Robinson540Project

November 23, 2025

1 Project Milestone 2 & 3

Derrick Robinson DSC 540 4/20/2025

[2]:
```
import pandas as pd

# Load the Excel file
df = pd.read_csv(r'C:\Users\rer12\Downloads\Most-Recent-Cohorts-Institution.
 ↪csv')
```

```
C:\Users\rer12\AppData\Local\Temp\ipykernel_105180\3667882468.py:4:
DtypeWarning: Columns (9,1537,1540,1542,1606,1608,1614,1615,1619,1620,1621,1622,
1623,1624,1625,1626,1627,1628,1629,1690,1692,1697,1700,1725,1726,1727,1728,1729,
1743,1815,1816,1817,1818,1823,1824,1830,1831,1879,1880,1881,1882,1883,1884,1885,
1886,1887,1888,1889,1890,1891,1892,1893,1894,1895,1896,1897,1898,1909,1910,1911,
1912,1913,1957,1958,1959,1960,1961,1962,1963,1964,1965,1966,1967,1968,1969,1970,
1971,1972,1973,1974,1975,1976,1983,1984,2376,2377,2403,2404,2495,2496,2497,2498,
2499,2500,2501,2502,2503,2504,2505,2506,2507,2508,2509,2510,2511,2512,2513,2514,
2515,2516,2517,2518,2519,2520,2521,2522,2523,2524,2525,2526,2527,2528,2529,2530,
2958,3215,3231,3235,3236) have mixed types. Specify dtype option on import or
set low_memory=False.
  df = pd.read_csv(r'C:\Users\rer12\Downloads\Most-Recent-Cohorts-
Institution.csv')
```

Step 1: Rename Columns – Cleaned up column headers to make them more readable and consistent with analysis best practices.

[4]:
```
df.rename(columns={
    'INSTNM': 'institution_name',
    'CITY': 'city',
    'STABBR': 'state',
    'ADM_RATE': 'acceptance_rate',
    'COSTT4_A': 'avg_cost',
    'C150_4': 'grad_rate',  # <-- updated to match actual column name
    'SAT_AVG': 'sat_avg',
    'DEBT_MDN': 'median_debt'
}, inplace=True)
```

Step 2: Fix Casing – Ensured institution names and city names are in Title Case for consistency in merging later

[6]:
```
df['institution_name'] = df['institution_name'].str.title()
df['city'] = df['city'].str.title()
```

Step 3: Format Percentages – Converted rates (originally decimals like 0.52) to percentage format (e.g., 52.0).

[8]:
```
df['acceptance_rate'] = df['acceptance_rate'] * 100
df['grad_rate'] = df['grad_rate'] * 100
```

Step 4: Remove Duplicates – Dropped exact duplicate institutions (based on name and state) to prevent inflated metrics.

[10]:
```
df.drop_duplicates(subset=['institution_name', 'state'], inplace=True)
```

Step 5: Remove Outliers – Filtered out any schools with suspiciously low or high SAT scores outside the realistic 600–1600 range

[12]:
```
df = df[(df['sat_avg'] >= 600) & (df['sat_avg'] <= 1600)]
```

[13]:
```
df_cleaned = df[['institution_name', 'city', 'state', 'avg_cost',␣
 ↪'acceptance_rate', 'grad_rate', 'sat_avg', 'median_debt']]
df_cleaned.head()
```

[13]:

	institution_name	city	state	avg_cost	acceptance_rate	grad_rate	sat_avg	median_debt
0	Alabama A & M University	Normal	AL	23167.0	68.40	26.78	920.0	16600
1	University Of Alabama At Birmingham	Birmingham	AL	26257.0	86.68	64.42	1291.0	15832
3	University Of Alabama In Huntsville	Huntsville	AL	25777.0	78.10	62.95	1259.0	13905
4	Alabama State University	Montgomery	AL	21900.0	96.60	27.73	963.0	17500
5	The University Of Alabama	Tuscaloosa	AL	31024.0	80.06	72.76	1304.0	17986

2 Ethical Implications

During the data wrangling process for the College Scorecard dataset, several changes were made to improve clarity and usability, including renaming columns, formatting percentage fields, standardizing text casing, removing duplicates, and filtering out unrealistic SAT score values. While the dataset is publicly available from the U.S. Department of Education and thus complies with open data usage policies, there are still ethical considerations regarding data interpretation and manipulation. For example, by removing outliers in SAT scores or institutions with incomplete data, we risk excluding underrepresented schools or misrepresenting national trends. Additionally, assumptions were made in identifying outliers (e.g., setting SAT score thresholds between 600–1600), which could impact analysis if those assumptions don't align with actual student population norms. The data was sourced from a credible and official government portal, ensuring its accuracy

and legitimacy. However, since transformations involved subjective decisions, transparency and documentation of these steps are critical to maintaining trust and reproducibility. To mitigate ethical concerns, I ensured that all transformations were reversible, reproducible, and guided by clear, consistent logic rather than biased selection. Further, any insights drawn from the cleaned data will be carefully framed to avoid overgeneralizations or unsupported claims.

3 Project Milestone 3

[30]:
```
import pandas as pd

url = "https://www3.cs.stonybrook.edu/~skiena/skienarank/"
tables = pd.read_html(url)

# inspect the table
df = tables[0]

# step 1: Rename columns
df.columns = ['Combined Rank', 'University', 'Length Rank', 'Hits Rank', 'Page␣
↪Rank']
```

[32]:
```
# Step 2: Remove Leading/Trailing Whitespace from University Names
df['University'] = df['University'].str.strip()

# Step 3: Remove Duplicate Rows
df.drop_duplicates(inplace=True)
```

[34]:
```
# Step 4: Convert Rank Columns to Integer Type
rank_cols = ['Combined Rank', 'Length Rank', 'Hits Rank', 'Page Rank']
df[rank_cols] = df[rank_cols].astype(int)

# Step 5: Filter Out Invalid or Extreme Ranks (Outlier Check)
# Remove rows where any rank is greater than 1000 (arbitrary cap for demo)
df = df[(df[rank_cols] <= 1000).all(axis=1)]
```

[36]:
```
# Display the cleaned DataFrame
print(df.head())
```

```
   Combined Rank                       University  Length Rank  Hits Rank  \
0              1      University of Pennsylvania            1          6
1              2                  Yale University            5          3
2              3              Columbia University            8          5
3              4               Harvard University          219          1
4              5  Massachusetts Inst. of Technology           24          4

   Page Rank
0         13
1          6
```

```
2          5
3          1
4          3
```

4 Ethical Implications – Website Data

The primary changes made to the data involved renaming columns for clarity, stripping whitespace, removing duplicates, converting ranks to integers, and filtering extreme values. Since the dataset is publicly available on an academic website and involves non-sensitive institutional information, there are minimal legal or regulatory concerns. However, one risk is that modifying or filtering ranks—especially outliers—could introduce bias or misrepresent the performance of certain universities. Assumptions were made that extremely high rank values were invalid or anomalous, though this may not always be accurate. The data was sourced from Stony Brook University's faculty site, which is generally regarded as credible, and scraping HTML tables is permissible under fair use for academic research. To mitigate potential ethical concerns, all transformations were documented transparently, and the data was used solely for educational and analytical purposes, without altering the meaning or context of the original rankings.

[ ]: