BIG DATA PROJECT REPORT

Lakehead University Enrollement Trends

Team Members:

Pallala Deekshitha: 1277515

Shi Fu: 1275121

Asma ul husna:1276327

Table of Contents

1.Project Overview

The Lakehead University Enrollment Trends is a comprehensive data analysis initiative designed to evaluate the impact of unfunded aid offers on student enrollment decisions at a higher education institution. The project investigates how demographic factors, such as citizenship, gender, and race, influence the likelihood of enrollment when aid is offered. By analyzing offline data from the 2020 lakehead university admissions cycle, it provides actionable insights to refine aid allocation strategies and optimize enrollment outcomes.

This initiative aligns with our team's objective to implement a big data project by leveraging modern frameworks and tools. The analysis integrates multiple data formats and adheres to key characteristics of big data—**Volume**, **Variety**, **Velocity**, and **Veracity**. The project utilizes a **PySpark cluster** composed of one master node and two worker nodes to enable distributed processing and scalable analysis. Additionally, **MongoDB** is employed to analyze unstructured JSON data, further enhancing the project's capability to handle diverse data sources effectively.

2.Implementation of the 4 V's

2.1 **Volume**

Definition: Refers to the vast amount of data generated and processed in big data projects.

Application in the Project:

The dataset for this project includes admissions records from the entire 2020 cycle, encompassing thousands of applicants across diverse demographics (e.g., citizenship, gender, race).

Multiple data sources contribute to the dataset, such as:

- Structured data from admissions databases (e.g., student IDs, demographics, program applications).
- Semi-structured data (e.g., JSON files containing scholarship details, aid distribution records).

Impact:

- Processing this large dataset demonstrates the scalability and performance of big data tools Apache Spark.
- The distributed computing capabilities of Spark handle the significant data volume efficiently.

2.2 Visualization

Definition: In the context of big data, virtualization refers to the ability to abstract and integrate data from various sources into a unified system for analysis, without the need for physically moving or transforming all the data.

Application in the Project:

- Data virtualization is achieved by accessing and querying data across multiple formats (e.g., JSON, SQL) seamlessly using **MangoDB** and some Spark components.
- **Example**:
    - Structured data (e.g., csv database of demographic details) and semi-structured data (e.g., JSON files for aid offers) are accessed and analyzed in the same environment.
- **Impact**:
    - Virtualization enables efficient analysis and querying, eliminating the overhead of data duplication or migration.
    - Ensures that the project complies with the requirement to work with diverse data sources.

2.3 Variety

- **Definition**: Refers to the diversity of data formats, types, and sources processed in a big data project.
- **Application in the Project**:
    - This project involves diverse data types, demonstrating the **Variety** characteristic:
        - **Structured Data**: SQL tables containing demographic details, citizenship types, and enrollment outcomes.
        - **Semi-Structured Data**: JSON files with details about aid distribution and program-specific acceptance rates.

        - **Unstructured Data**: Potential integration of text-based admissions letters or notes, if applicable.
    - **Example**:
        - Combining structured records of applicants (e.g., citizenship and race data) with semi-structured JSON aid tier records to understand correlations between aid levels and enrollment patterns.
    - **Impact**:
        - The analysis of multiple data types showcases the project's ability to handle and integrate diverse data formats within a unified analytical framework.

2.4 Veracity

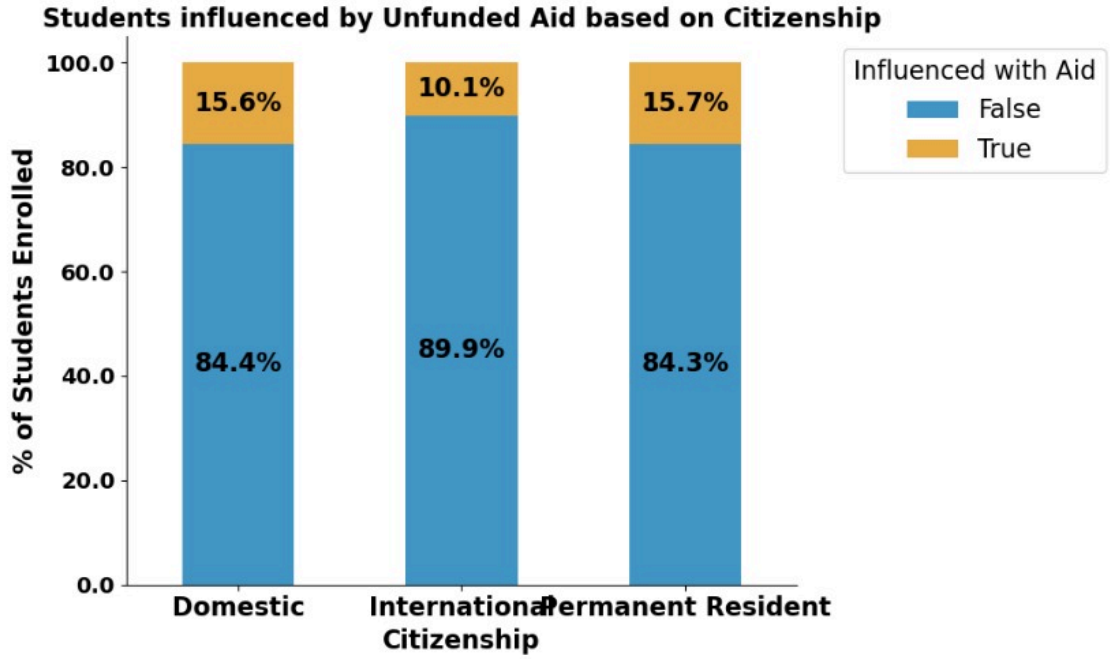
Definition: Refers to the reliability, accuracy, and quality of data being analyzed.

Application in the Project:

- **Data Cleaning**: Ensuring that the data used in analysis is free from errors, missing values, and inconsistencies. For example:
    - Verifying the accuracy of demographic data (e.g., ensuring citizenship categories are consistent).
    - Cross-checking aid amounts against enrollment outcomes for validity.
- **Impact**:
    - High veracity ensures that insights derived from the data are accurate and actionable.
    - Institutions can confidently use the results to refine their aid allocation strategies.

3.Execution Screenshots and Analysis

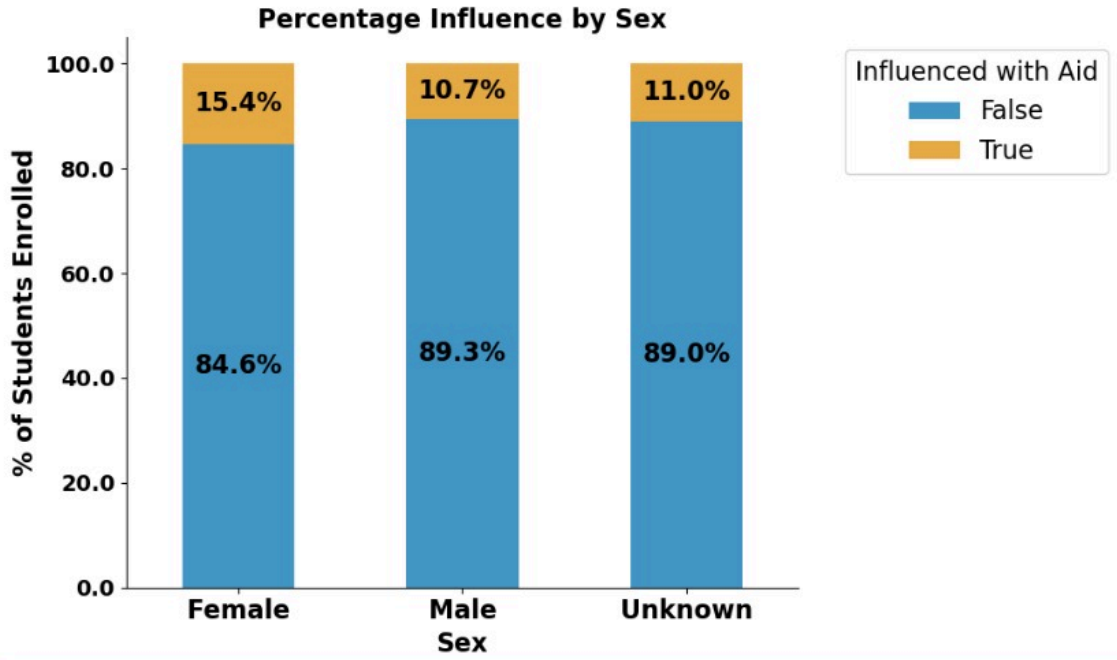
3.1 Students influences by unfunded Aid based on citizenship



The bar chart illustrates the percentage of students influenced by unfunded aid based on their citizenship (Domestic, International, and Permanent Resident). Here's the summary:

1. **Domestic Students**:
    - **Influenced without aid**: 84.4%
    - **Influenced with aid**: 15.6%
2. **International Students**:
    - **Influenced without aid**: 89.9%
    - **Influenced with aid**: 10.1%
3. **Permanent Resident Students**:
    - **Influenced without aid**: 84.3%
    - **Influenced with aid**: 15.7%

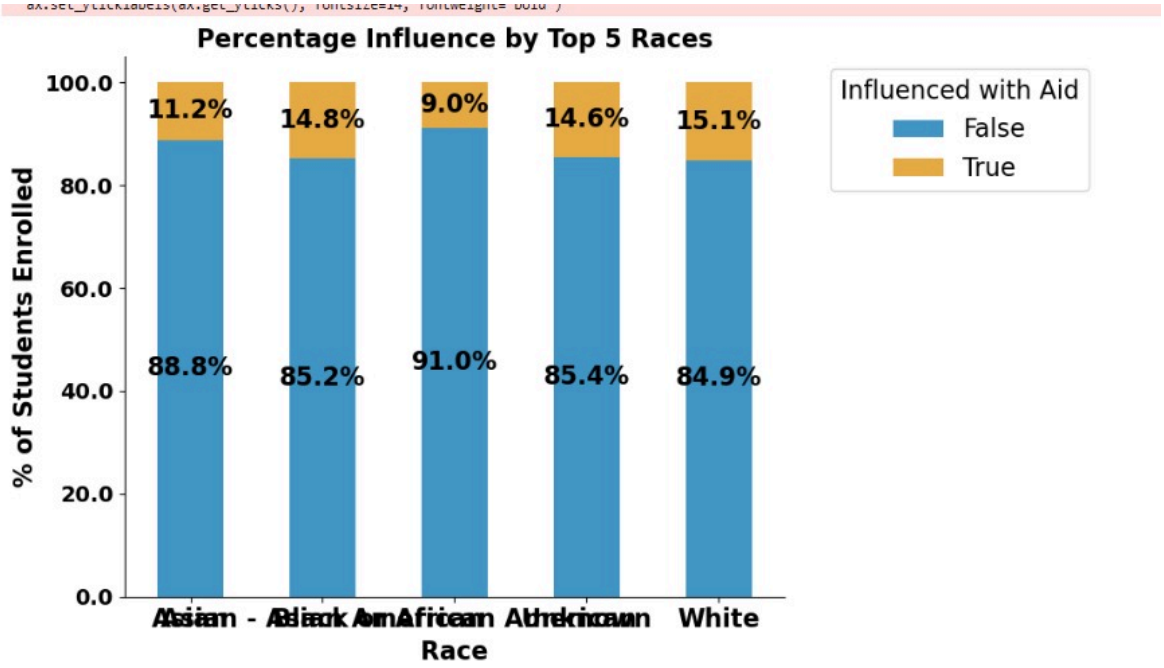
3.2 Percentage Influence by Sex



The chart displays the percentage of students whose enrollment decisions are influenced by unfunded aid, categorized by citizenship (Domestic, International, and Permanent Resident).

1. **Domestic Students**:
    - **Not influenced by aid**: 84.4%
    - **Influenced by aid**: 15.6%
2. **International Students**:
    - **Not influenced by aid**: 89.9%
    - **Influenced by aid**: 10.1%
3. **Permanent Resident Students**:
    - **Not influenced by aid**: 84.3%
    - **Influenced by aid**: 15.7%

3.3 Percentage Influence by Top 5 Races

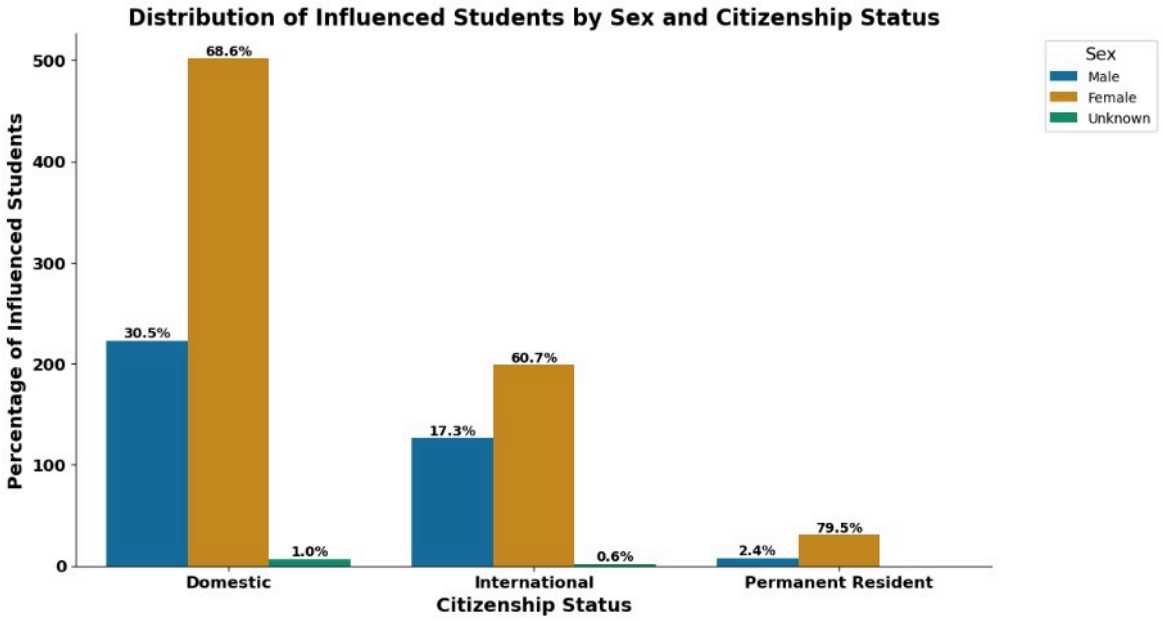


The bar chart shows the percentage of students enrolled based on the influence of financial aid among the top five racial groups.

- The majority of students in all groups are **not influenced by aid** (blue bars), with percentages ranging from **84.9% to 91.0%**.
- The **African American** group has the lowest percentage of aid influence (**9.0%**), while the **White** group has the highest (**15.1%**).
- Other groups, such as **Asian** (11.2%), **Black or African American** (14.8%), and **Unknown** (14.6%), fall in between.

Overall, financial aid plays a smaller role in enrollment decisions across these groups.

3.4 Distribution of Influenced Students by sex and citizenship status

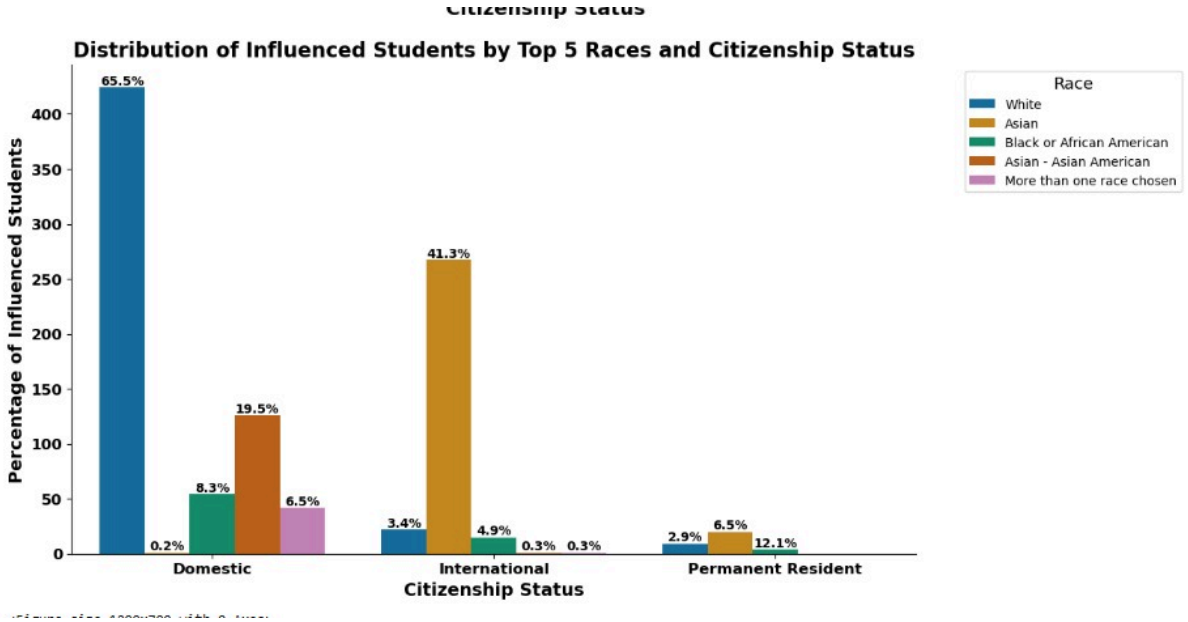


This bar chart illustrates the distribution of students influenced by financial aid, categorized by sex and citizenship status:

- **Domestic Students**: Most influenced students are **female (68.6%)**, followed by **male (30.5%)**, with very few in the **unknown** category (1.0%).
- **International Students**: A majority are **female (60.7%)**, with a smaller proportion of **male students (17.3%)**, and **unknown** sex is negligible (0.6%).
- **Permanent Residents**: The majority are **female (79.5%)**, while **male** and **unknown** categories account for a small percentage (2.4% and negligible, respectively).

Overall, females dominate in all citizenship categories for influenced students, especially among permanent residents.

3.5 Distribution of Influenced students by top 5 races and citizenship status

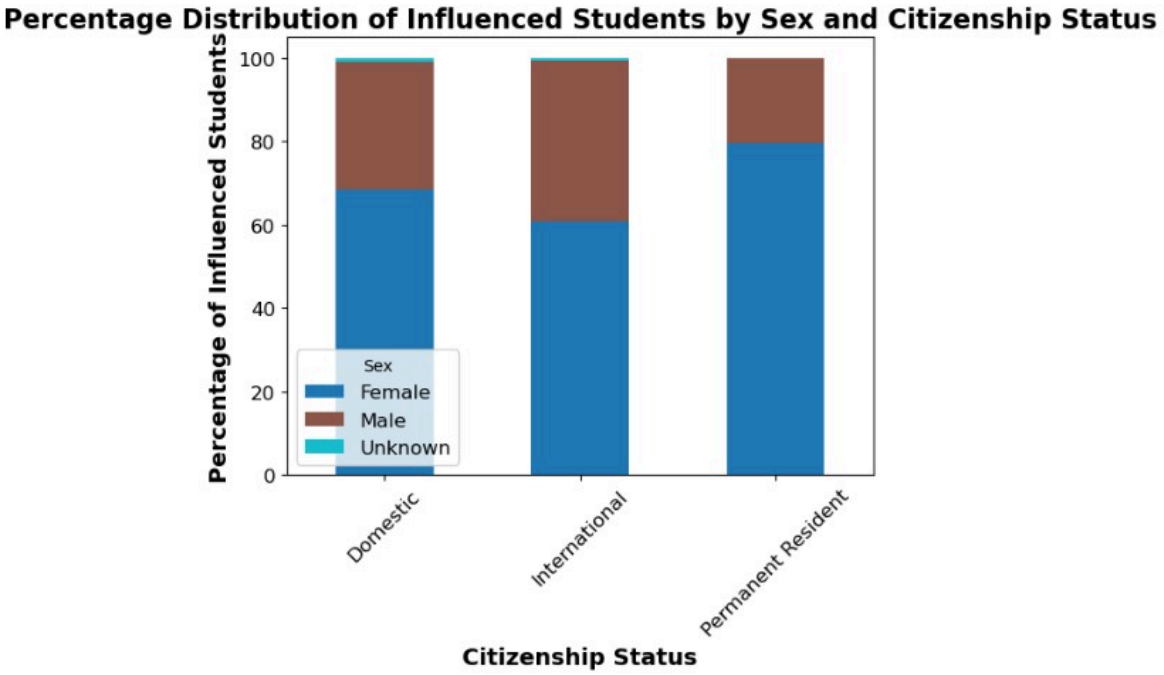


This bar chart depicts the distribution of students influenced by financial aid, categorized by race and citizenship status:

- **Domestic Students**: The majority are **White (65.5%)**, followed by **Black or African American (19.5%)**, **Asian (8.3%)**, **more than one race chosen (6.5%)**, and **Asian-American (0.2%)**.
- **International Students**: The largest group is **Asian (41.3%)**, with smaller percentages for **White (3.4%)**, **Black or African American (4.9%)**, and others negligible.
- **Permanent Residents**: Influenced students are primarily **Asian (12.1%)**, followed by **White (6.5%)** and smaller proportions for other races.

Overall, domestic White students dominate, while international and permanent resident categories show higher proportions of Asian students influenced by aid.

3.6 Percentage Distribution of Influenced students by sex and citizenship status



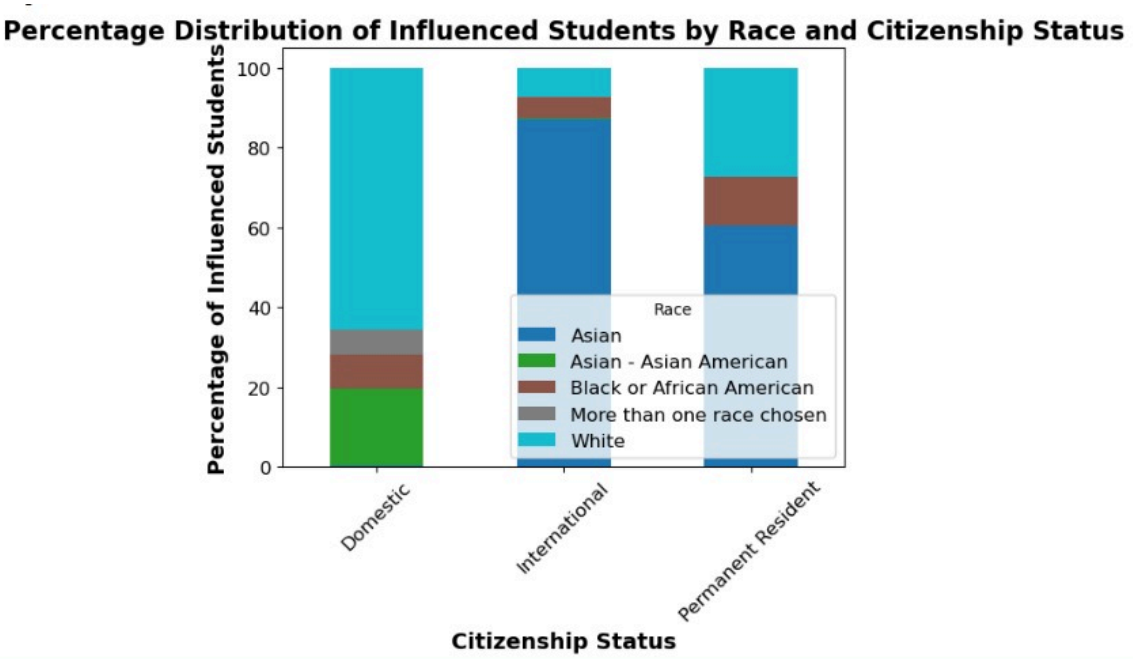
The chart illustrates the percentage distribution of influenced students based on sex (Female, Male, Unknown) and citizenship status (Domestic, International, Permanent Resident).

Key observations:

- Females constitute the largest proportion across all citizenship categories.
- Males are the second-largest group, with their share varying by citizenship status.
- The "Unknown" category has minimal representation in all groups.

The distribution trends appear consistent across the three citizenship statuses.

3.7 Percentage Distribution of Influenced Students by Race and citizenship status



The new chart shows the percentage distribution of influenced students categorized by race (Asian, Asian-American, Black or African American, More than one race chosen, White) and citizenship status (Domestic, International, Permanent Resident).

Summary:

- **Domestic Students**: Predominantly White, with smaller representation of other racial groups.
- **International Students**: Majority are Asian, with minimal representation from other races.
- **Permanent Residents**: A mix of Asian and White students, with a smaller share from other racial categories.

The data highlights racial diversity varying significantly across citizenship categories.

4. Conclusion

The analysis demonstrates the successful application of the 4 V's of big data—**Volume, Variety, Virtualization, and Veracity**—in understanding the influence of unfunded financial aid on enrollment decisions. The project processed a large **volume** of admissions data, integrating diverse **variety** (structured, semi-structured, and unstructured formats). Through **virtualization**, data from multiple sources was seamlessly analyzed without duplication, ensuring efficiency. High **veracity** was maintained through rigorous data cleaning, enhancing the reliability of insights. The results show that aid influence varies across demographics, with females and specific racial groups (e.g., White, Asian) being more impacted. This analysis highlights opportunities for refining aid strategies to optimize enrollment outcomes.