**Memory Document**

**Title:** University Rankings Analysis (2014-2024)

**1. Project Overview**

The University Rankings Analysis project aimed to examine the global rankings of universities from 2014 to 2024, focusing on identifying patterns, trends, and key drivers of ranking changes over this period. The data was sourced from the Center for World University Rankings (CWUR), providing a rich dataset with 11,000 data points, including metrics such as World Rank, National Rank, Research Rank, Education Rank, Employability Rank, and Faculty Rank. This analysis sought to investigate whether certain factors like research output and faculty quality drive global ranking positions and to explore the impact of economic and regional factors on these rankings.

The project also set forth three hypotheses:

1. Research and employability ranks are primary elements of global university rankings.
2. Top university rankings exhibit relative stability over time.
3. Universities in developing countries show an upward trend in rankings compared to those in developed nations.

**2. Data Cleaning and Preprocessing**

One of the major challenges in this project was data cleaning. Each year’s dataset contained unique inconsistencies, particularly in the way missing values or threshold rankings were represented. For example, in some years, missing values were marked as '-', while in other years, values that exceeded a certain threshold were marked as '>200' or '200+'. This inconsistency meant that a single cleaning method could not be applied uniformly across all datasets.

To address this, I used a separate Jupyter Notebook for each year, implementing custom data-cleaning techniques to handle these inconsistencies. In these notebooks:

* All data ranks were initially recorded as strings in the dataset. I converted them to numeric values to enable accurate analysis and statistical computation.
* Missing values were initially replaced with NaN, and later, depending on specific conditions, were filled with the maximum value of the data plus one.
* Threshold values such as '>200' and '200+' were converted to numeric values and were filled with the maximum value of the data plus one (for example 201).
* After each dataset was cleaned individually, I merged them into a comprehensive dataset covering the entire period from 2014 to 2024. This merged dataset allowed for longitudinal analysis and trend visualization.

Additionally, the dataset included a "Score" metric, which was not used in this study. The scores showed significant variability across the years, with minimum scores around 44 between 2014 and 2017, increasing to approximately 75 from 2018 to 2024. Due to this inconsistency, scores were excluded from the analysis to avoid potential distortions in the results.

**3. Techniques and Methodology**

The analysis involved several techniques and tools, including exploratory data analysis, statistical analysis, and various visualizations.

* **Exploratory Data Analysis (EDA)**: I used histograms to examine the distributions of each ranking metric (World, Research, Education, Employability, Faculty) across different years. This helped in understanding where most universities were concentrated within each ranking category and identifying outliers or notable patterns.
* **Visualization Techniques**:
  + **Spider Plots** were employed to illustrate the ranking spread among top countries (e.g., the USA, China, the UK, Japan). These plots highlighted the USA's dominant presence and China's rapid growth in rankings.
  + **Trend Lines** displayed changes over time for the top 20 universities, focusing on ranking shifts and identifying institutions that rose or fell significantly in rank.
  + **Line Graphs and Histograms** were used to depict the frequency and distribution of universities across various ranking categories over time.
* **Statistical Analysis**: I calculated correlations among the different ranking factors (World Rank, Research Rank, Faculty Rank, etc.) to identify key drivers for high global rankings. The results supported the hypothesis that Research and Faculty metrics are primary contributors to a university’s World Rank, while other metrics such as Education and Employability played a moderate role.
* **Hypothesis Testing**: The hypotheses were tested based on observed trends and statistical correlations. For instance, by observing the stability in rankings among top universities and analyzing correlations, I was able to validate or refute each hypothesis based on empirical data.

**4. Challenges and Difficulties Encountered**

The project presented several challenges, particularly in data handling, methodology alignment, and merging different datasets.

* **Data Inconsistencies Across Years**: Handling each year's dataset separately was necessary due to the unique ways missing and threshold values were recorded. This required individualized attention and verification, as a single standardized cleaning method would have led to inaccuracies. Each dataset demanded a tailored approach, which was labor-intensive but crucial for accuracy.
* **Interpreting Correlations and Causality**: While correlations provided insights into the relationship between ranking factors, interpreting these as causative required caution. External factors, such as funding or government policy changes, may influence rankings indirectly, which couldn’t be captured through the data alone.
* **Visualizing a Large Dataset**: Representing trends across 11,000 data points posed a visualization challenge. I had to carefully select the visual types and scale the data to ensure readability without oversimplifying the information.

**5. Key Findings and Insights**

The analysis provided several meaningful insights into the dynamics of university rankings over the 2014-2024 period:

* **Stability of Top Universities**: The top 5 universities showed minimal changes in their ranking positions, highlighting the dominance of these institutions over the decade. This finding supports the hypothesis that top university rankings are relatively stable, especially among the elite tier.
* **Primary Drivers of Rankings**: Research Rank and Faculty Rank emerged as the strongest predictors of high World Rank, indicating that universities investing in faculty quality and research output are more likely to rank higher globally. Education Rank and Employability Rank had a moderate impact, supporting the first hypothesis that research and employability are crucial to global rankings.
* **Regional Growth Patterns**: Developing countries, particularly China, Malaysia, and Iran, showed significant upward trends in university rankings. This growth underscores the investments these countries are making in higher education, allowing them to compete with traditionally high-ranking institutions from developed nations. This trend aligns with the third hypothesis regarding the rise of universities in developing countries.
* **Impact of National Context**: The National Rank metric demonstrated weak correlations with global ranking factors, suggesting that a university’s standing within its own country does not directly impact its global rank. This points to the influence of global metrics over national ones in shaping international perceptions.

**6. Data Sources and Tools**

* **Primary Dataset**: Center for World University Rankings (CWUR) for years 2014-2024, covering a wide array of metrics.
* **Tools Used**:
  + **Python Libraries**: Numpy and Pandas for data manipulation, Matplotlib and Seaborn for visualization.
  + **Jupyter Notebooks**: Separate notebooks were used to clean each year’s data, ensuring flexibility in handling year-specific issues.

**7. Future Directions**

This project opens up several avenues for further exploration:

* **Comparative Analysis of Ranking Systems**: Expanding the analysis to include rankings from other systems, such as THE, QS, and ARWU, would allow a more comprehensive understanding of discrepancies and patterns across ranking methodologies.
* **Economic and Institutional Analysis**: Examining how factors like GDP, education spending, and the public vs. private distinction influence rankings could provide additional insights.
* **Impact of University Types**: Analyzing how public, private, research-focused, and teaching-focused institutions perform differently in rankings could reveal nuanced dynamics within the academic landscape.
* **Regional and Cultural Factors**: Investigating the impact of language, historical context, and cultural factors on rankings could highlight biases or advantages in global academic standings.