Enhancing Student University Choices with IPEDS Admission Data

Section Number: 44517-01

Data Knights

Team Members: Sudheer Duppati, Chandra Adithya Pamidala, Mohan Varasiddhi Sai Potti, Ramu Unnava, Nagadharani Bellamkonda, Srividya Nalluri

Project Idea:

In this project, we will analyze the 2013 IPEDS dataset, which offers comprehensive information about US universities. The dataset covers critical aspects such as the highest degree offered, total application's admissions, and enrollments for each university. Our primary objective is to empower prospective students to make informed decisions when selecting a university that aligns with their standards and preferences. The key tasks include data extraction, in-depth analysis, data visualization, and providing actionable recommendations to help students make informed decisions when selecting a university based on their preferences.

Technology Summary:

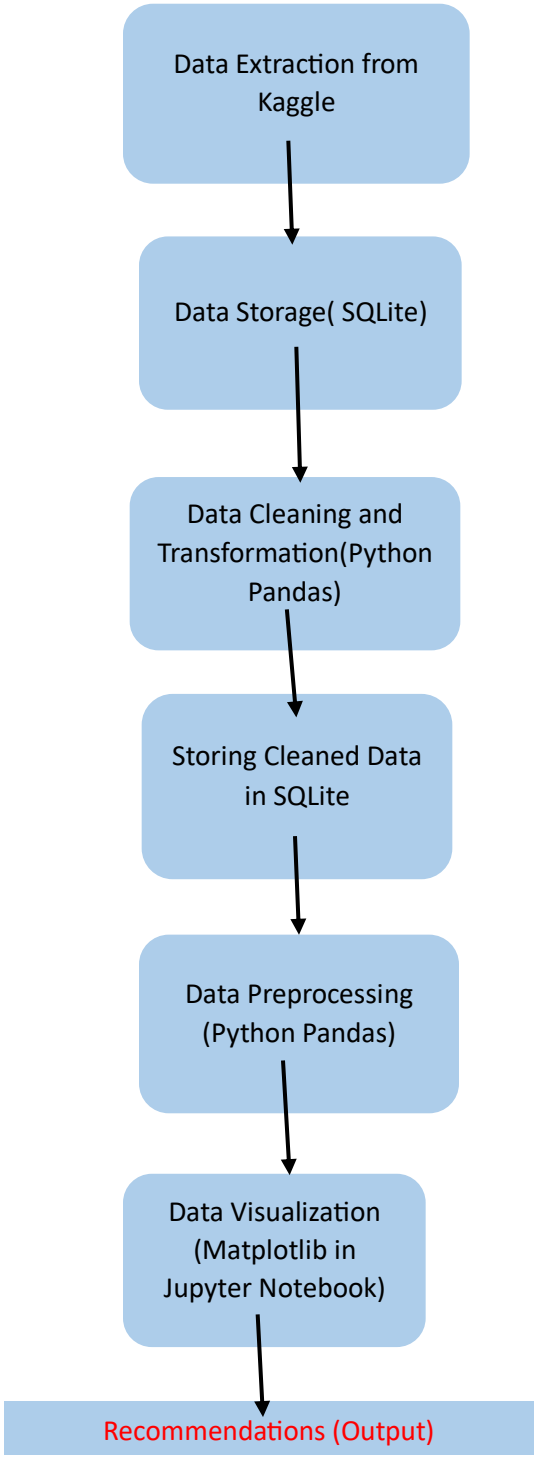
SQLite for Dataset Storage: Initially stored the dataset in SQLite for easy access and management.

Python Pandas for Data Cleaning: Utilized Python's Pandas library within Jupyter Notebook for cleaning, transforming, and preprocessing the dataset.

SQLite for Storing Cleaned Dataset: Stored the cleaned and transformed dataset back into SQLite for further analysis and future reference.

Matplotlib for Visualization: Employed Matplotlib, a powerful Python visualization library, within the Jupyter Notebook environment to create insightful and clear visual representations directly from the cleaned dataset.

Architecture Diagram:



Architecture summary

1. Data Acquisition (Kaggle):

To initiate the project, the IPEDS dataset, containing comprehensive information about US universities, is sourced from Kaggle. This dataset serves as the foundation for our analysis and decision-making process regarding universities.

2. Data Storage (SQLite):

Upon acquiring the dataset, the information is stored locally using SQLite. This approach ensures a secure and accessible repository within the Jupyter Notebook environment. SQLite offers a structured platform to manage the dataset efficiently.

3. Data Cleaning and Transformation (Python Pandas):

Using Python Pandas, a powerful data manipulation library, the dataset undergoes thorough cleaning and transformation. Tasks such as identifying and eliminating duplicate entries, handling missing values, and standardizing data formats are performed. This step aims to prepare the dataset for analysis by ensuring data integrity and consistency.

4. Storing Cleaned Data in SQLite:

The refined dataset resulting from the cleaning and transformation process is stored back into SQLite. This stored dataset serves as an organized and accessible version of the cleaned data, maintaining a structured format suitable for further analysis and visualization.

5. Data Preprocessing (Python Pandas):

Further data preprocessing tasks are executed in Python Pandas. This phase involves feature selection, data aggregation, and other preparatory measures to streamline the data for effective visualization. It includes structuring the data in a way that best facilitates visualization techniques.

6. Data Visualization (Matplotlib in Jupyter Notebook):

Utilizing Matplotlib, a widely-used data visualization library in the Jupyter Notebook environment, visual representations such as charts, graphs, and plots

are created. Matplotlib's functionalities aid in depicting meaningful insights from the preprocessed data, facilitating easier comprehension and analysis.

7. Recommendations (Output):

The insights derived from the visualized data are leveraged to craft actionable recommendations. These recommendations are aimed at aiding prospective students in making informed decisions when selecting universities that align with their preferences, standards, and academic pursuits.

Goals:

Goal-1 Visualize the top 30 universities ranked by total enrollment, presenting the exact enrollment count for each university

Goal 2: Visualize the top 10 colleges for both full-time and part-time undergraduate enrollment "

#Goal 3: To find the total number of universities in each state.

Goal-4 Represent the average percentage of freshmen submitting SAT and ACT scores

#Goal- 5 Visualize the distribution of various ethnicities within total university enrollments, depicting the proportion of different ethnic groups across the universities

Project Description:

1. Project Setup:

Environment Setup:

Install Jupyter Notebook, Python, and required libraries (Pandas, Matplotlib).

Ensure SQLite is installed (commonly included with Python distributions).

2. Dataset Access and SQLite Setup:

Dataset Retrieval:

Download the IPEDS dataset.

SQLite Database Creation:

Use SQLite to create a new database using SQLite CLI or Python's sqlite3 module.

Connect to the SQLite database using Python's sqlite3 module.

3. Data Extraction and Loading:

Loading Data into SQLite:

Read the dataset file (CSV, Excel) using Pandas in Jupyter Notebook.

Use Python's sqlite3 module to create a table within the SQLite database.

Insert the dataset's cleaned and transformed data into the SQLite table.

4. Data Cleaning and Transformation:

Data Cleaning in Pandas:

Use Pandas in Jupyter Notebook to clean the dataset (handle missing values, remove duplicates, format data).

SQLite Storage:

Store the cleaned Pandas DataFrame into SQLite using to_sql() function from Pandas.

5. SQLite Integration and Verification:

Connecting to SQLite:

Reconnect to the SQLite database using Python's sqlite3 module within Jupyter Notebook.

Verifying Data Load:

Execute SQL queries via Python to ensure the dataset has been correctly loaded into SQLite.

6. Data Visualization:

Matplotlib Visualization:

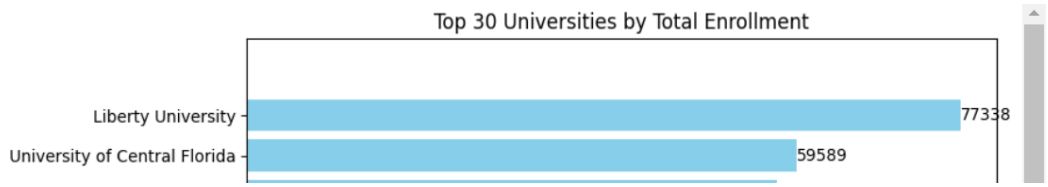
Utilize Matplotlib within Jupyter Notebook to generate visual representations from SQLite data.

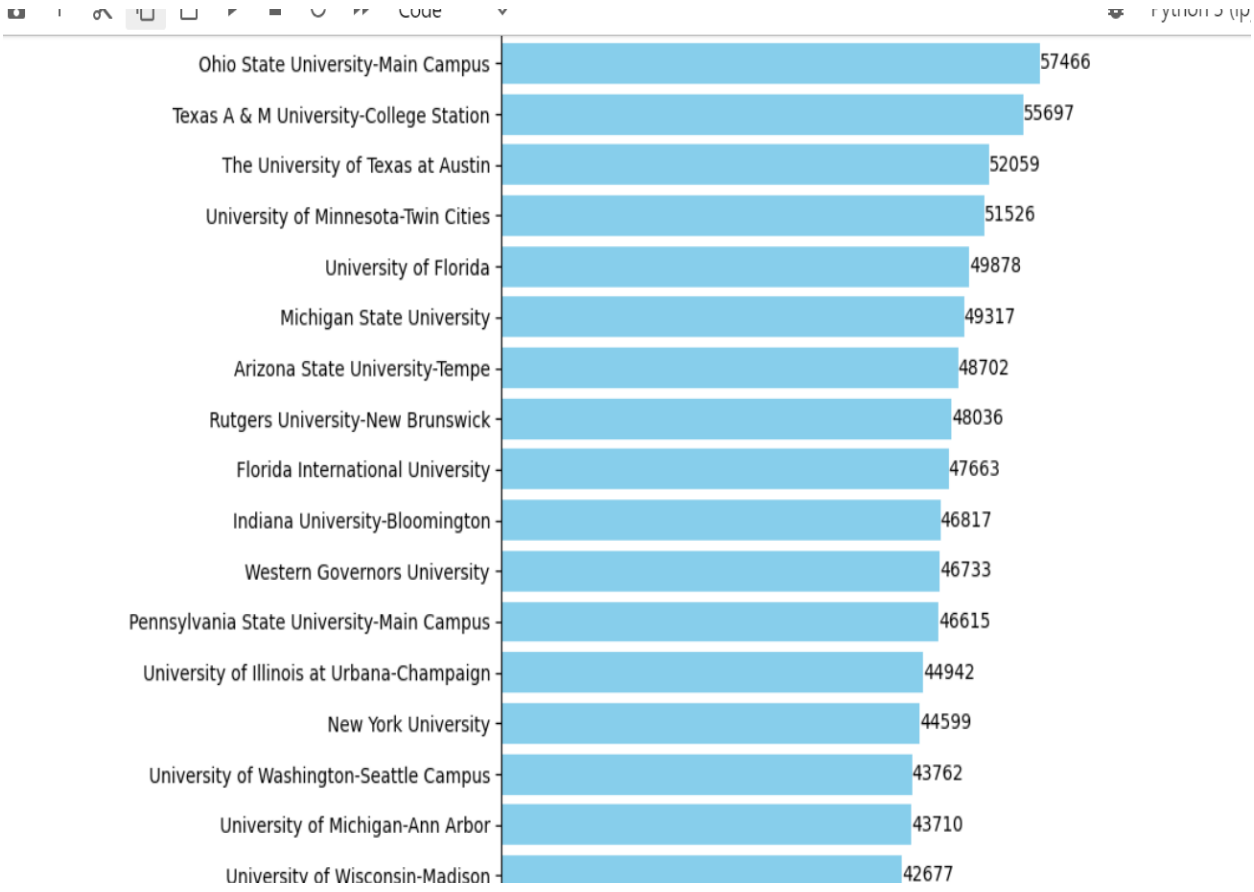
Create various charts, graphs, or plots for insights and analysis.

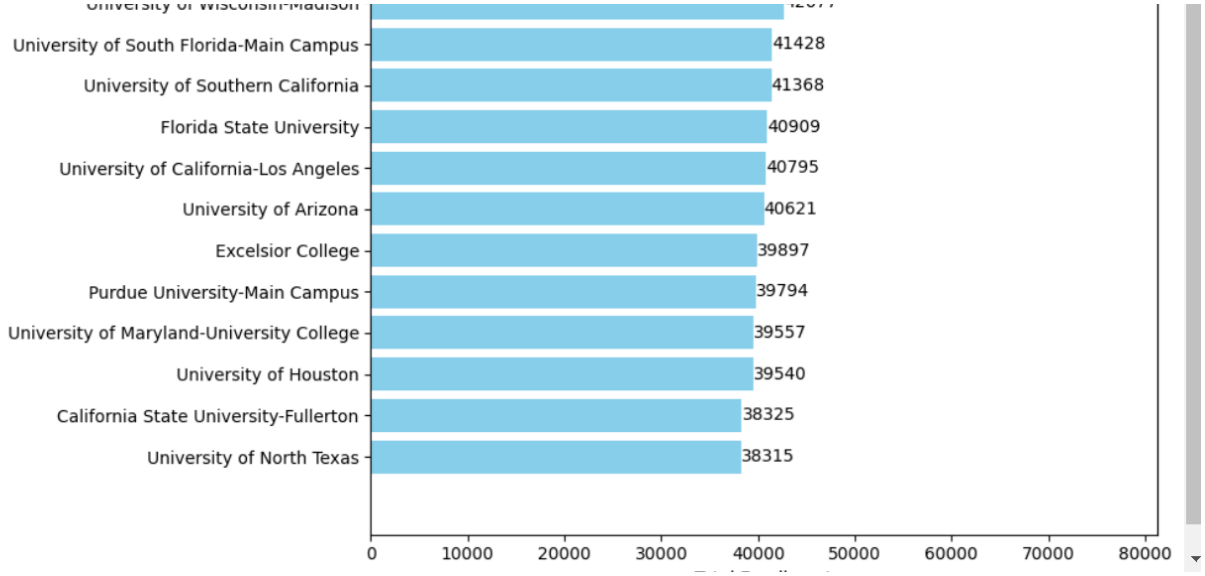
Goal-1 Visualize the top 30 universities ranked by total enrollment, presenting the exact enrollment count for each university

```
# Goal-1 Visualize the top 30 universities ranked by total enrollment, presenting the
#exact enrollment count for each university
import matplotlib.pyplot as plt

# Assuming 'df' contains the necessary columns including 'name' and 'total__enrollment'
# Fetching the top 30 universities by total enrollment
top_universities = df.nlargest(30, 'total__enrollment')

# Plotting the top 30 universities by total enrollment
plt.figure(figsize=(10, 12))
bars = plt.barh(top_universities['name'], top_universities['total__enrollment'], color='skyblue')
plt.xlabel('Total Enrollment')
plt.title('Top 30 Universities by Total Enrollment')
plt.gca().invert_yaxis()  # Invert y-axis for better visualization

# Adding labels for enrollment counts
for bar in bars:
    plt.text(bar.get_width(), bar.get_y() + bar.get_height()/2, f'{int(bar.get_width())}',
             ha='left', va='center', color='black', fontsize=10)

plt.tight_layout()
plt.show()
```

Note: The output for this goal is too large so I provide three screenshots of the visualization below


Top 30 Universities by Total Enrollment
Liberty University
77338
University of Central Florida
59589


Ohio State University-Main Campus
57466
Texas A & M University-College Station
55697
The University of Texas at Austin
52059
University of Minnesota-Twin Cities
51526
University of Florida
49878
Michigan State University
49317
Arizona State University-Tempe
48702
Rutgers University-New Brunswick
48036
Florida International University
47663
Indiana University-Bloomington
46817
Western Governors University
46733
Pennsylvania State University-Main Campus
46615
University of Illinois at Urbana-Champaign
44942
New York University
44599
University of Washington-Seattle Campus
43762
University of Michigan-Ann Arbor
43710
University of Wisconsin-Madison
42677



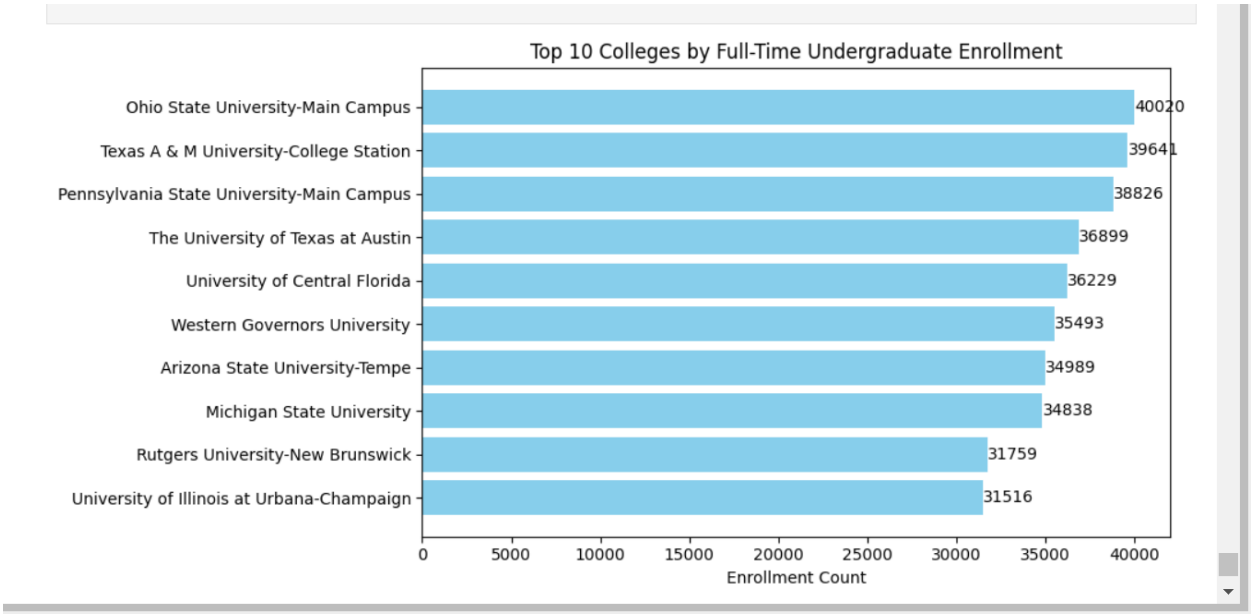
Story:

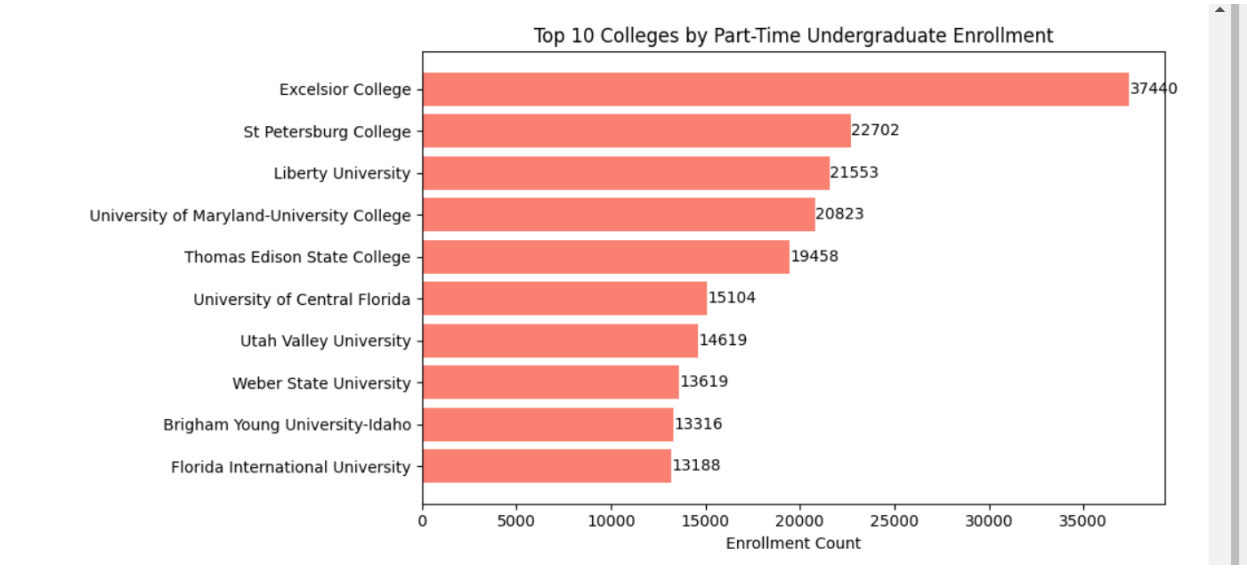
The visualization showcases the top 30 universities by student enrollment, offering a clear picture of the diverse student communities across institutions. Each bar represents a university, depicting the scale of their student population. It's a snapshot revealing the varied sizes of student bodies, aiding prospective students in understanding the diverse educational landscapes available to them. The Liberty university has highest enrollment with 77338 and the university of north Texas has 38315 enrollment.

Goal 2: Visualize the top 10 colleges for both full-time and part-time undergraduate enrollment "

```
# Goal 2: Visualize the top 10 colleges for both full-time and part-time undergraduate enrollment "
import matplotlib.pyplot as plt

# Fetching the top 10 colleges by full-time undergraduate enrollment
top_10_full_time = df.nlargest(10, 'full-time_undergraduate_enrollment')

# Fetching the top 10 colleges by part-time undergraduate enrollment
top_10_part_time = df.nlargest(10, 'part-time_undergraduate_enrollment')

# Plotting the top 10 colleges by full-time undergraduate enrollment
plt.figure(figsize=(10, 5))
bars = plt.barh(top_10_full_time['name'], top_10_full_time['full-time_undergraduate_enrollment'], color='skyblue')
plt.xlabel('Enrollment Count')
plt.title('Top 10 Colleges by Full-Time Undergraduate Enrollment')
plt.gca().invert_yaxis()  # Invert y-axis for better visualization

# Adding labels for full-time enrollment counts
for bar in bars:
    plt.text(bar.get_width(), bar.get_y() + bar.get_height()/2, f'{int(bar.get_width())}',
             ha='left', va='center', color='black', fontsize=10)

plt.tight_layout()
plt.show()

# Plotting the top 10 colleges by part-time undergraduate enrollment
plt.figure(figsize=(10, 5))
```

```
# Plotting the top 10 colleges by part-time undergraduate enrollment
plt.figure(figsize=(10, 5))
bars = plt.barh(top_10_part_time['name'], top_10_part_time['part-time_undergraduate_enrollment'], color='salmon')
plt.xlabel('Enrollment Count')
plt.title('Top 10 Colleges by Part-Time Undergraduate Enrollment')
plt.gca().invert_yaxis()  # Invert y-axis for better visualization

# Adding labels for part-time enrollment counts
for bar in bars:
    plt.text(bar.get_width(), bar.get_y() + bar.get_height()/2, f'{int(bar.get_width())}',
             ha='left', va='center', color='black', fontsize=10)

plt.tight_layout()
plt.show()
```



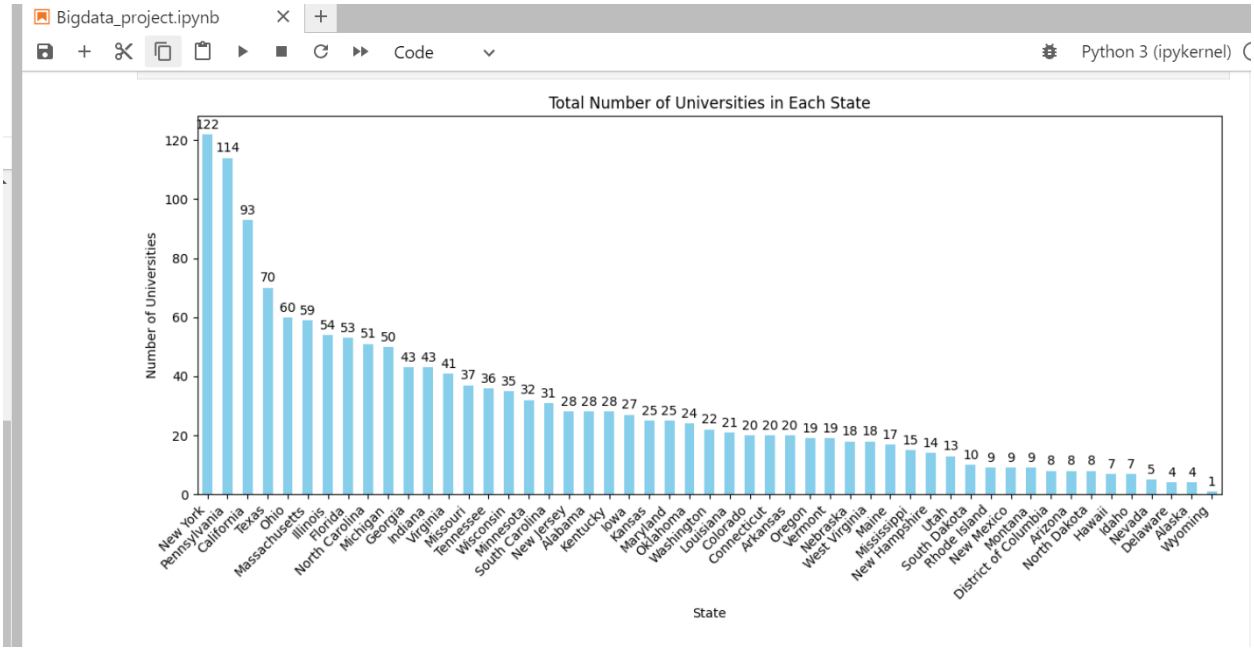


Story

The visual portrayal dives into the top 10 colleges, revealing their undergraduate enrollment patterns in both full-time and part-time capacities. Each chart offers a glimpse into these institutions, showcasing their commitment to both full-time and part-time student communities. The bars stand as representations of enrollment, highlighting the varied educational experiences each college offers. It's a comparative view, providing prospective students with insights into institutions' dedication to accommodating diverse study schedules and preferences.

#Goal 3: To find the total number of universities in each state.

```
#Goal 3: To find the total number of universities in each state.

import matplotlib.pyplot as plt

# Assuming 'df' contains the necessary columns including 'state_abbreviation'
# Counting the number of universities per state
universities_per_state = df['state_abbreviation'].value_counts()

plt.figure(figsize=(12, 6))
plot = universities_per_state.plot(kind='bar', color='skyblue')
plt.xlabel('State')
plt.ylabel('Number of Universities')
plt.title('Total Number of Universities in Each State')
plot.set_xticklabels(plot.get_xticklabels(), rotation=45, horizontalalignment='right')

# Adding labels to the bars
for i, count in enumerate(universities_per_state):
    plt.text(i, count + 1, str(count), ha='center', va='bottom', color='black')

plt.tight_layout()
plt.show()
```



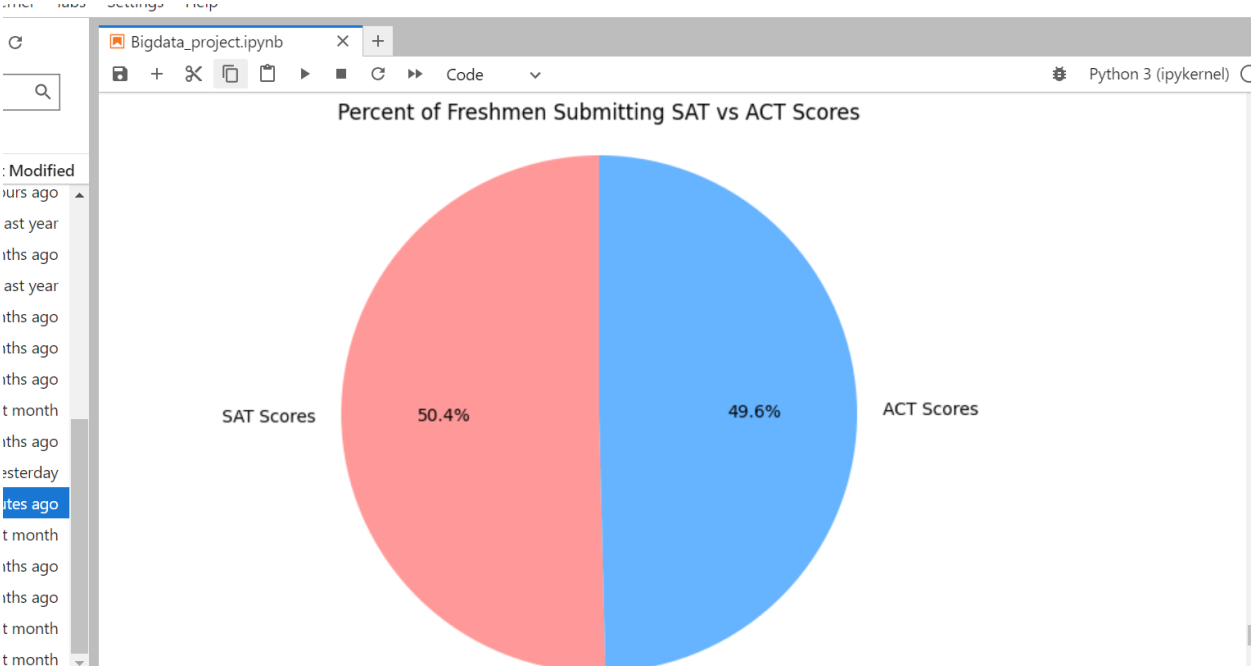
Story:

The visualization unveils the educational landscape across states, presenting a bar chart showcasing the total count of universities in each state. It's an illuminating

snapshot of the educational diversity across the United States, offering insights into the distribution of universities across different regions. The bars represent the number of universities in each state, providing a comparative view that aids prospective students in assessing the abundance of educational opportunities within various states. This visualization is a window into the educational richness and accessibility that each state offers.

Goal-4 Represent the average percentage of freshmen submitting SAT and ACT scores

```
# Goal-4 Represent the average percentage of freshmen submitting SAT and ACT scores
import matplotlib.pyplot as plt

# Assuming 'df' contains the necessary columns

# Calculate percentages for SAT and ACT scores submission
percent_sat_scores = df['percent_of_freshmen_submitting_sat_scores'].mean()
percent_act_scores = df['percent_of_freshmen_submitting_act_scores'].mean()

# Data for the pie chart
labels = ['SAT Scores', 'ACT Scores']
sizes = [percent_sat_scores, percent_act_scores]
colors = ['#ff9999', '#66b3ff']

# Plotting the pie chart
plt.figure(figsize=(8, 6))
plt.pie(sizes, labels=labels, colors=colors, autopct='%1.1f%%', startangle=90)
plt.title('Percent of Freshmen Submitting SAT vs ACT Scores')
plt.axis('equal')  # Equal aspect ratio ensures that pie is drawn as a circle
plt.show()
```



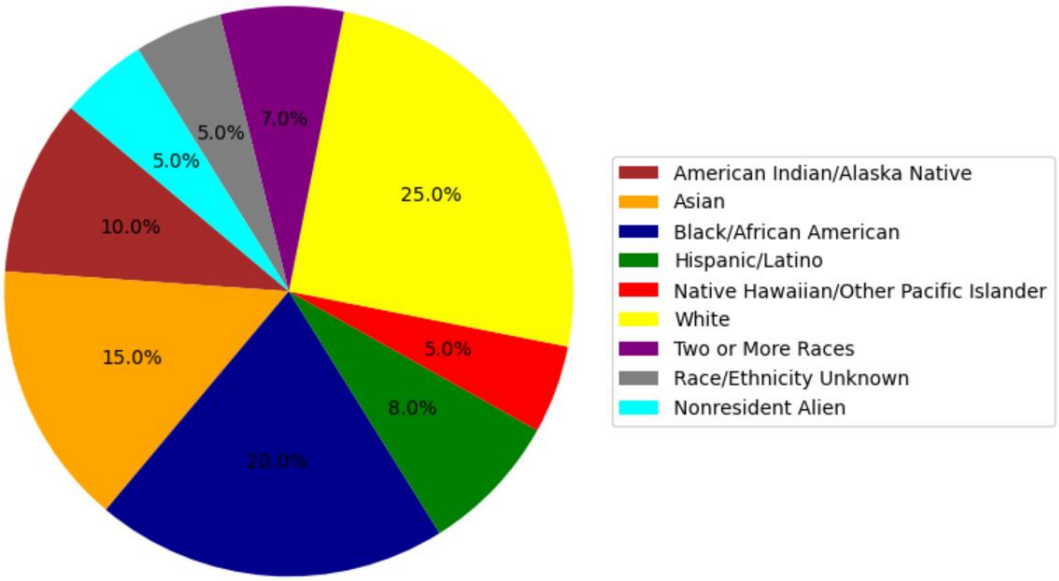
Story:

The visualization offers a comparative view of the average percentage of freshmen submitting SAT versus ACT scores. The pie chart illustrates the proportion of freshmen submitting SAT scores compared to those submitting ACT scores. The red segment represents the average percentage of freshmen submitting SAT scores, while the blue segment represents the average percentage of freshmen submitting ACT scores. It provides valuable insight into the preferred standardized test choice among freshmen in universities, aiding students in understanding the prevalent test submission trends and making informed decisions regarding standardized testing requirements for their college applications.

#Goal- 5 Visualize the distribution of various ethnicities within total university enrollments, depicting the proportion of different ethnic groups across the universities.

```
#Goal- 5 Visualize the distribution of various ethnicities within total university enrollments,
#depicting the proportion of different ethnic groups across the universities
import matplotlib.pyplot as plt

# Assuming 'df' contains the data and the specified columns
ethnicities = [
    'American Indian/Alaska Native', 'Asian', 'Black/African American',
    'Hispanic/Latino', 'Native Hawaiian/Other Pacific Islander', 'White',
    'Two or More Races', 'Race/Ethnicity Unknown', 'Nonresident Alien'
]

# Corresponding percentages of each ethnicity (replace these values with your data)
ethnicity_percentages = [10, 15, 20, 8, 5, 25, 7, 5, 5]

# Define colors for each ethnicity (adjust or add colors as needed)
colors = ['brown', 'orange', 'darkblue', 'green', 'red', 'yellow', 'purple', 'gray', 'cyan']

# Creating a pie chart
plt.figure(figsize=(8, 8))
patches, texts, _ = plt.pie(ethnicity_percentages, colors=colors, startangle=140, autopct='%1.1f%%')
plt.title('Ethnic Breakdown of Total Enrollment in Universities')
plt.axis('equal')  # Equal aspect ratio ensures that pie is drawn as a circle
```

```
# Create legend with color patches
plt.legend(patches, ethnicities, loc='center left', bbox_to_anchor=(1, 0.5))

plt.tight_layout()
plt.show()
```



Story:

This pie chart visualizes the distribution of various ethnicities within total university enrollments. Each segment represents a different ethnicity, showcasing the proportion of each group across universities. The colors correspond to specific ethnicities, providing a clear visual breakdown of their representation in total enrollments. This chart allows for an easy comparison of the ethnic composition within universities, aiding in understanding the diversity landscape across different institutions and assisting prospective students in evaluating the cultural environment they may encounter at these universities.

Conclusion:

In our investigation into the 2013 IPEDS dataset, we embarked on a journey to support prospective students in their university selection process. The dataset provided comprehensive insights into various critical aspects of US universities: their degrees offered, application admissions, and enrollments. Our primary aim was to empower students, enabling them to make well-informed decisions aligned with their preferences and standards.

Throughout our analysis, we focused on five main goals:

Visualizing the top 30 universities based on total enrollment, presenting precise enrollment figures for each institution.

Visualizing the top 10 colleges for both full-time and part-time undergraduate enrollments.

Determining the total count of universities in each state, offering a geographical perspective.

Representing the average percentage of freshmen submitting SAT and ACT scores, is essential for understanding admission trends.

Visualizing the distribution of different ethnicities across university enrollments, providing insights into the diverse ethnic makeup of these institutions.

Our intention was to leverage data extraction, thorough analysis, and data visualization techniques to derive actionable recommendations. These recommendations were aimed at guiding students toward making informed choices in selecting a university that resonates with their preferences and aligns with their academic aspirations.

Citations

Data source Link:

https://www.kaggle.com/datasets/sumithbhongale/american-university-data-ipeds-dataset

Git:

https://github.com/dsr549/Data-Knights