Unearthing the Patterns in CU's Graduate Admission Process

Niranjan Cholendiran, Santhosh Pattamudu Manoharan, Saaijeesh Sottalu Naresh, Abiram Vyas
Department of Data Science, University of Colorado Boulder, United States.

Abstract

Over the last few years, we have seen a drastic increase in the number of applications that the universities in the USA have received. The field of Data Science has especially boomed during the latter part of the previous decade and has continued to grow at a steady pace. This report focuses on some of the trends seen in student applications over the last ten years and attempts to conclude some of the findings through various hypothesis-testing methodologies. The dataset used to generate the findings in the report is provided through the Institutional Research, Office of Data Analytics at The University of Colorado Boulder. The obtained data is cleaned, and exploratory data analysis is done on the same to observe the trends in the data as well as some of the underlying patterns. We finally observed that the admission ratio at the University of Colorado Boulder is influenced by various factors such as the applicant's residency status and submission of standardized test scores- which was used to estimate the trends of the entire United States admission process.

1. Introduction

The goal of the report is to gain an in-depth knowledge about the master's degree admissions at The University of Colorado Boulder by analyzing the application trends and admission ratios. With this motive we have tried to answer some of the key questions regarding the applicant trends and course popularity which is primarily driven by the rapid development in technology [1]. By evaluating the historical data, we aim to estimate the consistent growth in the application pool. This has the potential to have a massive impact on universities to facilitate courses to meet the evolving academic and industry demands.

Nowadays, students are skeptical about choosing their courses and how to prepare their applications to secure admission. We have explored a holistic review process by gathering some insights from the admission data.

The Institutional Research, Office of Data Analytics at CU Boulder currently maintains a Tableau workbook that shows the information about graduate applicants, their admit status, and newly enrolled students at CU Boulder from 2011 through the most recent year. This workbook is a live

workbook, and the data is updated each fall with the latest application season. The data fed into the Tableau workbook is from CU Boulder's student information system. Over the various academic years, there have been situations where the students are forced to take a break away from CU. This could be due to academic probation, a semester break, or a dropout from the program itself. Aside from this, more emphasis on exchange programs has been placed. The data procured currently do not account for the two scenarios mentioned and solely focus on the admission process for full-time enrolment. The data is not available before 2011.

Ever since COVID-19, there has been a trend seen where mandatory declaration of standardized scores has been waived. This has led to a huge influx of applicants for the plethora of programs CU Boulder has to offer. Also, 2021 saw the introduction of Data Science as a graduate program which further increased the number of applicants. This is in line with the demand for Data Science graduates in the field. The dataset obtained contains a mix of students aiming to pursue a higher-level education - Masters, PhD, or Law. For our report, emphasis would be placed solely on the applicants who had applied for the master's program.

For presenting our findings in this report, the report has been split into multiple parts, each serving a specific purpose. The following section covers a detailed overview of the data. Followed by which, some of the pre-processing steps that were performed to get the data to an analysis-ready state were covered along with some insights from the data. The subsequent section presents two hypothesis tests that were conducted using statistical methods in R Programming. It describes the null and the alternate hypothesis for the statement whose validity was tested. The next section focuses on the results and conclusions obtained from the exploratory data analysis as well as the hypothesis testing phase. Finally, the report discusses some of the challenges and outlines the future work of this paper.

2. Data

As mentioned in Section 1, the graduate applicant data is provided as the backend data source to populate the Tableau Workbook. This data is obtained from the Institutional Research, Office of Data Analytics at CU Boulder. The Tableau Workbook is available to the public and can be downloaded and analyzed as well.

As far as the data structure is concerned, a possible bias in the data could be the volume of applications from international students. Over the last few years, there has been a large inflow of applicants from outside the USA, which eventually under-represents the resident applicants. This might skew the outcome of our exploratory data analysis towards the international students but in further sections, there is an attempt to prove a particular hypothesis in favor of resident students [2].

The data that is made available is present from 2011 and no data before 2011 is given. The data does not provide any applicant's personal information but instead restricts it to the applicant's undergraduate demographic information such as state, location, city, and country of their undergraduate institution. There are quite a few critical columns that are part of the data that would be used for further exploratory data analysis and potential hypothesis testing. Some of the key columns are as follows -

- **Admit_Status** - Status of the application (admitted & enrolled, admitted & not enrolled, applied & not admitted)
- **Level** - level of graduate studies (Master, Doctoral, Law)
- **Residency** - indicates residency of Colorado (Resident, Non-Resident)
- **Bach_GPA** - undergraduate GPA
- **GRE2011 Quant** - Quant score in GRE (nulls indicate that the score has not been declared)
- **GRE2011 Verbal** - Verbal score in GRE (nulls indicate that the score has not been declared)
- **Race Ethnicity** - Applicant's race and ethnicity (African American, Asian, American Indian, Hispanic/Latino, Native Hawaiian, White)
- **Major** - Graduate major the applicant wishes to pursue.

Most of the analysis specified in this report involves the use of the above list of columns.

3. Methods

3.1 Data Preprocessing

In Streamlining the dataset, we aimed to optimize the relevance by performing basic data cleaning tasks. We started by selecting features after a meticulous review, which led us to retain only 11 features that posed great significance for our research. Then, we renamed all the column names to lowercase to increase the readability of the code. After that, we conducted a set of examinations to confirm the presence of any missing data points, and to ensure the completeness of the dataset.

Furthermore, we have evaluated the datatypes of all the features to ensure uniformity in the dataset which is helpful for further transformation. Also, we have encountered 12 duplicated rows and handled them to avoid redundancy in the data. Finally, since our research is based only on the master's degree, we have applied the filtered data only to the "Master" level of study. These were the sequential processes that helped us to go forward with our research.

3.2 Exploratory Data Analysis

The goal of the exploratory data analysis is to understand and unearth some of the underlying patterns and insights that can be obtained from the applicant dataset. This analysis could be vast with a variety of features being a part of the analysis. For the scope of this report, the primary focus would be the columns specified in the previous section.

3.2.1 Trend of applicants over time

The goal of this analysis is to understand the trend of applicants over time. The assumption is that over the last 12 years, the number of applicants has steadily increased albeit a small dip during the covid year of 2020. The following insights were generated from the data for this analysis:

- The total volume of applications has steadily increased since 2011 (Fig 1).
- From the year 2011, there has been a gradual increase in the trend of applicants. 2019 - a small dip - could have been considered as a one-time dip (macroeconomic factor), and was expected to pick up the following year, but due to COVID-19 the applicants remained the same as in 2018 with borders opening towards early 2021 [3].
- A huge spike in the number of applicants has been seen in the years 2021 and 2022 post which the numbers decreased for 2023. This can be because of economic recession across countries.

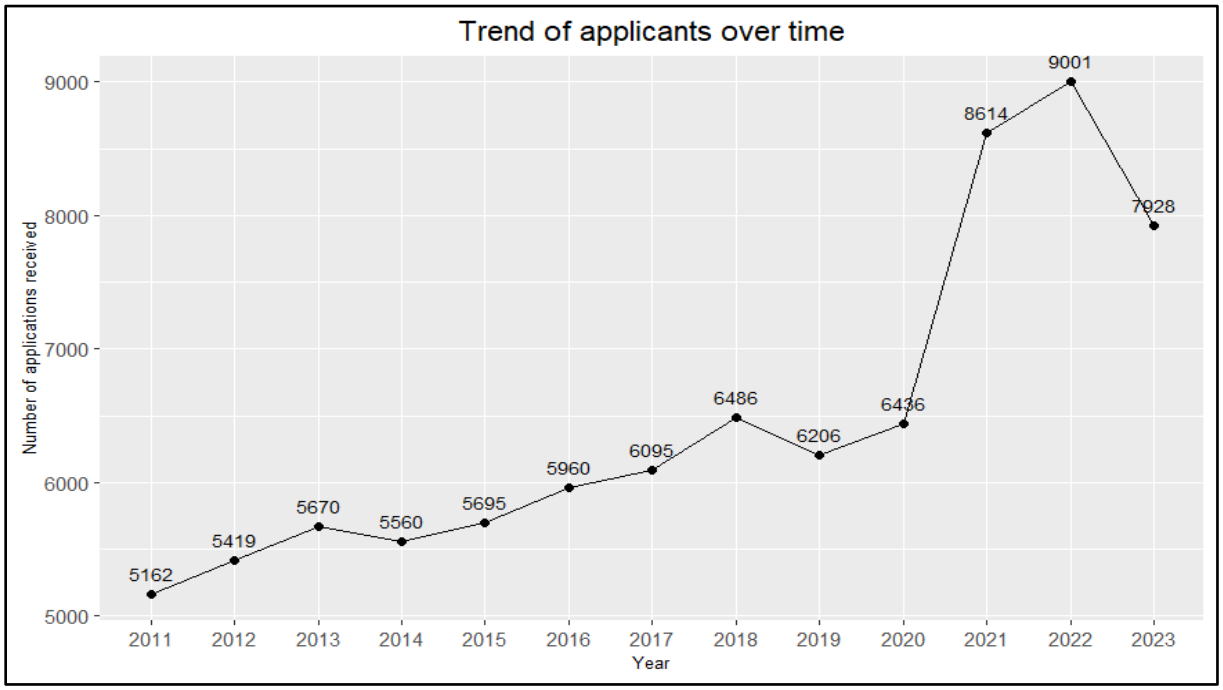

Fig 1: Number of applications received since 2011.

- In Fig 2., the admitted ratio is defined as the ratio of applications that were admitted, and the enrolled ratio is defined as the ratio of admitted applications that were enrolled.

- The increase in several applications correlates with the admitted ratio- meaning CU Boulder has been admitting more students as the number of applications increases. This could indicate one of the following - the number of seats for a course has increased or there has been a gradual increase in the number of graduate courses offered by the university.
- The increasing trend of admission ratio might be because the university decides to compensate for the decreasing enrollment ratio to attain enough enrollments.

Both the exploratory data analysis for this section are presented together and the points mentioned above are clearly seen through the trends and patterns. This serves as a good benchmark to analyze how the applicant traffic has been at CU and the enrolment ratio for the set of admits given to applicants across various graduate programs.

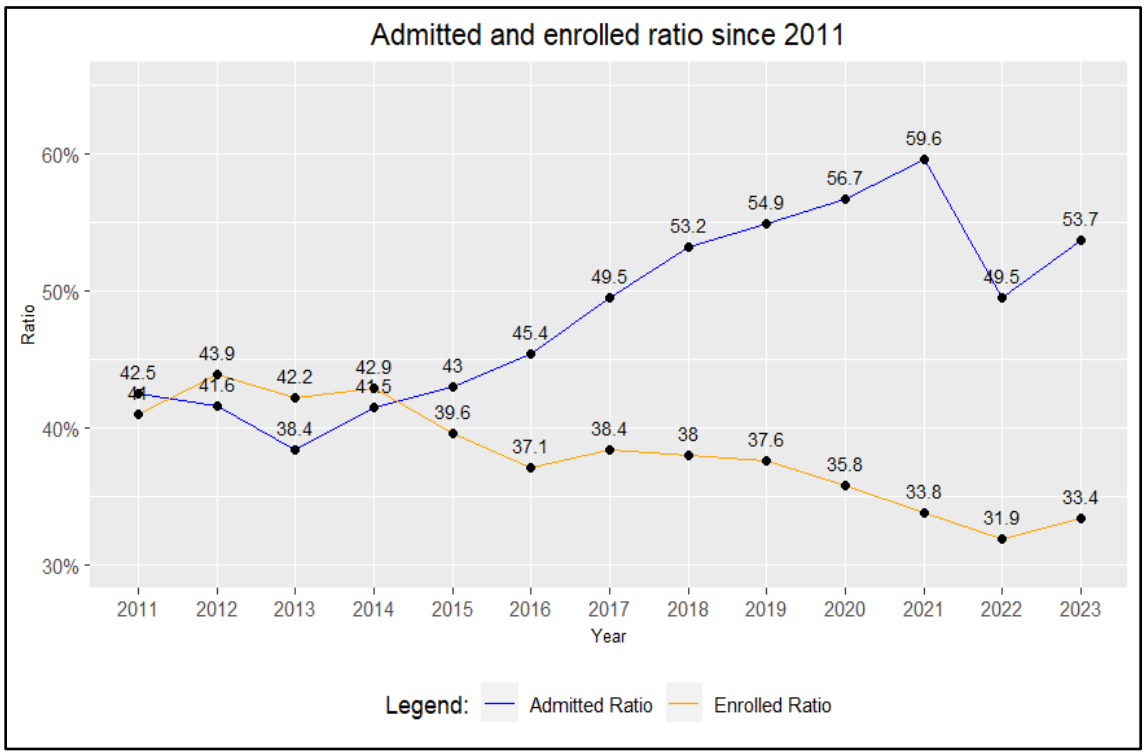

Fig 2: Admitted and enrolled ratio since 2011

3.2.2 Course Popularity at CU

CU is known for its great curriculum for programs offered by its aerospace engineering, electrical, and geography departments. On the contrary, the world is moving towards a technology-driven state which makes courses such as computer science and business analytics have high demand. Furthermore, there have been courses that have been launched in the last 5 years, e.g., data science,

that have quickly gained popularity. Hence, this EDA is done to understand the course popularity at CU.

Students at CU Boulder had a diverse range of 73 majors to choose from in the academic year 2023. However, the distribution of applications across these majors was not uniform, with only ten majors accounting for 64.6% of all applications received (Fig 3).

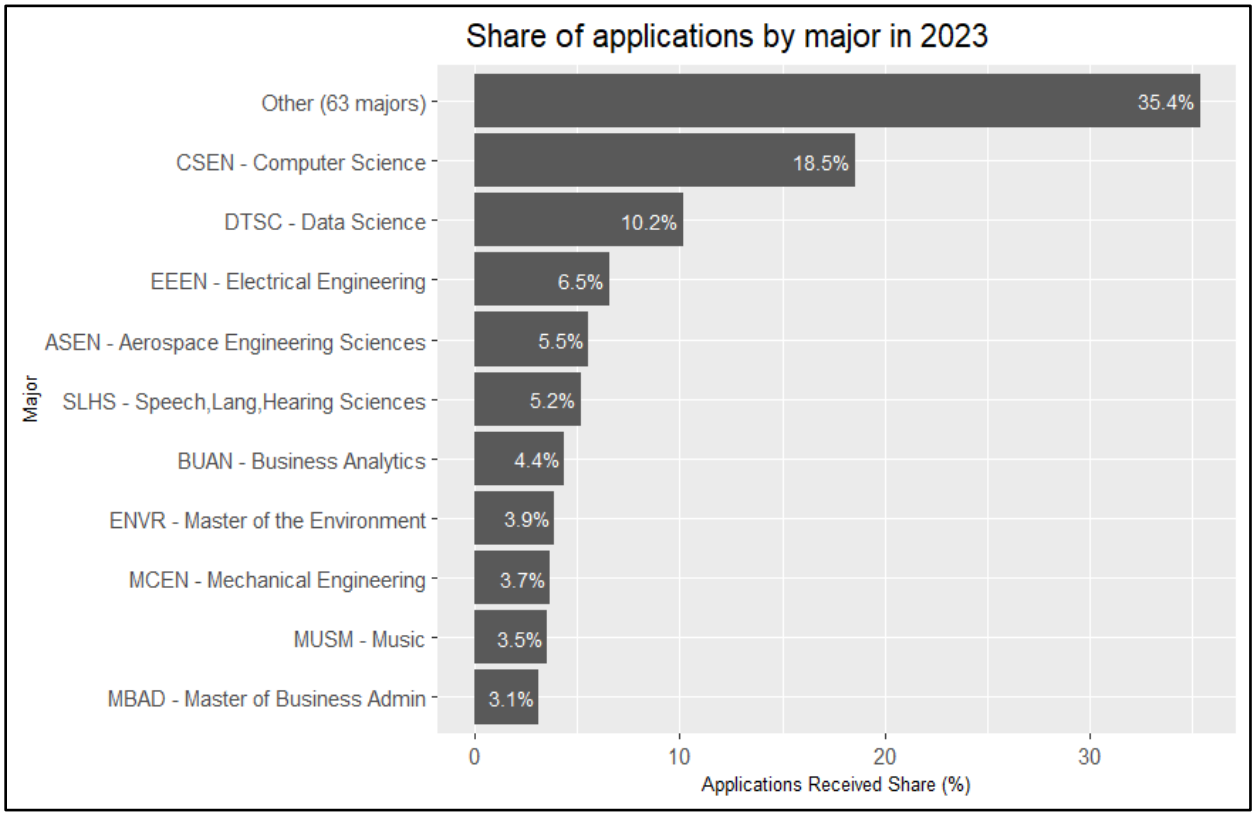

Fig 3: Share of applications by major in 2023

As expected, Computer Science stands out as the most sought-after major, receiving the highest number of applications. This popularity can be attributed to the field's relevance and job market demand. Surprisingly, Data Science, introduced as a major in 2021, quickly gained traction and became one of the top contributors to the application pool. Additionally, Electrical Engineering emerged as another major with a significant number of applications, demonstrating its popularity among prospective students.

3.2.3 Influence of standardized scores on admission status

Universities in the USA have reduced the mandate on the submission of standardized test scores (GRE and GMAT) for applicants thereby looking purely at the profile of the student (undergraduate GPA, recommendations, work experience, etc.).

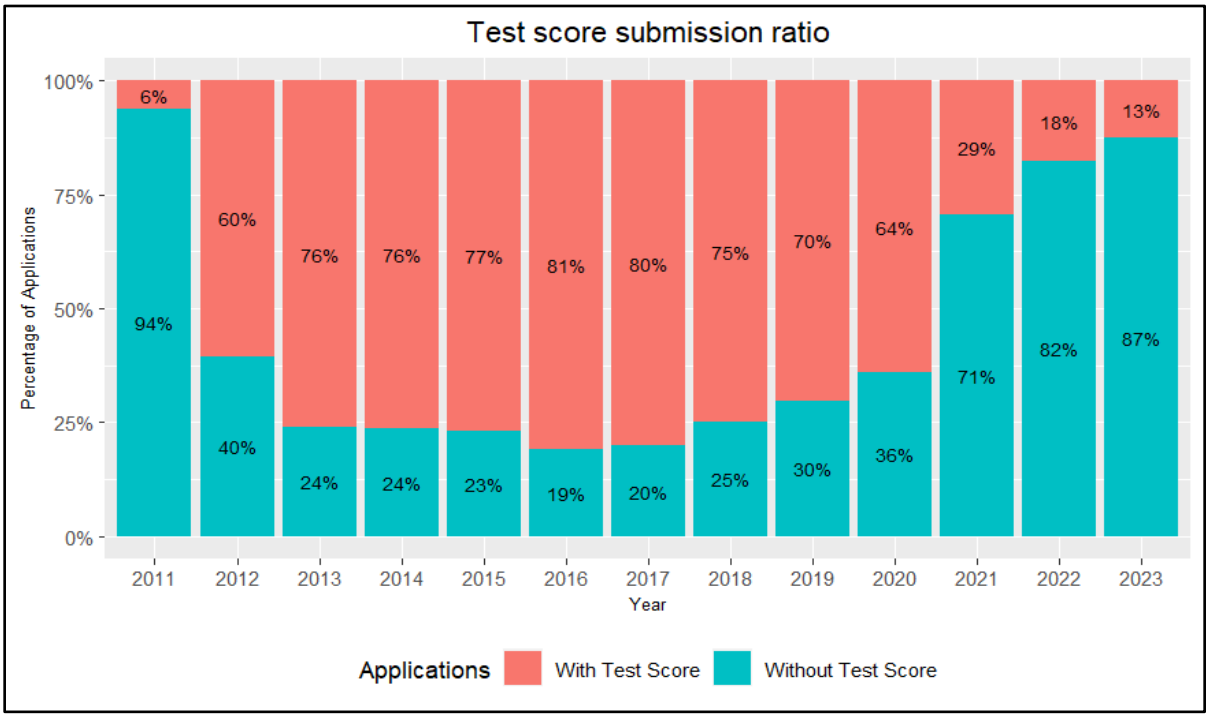

Fig 4: Ratio of applications received split by test scores.

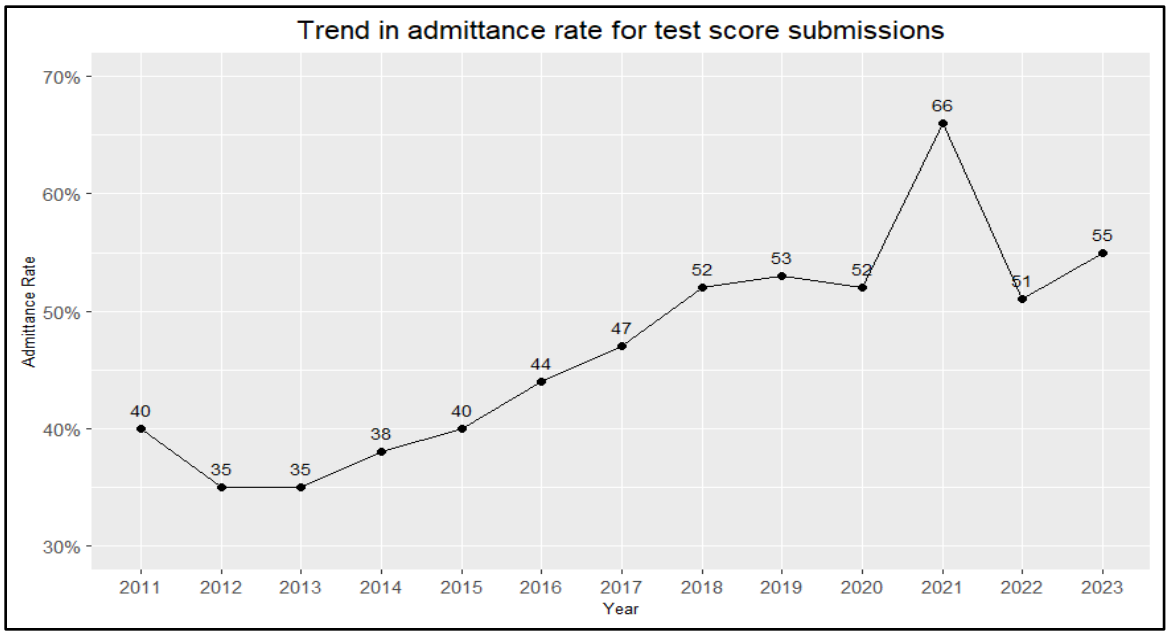

Fig 5: Admittance rate for submitting standardized test scores.

However, despite there being an option to not submit the standardized scores, applicants still go ahead and submit the same. This EDA is done to understand whether there exists an influence of these scores on the admission status of the applicant.

The following insights were generated from the data for this analysis:

- The proportion of candidates who provided their test scores remained consistently elevated until 2020, coinciding with the period when numerous programs required their submission, and post-2021, there was a notable decline in the proportion of applicants presenting standardized test scores (Fig 4).
- Nevertheless, with the shift to optional standardized test score submissions, the acceptance rate for applicants who chose to submit them increased, particularly in 2021 (Fig 5). Various factors may contribute to this trend, with one possibility being that students who voluntarily submitted test scores had a comparatively stronger profile.
- Whether the observed upward trend in the acceptance rate of applicants who submitted test scores is statistically significant is explored through a hypothesis test in the subsequent section of the article.

4. Hypothesis Testing

4.1. Hypothesis 1

In the first hypothesis, we tested if the proportion of residents getting admitted is more than non-residents for master's degrees in the U.S.

For this hypothesis test, the population, that was estimated, is the entire United States university admission for master's courses using the University of Colorado Boulder's admission data as the sample. A two-sample hypothesis test has been performed with the null hypothesis as the ratio of admitted resident applicants is less than or on par with the ratio of non-resident applicants and the opposite of it is defined as the alternate hypothesis [4].

$$H_0: P_1 \leq P_2$$
$$H_a: P_1 > P_2$$

Where is the ratio of admitted (admitted/applied) resident applicants in the population and is the ratio of admitted non-resident applicants in the population?P_2 is the ratio of admitted non-resident applicants in the population.

We have collected the following two samples for the analysis:

1. Residents' applications received in the University of Colorado Boulder since 2011 with the sample size (n_1) as 13,016 records.
2. Non-resident applications received in the University of Colorado Boulder since 2011 with the sample size (n_2) as 69,823 records.

The sample sizes are large enough to safely assume that the sample proportions (if multiple samples were taken and their proportions were plotted) would follow a normal distribution according to the Central Limit Theorem., are large enough to safely assume that the sample

proportions (if multiple samples were taken and their proportions are plotted) would follow a normal distribution according to the Central Limit Theorem.

The admission rates in the samples are as follows:

- $\widehat{P_1}$- 0.75
- $\widehat{P_2}$- 0.45

Using the above test statistics, a two-sample Student's t-test was conducted (since the sample size is high, Student's t-test can be used as a replacement for the z-test) in R to test the hypothesis.

4.2. Hypothesis 2

For the second hypothesis, we tested if the standardized test scores (GRE/GMAT) influence the admission conversion rate post-covid in the United States master's admission process.

For this hypothesis test, similar to the previous one, the population estimated was the entire United States university admission for master's courses using the University of Colorado Boulder's admission data as the sample. A two-sample hypothesis test has been performed with the null hypothesis as the admission ratio of the residents who submitted the standardized test scores pre-COVID is greater than or on par with the admission ratio pre-COVID at the University of Colorado Boulder. The opposite of it is defined as the alternate hypothesis:

$$H_0 : P_1 \geq P_2$$
$$H_a : P_1 < P_2$$

Where is the admission ratio (admitted/applied) of applicants who submitted the test scores pre-covid and is the admission ratio of applicants who submitted the test scores post-covid?P_2 is the admission ratio of applicants who submitted the test scores post-covid.

We have collected the following two samples for the analysis:

1. Applications received in the University of Colorado Boulder along with standardized test scores pre-COVID with the sample size (n_1) as 35,403 records.
2. Applications received in the University of Colorado Boulder along with standardized test scores post-COVID with the sample size (n_2) as 5,126 records.

The sample sizes are large enough to safely assume that the sample proportions (if multiple samples were taken and their proportions were plotted) would follow a normal distribution according to the Central Limit Theorem., are large enough to safely assume that the sample proportions (if multiple samples were taken and their proportions are plotted) would follow a normal distribution according to the Central Limit Theorem.

The admission rates in the samples are as follows:

- $\widehat{P_1}$- 0.44
- $\widehat{P_2}$- 0.59

Using the above test statistics, a two-sample Student's t-test was conducted (since the sample size is high, Student's t-test can be used as a replacement for the z-test) in R to test the hypothesis.

5. Results

5.1 Hypothesis Test 1

In the first hypothesis, we decided the level of significance (α) as 5% and plotted the confidence interval for the difference of sample proportions $\widehat{P_1}$ and $\widehat{P_2}$ (Fig 6). The outcome states that 95% the of time the difference in population proportions (P_1 and P_2) will fall in the interval between 0.293 and 1.000. However, according to our null hypothesis, the difference should be less than or P_2ero.) will fall in the interval between 0.293 and 1.000. However, according to our null hypothesis, the difference should be less than or equal to zero.

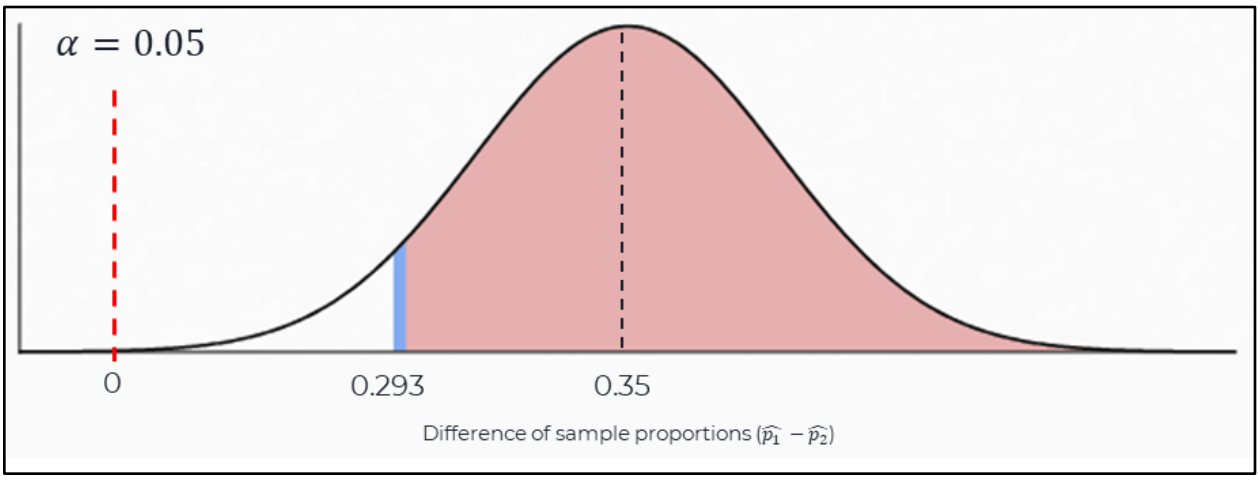

Fig 6: 95% confidence interval in the hypothesis test 1

Hence, we reject the null hypothesis and conclude that the admission ratio is not lesser or equal to the admission ratio of non-residents in the United States university's master's applications.

5.2 Hypothesis Test 2

Like the first hypothesis, in the second hypothesis,α is set as 5% and plotted the confidence interval for the difference in sample proportions ($\widehat{P_1}$ and $\widehat{P_2}$) (Fig 7). Which states that 95% of the time, the difference in population proportions ($P_1$and P_2) will fall in the interval between -1.000 and -0.146.

However, according to our null hypothesis, the difference should be greater than oP_2 equal to zero.) will fall in the interval between -1.000 and -0.146. However, according to our null hypothesis, the difference should be greater than or equal to zero.

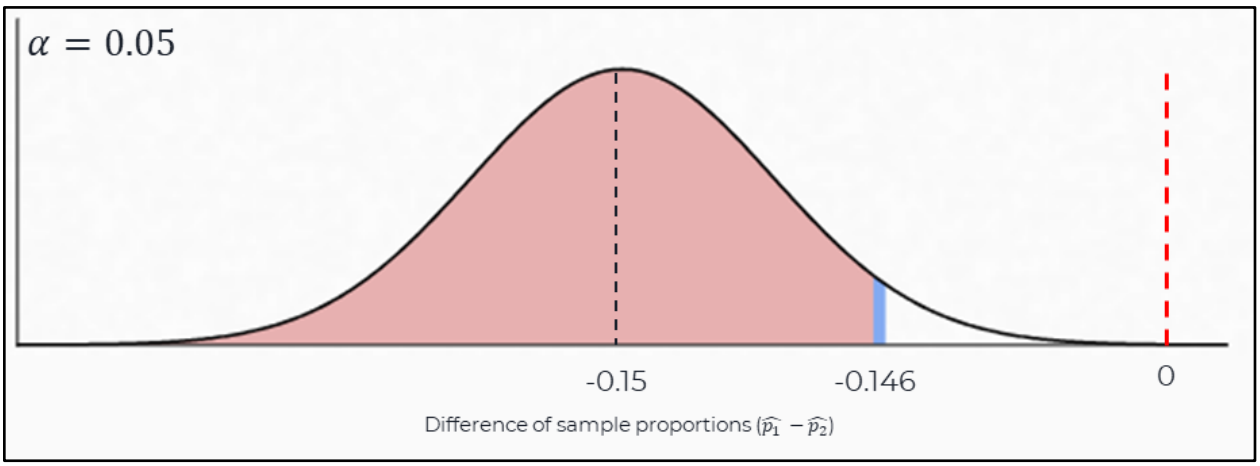

Fig 7: 95% confidence interval in the hypothesis test 2

Hence, we reject the null hypothesis and conclude that the admission ratio for students who submit the standardized test scores pre-COVID is not higher or equal to the admission ratio of students who submit the standardized test scores pre-COVID in the United States university's master's applications.

However, this hypothesis test does not convey that the standardized test scores have the highest influence, nor it is the only decider in the admission decision. Rather it states that the assumption that the test scores do not have any influence on admission is false.

6. Conclusion

By examining the master's degree admission data, we got interesting trends that shape higher education. We have confirmed that there's a consistent increase in applicant numbers over the period with the data indicating that the interest moves towards pursuing advanced degrees. Particularly, Data Science, Computer Science, Electrical Engineering, and Aerospace stand out as the most preferred courses in the application pool.
Even though most of the Universities in the U.S. have made standardized test scores submission optional after the COVID global pandemic, they still influence the admissions process. This shows that the significance of the standardized test score still has some impact on admission.

Additionally, we have confirmed there is an interesting trend in admission ratios where residents of Colorado have higher admissions rates than non-resident applicants using two sample hypotheses. These insights can be utilized by the universities in facilitating more in-demand

courses which attract more applicants to CU. Also, students can make use of them in choosing courses that are of interest that have higher demand in the market.

6. References

[1] Debopam Bhattacharya, Shin Kanaya, Margaret Stevens; Are University Admissions Academically Fair? The Review of Economics and Statistics 2017; 99 (3): 449–464.

[2] Klayman, J., & Ha, Y.-w. (1987). Confirmation, disconfirmation, and information in hypothesis testing. Psychological Review, 94(2), 211–228.

[3] A. I. Gufroni, P. Purwanto, F. Farikhin, A. Wibowo and B. Warsito, "Exploratory Data Analysis To Identify The Most Important Feature Of University Admission Test Criteria Using Random Forest And Neural Network Algorithm," 2021 5th International Conference on Informatics and Computational Sciences (ICICoS), Semarang, Indonesia, 2021, pp. 1-5, doi: 10.1109/ICICoS53627.2021.9651757.

[4] Van Heesch, M. M., Bosma, H., Traag, T., & Otten, F. (2012). Hospital admissions and school dropout: A retrospective cohort study of the ‘selection hypothesis’. European Journal of Public Health, 22(4), 550-555.