Final Project Report: Analyzing Erasmus Program Funding Patterns and Predicting Future Allocations

Introduction

The Erasmus program, launched in 1987, is a European Union initiative aimed at supporting education, training, youth, and sport in Europe. With 27 EU Member States and 6 non-EU associated countries, the primary objective of the program is to promote transnational learning mobility and cooperation between organizations and policymakers. This project analyzes historical funding data from the Erasmus program from 2014 to 2019 to identify trends and predict future allocations to different countries.

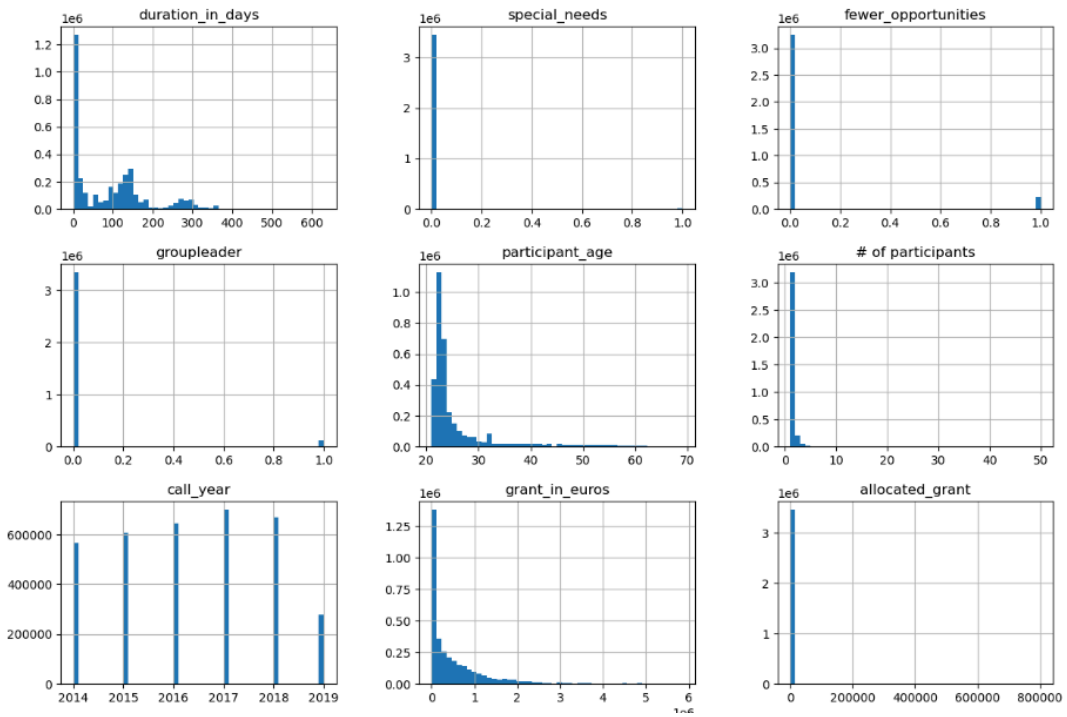
Problem Statement

Based on Erasmus funding from 2014 to 2019, we aim to analyze previous patterns and factors to identify trends and predict which countries are most likely to receive future allocations.

Approach

Data Preparation

1. Data Loading and Merging: We combined projects and funding data to create a comprehensive dataset and review the histograms.



2. Missing Data Handling:

- Imputation using relevant field information.
- Mean imputation.
- Proportional imputation based on known values.
- Imputation using means of similar categories.

3. Data Normalization: Categorical data were normalized, and binary columns were created as needed.

4. Datatypes: Revised and converted to ensure consistency.

5. Date Consistency: Date columns were verified for errors and inconsistencies.

6. Duplicates Removal: Duplicate records were identified and removed.

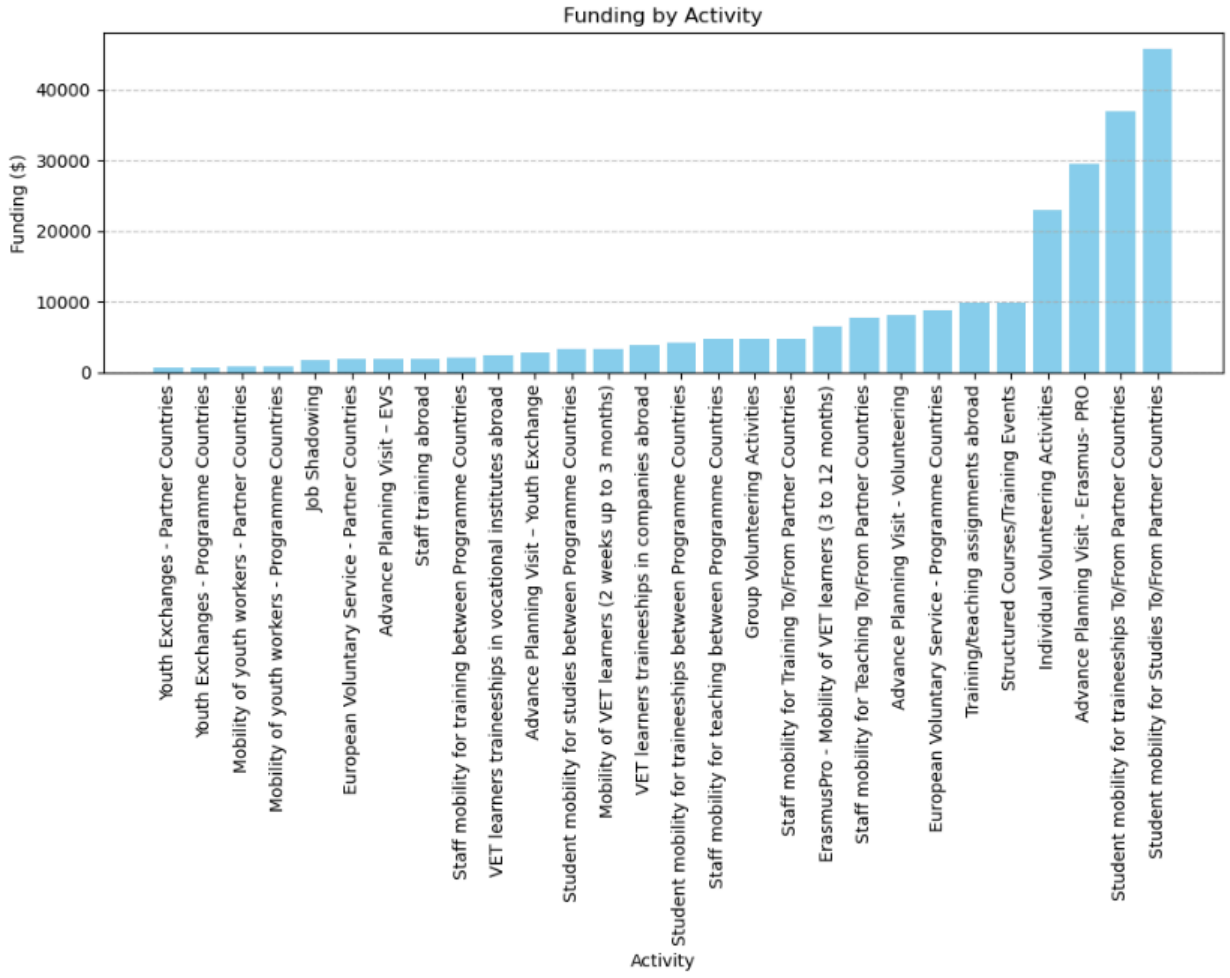
7. Feature Engineering:

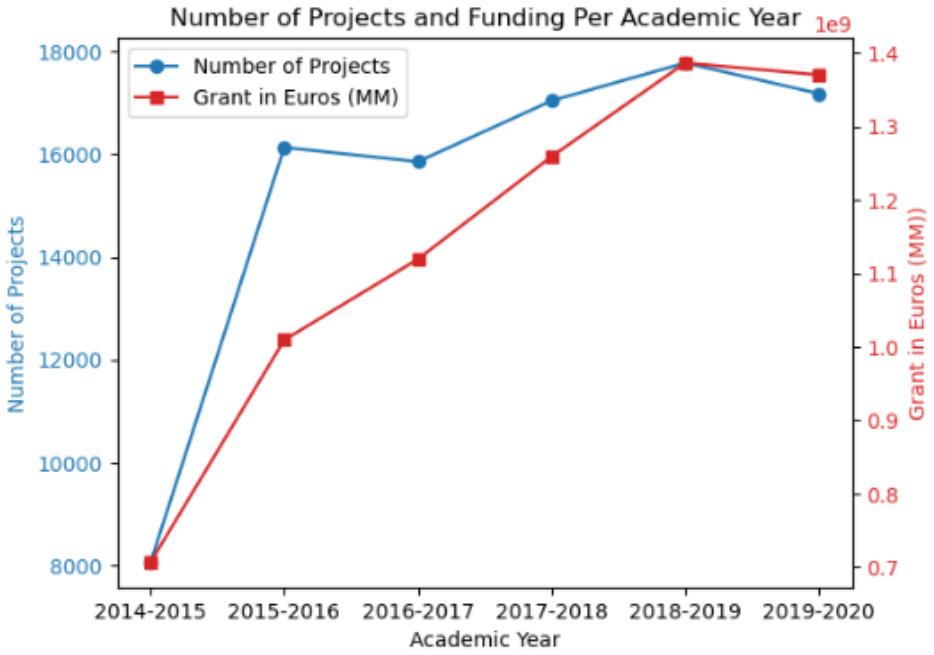
- Created `participant_id` and `allocated_grant` features.
- Aggregated funding, participant counts, project durations, etc., by country and year.
- Introduced lag features and rolling statistics for funding and project counts.
- Formulated a multi-year target variable: `cumulative_future_allocations`.

Exploratory Data Analysis (EDA)

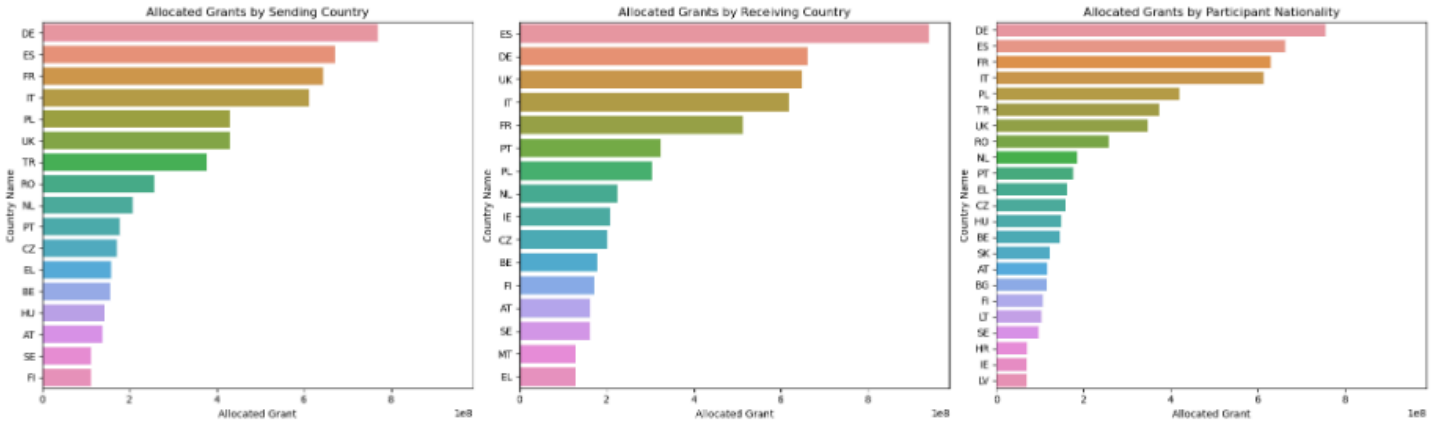
The EDA uncovered several important trends:

- Funding increased over the years, with a significant focus on projects related to student study/traineeship.

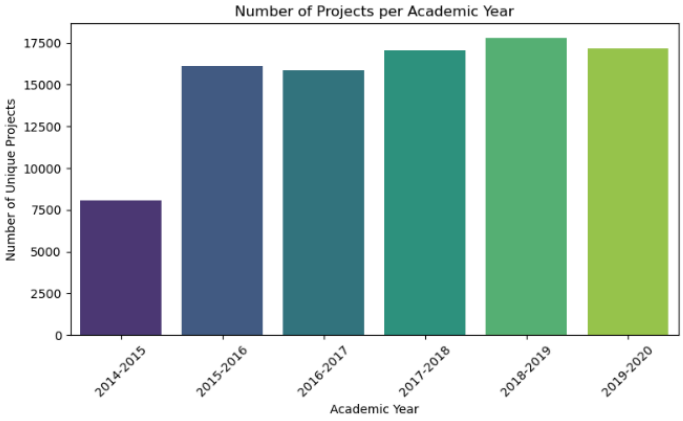




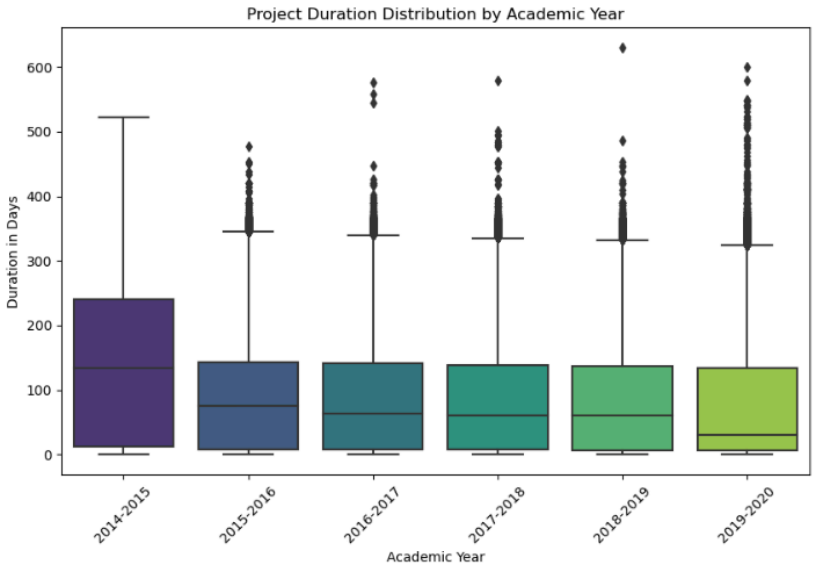
- Germany and Spain received the most funding, followed by France and Italy, with Poland, Portugal, the UK, Turkey, and Romania also being key recipients.



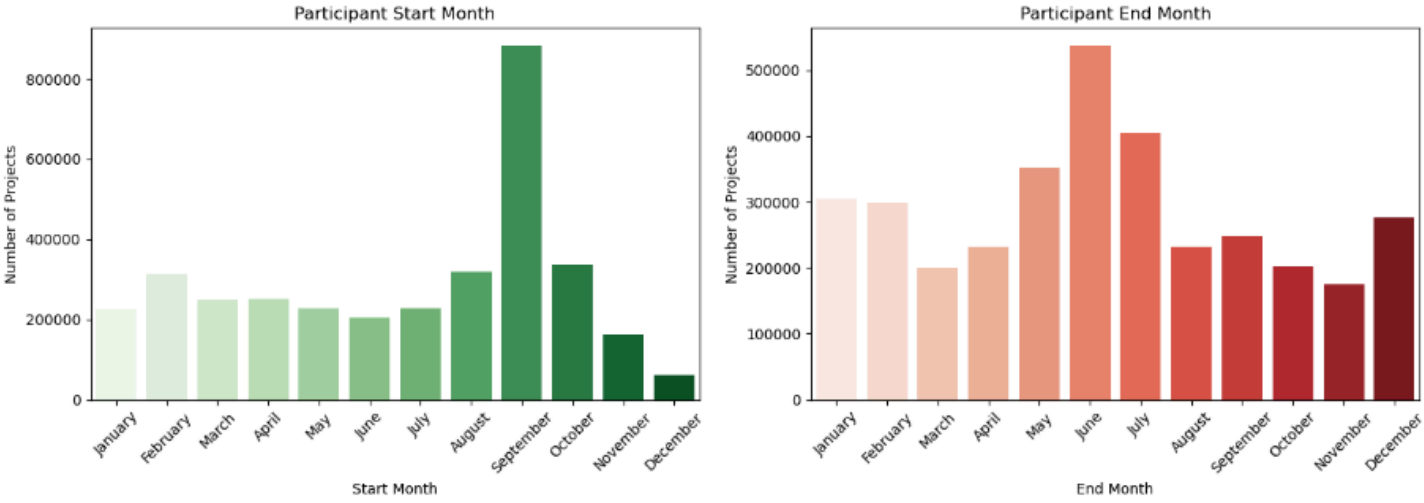
- The number of projects increased after an initial slow start in the first year.



- Project durations were longer in the initial year and then stabilized, but saw a decrease in 2019-2020.



- Participant observations indicated activity clusters starting in September and ending in June, aligning with the academic year. A higher proportion of females participated, with the average participant age being between 20-29 years.



Modeling Preparation

1. Feature Engineering: Derived features for model input, such as country-year aggregations, lag features, rolling statistics, and cumulative future allocations.

2. Categorical Encoding: Hot-encoded categorical variables into a numerical format.

3. Data Normalization: Standardized features to ensure comparable scales.

4. Data Splitting: Applied TimeSeriesSplit to account for the temporal nature of the data.

Model Definition

Per the Modeling Rubric, two models were defined: Linear Regression and Random Forest Regressor.

- **Linear Regression**: A simple linear model to understand baseline performance.
- **Random Forest Regressor**: A more complex ensemble model for potential performance improvement.
- **Hyperparameter Tuning**: Planned to use RandomizedSearchCV for tuning hyperparameters.
- **Model Evaluation**: Intended to use mean squared error (MSE) and R^2 as evaluation metrics.

Challenges Faced

Memory limitations restricted the ability to run the final analysis. Multiple strategies were explored, including:

- Running on Google Colab.
- Utilizing Dask.
- Converting data to sparse format.
- Minimizing hyperparameters and tuning strategies.

Ultimately, even these approaches did not resolve the memory constraints.

Recommendations

Despite the inability to run the final analysis, we can propose recommendations on how the client could utilize the results if a conclusive analysis had been possible:

1. Strategic Allocation Planning:

- By identifying which countries are most likely to need or receive future funds, the Erasmus program can strategically plan its budget allocations. Countries receiving consistent funding could be highlighted for more detailed reviews to ensure alignment with Erasmus objectives.
- Recognizing underfunded countries with potential growth can lead to proactive support and development initiatives, ensuring a balanced distribution of educational opportunities across Europe.

2. Targeted Program Development:

- With insights into participant demographics and project types, Erasmus can design targeted programs addressing specific needs, such as increasing male participation or supporting countries with lower project durations.
- By understanding the funding patterns and areas of high impact, Erasmus can develop specialized workshops, training sessions, or partnerships tailored to enhance the projects' effectiveness in those regions.

3. Policy and Engagement Strategies:

- Leveraging the trends and predictive analytics, policymakers can be better equipped to advocate for continuous or increased funding in certain regions, potentially adjusting policy to support high-engagement programs.
- Engagement strategies could be refined by focusing marketing and informational campaigns in countries identified to have growing or high participant interest . Highlighting successful projects from previous years can serve as inspiration and motivation for new participants and organizations.

Conclusion

Although we faced technical limitations that prevented us from performing the final analysis, the preparatory work and exploratory analysis provided valuable insights into Erasmus program funding and participation trends. By addressing memory constraints, future iterations of this analysis could provide far more detailed and actionable insights.

By formalizing these steps, Erasmus can harness data-driven decision-making to support and expand transnational education, training, youth, and sport initiatives effectively.