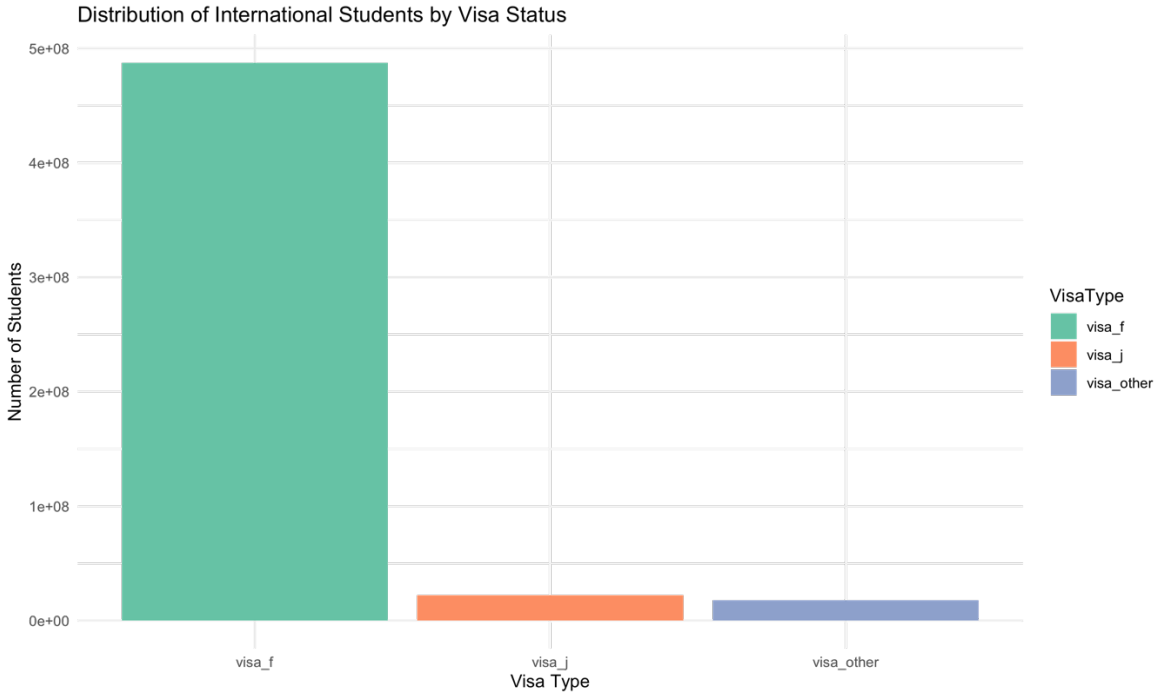
- The maximum number of international students by funding source is 345,358,272 for International Funding Sources.
- The minimum number of international students by funding source is 177,805,968 for US Funding Sources.
- This analysis helps identify the primary sources of funding for international students.

7. Top Universities Analysis:



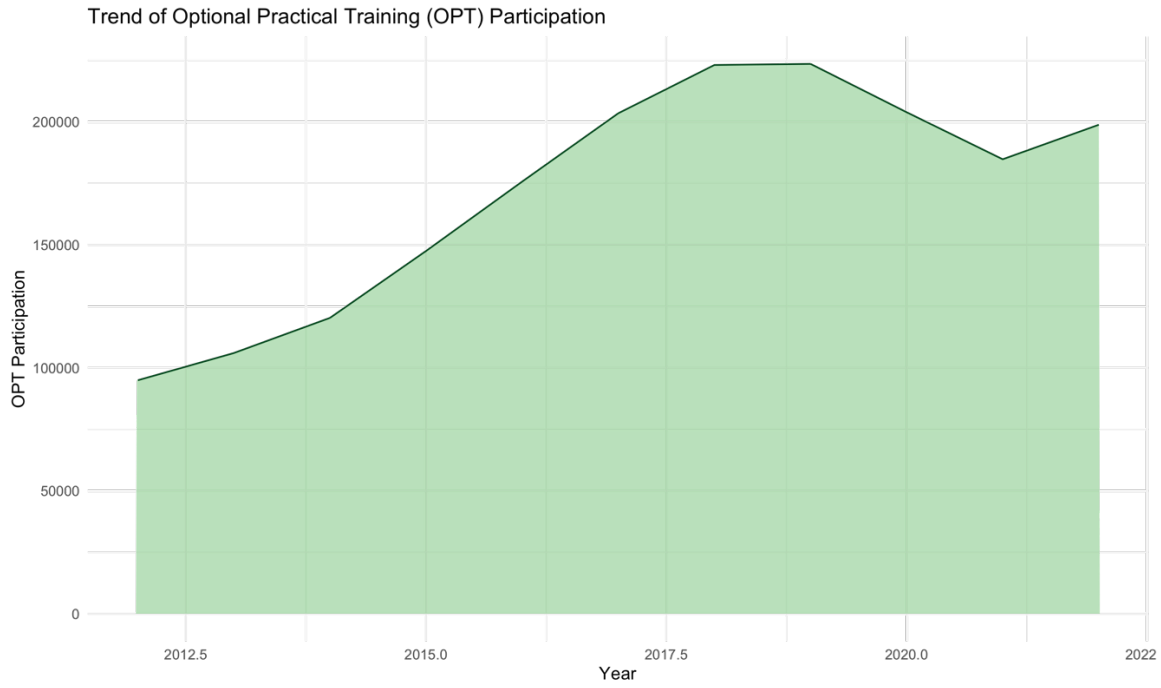
- The top university for international students is US College or University with 95,501,088 students.
- The university with the least international students is Wayne State University with 3,435,984 students.
- This analysis identifies the most popular universities among international students.

8. Visa Status Analysis:



- The maximum number of international students by visa type is 487,381,152 for F Visa.
- The minimum number of international students by visa type is 17,948,304 for Other Visa.
- This analysis provides insights into the visa status of international students.

9. Optional Practical Training (OPT) Analysis:



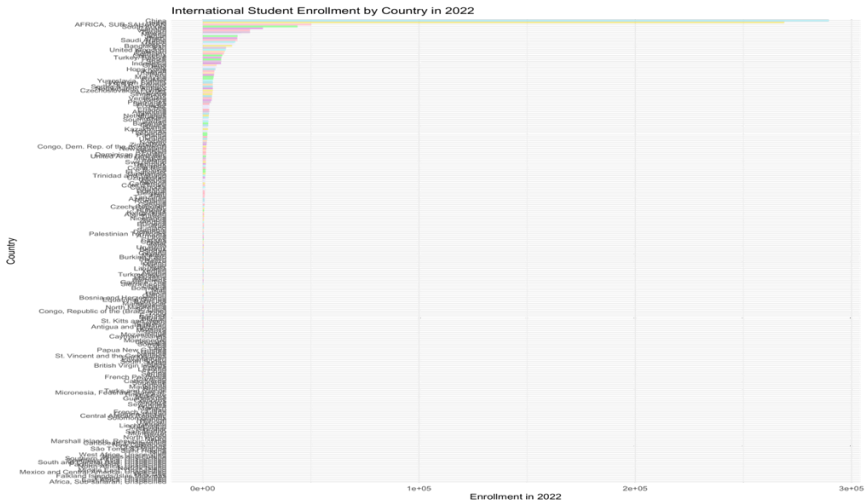
- The maximum number of international students in OPT is 223,539 in 2019.
- The minimum number of international students in OPT is 94,919 in 2012.
- This analysis helps understand the participation of international students in OPT programs.

MET CS 555 | TERM PROJECT

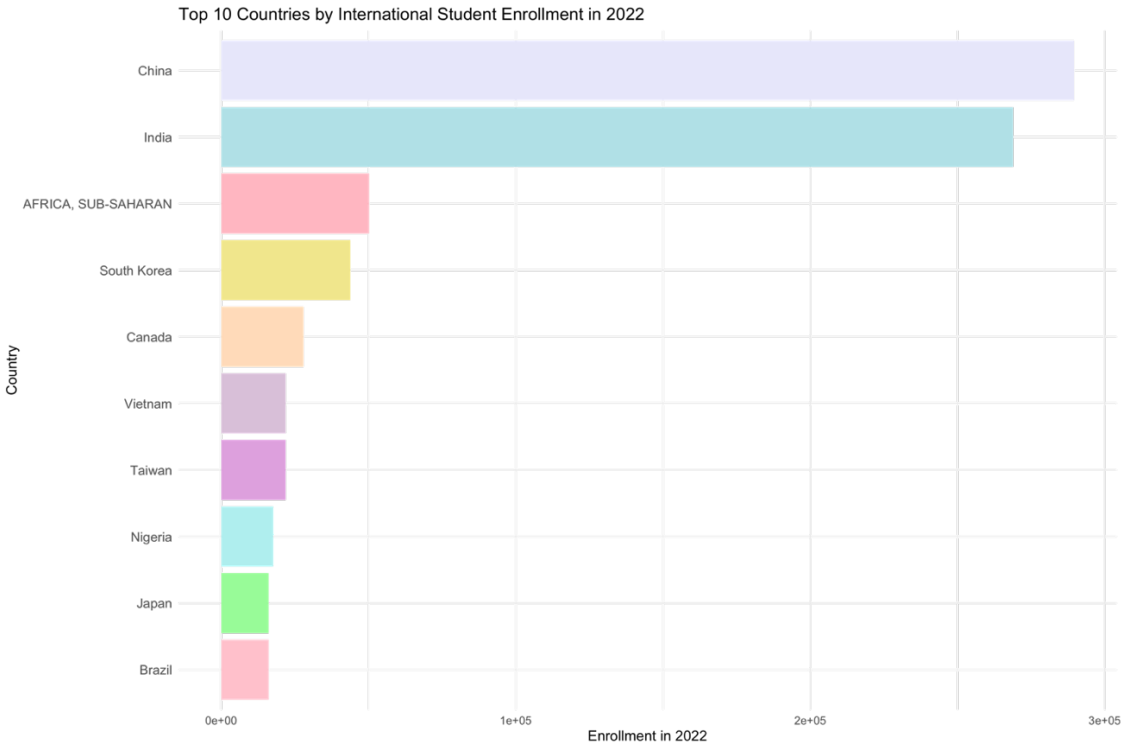
In addition to the analysis of the Combined Data dataset, an exploratory data analysis was conducted on the Country Distribution dataset to gain insights into the distribution of international students in the United States by country of origin.

The following analyses were performed:

1. Top Contributing Countries:



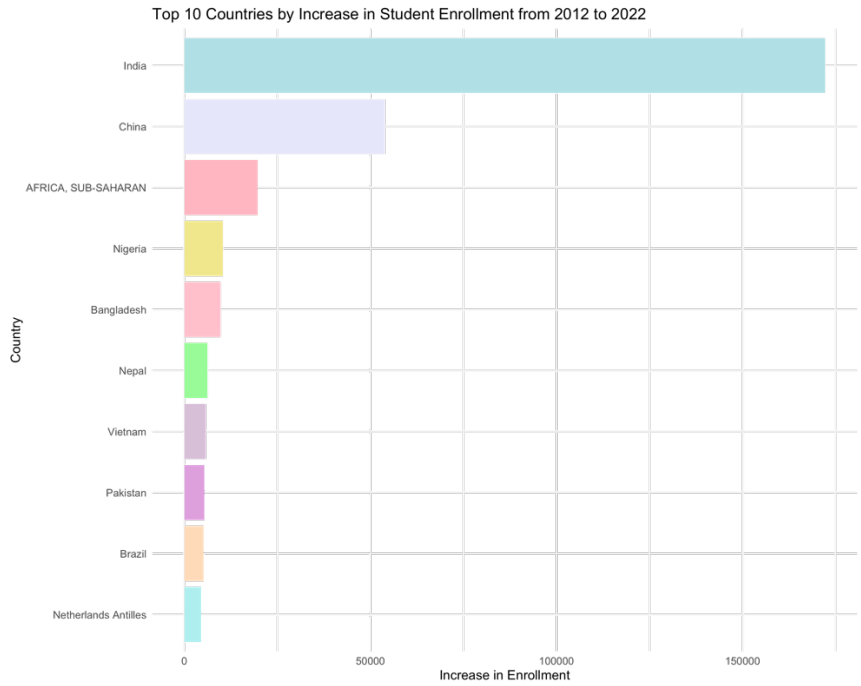
As this isn't readable or a good way to represent data, I created a separate horizontal bar chart was generated to focus on the top 10 countries contributing the highest number of international students in 2022.



- Country with the highest enrolment in 2022: China.
- Country with the lowest enrolment in 2022: Equatorial Guinea

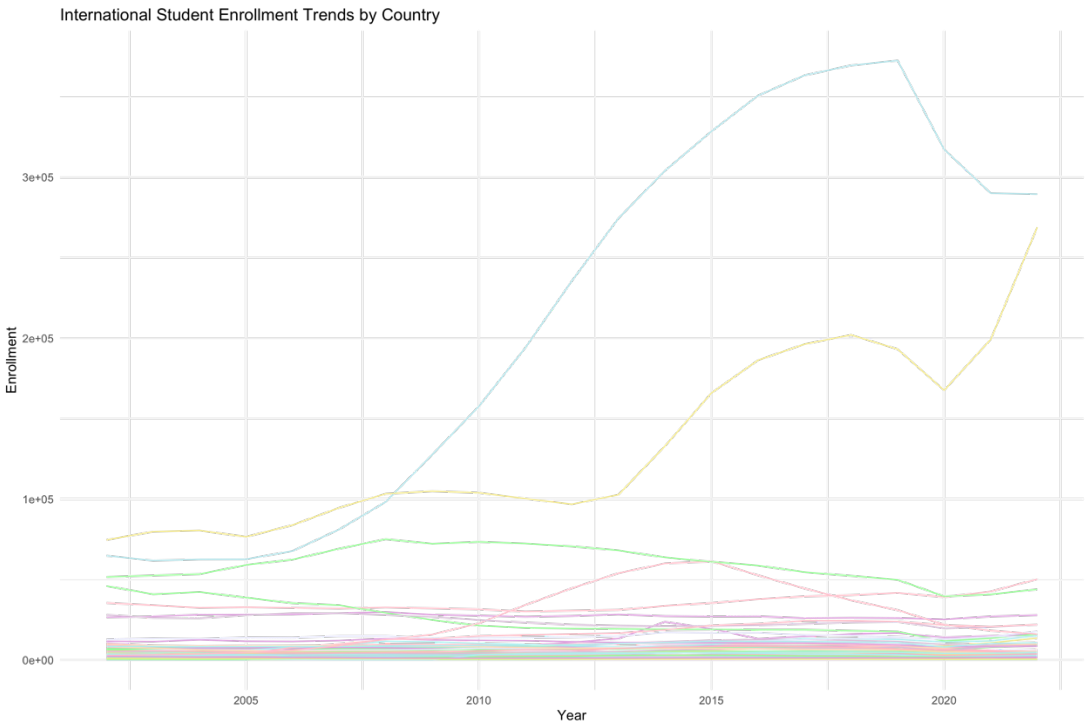
MET CS 555 | TERM PROJECT

2. Enrolment Changes from 2012 to 2022:



- The country with the biggest increase from 2012 to 2022 was India, with an increase of 172,169 students.
- The country with the biggest decrease from 2012 to 2022 was Saudi Arabia, with a decrease of 28,577 students.

3. Enrolment Trends by Country:



- A line plot was generated to visualize the international student enrolment trends by country over the years.
- The plot allows for the comparison of enrolment patterns across different countries and helps identify any notable trends or changes.

MET CS 555 | TERM PROJECT

Conclusion from EDA:

1. United States remains a top destination for international students, with a significant number of students coming from countries like China and India. In 2022, China had the highest enrolment of international students in the US, while India experienced the biggest increase in enrolment from 2012 to 2022.
2. Graduate programs consistently attract a substantial number of international students, indicating a strong interest in advanced studies and research opportunities in the United States.
3. Business/Management and Engineering are the most popular fields of study among international students, suggesting a high demand for these programs and potential career opportunities in these sectors.
4. Personal and family funding is the primary source of financial support for international students, highlighting the importance of financial considerations in the decision-making process.
5. The demographic analysis reveals a consistent growth in both female and male international student enrolment over the years. However, male students constitute a larger proportion of the international student population compared to female students.
6. The country-wise analysis shows variations in enrolment patterns across different countries. China and India consistently rank as the top sending countries of international students to the United States.
7. The United States has experienced a significant increase in the number of international students participating in Optional Practical Training (OPT) programs over the years.
8. The top universities hosting international students in the United States include New York University, University of Southern California, Columbia University, and Northeastern University.
9. F-1 visas are the most common visa type used by international students to study in the United States, followed by J-1 visas.

CORRELATION ANALYSIS:

To further investigate the relationships between variables in the Combined Data dataset, a correlation analysis was performed.

The following steps were taken:

1. Correlation Matrix:
   - A correlation matrix was computed using the cor() function in R, considering only the numeric columns in the dataset.
   - The correlation matrix provides a comprehensive view of the pairwise correlations between all variables.
2. Heatmap Visualization:
   - A heatmap was created using the pheatmap() function from the pheatmaplibrary to visualize the correlation matrix.
   - The heatmap uses a color scale similar to the Seaborn library in Python, with blue representing negative correlations, white representing no correlation, and red representing positive correlations.
   - The heatmap allows for the identification of patterns and clusters of highly correlated variables.

Based on the correlation analysis, here are the key findings:

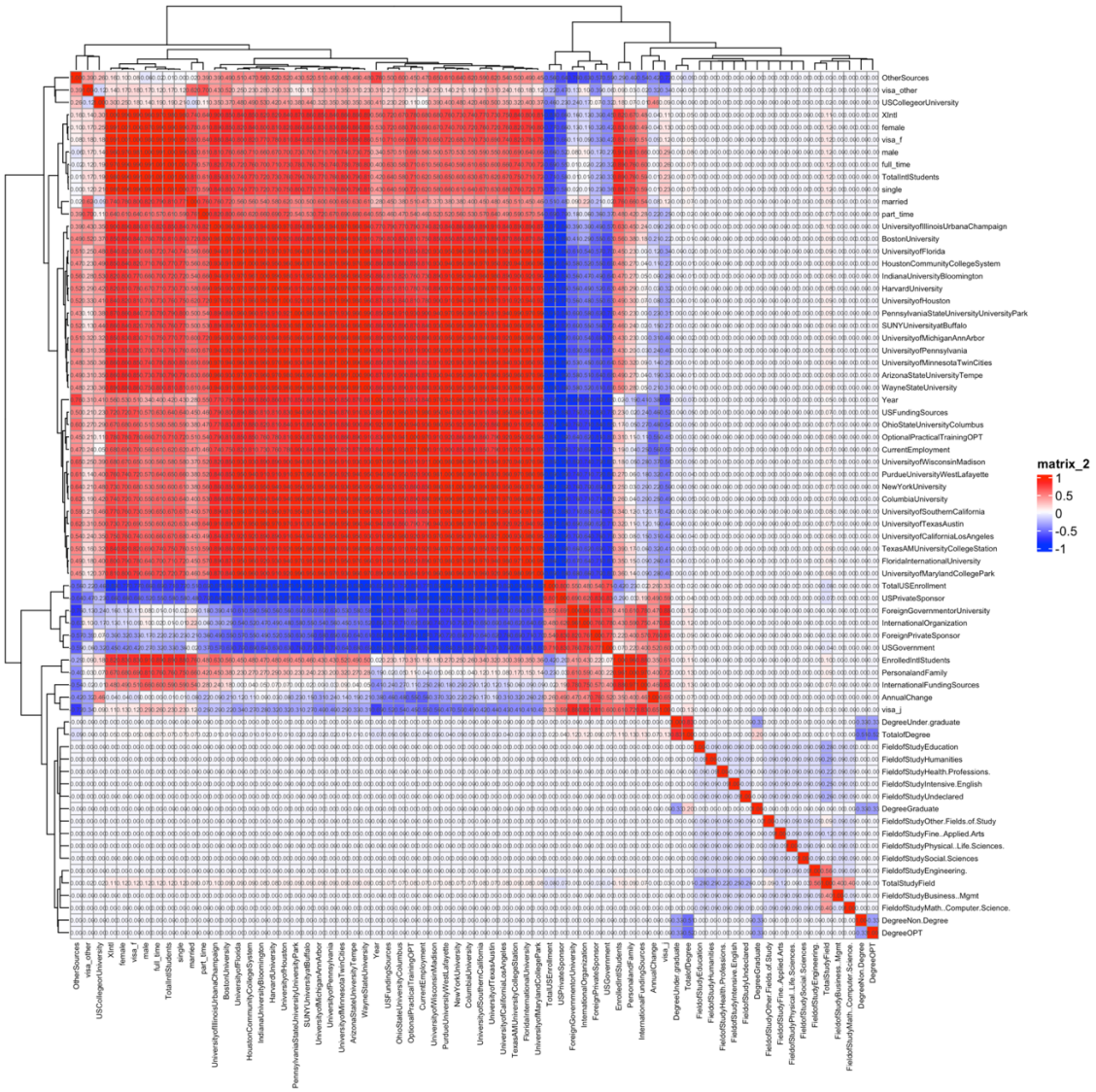
Most Significant Correlations:

- There is a very strong positive correlation (0.998) between the full time students and singles, indicating that a large proportion of full-time international students are single.
- Singles and Total number of International Students are highly positively correlated (0.998), suggesting that a higher number of single students is associated with a larger total international student population.
- Full time students are also strongly positively correlated (0.998) with ' Total number of International Students, implying that full-time enrolment is a significant driver of the total international student population.

MET CS 555 | TERM PROJECT

Least Significant Correlations:

- J visa and Field of Study (Undeclared) have a correlation close to zero, indicating no significant relationship between these two factors.
- Field of Study (Other) has a very low correlation with part time students enrolment and the university Purdue University West Lafayette, suggesting these factors are largely independent.
- The Field Of Study (Health Professions) has a near-zero correlation with the US Government funding source, implying that these two factors are not strongly associated.



In summary, the correlation analysis highlights the strong positive relationships between full-time enrolment, single marital status, and the total international student population. It also reveals several variables with negligible or no significant correlations, indicating their relative independence from other factors in the dataset.

MET CS 555 | TERM PROJECT

Modelling

As part of the modelling process, I started by splitting the data into train, test, and validation sets. This step is crucial to ensure that the models are evaluated on unseen data and to prevent overfitting, which can lead to poor generalization performance.

I decided to use an 80-15-5 split, where 80% of the data was allocated for training the models, 15% for testing, and the remaining 5% for validation.

Note: I have only selected the first 44 columns for modelling purposes, as these columns contained the relevant predictor variables and the target variable (Total International Students). This step helped reduce the dimensionality of the data and potentially improve the models' performance and interpretability.

The modelling process involved several steps, each with a specific purpose and rationale.

1. **Multiple Linear Regression:** I started with a Multiple Linear Regression model as a baseline approach. This model assumes a linear relationship between the target variable (TotalIntlStudents) and the predictor variables. The advantage of this method is its simplicity and interpretability, making it a standard choice for initial exploration. The summary of the Multiple Linear Regression model showed that several variables were statistically significant, indicating their potential influence on the target variable. However, the model suffered from multicollinearity issues, as evident from the singularities reported in the summary.

   Evaluation Metrics:

   - Test MSE: 1.133371e-30
   - Validation MSE: 7.097581e-31
   - Test RMSE: 1.064599e-15
   - Validation RMSE: 8.424714e-16
   - Test R-squared: 1.0
   - Validation R-squared: 1.0

   While the Multiple Linear Regression model achieved perfect R-squared values on both the test and validation sets, it is important to note that these results may be unreliable due to the "essentially perfect fit" warning. This warning indicates that the model has overfit the data, potentially leading to poor generalization performance on unseen data.

2. **Ridge Regression:** To address the multicollinearity issue observed in the Multiple Linear Regression model, we employed Ridge Regression. Ridge Regression is a regularized version of linear regression that adds a penalty term to the cost function, which helps to shrink the coefficients of correlated predictors towards zero, reducing the impact of multicollinearity.

   Evaluation Metrics:

   - Test MSE: 0.0008582452
   - Validation MSE: 0.0009234463
   - Test RMSE: 0.02929582
   - Validation RMSE: 0.03038826
   - Test R-squared: 0.9991036
   - Validation R-squared: 0.9991211

Ridge Regression performed well, with high R-squared values on both the test and validation sets, indicating a good fit to the data. However, it retained all the predictors in the model, making it potentially less interpretable and prone to overfitting.

Selected Features: All features were selected by Ridge Regression.

3. **Lasso Regression:** As an alternative to Ridge Regression, we employed Lasso Regression. Unlike Ridge Regression, Lasso Regression can effectively perform feature selection by forcing some coefficients to become exactly zero, resulting in a more interpretable model.

   Evaluation Metrics:

   - Test MSE: 0.0009855033
   - Validation MSE: 0.0009134396
   - Test RMSE: 0.03139273
   - Validation RMSE: 0.03022316
   - Test R-squared: 0.9989707
   - Validation R-squared: 0.9991306

   Lasso Regression achieved comparable performance to Ridge Regression while providing a more parsimonious model by selecting only a subset of the most relevant features.

   Selected Features: (Intercept), XIntl, female, male, single, full_time, visa_f

4. **Recursive Feature Elimination (RFE) with Random Forests:** To further explore feature selection and potentially improve model performance, we employed Recursive Feature Elimination (RFE) using Random Forests. Random Forests are an ensemble learning method that combines multiple decision trees, making them robust to overfitting and capable of capturing complex non-linear relationships. The RFE approach iteratively fits Random Forest models and eliminates the least important features based on their importance scores. This process is repeated until the desired number of features is reached, or the model performance no longer improves.

   Evaluation Metrics:

   - Test MSE: 1.874777e-30
   - Validation MSE: 1.697358e-30
   - Test RMSE: 1.369225e-15
   - Validation RMSE: 1.302827e-15
   - Test R-squared: 1.0
   - Validation R-squared: 1.0

   The RFE model achieved perfect R-squared values on both the test and validation sets, indicating an excellent fit to the data. However, it is crucial to interpret these results with caution, as they may be indicative of overfitting, similar to the Multiple Linear Regression model.

   Selected Features: AnnualChange, USPrivateSponsor, male, TotalUSEnrollment, single, full_time, XIntl, visa_f, female, USGovernment, ForeignPrivateSponsor, CurrentEmployment, OptionalPracticalTrainingOPT, visa_other, part_time, USCollegeorUniversity, ForeignGovernmentorUniversity, USFundingSources

Based on the evaluation metrics, the Ridge Regression and Lasso Regression models emerged as the most promising candidates, as they achieved high R-squared values while avoiding potential overfitting issues observed in the Multiple Linear Regression and RFE models.

The choice between Ridge Regression and Lasso Regression depends on the specific requirements of the problem at hand. If interpretability and feature selection are priorities, Lasso Regression may be preferred due to its ability to produce a more parsimonious model by setting some coefficients to zero. However, if retaining all the predictors is desirable, Ridge Regression may be a better choice, as it shrinks the coefficients but does not eliminate them entirely.

Answer to my research questions.

Based on the research questions stated and the analysis performed, here are the answers:

a) What are the key factors that significantly influence the number of international students coming to the USA?

The key factors identified through the modelling process are:

- Percentage of international students
- Gender variables: female and male
- Enrolment status: Single and Full time
- Visa status: F visa
- Annual Change (year-over-year percentage change in international students)
- Total enrolment in the USA: Total US Enrolment
- Funding sources: US Private Sponsor, US Government, Foreign Private Sponsor, Foreign Government or University, and US College or University.
- Optional Practical Training OPT (OPT participation)

b) Which countries contribute the most and least to the international student population in the USA in 2022? Also, which countries have shown a significant increase or decrease?

- The country contributing the most international students in 2022 was China.
- The country with the lowest non-zero enrolment in 2022 was Equatorial Guinea.
- The country that showed the biggest increase from 2012 to 2022 was India, with an increase of 172,169 students.
- The country with the biggest decrease from 2012 to 2022 was Saudi Arabia, with a decrease of 28,577 students.

c) Which fields of study and educational levels are most popular among international students in the USA?

- The most popular fields of study among international students were Business/Management and Engineering.
- The degree level with the highest enrolment was Graduate.

d) Which institute is most popular among international students?

- The university with the highest enrolment of international students was the University of Southern California.
- The university with the lowest enrolment was Wayne State University.

e) What is the predicted trend of international students in the USA for the next 5-10 years?

While the analysis does not provide a specific predicted trend for the next 5-10 years, the Lasso Regression model, being one of the best-performing models, can be used to make future predictions by inputting new data for the selected predictor variables. The model's feature selection capabilities suggest that factors such as percentage of international students, gender, enrolment status, visa type, annual change, total US enrolment, and funding sources are likely to play a significant role in shaping the future trends of international student enrolment in the USA.

In conclusion, this research has shed light on the key factors influencing international student enrolment in the USA, the countries contributing the most and least to the international student population, the most popular fields of study and educational levels, and the leading institutions attracting international students. The insights gained from the exploratory data analysis and the predictive modelling techniques can inform strategic decision-making and policy formulation to support and enhance the experiences of international students in the USA.