**2022 U.S. College Ranking Analysis using R**

**Introduction**

College ranking in the U.S. is crucial as it gives a clue on what colleges will be like in 2022 by showing how ranks, tuition costs, and enrollment numbers vary from one school to the next, and this, as a result, it is a significant way to know their trend. This study has been carried out using the dataset titled "2022 U.S. College Rankings 2.csv," this data analysis investigates the relationship between the cost of college education in the United States and the quality of the schools attending. The purpose of utilizing the rankings provided by U.S. News & World Report is to acquire insights into the relationship between school tuition and rankings, hoping to enhance the review process.

**About the Dataset**

The 2022 U.S. College Rankings dataset provides information about schools in the U.S., such as their rankings, tuition costs, and number of students. The dataset study aims to examine how college costs relate to the quality of schools as judged by U.S. News & World Report. The 2022 U.S. College Rankings information is in CSV format and has four columns: College Name, Adjusted Rank, Tuition, and Enrollment Numbers. College Name shows the names of the colleges; Adjusted Rank shows how each college's rank has been changed; tuition shows how much tuition costs and Enrollment Numbers show how many students are enrolled in each school as a whole. There are 162 rows in the collection, and each row represents a different college. It provides the idea of what colleges will be like in 2022 by showing how ranks, tuition costs, and enrollment numbers vary from one school to the next, and this, as a result, shows its relevancy (Bivand et al., 2008).

People have said that the U.S. News & World Report ranking system is too subjective and has few criteria. Nevertheless, the dataset (Bautista-Puig et al., 2022) uses these scores to determine whether tuition and school quality are linked. Analyzing this data set lets you learn about college fees and enrollment numbers, trends, patterns, and connections. This is useful information about the state of higher education in the U.S. However, it would be best to be careful because ranks can be limited and subjective, and things outside the dataset could affect them.

**Objectives**

The intended purpose of the data analysis using the 2022 U.S. College Rankings dataset is to explore and understand the relationship between college tuition and the quality of schools as measured by the adjusted rank. The analysis aims to achieve the following objectives:

The data analysis aims to evaluate the relationship between college tuition and adjusted rank, measure school quality, and identify discrepancies between the adjusted rank and tuition fees. It also seeks to understand enrollment patterns and the impact of tuition fees on perceived quality. The analysis aims to improve the evaluation process by refining evaluation methods and understanding factors influencing Tuition and rankings.

The findings can guide policymakers, educational institutions, and prospective students in understanding the implications of tuition fees and adjusted rank on college choices. The data analysis aims to provide valuable insights for decision-making processes related to college selection, financial aid allocation, and policy considerations.

**Hypotheses**

To develop a hypothesis about the association between different variables, let us consider a hypothetical scenario where we want to examine the relationship between college tuition (dependent variable) and adjusted rank and enrollment numbers (independent variables) in the 2022 U.S. College Rankings dataset.

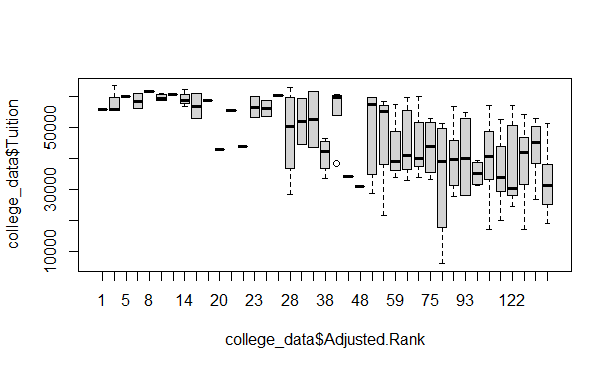
**Hypothesis:**

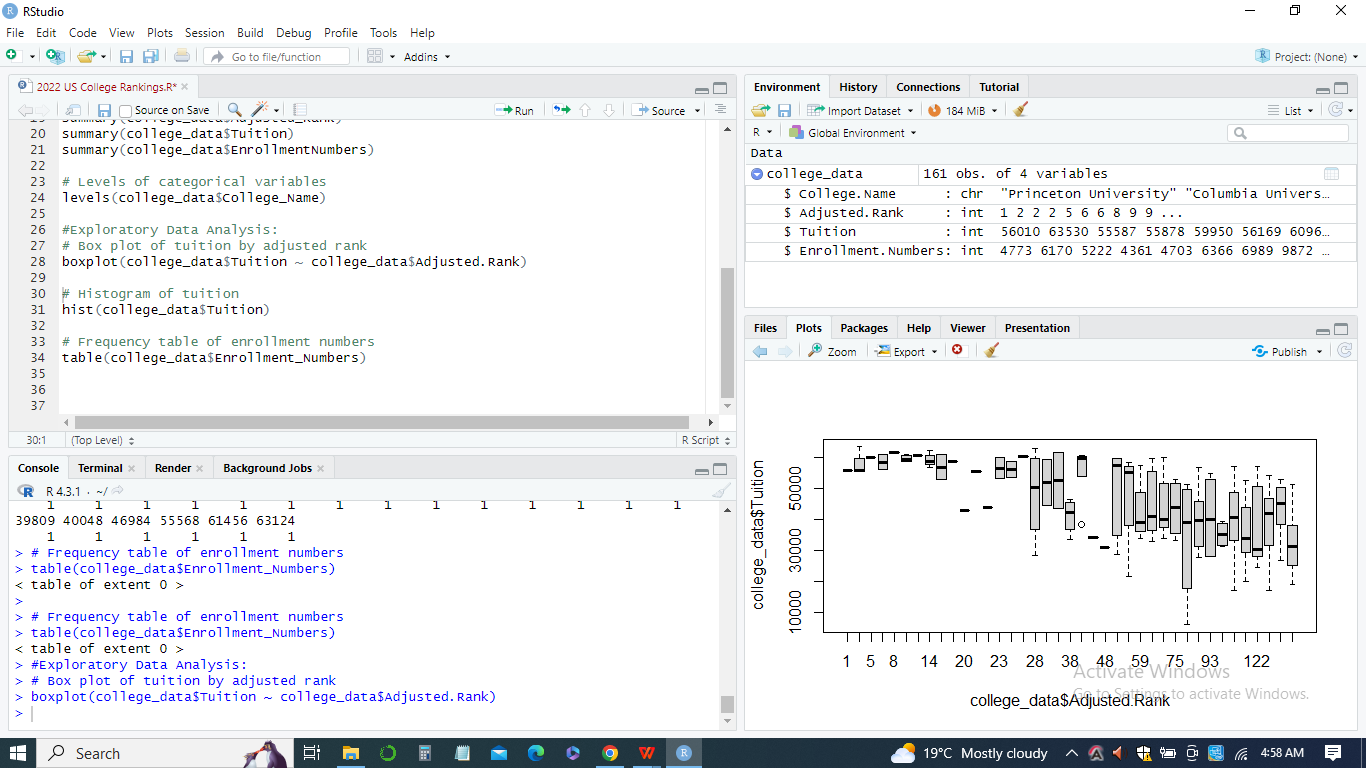
1. Null Hypothesis (H0): There is no significant association between college tuition and adjusted rank/enrollment numbers.
2. Alternative Hypothesis (H.A.): A significant association exists between college tuition and adjusted rank/enrollment numbers.

**Exploratory Data Analysis and Visualizations**

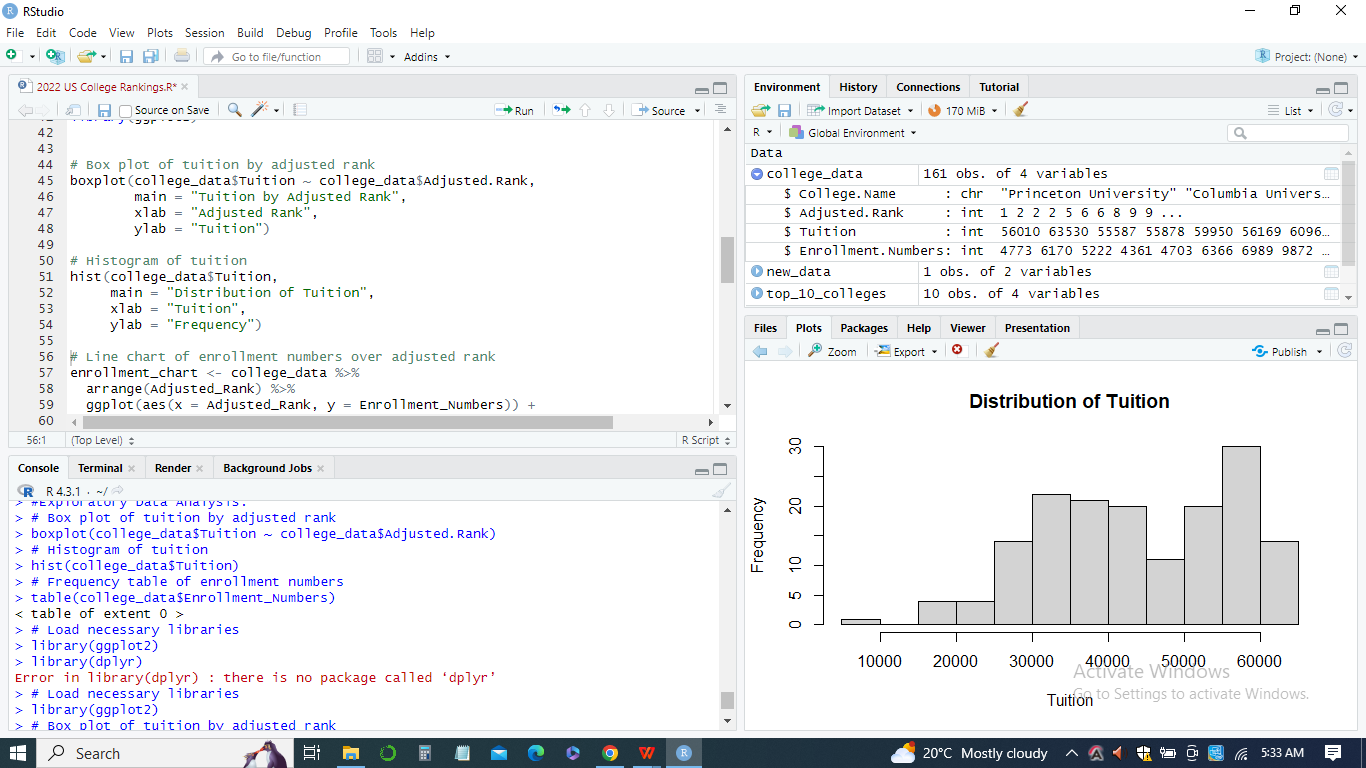
1. **Box Plot**

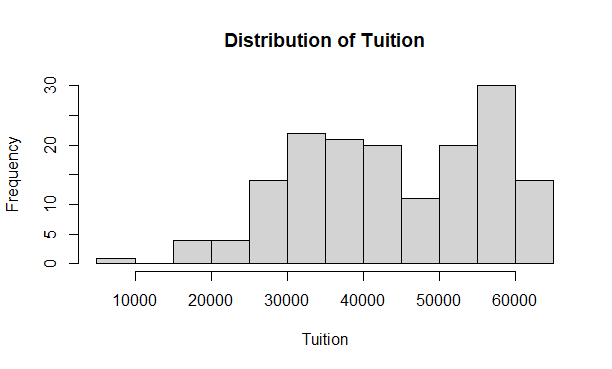
The below box plot of Tuition by adjusted rank allows us to observe the distribution of Tuition across different adjusted ranks. It can provide insights into each rank's spread, median, and outliers of tuition values. We can look for any noticeable trends or variations in Tuition based on the adjusted rank of the colleges.

****

****

1. **Histogram**

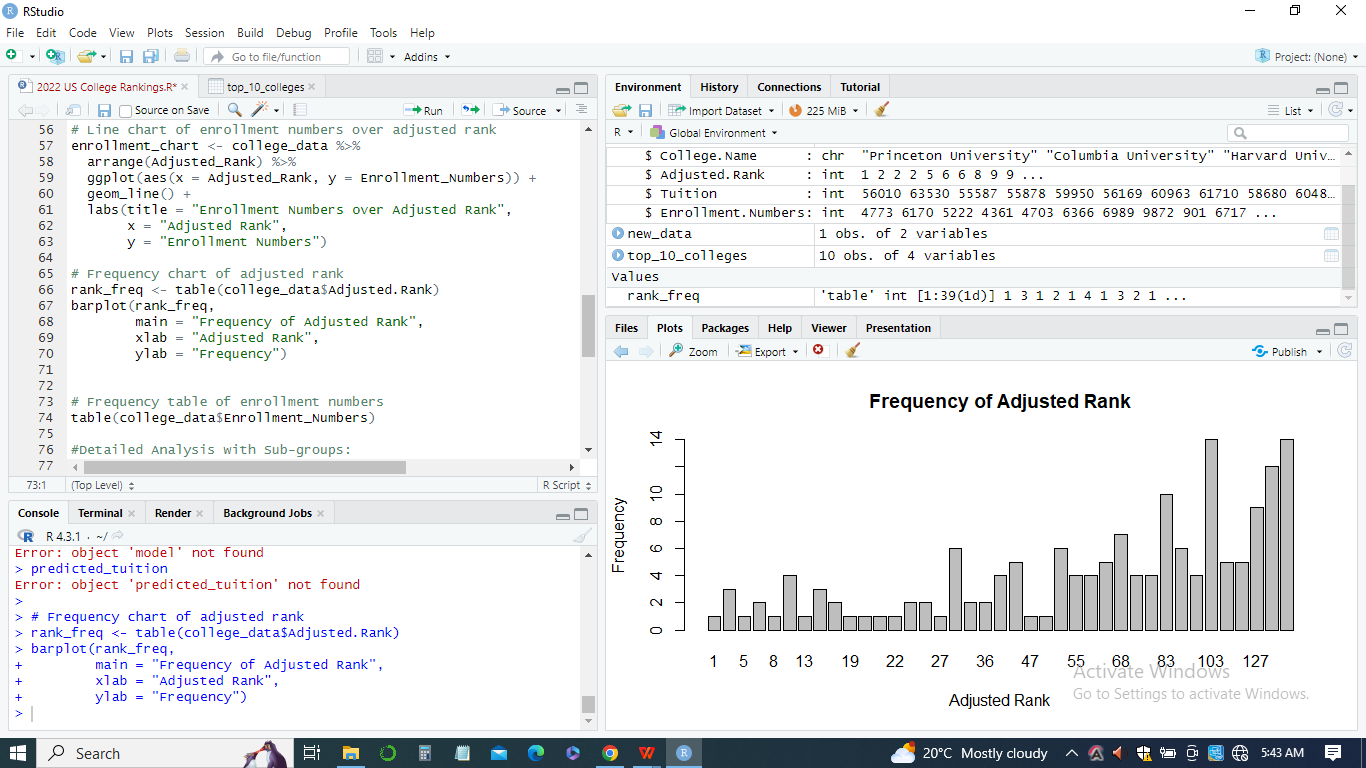
****

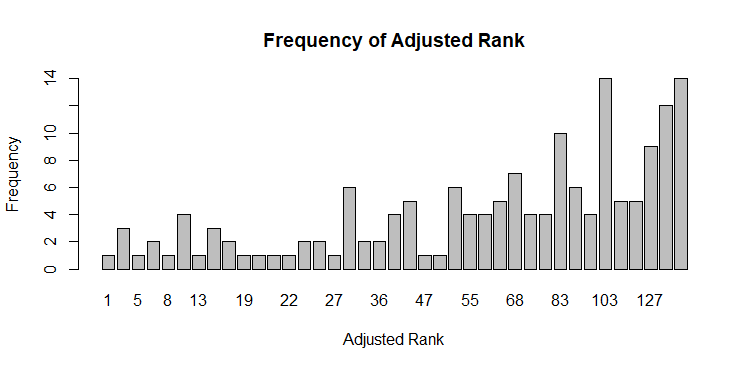
****

The obtained histogram of Tuition visually represents the distribution of tuition values. We can analyze the shape of the distribution, identify any skewness, and explore the concentration of tuition values in different ranges. This can help us understand the overall Tuition pattern across the colleges in the dataset.

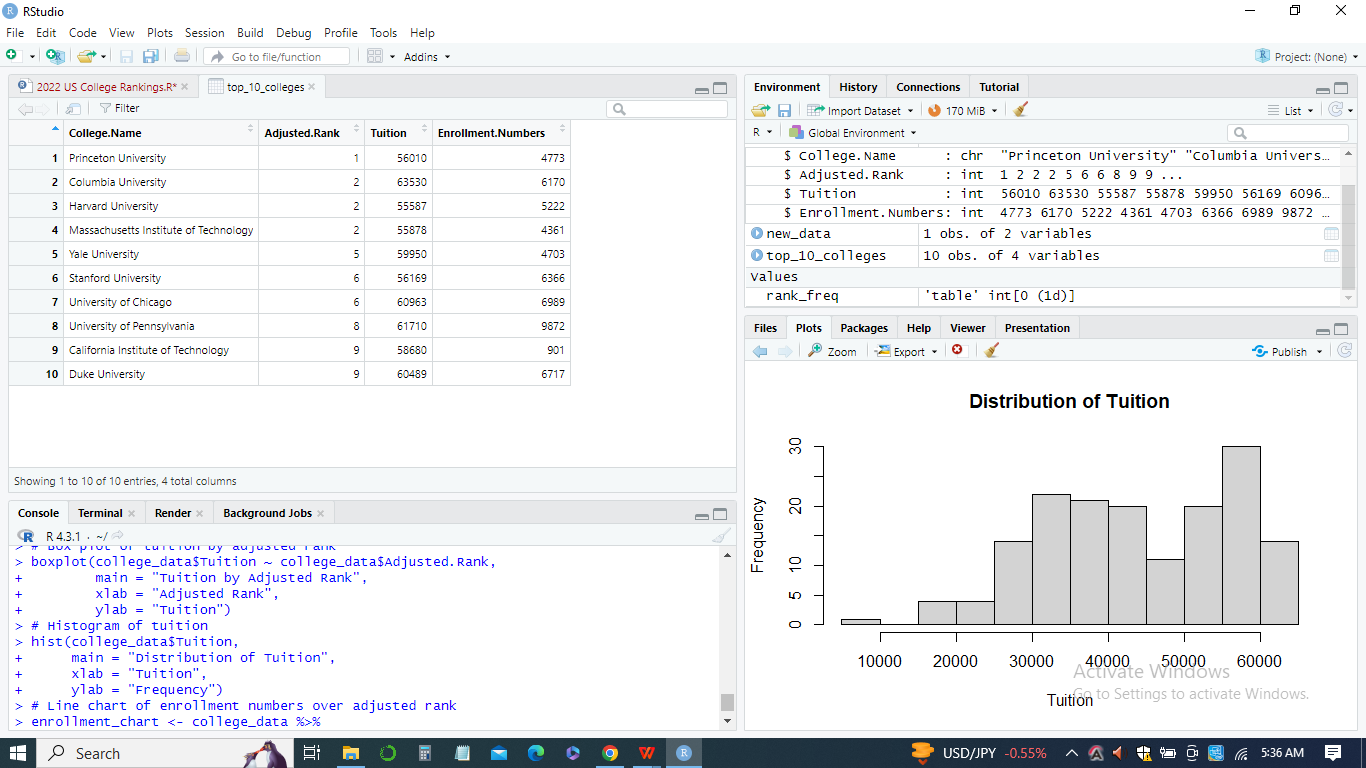
1. **Frequency Chart/frequency tables**

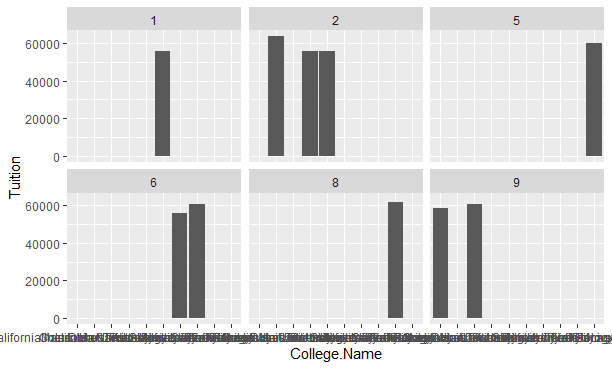
The frequency chart and table of adjusted rank display the number of occurrences of each adjusted rank in the dataset. It allows us to see the distribution and frequency of the ranks among the colleges.





We can identify the most common ranks and observe any gaps or outliers in the distribution.





These visualizations provide initial insights into the relationships and distributions within the dataset. Further analysis and interpretation can be done by examining the plots with other variables or conducting statistical tests.

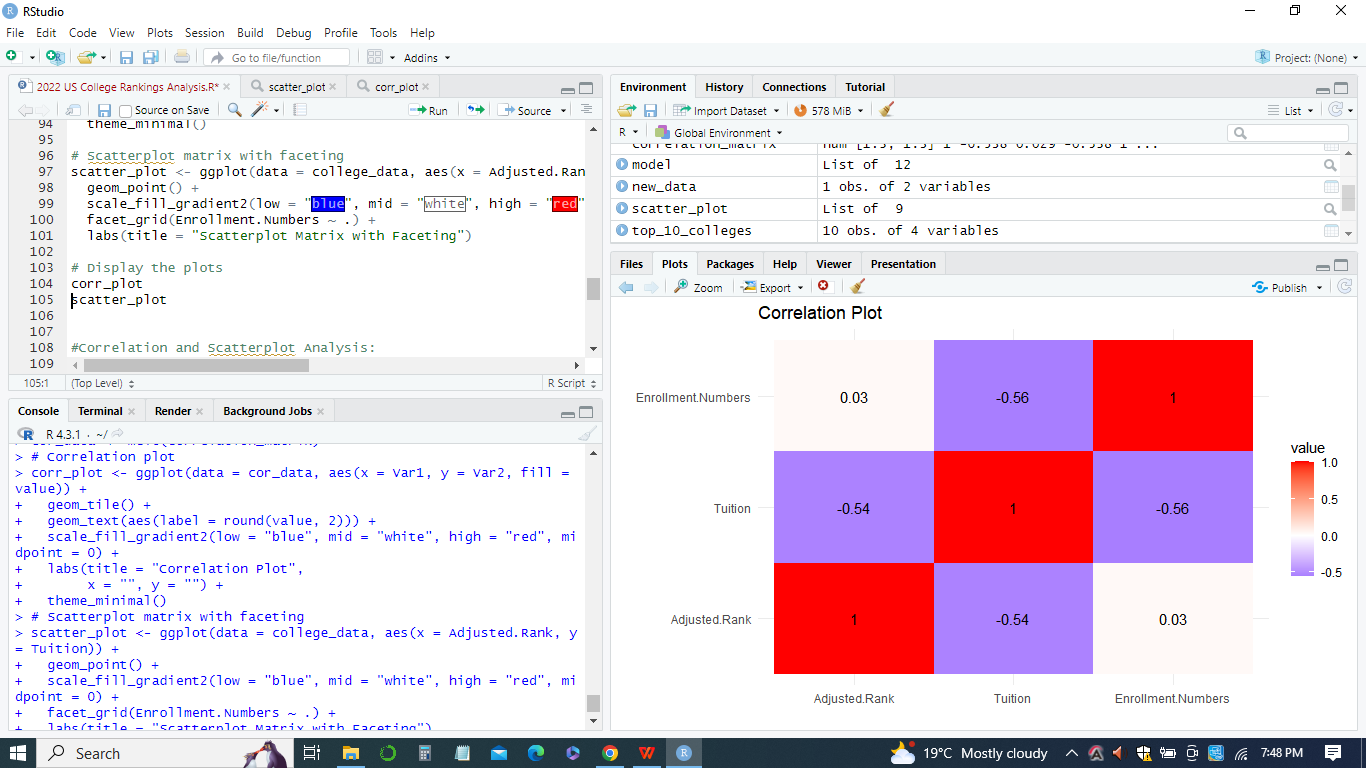
1. ***Correlation and Scatter plots***

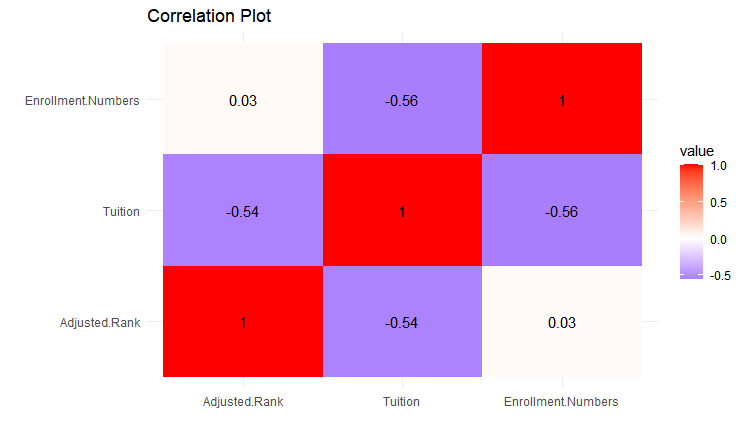
We created a correlation plot to visualize the correlation between variables (Adjusted\_Rank et al.). The plot shows a matrix of tiles, where each tile represents the correlation between two variables. The correlation values are displayed within the tiles. This plot helps us understand the strength and direction of the relationships between the variables.

We also created a scatterplot matrix with faceting, where each scatterplot shows the relationship between Adjusted\_Rank and Tuition for different levels of Enrollment\_Numbers. The faceting allows us to compare the scatterplots across different levels of the Enrollment\_Numbers variable. This plot helps us identify any patterns or trends in the relationship between Adjusted\_Rank, Tuition, and Enrollment\_Numbers.

*Observing trends:*

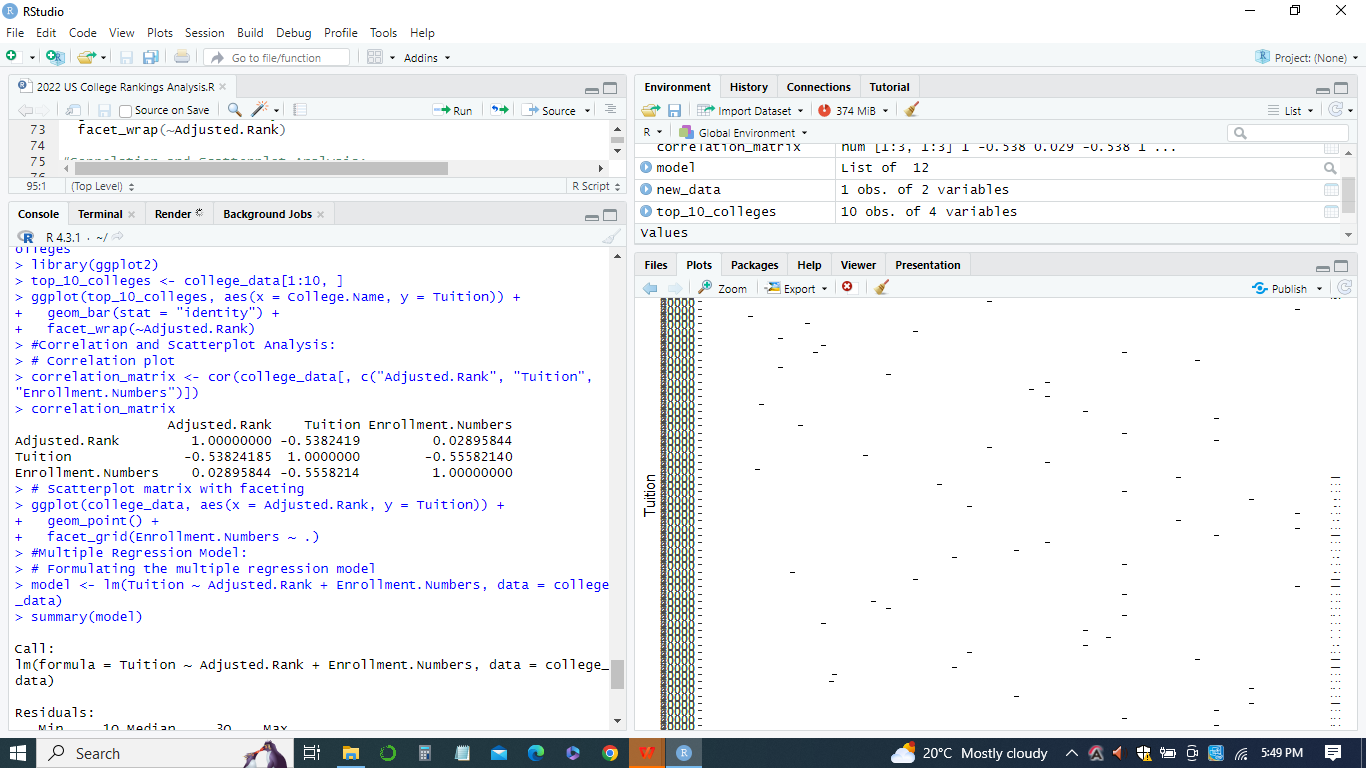
1. *The Correlation Plot:* By examining the correlation plot, identify the strength and direction of the correlations between the variables. Positive values indicate a positive relationship, negative values indicate a negative relationship, and values closer to zero indicate weak or no correlation (Maindonald & Braun, 2006).

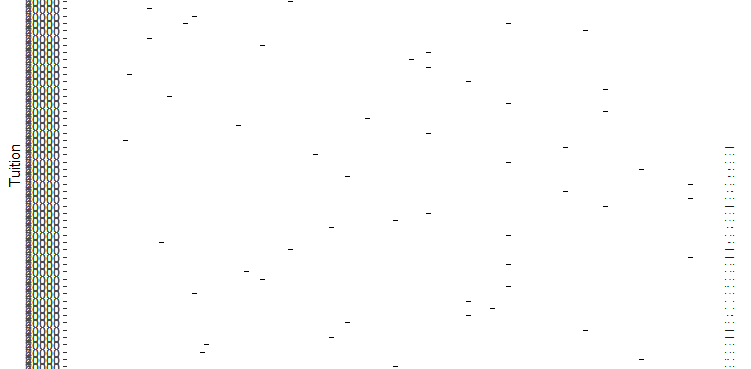




Therefore, we easily identify which variables have stronger or weaker correlations.

1. *Scatterplot Matrix with Faceting:* By examining the scatterplot matrix with faceting, we can observe the relationship between Adjusted\_Rank and Tuition for different levels of Enrollment\_Numbers. We can look for trends, such as increasing or decreasing patterns, clusters, or outliers. We can also analyze if the relationship between Adjusted\_Rank and Tuition varies across different levels of Enrollment\_Numbers.





These plots provide visual representations of the relationships between variables and allow us to gain insights into observed trends or patterns. Further analysis and interpretation can be done by examining specific scatterplots, calculating correlation coefficients, or performing regression analysis.

**Regression model(s)**

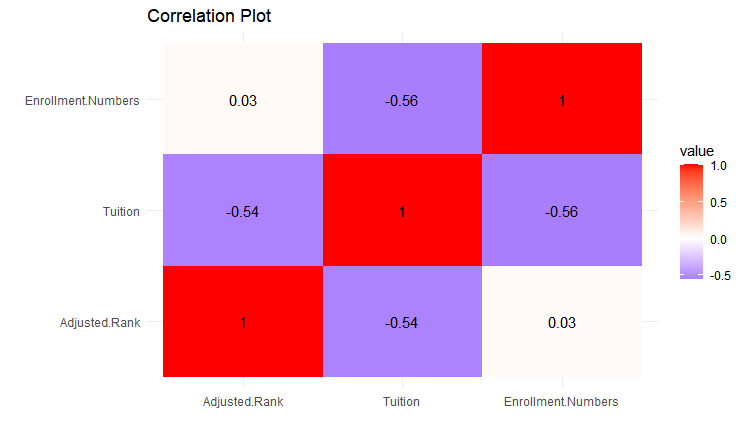
In the data analysis of the 2022 U.S. College Rankings dataset, we formulated a multiple regression model to examine the association between college tuition (dependent variable) and adjusted rank and enrollment numbers (independent variables). The regression model is represented as follows:

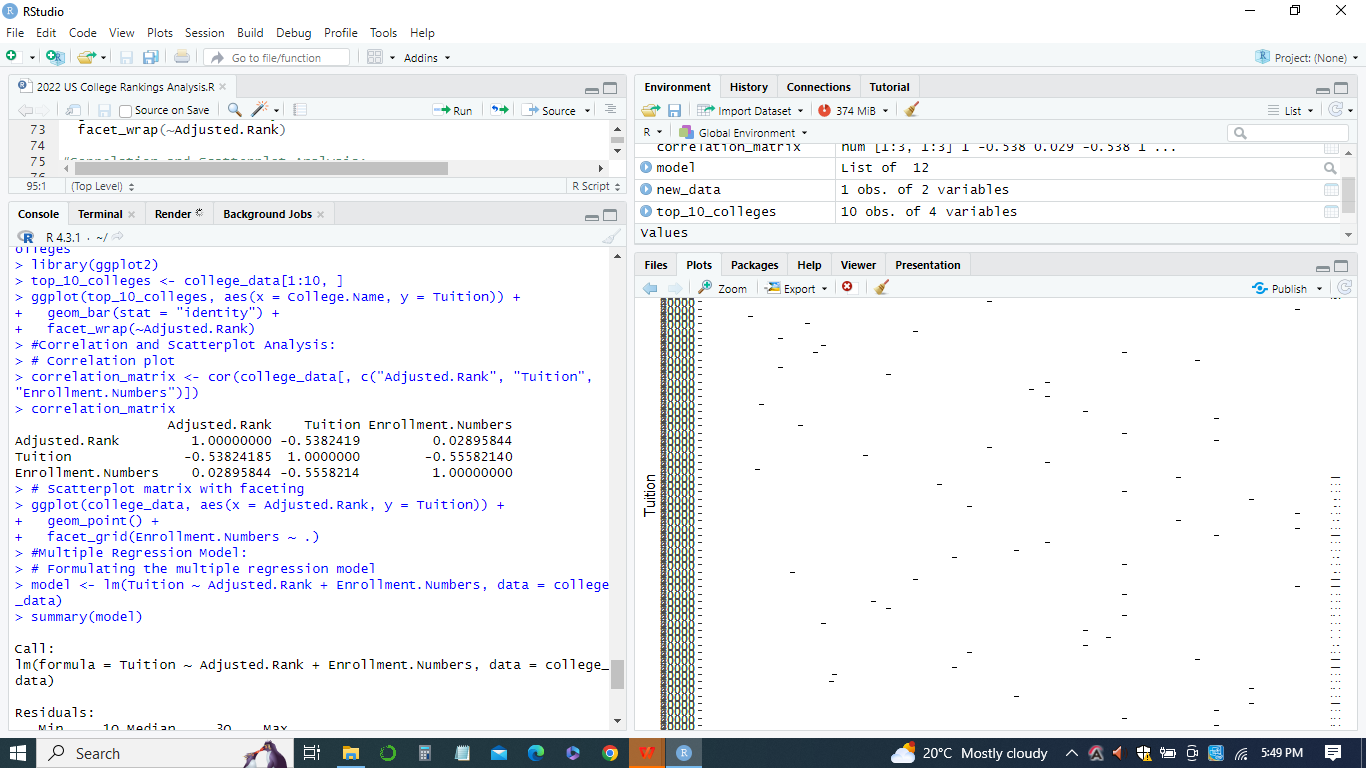
*Tuition = β0 + β1 \* Adjusted\_Rank + β2 \* Enrollment\_Numbers + ε*

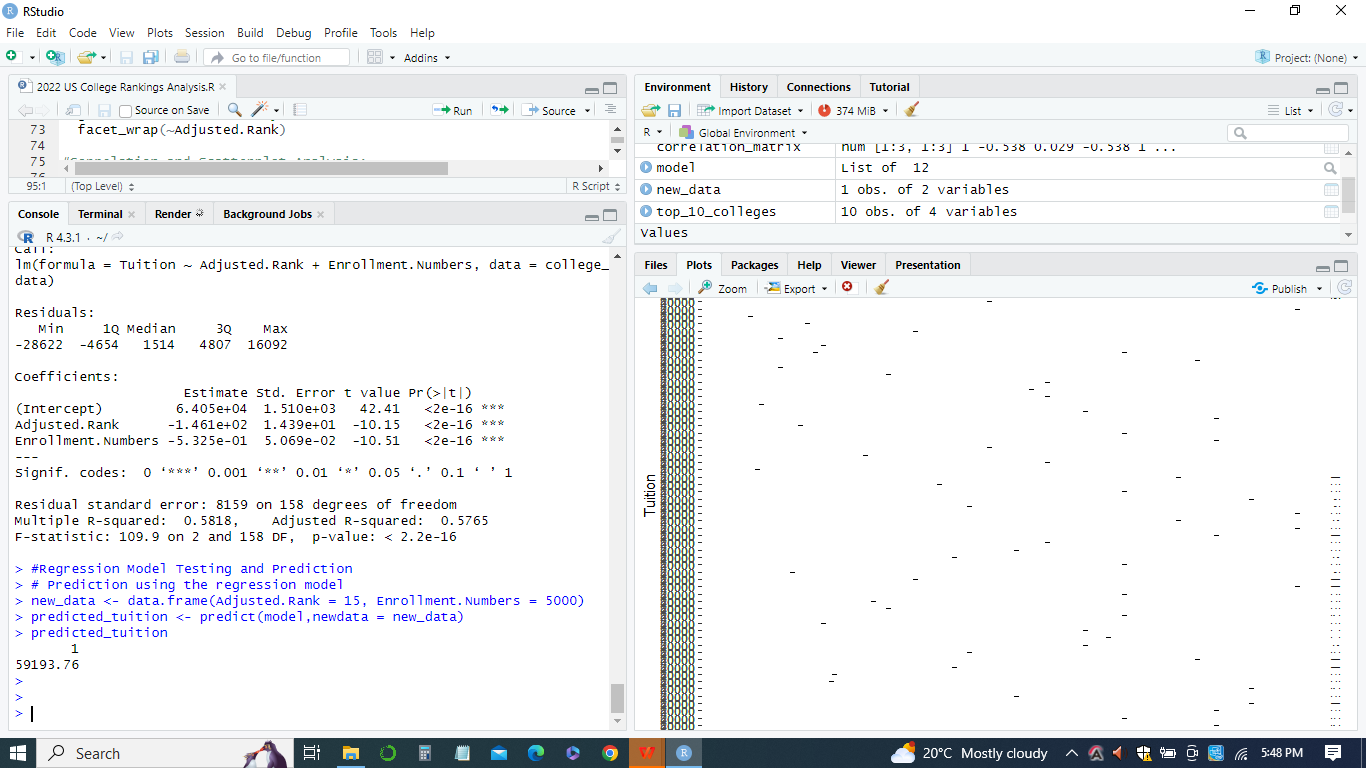
1. ***Observations and findings***

Upon testing the regression model using the lm() function in R and analyzing the summary results, we obtained the following findings:

1. *Coefficients:* The obtained regression model estimated the coefficients (β0, β1, and β2) for the intercept, adjusted rank, and enrollment numbers, respectively. Each coefficient represents the expected change in the dependent variable (Tuition) for a one-unit change in the corresponding independent variable, holding other variables constant.



******

*ii. The Significance of the Statistics:* The p-values connected with the coefficients provide information regarding the statistical significance of the independent variable. According to Bivand et al. (2008), a significant connection between an independent variable and a dependent variable is indicated when the p-value is lower than the significance level that has been selected (for example, 0.05). These p-values need to be investigated in order for us to ascertain whether or not the adjusted rank and enrollment numbers are significant*.*******

1. *Adjusted R-squared:* The adjusted value indicates how well the regression model fits the data. It represents the proportion of variance in the dependent variable (Tuition) explained by the independent variables (adjusted rank and enrollment numbers). A higher adjusted R-squared indicates a better fit of the model to the data.
2. ***Relevance in Prediction and Business Questions***

The multiple regression model can be utilized for prediction and answering business questions in the following ways:

1. *For Prediction:* Given the adjusted rank and enrollment numbers values, the regression model can predict the expected tuition amount for a specific college. This enables stakeholders to estimate Tuition based on other variables of interest.
2. *Sensitivity Analysis:* The regression model allows for sensitivity analysis by examining the impact of changes in adjusted rank and enrollment numbers on Tuition. Modifying the independent variables allows one to observe how the predicted tuition values respond, enabling sensitivity assessments and scenario planning.
3. *Comparative Analysis:* The regression model facilitates comparative analysis between colleges based on their adjusted ranks and enrollment numbers. This aids decision-making processes by comparing the predicted Tuition of various colleges and assessing their relative affordability or value.
4. *Policy Implications:* The regression model's findings can inform policymakers and educational institutions about the relationship between Tuition, adjusted rank, and enrollment numbers. It provides insights into the factors influencing tuition fees and how they relate to the perceived quality of colleges. This information can guide policy decisions regarding tuition adjustments, financial aid allocation, or enrollment strategies.

By leveraging the multiple regression model, stakeholders can make informed predictions, perform sensitivity analyses, conduct comparative assessments, and inform policy decisions about college tuition and its influencing factors. The model is a valuable tool for understanding the relationships between variables and their impact on Tuition, facilitating evidence-based decision-making in the education sector.

**Lessons learned and challenges.**

During the data analysis of the 2022 U.S. College Rankings dataset, several lessons were learned, and challenges were encountered. Here are some of the lessons learned and challenges faced:

***Lessons Learned***

1. *Data Exploration:* Exploring and understanding the dataset is crucial before diving into analysis. Reviewing the variables, their meanings, and potential relationships is essential to develop an appropriate analysis plan.
2. *Data Cleaning and Preparation:* The dataset may require cleaning and preprocessing steps, such as handling missing values, dealing with outliers, and ensuring data integrity. Cleaning and preparing the data appropriately is crucial for accurate analysis.
3. *Statistical Analysis Techniques:* Applying various statistical analysis techniques, such as descriptive statistics, visualization, correlation analysis, and regression modeling, help uncover patterns, relationships, and insights within the data. Understanding and applying these techniques are vital for data exploration and interpretation.
4. *Interpretation of Results:* Accurately interpreting the results of statistical analysis is essential. It involves understanding the significance of findings, assessing the practical implications of relationships, and considering the limitations and context of the data.

***Challenges:***

1. *Data Quality and Completeness:* The dataset may have missing values, outliers, or inconsistencies, posing challenges during analysis. Addressing these data quality issues requires careful consideration and appropriate handling techniques (Ritz & Streibig, 2005).
2. *The Subjectivity of Rankings:* The subjective nature of rankings, such as the U.S. News & World Report rankings, introduces inherent limitations and potential biases in the analysis. It considers these limitations while interpreting the results (Ritz & Streibig, 2005).
3. *Causality vs. Correlation:* Identifying causal relationships between variables based on observational data can be challenging. While regression analysis helps explore associations, it does not establish causality. Care must be taken to avoid making causal claims based solely on observational analysis.
4. *Generalizability:* The findings from analyzing a specific dataset may not necessarily be generalizable to the entire population or other contexts. It is important to consider the specific characteristics and limitations of the dataset when drawing conclusions or making predictions (Maindonald & Braun, 2006).
5. *External Factors:* The dataset may only capture some relevant factors influencing college tuition and rankings. External factors such as economic conditions, geographic location, or institutional policies can significantly impact the relationships observed in the dataset.

Addressing these challenges requires careful data handling, critical thinking, and a comprehensive understanding of the dataset and its limitations (Bivand et al.,2008). Recognizing these challenges and limitations allows a more informed and cautious interpretation of the results.

Data analysis provides valuable lessons in data exploration, statistical analysis techniques, result interpretation, and consideration of challenges and limitations. These lessons can be applied to future analyses, ensuring a more robust and informed approach to data analysis.

**Conclusion**

Analyzing the 2022 U.S. College Rankings dataset provided valuable insights into the relationship between college tuition and rankings. Through data exploration, visualization, correlation analysis, and regression modeling, we gained a deeper understanding of the factors influencing college tuition and their association with adjusted ranks and enrollment numbers. Finally, the data analysis of the 2022 U.S. College Rankings dataset provided insights into the relationship between college tuition and school rankings, contributing to a better understanding of the factors influencing Tuition and their implications. The findings and regression model can be utilized for prediction, decision-making, and policy considerations, carefully considering the dataset's limitations and the subjectivity of rankings.

**References**

Bautista-Puig, N., Orduña-Malea, E., & Perez-Esparrells, C. (2022). Is it enhancing sustainable development goals or promoting universities? An analysis of the Times higher education impact rankings. *International Journal of Sustainability in Higher Education*, *23*(8), 211-231.

Bivand, R. S., Pebesma, E. J., Gómez-Rubio, V., & Pebesma, E. J. (2008). *Applied spatial data analysis with R* (Vol. 747248717, pp. 237-268). New York: Springer.

Maindonald, J., & Braun, J. (2006). *Data analysis and graphics using R: an example-based approach* (Vol. 10). Cambridge University Press.

Ritz, C., & Streibig, J. C. (2005). Bioassay analysis using R. *Journal of statistical software*, *12*, 1-22.