EXPLORATORY DATA ANALYSIS ON MBA ADMISSION DATASET

INTRODUCTION

This MBA application dataset provides a comprehensive view of 6,194 applicants, capturing various demographic, academic, and professional attributes relevant to MBA admissions. Each record includes a unique application ID, gender, and international status, along with GPA, undergraduate major, and race, although some values are missing for the race attribute. Academic metrics include the applicant's GPA and GMAT scores, which offer insight into their academic readiness for the program. Additionally, demographic data such as gender and international status may support analysis of diversity in the applicant pool.

Professional experience in this dataset is represented by the work_exp and work_industry columns, detailing each applicant's years of experience and their industry background. The admission column, which originally contained data for only 1,000 applicants and had over 5,000 missing values, has been filled with the value deny to maintain data quality and focus on other key attributes. With this refined dataset, analysis can now focus on understanding trends in academic and professional profiles, as well as assessing diversity within the applicant pool based on factors like gender, race, and international status. This approach supports various analytical objectives, such as profiling typical applicant backgrounds and exploring demographic trends within the MBA application and admission process.

AIM

The focus of this Exploratory Data Analysis (EDA) on the MBA application dataset is to profile the applicant pool by exploring demographic, academic, and professional attributes, including GPA, GMAT scores, work experience, and undergraduate major, to uncover common applicant characteristics. Additionally, diversity metrics such as gender, race, and international status will help illustrate the variety within the applicant pool. The admission column, which indicates whether applicants were admitted, denied, or placed on a waiting list, enables analysis of factors linked to admission outcomes, providing insights into the academic and professional profiles of candidates pursuing an MBA and the diversity of backgrounds represented.

COLUMNS

- application_id: Unique identifier for each application
- gender: Applicant's gender (Male, Female)
- international: International student (TRUE/FALSE)
- gpa: Grade Point Average of the applicant (on 4.0 scale)
- major: Undergraduate major (Business, STEM, Humanities)
- race: Racial background of the applicant (e.g., White, Black, Asian, Hispanic, Other)
- gmat: GMAT score of the applicant (800 points)
- work_exp: Number of years of work experience (Year)
- work_industry: Industry of the applicant's previous work experience (e.g., Consulting, Finance, Technology, etc.)
- admission: Admission status (Admit, Waitlist)
- Mark_cat: gmat score is categorised into 3 (high, average, below average)
- exp_cat: Year of experience is categorized into 3 (entry level, intermediate, senior)
- normalized_gpa
- normalized_gmat
- combined_normalized_score : sum of normalized_gpa and normalized_gmat

OBJECTIVES

The objectives of conducting EDA on this MBA application dataset are:

1. Profile Applicant Characteristics: Analyze distributions of academic (GPA, GMAT) and professional (work_exp, work_industry) features to understand the typical qualifications and industry backgrounds of applicants.

2. Examine Diversity and Inclusivity: Assess demographic variables such as gender, race, and international status to gain insights into the diversity of the applicant pool, which can inform discussions on inclusivity within the program.

3. Identify Patterns and Trends in Academic and Professional Backgrounds: Explore relationships between key variables (e.g., between GPA and GMAT or work experience and industry) to uncover common pathways and profiles among applicants.

4. Handle and Address Missing Values: Review and manage missing values, particularly in columns with a large proportion of NaNs, to ensure a clean dataset for further analysis and avoid biases.

5. Generate Insights for Future Admissions Strategy: Provide data-driven insights into applicant profiles and trends, helping to inform admissions criteria and strategy for targeting diverse, high-potential applicants.

These objectives support a thorough understanding of the applicant pool, with insights into both individual profiles and overall trends.

DATA OVERVIEW

application_id	int64
gender	object
international	bool
gpa	float64
major	object
race	object
gmat	float64
work_exp	float64
work_industry	object
admission	object

Mark_cat	object
exp_cat	object
normalized_gpa	float64
normalized_gmat	float64
combined_normalized_score	float64

ANALYSIS

This MBA application dataset contains detailed information on 6,194 applicants, capturing various demographic, academic, and professional attributes relevant to admissions decisions. The dataset initially included 10 columns: six of object data type, one integer, six floats, and one boolean. Using df.info(), the structure of the Data Frame was quickly reviewed to understand data types and missing values. Additionally, df.describe() provided a statistical summary of the numerical columns, aiding in understanding the data distribution. Checks for null and duplicate values revealed that the race column contained 1,842 missing values, while the admission column had 5,194 missing values, though no duplicate records were identified in the dataset. The 'race' column's null values have been replaced with 'Unknown,' and the 'admission' column, containing over 5,000 entries, has been replaced with 'deny'. The 'Mark_cat' column has been created by categorizing GMAT scores into 'High' for scores of 700 and above, 'Average' for scores of 570 and below, and 'Below Average' for scores in between.

Code Implementation in Pandas

1. FEATURE ENGINEERING

```
df['Mark_cat'] = df['gmat'].apply(lambda x:'High' if x>=700 else 'Average' if x<=570 else 'Below Average')

df['exp_cat'] = df['work_exp'].apply(lambda x:'Senior' if x>=7 else 'Intermediate' if x<=4 else 'Entry Level')

df['normalized_gpa'] = (df['gpa'] - df['gpa'].min()) / (df['gpa'].max() - df['gpa'].min())
```

```
df['normalized_gmat'] = (df['gmat'] - df['gmat'].min()) / (df['gmat'].max() - df['gmat'].min())

df['combined_normalized_score'] = df['normalized_gpa'] + df['normalized_gmat']
```

2.Value counts of gender

```
df['gender'].value_counts()
```

3. Value counts of the columns in object data type

```
for col in df:
 if df[col].dtype=='object':
   print(df[col].value_counts())
   print("")
   print()
```

4.Value counts of the international students

```
df['international'].value_counts()
```

5. Min and max gpa

```
df['gpa'].max()

df['gpa'].min()
```

6. Min and max work experience

```
df['work_exp'].max()

df['work_exp'].min()
```

7. Average gmat score for females and males

```
df[df['gender'] == 'Female']['gmat'].mean()

df[df['gender'] == 'Male']['gmat'].mean()
```

8. Highest Work Experience Recorded for Each Gender

```
df[['gender','work_exp']].groupby('gender').max()
```

9. Average GMAT Score for Each Major

```
average_gmat_by_major=df[['major','gmat']].groupby('major').mean()
```

10. GMAT Scores Vary with Different Levels of Work Experience.

```
correlation_work_exp_gmat =
df[['work_exp','gmat']].groupby('work_exp').mean()
```

11. Average GMAT Score by Work Industry

```
avg_gmat_by_industry =
df[['work_industry','gmat']].groupby('work_industry').mean()
```

12. Distribution of Different Work Industries Across Gender

```
industry_by_gender =
df[['work_industry','gender']].groupby('gender').value_counts()
```

13. Gender Distribution by GMAT Category

```
gender_dist_by_Mark_cat =
df[['Mark_cat','gender']].groupby('gender').value_counts()
```

14. Occurrences of Each Unique Combination of Admission and Exp_Cat Values

```
df[['admission','exp_cat']].groupby('admission').value_counts()
```

15. ADMISSION STATUS ACROSS VARIOUS DEMOGRAPHIC AND CATEGORICAL FEATURES

```
lst=['gender','major','race','work_industry','Mark_cat','exp_cat']
 for i in lst:
        print(df[[i,'admission']].groupby(i).value_counts())
        print('\n')
        print('------------------------')
```

16. TOP 10 HIGHEST VALUES IN THE COMBINED_NORMALIZED_SCORE

```
top_10_scores =
df['combined_normalized_score'].sort_values(ascending=False).head(10)
```

17. BOXPLOT

```
for col in df:
        if df[col].dtype=='int' or df[col].dtype=='float':
                print(col.capitalize())
                print(f"Mean:{df[col].mean()}")
                print(f"Median:{df[col].median()}")
                print(f"Minimum of {col}:{df[col].min()}")
                print(f"Maximum of {col}:{df[col].max()}")
                print(f"Variance of {col}:{df[col].var()}")
                print(f"Standard deviation:{df[col].std()}")
                print()
                plt.figure(figsize=(8,5))
                plt.boxplot(df[col])
```

```
plt.title(col)
plt.show()
print("_________")
print()
```

18. COUNTPLOT

```
for col in df:
    if df[col].dtype=='object':
        sns.countplot(data=df,x=df[col])
        plt.grid()
        plt.xticks(rotation=90)
        plt.show()
```

19. HISPLOT

- ```
  sns.histplot(data=df,x='gpa',bins=5,hue='major')
  plt.xticks(rotation=70)
  plt.show()
  ```

- ```
  sns.histplot(data=df,x='gmat',bins=5,hue='major')
  plt.xticks(rotation=70)
  plt.show()
  ```

20. PIEPLOT

- ```
  plt.pie(df['major'].value_counts(),labels=df['major'].unique(),autopct="%1.0f%%")
  plt.title("major value counts")
  plt.plot()
  ```
  ```

- plt.pie(df['work_industry'].value_counts(),labels=df['work_industry'].unique(),autopct="%1.0f%%")
  plt.title("work_industry value counts")
  plt.plot()

21. HEATMAP

df.corr(numeric_only=True)

sns.heatmap(df.corr(numeric_only=True),annot=True)

22. BAR PLOT

- plt.figure(figsize=(10, 6))

  sns.barplot(data=df, x='combined_normalized_score', y='admission')

  plt.title('Combined Normalized Score by Admission')

  plt.xlabel('Combined Normalized Score')

  plt.ylabel('Admission')

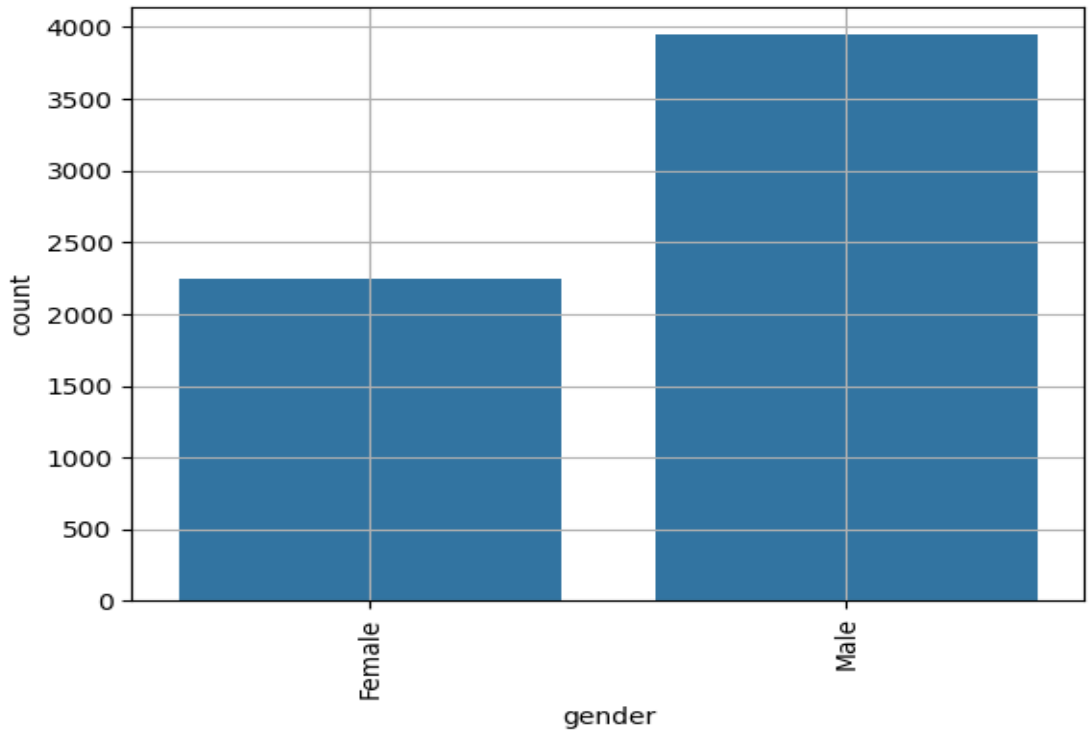
  plt.show()

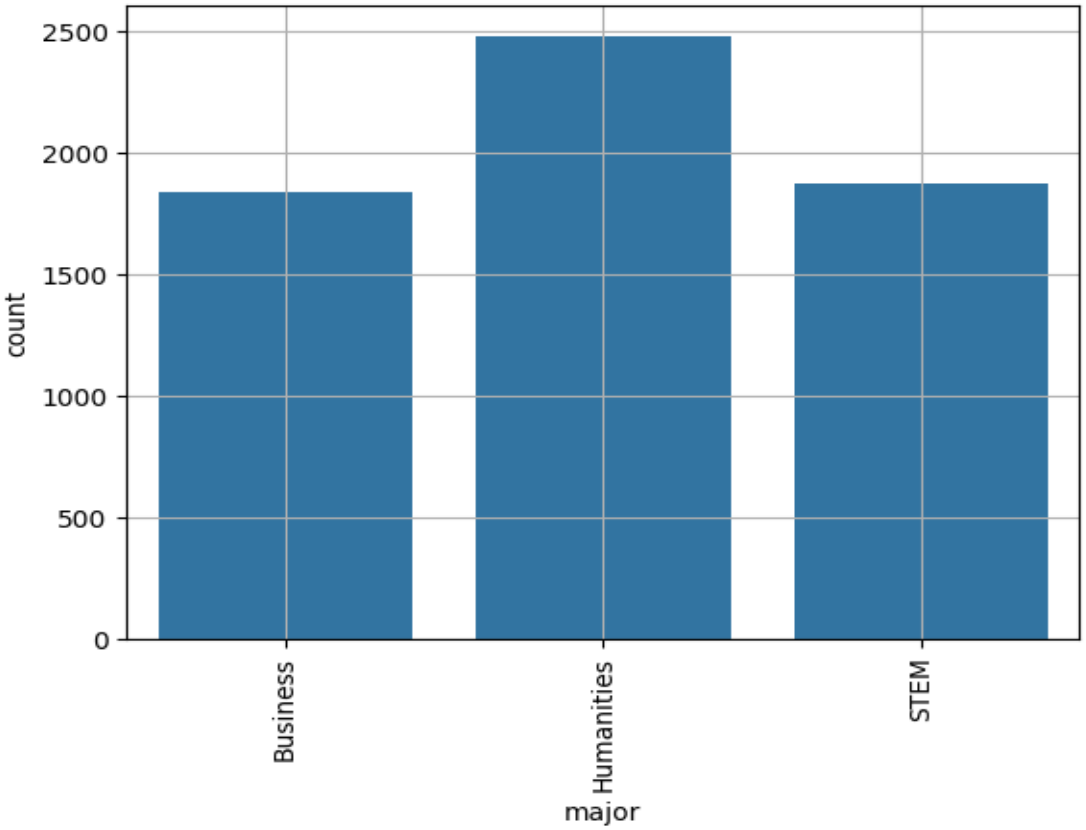
- plt.figure(figsize=(10, 6))
  sns.barplot(data=df, x='gpa', y='admission')
  plt.title('GPA by Admission')
  plt.xlabel('GPA')
  plt.ylabel('Admission')
  plt.grid()
  plt.show()

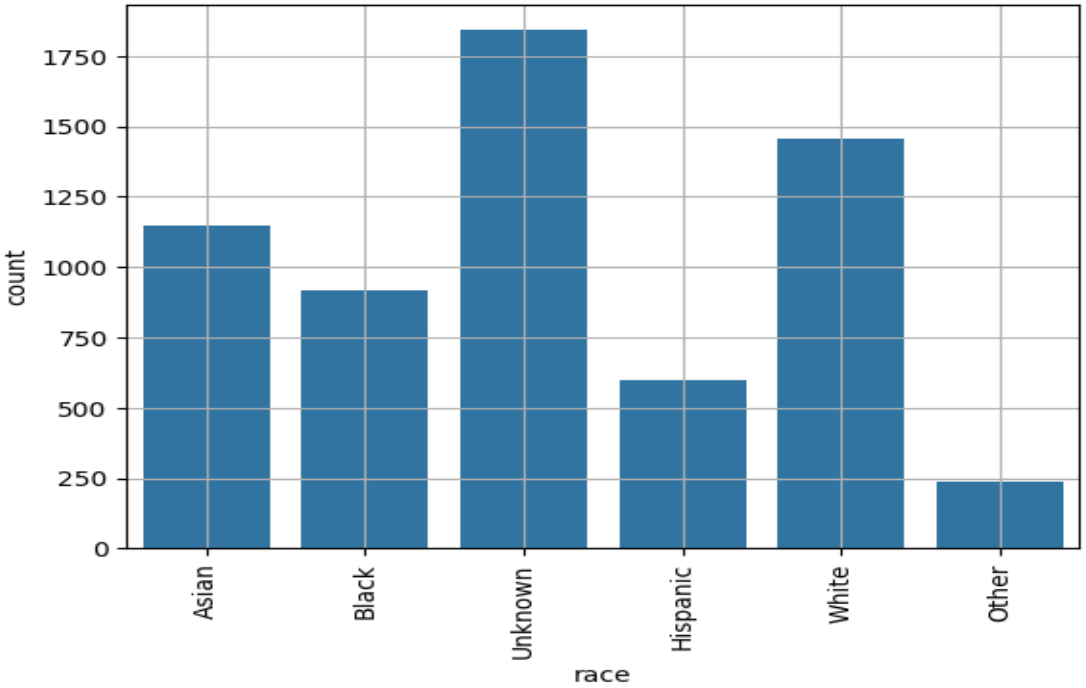
- plt.figure(figsize=(10, 6))
  sns.barplot(data=df, x='gmat', y='admission')
  plt.title('GMAT by Admission')
  plt.xlabel('GMAT')
  plt.ylabel('Admission')
  plt.grid()
  plt.show()

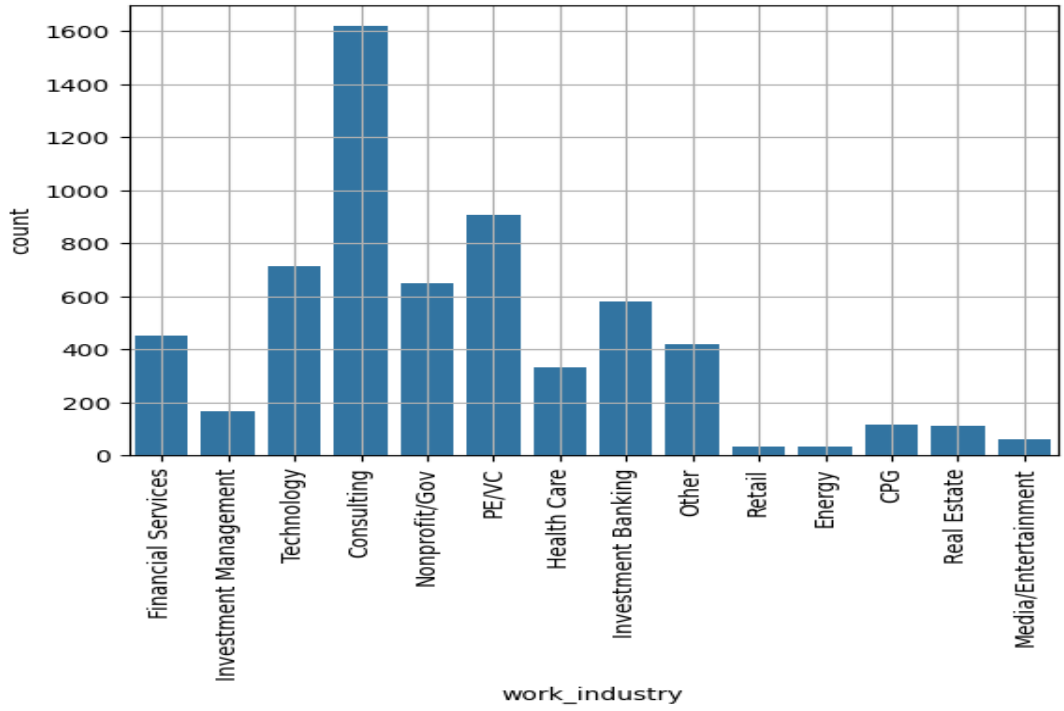
- plt.figure(figsize=(10, 6))
  sns.barplot(data=df, x='combined_normalized_score', y='major')
  plt.title('Combined_normalized_score by Major')
  plt.xlabel('Combined_normalized_score')
  plt.ylabel('Major')
  plt.grid()
  plt.show()

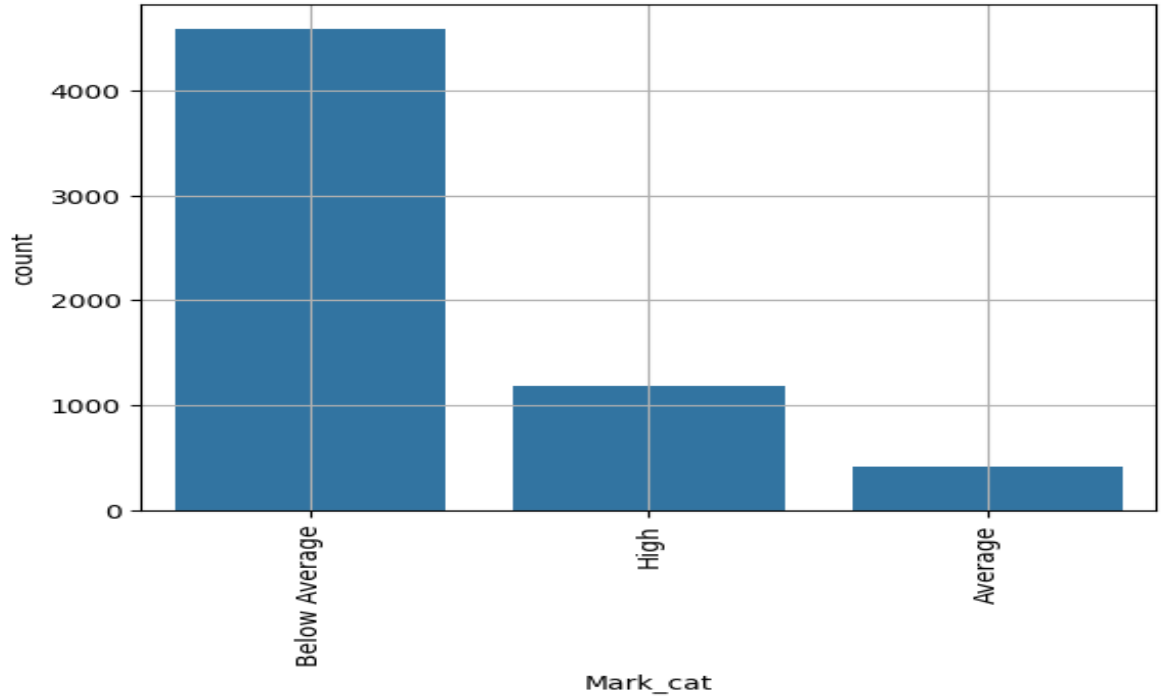
COUNTPLOT

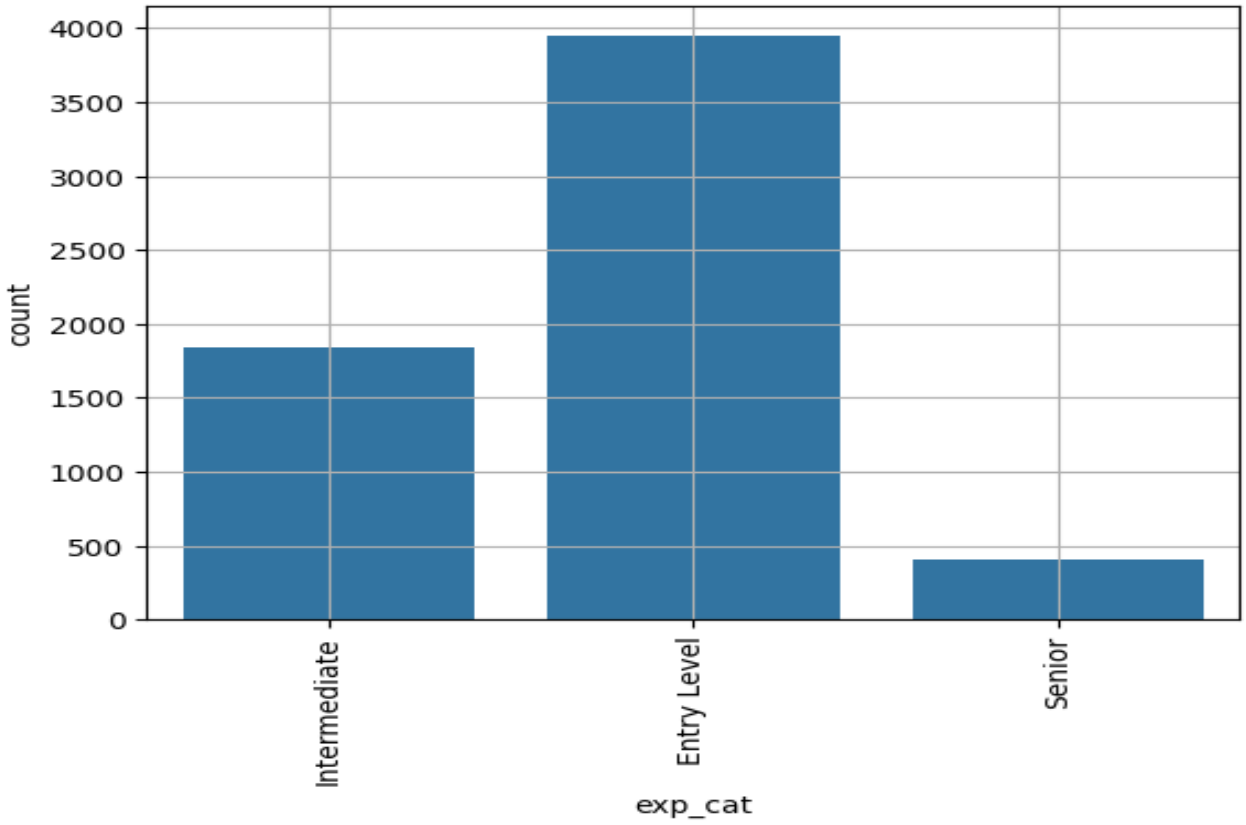



2500
2000
1500
1000
500
0
count
Business
Humanities
STEM
major

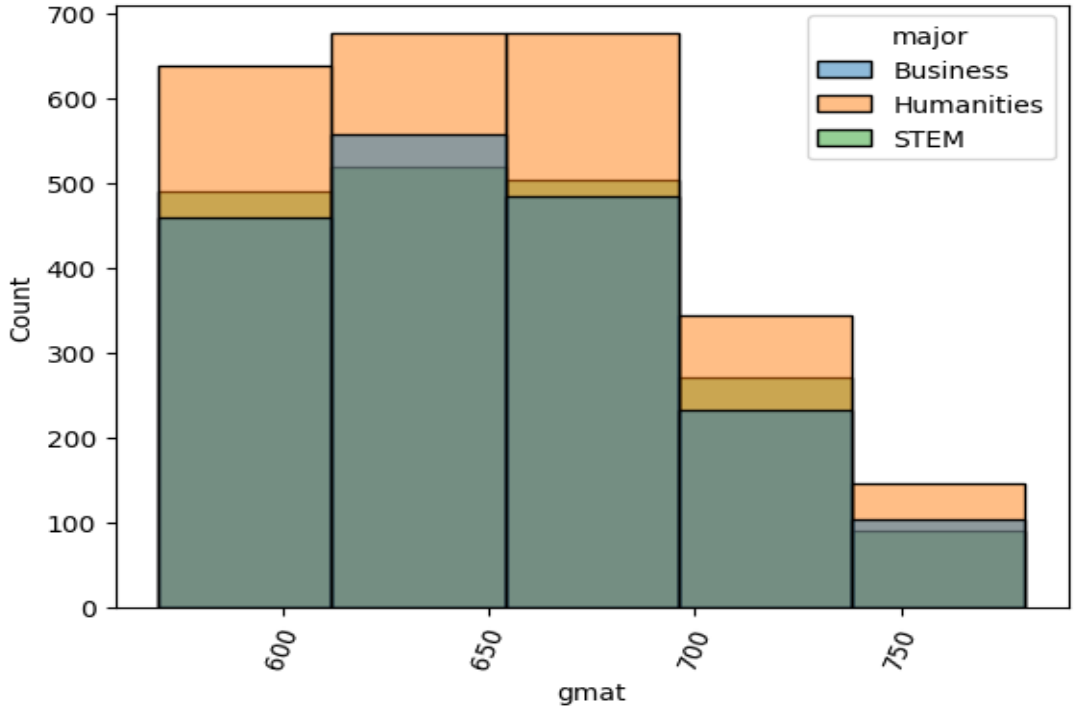

1750
1500
1250
1000
750
500
250
0
count
Asian
Black
Unknown
Hispanic
White
Other
race

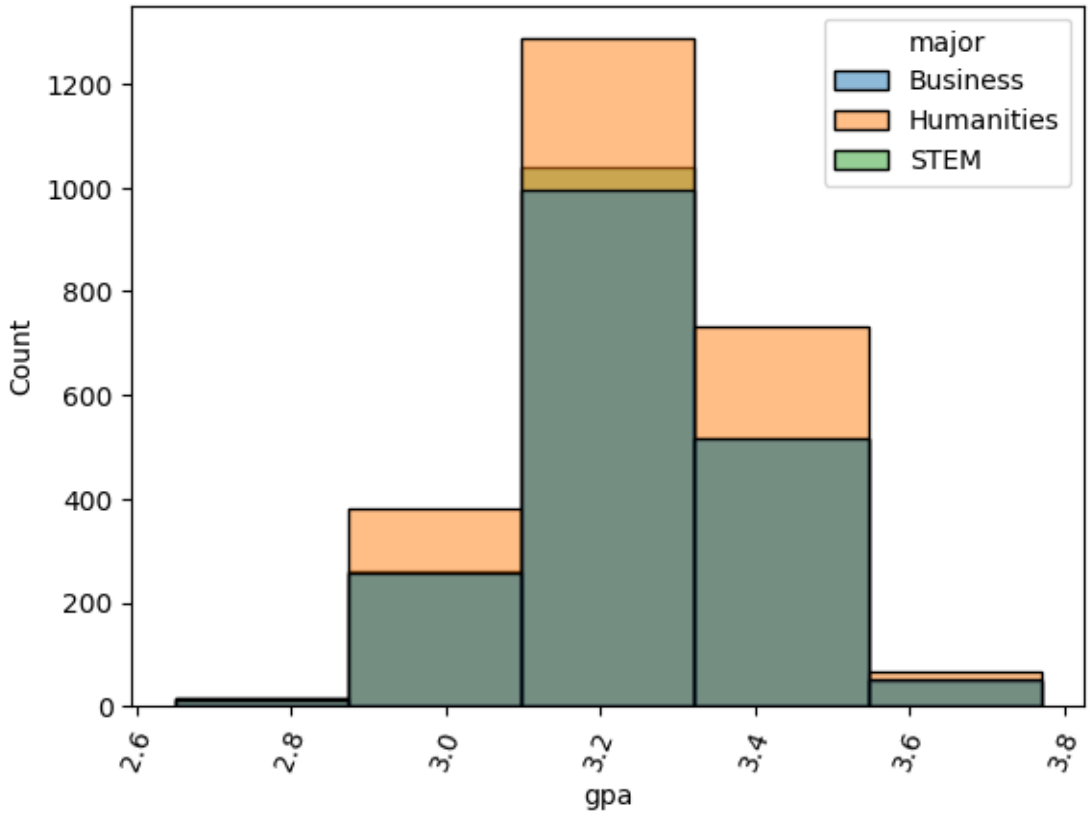

count
1600
1400
1200
1000
800
600
400
200
0
Financial Services
Investment Management
Technology
Consulting
Nonprofit/Gov
PE/VC
Health Care
Investment Banking
Other
Retail
Energy
CPG
Real Estate
Media/Entertainment
work_industry


count
4000
3000
2000
1000
0
Below Average
High
Average
Mark_cat

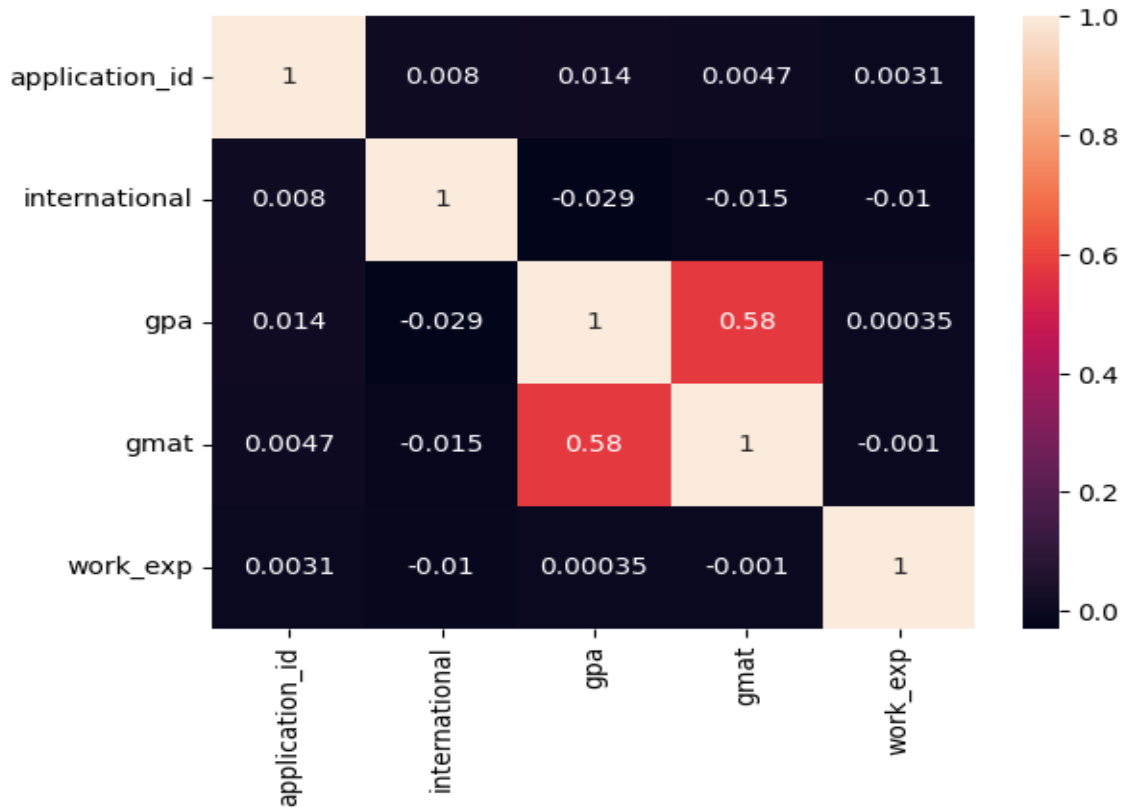

4000
3500
3000
2500
2000
1500
1000
500
0
count
Intermediate
Entry Level
Senior
exp_cat

HISTPLOT

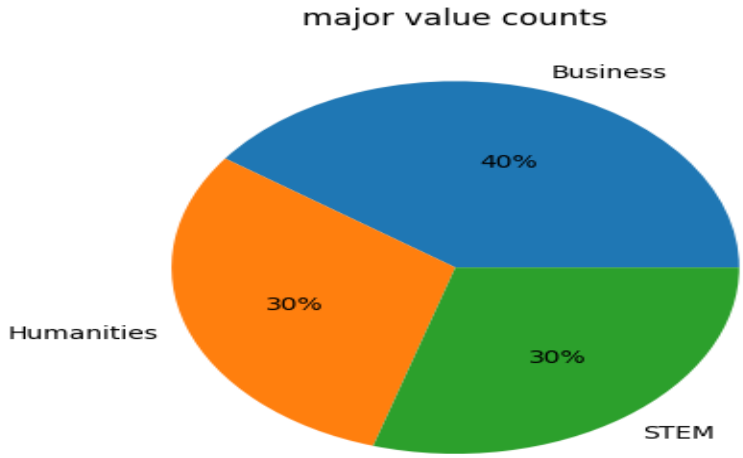

major
Business
Humanities
STEM
Count
700
600
500
400
300
200
100
0
600
650
700
750
gmat

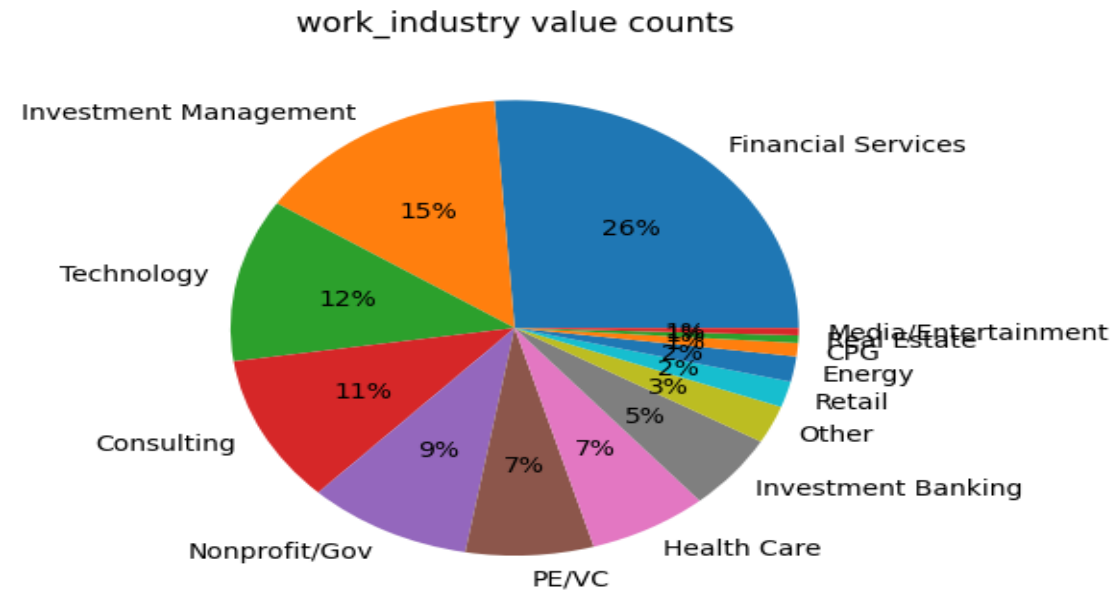

major
Business
Humanities
STEM
Count
1200
1000
800
600
400
200
0
2.6
2.8
3.0
3.2
3.4
3.6
3.8
gpa

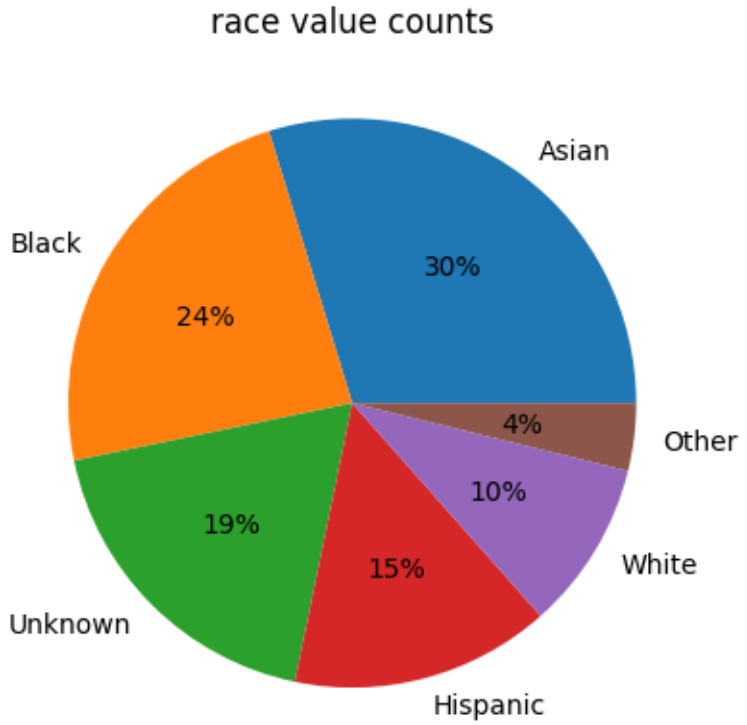
HEATMAP



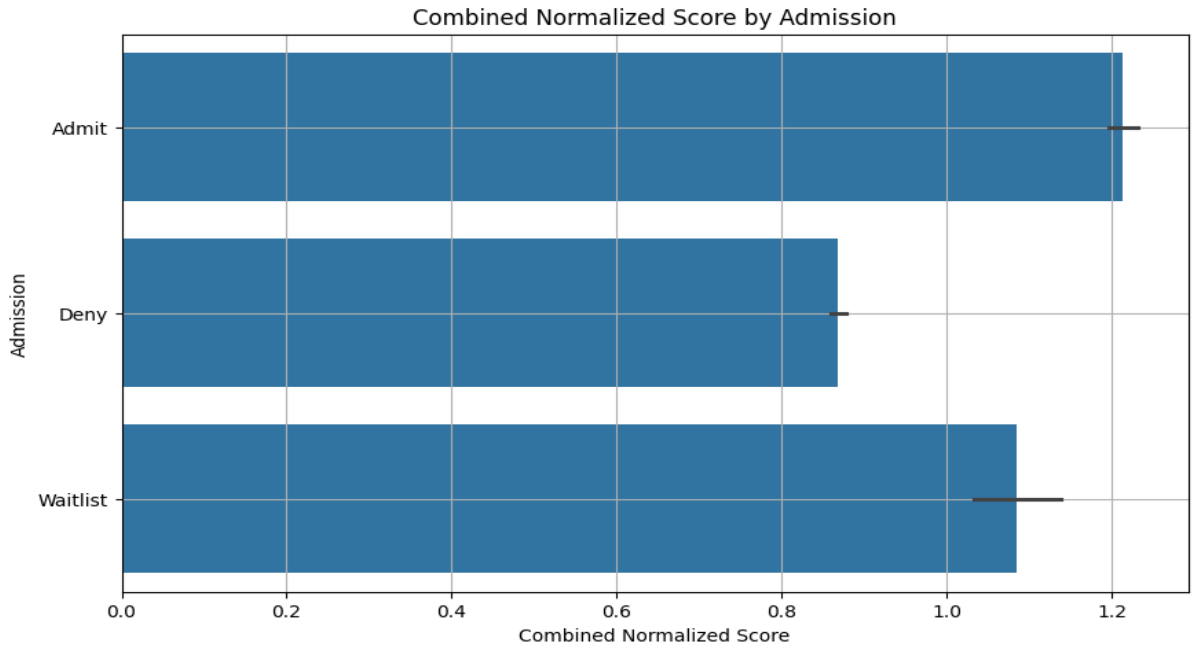
PIEPLOT

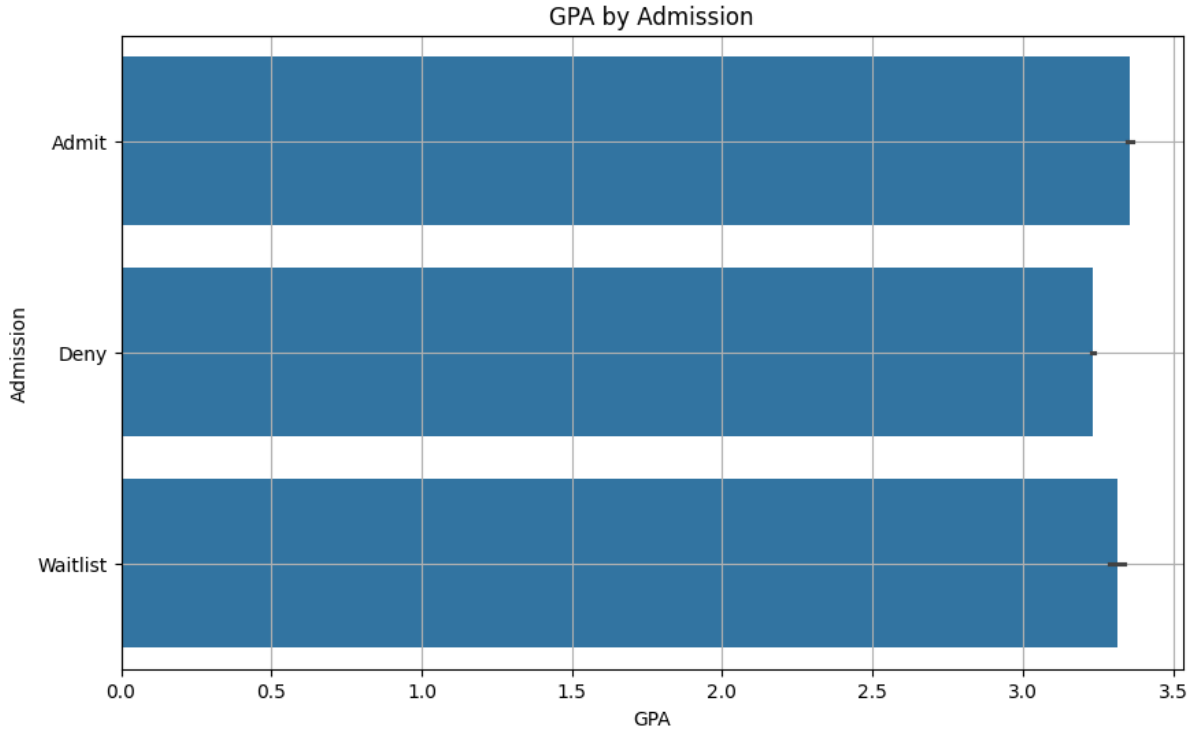


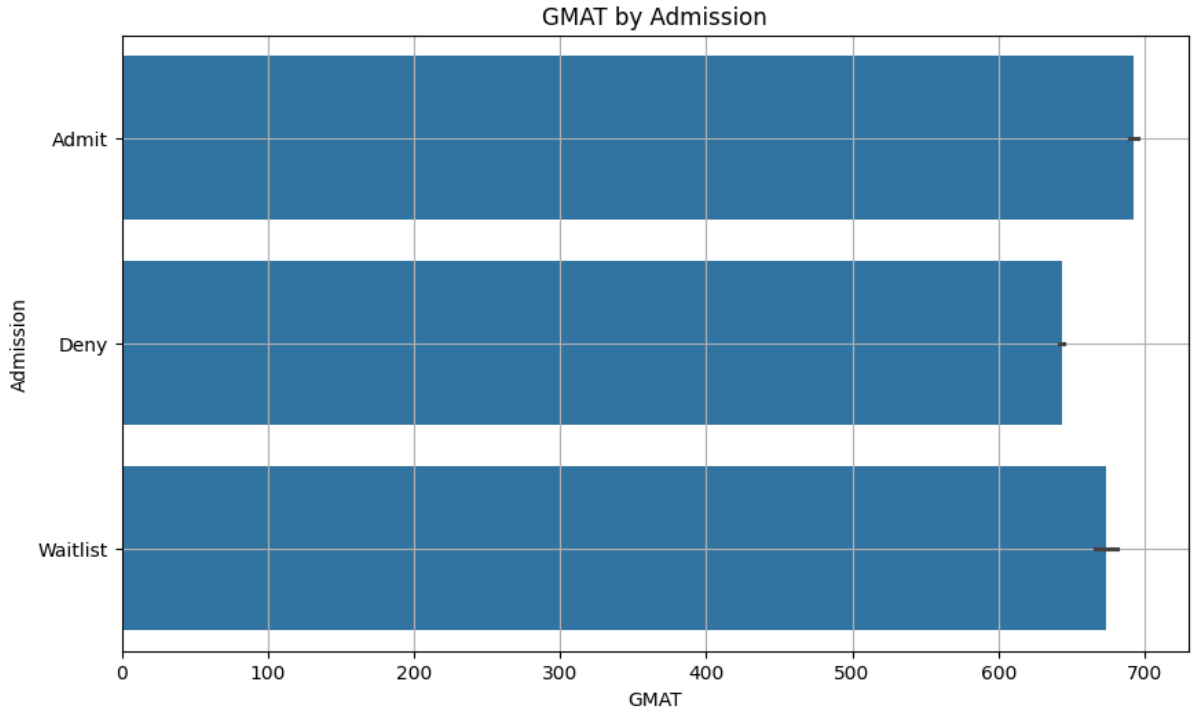

work_industry value counts
Investment Management
Financial Services
15%
26%
Technology
12%
11%
Consulting
9%
7%
7%
Nonprofit/Gov
PE/VC
Health Care
5%
Investment Banking
3%
Other
2%
Retail
2%
Energy
CPG
Real Estate
Media/Entertainment

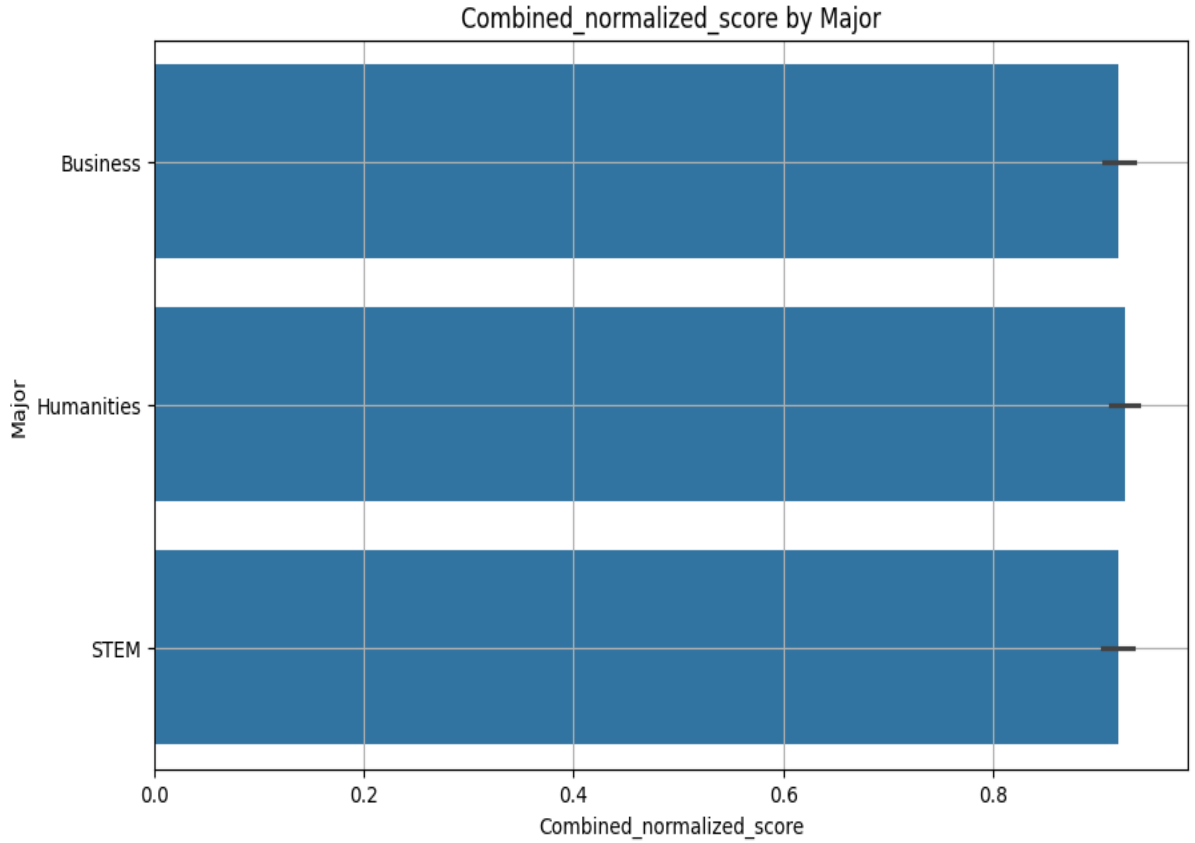

race value counts
Asian
30%
Black
24%
4%
Other
10%
19%
White
15%
Unknown
Hispanic

BARPLOT


Combined Normalized Score by Admission
Admit
Deny
Waitlist
Admission
0.0
0.2
0.4
0.6
0.8
1.0
1.2
Combined Normalized Score


GPA by Admission
Admit
Deny
Waitlist
Admission
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
GPA


GMAT by Admission
Admit
Deny
Waitlist
Admission
0
100
200
300
400
500
600
700
GMAT


Combined_normalized_score by Major
Business
Humanities
STEM
Major
0.0
0.2
0.4
0.6
0.8
Combined_normalized_score

CONCLUSION

1. The dataset reveals a gender disparity, with **3,943 males** (approximately **63.7%**) and **2,251 females** (approximately **36.3%**)

2. The **GPA** range (2.65 to 3.77) indicates that most applicants have maintained a relatively good academic standard.

3. The **distribution of majors** shows a strong inclination toward **Humanities** (2,481 students), followed by **STEM** (1,875) and **Business** (1,838).

4. GMAT scores reveal a competitive applicant pool, with a **highest score of 780** , **minimum score of 570** and a **median score of 650**.

5. The presence of **1,842 international students** signifies a diverse applicant pool, which can enrich classroom discussions and group projects with varied cultural perspectives.

6. In the analysis of industry categories, there were **14 types of industries** represented, with the consulting industry having the highest number of applicants, while the energy industry had the fewest. A significant majority of both males and females were found in the consulting sector.

7. Additionally, when examining the race column, it was noted that there were **1,842 rows with NaN values**, which were subsequently replaced with "unknown." The revised counts revealed that the "unknown" category had the highest frequency, followed by the **"White" category with 1,456 applicants, "Asian" with 1,147, "Black" with 916, "Hispanic" with 596, and "Other" with 237 applicants.**

8. The applicants had a **minimum work experience of 1 year** and a **maximum of 9 years.**

9. A new column named **Mark_cat** was created to categorize applicants based on their performance, resulting in the following distribution**: 4,582 applicants were classified as "Below Average," 1,190 as "High," and 422 as "Average."**

10. The average GMAT scores are quite close between genders, with males scoring slightly higher than females.

11. The maximum work experience for males is greater than for females, suggesting potential differences in experience levels.

12. GMAT scores vary by major, with Humanities slightly outperforming Business and STEM.

13. The analysis of work experience indicates a potential correlation, where 1-year experience candidates have the highest average GMAT score, while scores decrease with more experience, except for 7 years, which sees an increase.

14. When analyzing by work industry, the Energy sector had the highest average GMAT score, while Real Estate had the lowest.

15. When analysis the gender with Mark_cat, it shows that both genders are facing academic challenges, with 63.2% of females and 73.5% of males in the "Below Average" category. In the "High" category, males make up 64.2% while females account for 35%, and in the "Average" category, males represent 64% compared to females' 36%, highlighting a notable gap in performance between genders.

16. Males have a much higher denial rate compared to females, with a total of 3,443 males denied admission compared to 1,751 females. Both genders have the same number of admitted applicants (450) and waitlisted individuals (50).

17. Humanities majors faced the highest denial rate (2,073), while Business majors had the highest number of admissions (270). Each major shows a trend where denials far exceed admissions.

18. The "Unknown" race category has the highest denial rate (1,534). Among the known racial categories, Asians have the highest number of denials (935) and the highest admissions (190).

19. Consulting has the highest number of denials (1,362) compared to admissions (224). Most industries show a trend of significantly more denials than admissions.

20. The "Below Average" academic performance category has the highest number of denials at 4,022, indicating that lower academic performance significantly affects admission outcomes. Additionally, entry-level candidates face the highest denial rate at 3,280 compared to 604 admissions, highlighting a general trend of denials across all experience levels, particularly for those with entry-level qualifications.

21. The most significant finding is the moderate positive correlation between GPA and GMAT scores, which suggests that high academic achievers tend to perform well on standardized tests. This insight can help in evaluating applicants for admission,

emphasizing the importance of academic performance as a predictor of future success.

22. The weak correlations involving international status and work experience imply that these factors may not play a substantial role in academic outcomes within this dataset, suggesting that admissions committees may focus more on GPA and GMAT scores when assessing candidates.

In Nutshell,

The dataset reveals important insights about admission procedures. There is a significant gender imbalance, with males making up 63.7% of applicants compared to 36.3% females. Most applicants have good academic performance, with GPAs between 2.65 and 3.77 and GMAT scores from 570 to 780, highlighting the importance of strong academics for admissions. The Humanities major is the most common, while Consulting has the highest number of applicants among industries. The "Below Average" performance category has the highest denial rate (4,022), and entry-level candidates face a notable denial rate of 3,280. The "unknown" race category also has the most denials (1,534), while Asians have the highest admissions (190) despite high denial rates (935). There is a moderate positive correlation between GPA and GMAT scores, indicating that academic performance is crucial for admission decisions. In summary, strong academic records are essential for prospective students, with a focus on GPAs and standardized test scores influencing admission outcomes.