

**Minor Project Report**

**on**

**[University Ranking and Accreditation Analysis]**

Submitted to

**LOVELY PROFESSIONAL UNIVERSITY**

in partial fulfilment of the requirements for the award of degree of

**Master of Computer Applications**

**Submitted By Supervised By**

**Mohd Shah Wasif Rafiq Sir**

**12223538**

**Anmol Gupta**

**12223484**

**LOVELY FACULTY OF TECHNOLOGY & SCIENCES**

**LOVELY PROFESSIONAL UNIVERSITY**

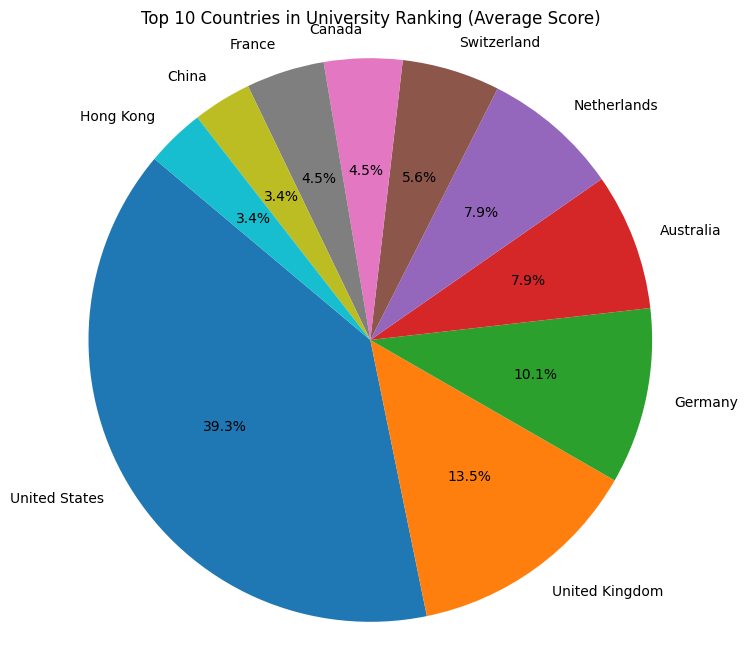
**PUNJAB**

**[November 2023]**

|  |  |  |
| --- | --- | --- |
| **Table of Content** | | |
| ***Sr. No.*** | ***Content*** | ***Page Index*** |
| 1 | Introduction about project | 1-3 |
| 2 | Project Modules and its description | 3-6 |
| 3 | Coding | 6-27 |
| 4 | Screenshot output | 27-28 |

**Introduction**

In this project we analyse the World University Rankings 2023. The Times Higher Education World University Rankings 2023 includes 1,799 universities across 104 countries and regions, making them the largest and most diverse university rankings to date.



This dataset includes the following 13 columns:

1. **University Rank**: The global ranking of the university.
2. **Name of University**: The name of the university.
3. **Location**: The country where the university is located.
4. **No of student**: The total number of students at the university.
5. **No of student per staff**: The ratio of students to staff at the university.
6. **International Student**: The percentage of students who are international.
7. **Female: Male Ratio**: The ratio of female to male students at the university.
8. **Overall Score**: The overall score of the university, presumably out of 100.
9. **Teaching Score**: The university's score for teaching, presumably out of 100.
10. **Research Score**: The university's score for research, presumably out of 100.
11. **Citations Score**: The university's score for citations, presumably out of 100.
12. **Industry Income Score**: The university's score for industry income, presumably out of 100.
13. **International Outlook Score**: The university's score for international outlook, presumably out of 100.

.

The goal of this project is to take a deep dive into the dataset and answer the following four questions:

Which country has the highest number of universities in the top 100 of the World University Rankings 2023

What is the university ranking distribution of the top 10 scoring Countries?

What is the correlation between each score and overall ranking? Which factor is the most important?

**Project Module**

1. **University Rank**: There are 162 unique university ranks in the dataset. The most frequent rank is 'Reporter', which appears 512 times.
2. **Name of University**: There are 2233 unique universities in the dataset. The most frequent university is the University of Oxford, which appears once.
3. **Location**: The universities are located in 116 unique locations. The most frequent location is the United States, which is the location of 173 universities in the dataset.
4. **No of student**: The number of students ranges from 17,080 to an unknown maximum value. The most frequent number of students is 17,080, which appears twice.
5. **No of student per staff**: The ratio of students to staff ranges from 0.4 to 232.2. The average ratio is approximately 19. The standard deviation is approximately 12.13, indicating a wide spread of the data.
6. **International Student**: There are 79 unique values for the percentage of international students. The most frequent value is 0%, which appears 336 times.
7. **Female:Male Ratio**: There are 87 unique values for the female-to-male ratio. The most frequent ratio is 58:42, which appears 95 times.
8. **OverAll Score**: There are 160 unique overall scores. The most frequent score is 18.4–24.3, which appears 305 times.
9. **Teaching Score**: The teaching score ranges from 11.6 to 94.8, with an average of approximately 27. The standard deviation is approximately 13.28, indicating a wide spread of the data.
10. **Research Score**: The research score ranges from 7.4 to 99.7, with an average of approximately 23. The standard deviation is approximately 16.76, indicating a wide spread of the data.
11. **Citations Score**: The citations score ranges from 0.8 to 100, with an average of approximately 48.5. The standard deviation is approximately 27.97, indicating a wide spread of the data.
12. **Industry Income Score**: The industry income score ranges from 36.9 to 100, with an average of approximately 47.1. The standard deviation is approximately 15.09, indicating a wide spread of the data.
13. **International Outlook Score**: The international outlook score ranges from 14.1 to 99.7, with an average of approximately 46.9. The standard deviation is approximately 22.58, indicating a wide spread of the data.

**Code**

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

import plotly.express as px

university= pd.read\_csv('World University Rankings 2023.csv')

university.head()

university.info()

missing\_values = university.isnull().sum()

print("Missing Values:\n", missing\_values)

missing\_values = university.isnull().sum()

print("Missing Values:\n", missing\_values)

university.drop\_duplicates(inplace=True)

university['Location'].value\_counts()

plt.figure(figsize=(12, 8))

sns.countplot(data=university, x='Location', order=university['Location'].value\_counts().index)

plt.title('Regional Distribution of Universities')

plt.xlabel('Region')

plt.ylabel('Count')

plt.xticks(rotation=90)

plt.show()

# Get the top ten countries with the most universities in the top 100

top\_100= university.head(100)

countries = top\_100['Location']

# Find rows with null values in the 'Location' column

null\_locations = top\_100[top\_100['Location'].isnull()]

# Display the rows with null values

print(null\_locations)

#fill in the missing country data

universities\_to\_countries = {

'Tsinghua University': 'China',

'Peking University': 'China',

'National University of Singapore': 'Singapore',

'Technical University of Munich': 'Germany',

'University of Hong Kong': 'Hong Kong',

'LMU Munich': 'Germany',

'KU Leuven': 'Belgium',

'Universität Heidelberg': 'Germany',

'Chinese University of Hong Kong': 'Hong Kong',

'McGill University': 'Canada',

'The University of Queensland': 'Australia',

'University of Manchester': 'United Kingdom',

'The Hong Kong University of Science and Technology': 'Hong Kong',

'Zhejiang University': 'China',

'UNSW Sydney': 'Australia',

'University of Science and Technology of China': 'China',

'University of Groningen': 'Netherlands',

'University of Bristol': 'United Kingdom',

'Leiden University': 'Netherlands',

'Yonsei University (Seoul campus)': 'South Korea',

'Hong Kong Polytechnic University': 'Hong Kong',

'Erasmus University Rotterdam': 'Netherlands',

'University of Glasgow': 'United Kingdom',

'McMaster University': 'Canada',

'University of Adelaide': 'Australia',

'City University of Hong Kong': 'Hong Kong'

}

# Iterate through the universities and assign the corresponding country

for university\_name, country in universities\_to\_countries.items():

top\_100.loc[top\_100['Name of University'] == university\_name, 'Location'] = country

# number of universities in each country

country\_counts = countries.value\_counts()

# top ten countries

top\_ten\_countries = country\_counts.head(10)

print(top\_ten\_countries)

# Convert 'OverAll Score' column to float using .loc for assignment

top\_100.loc[:, 'OverAll Score'] = top\_100['OverAll Score'].astype(float)

# Create a pie chart

plt.figure(figsize=(8, 8))

plt.pie(top\_ten\_countries, labels=top\_ten\_countries.index, autopct='%1.1f%%', startangle=140)

plt.title('Top 10 Countries in University Ranking (Average Score)')

plt.axis('equal')

# Show the pie chart

plt.show()

# top ten countries and their count of universities in the top 100

top\_ten\_countries\_df = top\_ten\_countries.reset\_index()

# Rename the columns

top\_ten\_countries\_df.columns = ['Country', 'University Count']

#create the barchart

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

bar\_chart = sns.barplot(x='Country', y='University Count', data=top\_ten\_countries\_df, palette='viridis')

bar\_chart.set\_title('Top 10 Countries with Universities in Top 100')

bar\_chart.set\_xlabel('Country')

bar\_chart.set\_ylabel('Number of Universities')

plt.xticks(rotation=45, ha='right')

plt.show()

# Filter the data to include only the top 10 countries

filtered\_data = top\_100[top\_100['Location'].isin(top\_ten\_countries.index)]

# Create a distribution box plot

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

sns.boxplot(x='Location', y='OverAll Score', data=filtered\_data)

plt.title('Distribution of University Rankings in Top 10 Countries')

plt.xticks(rotation=45)

plt.xlabel('Country')

plt.ylabel('OverAll Score')

plt.show()

# Scatterplot for Teaching Score vs. OverAll Score

plt.scatter(top\_100['Teaching Score'], top\_100['OverAll Score'])

plt.title('Teaching Score vs. OverAll Score')

plt.xlabel('Teaching Score')

plt.ylabel('OverAll Score')

plt.show()

# Scatterplot for Research Score vs. OverAll Score

plt.scatter(top\_100['Research Score'], top\_100['OverAll Score'])

plt.title('Research Score vs. OverAll Score')

plt.xlabel('Research Score')

plt.ylabel('OverAll Score')

plt.show()

# Scatterplot for Citations Score vs. OverAll Score

plt.scatter(top\_100['Citations Score'], top\_100['OverAll Score'])

plt.title('Citations Score vs. OverAll Score')

plt.xlabel('Citations Score')

plt.ylabel('OverAll Score')

plt.show()

# Scatterplot for Industry Income Score vs. OverAll Score

plt.scatter(top\_100['Industry Income Score'], top\_100['OverAll Score'])

plt.title('Industry Income Score vs. OverAll Score')

plt.xlabel('Industry Income Score')

plt.ylabel('OverAll Score')

plt.show()

# Scatterplot for International Outlook Score vs. OverAll Score

plt.scatter(top\_100['International Outlook Score'], top\_100['OverAll Score'])

plt.title('International Outlook Score vs. OverAll Score')

plt.xlabel('International Outlook Score')

plt.ylabel('OverAll Score')

plt.show()

top\_n = 50

top\_scores\_teaching\_rank = university.nlargest(top\_n, 'Teaching Score')

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

sns.barplot(x='Name of University', y='Teaching Score', data=top\_scores\_teaching\_rank)

plt.title(f'Top {top\_n} Universities by Teaching Scores')

plt.xlabel('University Name')

plt.xticks(rotation=90)

plt.ylabel('Teaching Score Rank')

plt.show()

top\_n = 50

top\_scores\_research\_rank = university.nlargest(top\_n, 'Research Score')

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

sns.barplot(x='Name of University', y='Research Score', data=top\_scores\_research\_rank, color='green')

plt.title(f'Top {top\_n} Universities by Research Scores')

plt.xlabel('University Name')

plt.xticks(rotation=90)

plt.ylabel('Research Score')

plt.show()

top\_n = 50

top\_scores\_citation = university.nlargest(top\_n, 'Citations Score')

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

sns.barplot(x='Name of University', y='Citations Score', data=top\_scores\_research\_rank, color='skyblue')

plt.title(f'Top {top\_n} Universities by Citations Scores')

plt.xlabel('Name of Universitys')

plt.xticks(rotation=90)

plt.ylabel('Citations Score')

plt.show()

**Screenshot Output**

