Exercise 6.1

Answers

Part 1: Reasons for Choosing this Dataset

The excel file titled **cwurData.csv** was selected for the analysis, because:

- It contains a comprehensive, structured dataset of global university rankings with multiple quantitative variables, making it suitable for data profiling, cleaning, and statistical analysis.
- The dataset is open, well-documented, and relevant for exploring questions about institutional performance, education quality, and global comparisons.
- Its structured format and rich variable set allow for meaningful data cleaning, descriptive statistics, and the development of insightful analytical questions.

The entire dataset was retrieved from the following source:

https://www.kaggle.com/datasets/mylesoneill/world-university-rankings

Part 2: Ethical Consideration

When analyzing global university rankings, it is crucial to approach the data with care and responsibility to ensure ethical standards are upheld. Key ethical considerations include:

1. **Data Source Transparency**: The CWUR ranking relies on publicly available data, but the specific sources and methodologies used to compile the data may not always be transparent. It is important to acknowledge any uncertainties or limitations in the data's origin and collection methods.
2. **Ranking Bias**: University rankings are often criticized for their inherent biases towards certain types of institutions or metrics (e.g., research-focused universities, English-speaking institutions). It is important to recognize and discuss these potential biases to avoid perpetuating unfair or inaccurate comparisons.
3. **Misinterpretation of Metrics**: The ranking metrics (e.g., quality of education, alumni employment) are complex and multifaceted. Oversimplifying or misinterpreting these metrics can lead to flawed conclusions about institutional performance.

4. **Impact on Institutions**: University rankings can have significant consequences for institutions, affecting their reputation, funding, and student enrollment. Analyses should be conducted responsibly to avoid unfairly stigmatizing or harming institutions.
5. **Cultural and Contextual Sensitivity**: Universities operate in diverse cultural and national contexts, and rankings may not fully capture the unique strengths and challenges of institutions in different regions. It is important to consider the broader cultural and socioeconomic factors that influence university performance.
6. **Transparency in Analysis**: Clearly communicate the methodologies, assumptions, and limitations of your analysis to avoid misleading or overstating findings. Ensure that the analysis is reproducible and transparent.

Part 3: Details about the Data

Data Source

The data is from the Center for World University Rankings (CWUR), which lists global university rankings based on various performance indicators. It covers multiple years and includes universities from different countries.

Data Collection

The dataset was collected from open sources and compiled on Kaggle, ensuring it is accessible and reliable for analysis.

Data Limitations

The dataset may have missing values and could reflect biases in ranking criteria or data collection methods.

Why this data

With my background as a student data consultant and my passion for a career in educational analytics, this dataset aligns perfectly with my interests and expertise.

Part 4: Questions to Explore

Ranking Trends:

- How have the rankings of specific universities changed over the years, and what factors might explain these trends?

- What is the correlation between overall score and individual ranking metrics (e.g., quality of education, publications, influence)?

Geographical Analysis:

- How do universities from different countries compare in the rankings, and what regional factors might influence these rankings?
- Are there specific countries or regions that consistently outperform others in certain ranking metrics?

Performance Indicators:

- Which factors (e.g., quality of faculty, research output) have the strongest correlation with a university's overall ranking?
- How do different universities balance their performance across various ranking indicators?

Impact of Funding:

- Is there a correlation between a country's expenditure on education (as a percentage of GDP) and the ranking of its universities?
- Do universities with higher scores in specific metrics also have higher levels of research funding or endowments?

Longitudinal Stability:

- Which universities have maintained consistently high rankings over the years, and what strategies might they have employed to achieve this stability?
- How resistant are university rankings to external factors such as economic changes, policy shifts, or global events?

Part 4: Data Cleaning Summary

1. **Handling Missing Values:**

- Filled missing values in numerical columns using the mean of each column to preserve sample size and statistical power. Specific columns imputed include quality_of_education, alumni_employment, quality_of_faculty, publications , influence, citations, broad_impact, patents, and score.

2. **Removing Duplicates:**

- Removed duplicate rows to ensure the dataset contains unique entries, preventing potential bias in analyses.

3. **Standardizing Text Data:**

- Standardized the country column by removing leading and trailing spaces to ensure consistent categorization and accurate aggregations.

4. **Examining Data Timeliness:**

- Verified the range of years in the year column to ensure the data covers the intended time frame for analysis and to identify any potential gaps in the temporal coverage.

5. **Verifying Data Types:**

- Checked each column for mixed data types to ensure proper data handling. No mixed-type columns were identified, ensuring that each column's data is uniformly typed for accurate processing.

6. **Outlier Management:**

- Identified outliers using box plots across all numerical columns. The score column was clipped to a maximum value of 100 to mitigate the impact of extreme outliers on statistical analyses, ensuring that no scores exceed the logical maximum.

7. **Saving the Cleaned Dataset:**

- Saved the cleaned and processed dataset to a new CSV file for easy access and further analysis. This ensures that all cleaning steps are preserved and the data is readily available for subsequent analytical tasks.