Veda Pick, Riley Plummer,

Claire Pfutzenreuter & Jack Lemke

**Data Wrangling Project Report**

**Happiness**

1. **Introduction**

A recent study found that 88% of college students reported their school life to be stressful (American Addiction Centers). These stress levels often stem from different factors in their lives like grades, involvement, living situation, etc. that then affect their satisfaction with their college experience. Our goal for this project is to use the responses from a survey about the ideal student life to determine which factors lead to a satisfactory student experience and whether there is a link between stressors and life satisfaction.

Looking beyond the scope of college students' life satisfaction, we sought to explore how happy people around the world are with their quality of life. The Happiness Report is a survey completed on a global scale that measures the happiness score of individuals within each country.

1. **Data**

This project uses three sources of data: a dataset of survey responses with information about student satisfaction and stress, a webpage containing the happiness rankings of each major, and Wikipedia information about happiness levels from different countries around the world. Our research questions include, but are not limited to:

* What affect does major choice have on student life satisfaction?
* Are student stress levels correlated with lower student life satisfaction?
* What are the leading factors of increased stress among students?
* Does a higher overall life satisfaction within a country relate to a higher student life satisfaction?

*2.1 Student Life Satisfaction Survey*

The first dataset we will use in this project is an existing dataset from Kaggle[[1]](#footnote-4377). It is a combination of two similar datasets that include questions and responses from students regarding their ideal student life and factors that affect their stress levels. There are 2958 total students that responded to multiple survey items and questions. The link included a zip file of a “responses” dataset and a “questions” dataset. We used most of the “responses” data and added some data from the “questions” dataset. If we felt as though a column had insignificant data, it was removed/not included. Some columns are simple “yes” or “no” responses and others are numeric or text. The dataset includes information about stress levels activities, involvement, and more.

*2.2 Major Rankings*

The second dataset is scraped from a webpage[[2]](#footnote-12957) containing the ranking of each major based on student satisfaction. From this site, we scraped each major and it’s corresponding ranking out of 10.

We then merged this data frame of majors and rankings with the original data frame of survey responses. Since the two data frames did not have any columns in common, we added a new column in the original survey data frame converting the department (current study area) of each student to the related major matching the new data frame. To make sure the majors in both data frames matched, we renamed the names of each major to abbreviations. Once the names of each major in both data frames were identical, we were able to horizontally merge the data and create a new column called “Major\_Rank” in the survey data frame.

*2.3 Nationality Rankings and Score*

Our Third dataset is scraped from the Wikipedia page titled “World Happiness Report 2019[[3]](#footnote-17395),” a publication from the United Nations Sustainable Development Network. The data consists of rankings of national happiness based on individual's ratings of their own life. Various quality of life measures are also included in the rankings. We then wrote a crawling script to retrieve information about each country and put it into a data frame, then cleaned the data frame to contain only the country and each country’s happiness ranking and score to correspond with our original merged dataset.

To merge the nationality data with the survey responses and major rankings, we renamed the country column to “Nationality.” Once the column names matched, we were able to horizontally merge the data, creating two new columns: “Nationality\_Ranking” and “Nationality\_Score”.

*2.4 Final Merged Dataset*

After merging all three datasets, 4 variables were added to the original dataset creating a final data frame of 25 variables and 2912 rows. Table 1 contains a description of each variable and BAIS3250HappinessFinalProject.R contains the code for scraping, cleaning, and merging the data.

1 <https://www.kaggle.com/shivamb/ideal-student-life-survey?select=survey_questions_meta.csv1>

2<https://www.topuniversities.com/university-rankings-articles/university-subject-rankings/which-subjects-have-happiest-students>

3 <https://en.wikipedia.org/wiki/World_Happiness_Report>

*Table 1 Data Dictionary*

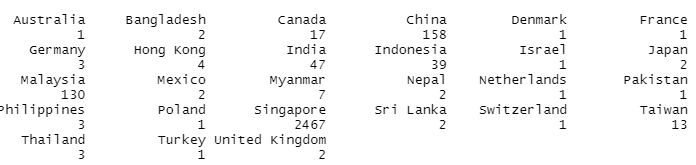
|  |  |  |
| --- | --- | --- |
| **Column** | **Type** | **Description** |
| Response\_ID | text | Unique ID for each student |
| Career | text | Graduate or Undergraduate |
| Citizenship | text | Student citizenship status |
| Nationality | text | Student nationality |
| Year\_of\_Study | numeric | The year of which the students are currently in university |
| Gender | text | Student Gender (Male/Female) |
| Department | text | Current student area or department of study |
| Housing\_Type | text | The type of housing the student resides in |
| Events\_Volunteered | numeric | The number of events the student has volunteered for |
| Events\_Participated | numeric | The number of events the student has participated in |
| Activities\_Interested | numeric | The number of activities the student is interested in |
| Activities\_Passionate | numeric | The number of activities the student is passionate about |
| Stress\_Level | numeric | The levels of stress the students feel (0-9) |
| Student\_Life\_Satisfaction | numeric | How satisfied students are with their student life (0-9) |
| Effort\_For\_Interaction | numeric | The amount of effort students put forth to interact with others (0-3) |
| Societies\_Participation | text | Student participation in society or interest group (Yes/No) |
| Clubs\_Participation | text | Student participation in any clubs (Yes/No) |
| Academic\_Stress | text | Student academic stress (Yes/No) (1/0) |
| Financial\_Stress | text | Student financial stress (Yes/No) (1/0) |
| Family\_Stress | text | Student family stress (Yes/No) (1/0) |
| Friend\_Stress | text | Student friendship stress (Yes/No) (1/0) |
| Major | text | Student’s major |
| Major\_Rank | numeric | The happiness ranking of each major (1-10) |
| Nationality\_Rank | numeric | The happiness ranking of each country (1-140) |
| Nationality\_Score | numeric | The happiness score of each country out of 10 |

1. **Analysis**

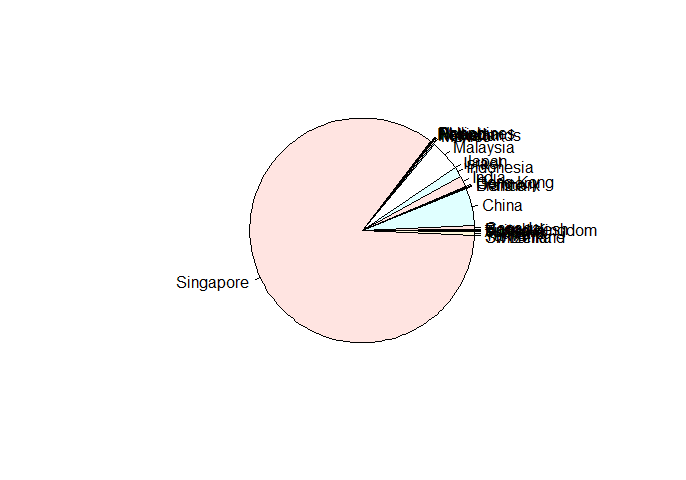
The goal of this project is to determine what factors may determine the stress levels, happiness levels, and life satisfaction of individuals across many nationalities. We initially conducted several visualizations and summary statistics to understand the distributions within our dataset variables. These visualizations allowed us to prepare for further analysis through several tests and statistical methodology.

*3.1 Observing Survey Respondent Nationalities*

What is the proportion of nationalities among our survey respondents? This is an important question to explore so we can know where our data is coming from and deal with any bias in our overall results from the data. After calculating a count table and proportion table, we found that many of our survey responses came from Singapore, along with a decent amount from China and Malaysia.



*Figure 1 : Proportion of survey responses by nationality*



*Figure 2: Pie chart representing survey respondents by nationality*

*3.2 Student Life Satisfaction Summary Statistics*

What is the distribution among our respondents when stating their Student Life Satisfaction? When generating summary statistics about Student Life Satisfaction, we found the following:

*Table 2 Student Life Satisfaction Summary*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | ***Min*** | ***Median*** | ***Mean*** | ***Max*** |
| *Student\_Life\_Satisfaction* | 0 | 6 | 5.777 | 9 |

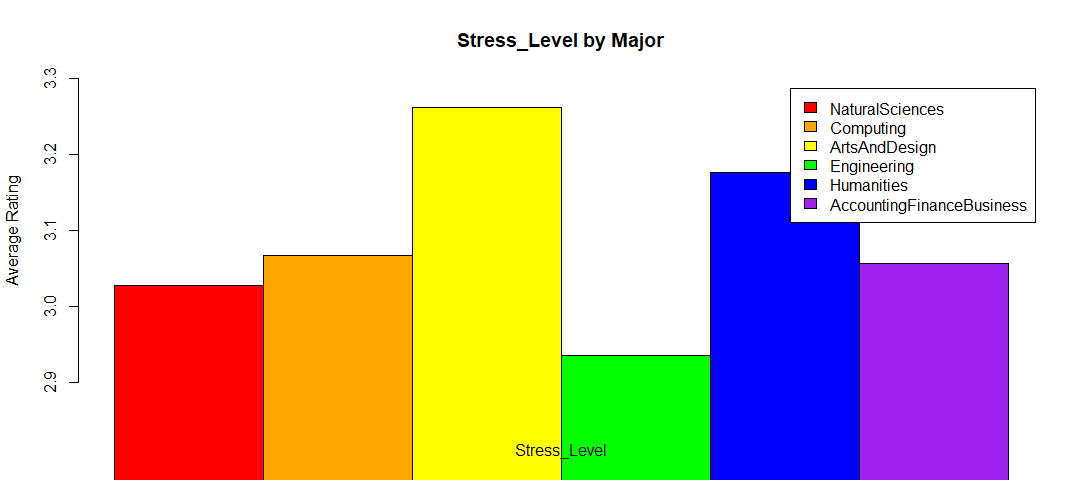
The average rating that respondents gave was a 5.777 out of 9. It is important to note that we rescaled this variable from its original state. The original scale was 0-3, but in order to relate it more accurately to our merged dataset, we transformed it to a level of 0-9 to better relate to the scale of country-by-country happiness we scraped (Nationality\_Score).

*3.3 Stress Levels and Student Life Satisfaction by Major*

Which majors have the highest stress levels and student life satisfaction rankings for each major? To answer this question, we created a summary table that calculates the Min, Median, Mean, and Max, and total amount of students. *Tables 2 and 3* display the resulting summary table and *Figures 3 and 4* display the resulting bar plots for the means.

*Table 3 Major Stress Level Summary*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ***Major*** | ***Min*** | ***Median*** | ***Mean*** | ***Max*** | ***Total Students*** |
| *Natural Sciences* | *0* | *3* | *3.027* | *9* | *808* |
| *Computing* | *1* | *3* | *3.067* | *9* | *135* |
| *Arts & Design* | *1* | *3* | *3.263* | *9* | *906* |
| *Engineering* | *0* | *3* | *2.935* | *9* | *573* |
| *Humanities* | *0* | *3* | *3.177* | *9* | *113* |
| *Accounting, Finance, & Business* | *1* | *3* | *3.057* | *9* | *369* |

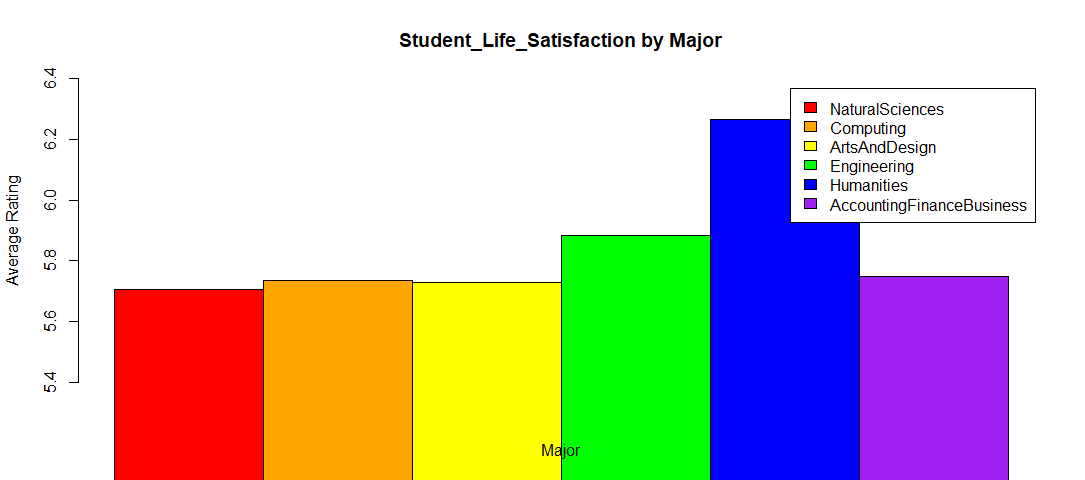


*Figure 3: Bar chart of stress level by major*

Our analysis found that Arts & Design majors have the highest stress levels among respondents, while Engineering majors have the lowest stress levels.

*Table 4 Major Student Life Satisfaction Summary*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ***Major*** | ***Min*** | ***Median*** | ***Mean*** | ***Max*** | ***Total Students*** |
| *Natural Sciences* | *0* | *6* | *5.707* | *9* | *808* |
| *Computing* | *0* | *6* | *5.733* | *9* | *135* |
| *Arts & Design* | *0* | *6* | *5.728* | *9* | *906* |
| *Engineering* | *0* | *6* | *5.885* | *9* | *573* |
| *Humanities* | *0* | *6* | *6.265* | *9* | *113* |
| *Accounting, Finance, & Business* | *1* | *6* | *5.748* | *9* | *369* |



*Figure 4: Bar chart of happiness rating by major*

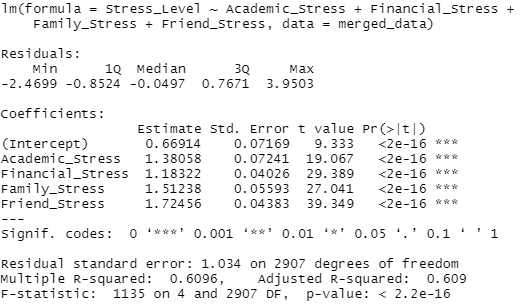
Our analysis found that Humanities majors have the highest student life satisfaction among respondents, while Natural Sciences majors have the lowest student life satisfaction.

3.4 *Are student stress levels correlated to student life satisfaction? What factors increase or decrease overall stress levels?*

When determining which factors within the dataset correlate with a lower or higher Student Life Satisfaction, we performed several tests among the variables present within the data. We found the correlation between Stress Level and Student Life Satisfaction to be –0.1167. This allowed us to conclude that respondent stress levels did not always correlate directly with the score they may have given for their overall student life satisfaction.

However, when breaking down each stressor and seeing their effects on overall stress levels, we found high significance. For Academic Stress, Financial Stress, Family Stress, and Friend Stress, survey respondents were prompted with a “Yes” or “No” option as a response. In order to perform a linear regression with our dataset, we needed to transform these character responses to a binary scale of 1 and 0.

Our linear model found that when Academic Stress is 1 (I.e., students answered “Yes”) in the dataset, overall Stress Level is increased by 1.38058. When Financial Stress is 1, overall Stress Level is increased by 1.18322. Family Stress and Friend Stress had the same effect on Stress Level at 1.51238 and 1.72456 respectively.



*Figure 5: Linear regression of types of stress in comparison to stress level*

3.5 *Is Student Life Satisfaction normally distributed?*

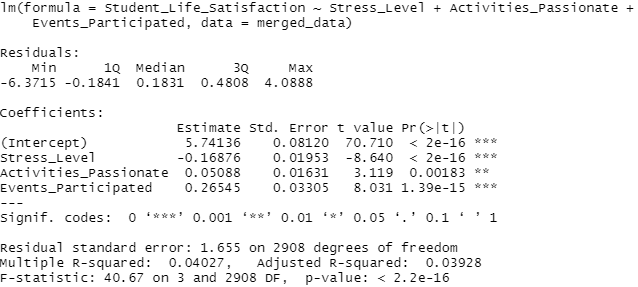
In order to see if Student Life Satisfaction is normally distributed within the data, we conducted a Shapiro-Wilk normality test and also visualized a histogram of scores given. The null hypothesis within our test was that the data was normally distributed among Student Life Satisfaction. Our p-value came out to 2.2e-16, which allowed us to reject the null hypothesis and assume that the data is not normally distributed. In order to visualize this conclusion, we created a histogram of the results:



*Figure 6: Histogram of student life satisfaction*

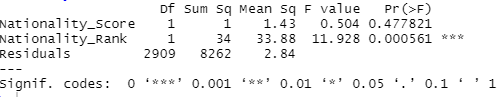
3.6 *Is Student Life Satisfaction related to student Stress Level and the number of activities the student is involved in?*

We ran a linear regression in order to determine how the student’s stress level, the number of activities the student is passionate about, and the events they participated in affect student life satisfaction. An important part of the figure below is the p-value. We can see that the p-value is extremely low, meaning that we can reject the null hypothesis and can confirm that these three factors influence student life satisfaction. Another important section of the figure is the “Estimate” column. Every increase in 1 in the coefficient affects student life satisfaction by the number in this “Estimate” column. For example, every increase of 1 in stress level decreases student life satisfaction by 0.16876. This shows that stress level negatively affects student life satisfaction while activities the student is passionate about and the amount of events they participate in positively affect student life satisfaction.



*Figure 7: Linear regression comparing student life satisfaction with activity involvement*

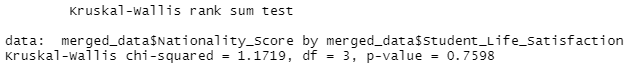
3.7 *Using ANOVA to determine if Nationality\_Score and Nationality Rank are related to Student Life Satisfaction*



*Figure 8: ANOVA testing the relation of nationality and nationality rank to student life satisfaction*

When using ANOVA, it is important to consider the p-values of the variables you are relating to the independent variable (in our case Student Life Satisfaction). We utilized a two-way Anova test and found the p-value for Nationality Rank is statistically relevant to the target variable since it is under .05. (.000561). Also, the larger the F-Value for a variable, the higher likelihood that the variation caused by the independent variable is real and relevant, and Nationality Rank has a high F-score. This shows that, although Nationality Score may not be directly related to Student Life Satisfaction, Nationality Rank may be.

We also conducted a one-way Kruskal-Wallis test with Nationality Score and Student Life Satisfaction, which has a null hypothesis of the mean ranks of the groups being the same. Our results below lead us to reject the null hypothesis that the average ranks between these two variables are the same.



*Figure 9: Comparing nationality score and student life satisfaction means*

We also concluded with a chi-squared test relating Nationality Rank and Student Life Satisfaction, along with a Shapiro test on Nationality Score to check for normal distribution. These additional tests and their analysis can be found in our .R script file.

1. **Conclusion**

This project was a combination of data utilized from Kaggle, that provided information about student satisfaction and stress with data pertaining to happiness levels from different countries around the world. We also incorporated a third data set into our project. This data set originated from a webpage that gave us the happiness rankings of each university major.

We performed multiple analyses to determine what factors contribute to a student's stress levels and happiness. We started with grouping our data by nationality and major to get a better understanding of the survey respondents. Using pie charts, summary tables, bar plots, linear regression, Shapiro-Wilk normality test, ANOVA, and a Kruskal-Wallis test, we found evidence to suggest that relationships are present between many of the variables in our dataset.

Our linear models found that academic, financial, family, and friend related stress all have a significant effect on a student’s stress levels. We found that friend-related stress had the largest effect on students and the amount of stress they had. We also found that stress levels negatively affect student life satisfaction, while activities the student is passionate about and the amount of events they participate in positively affect student life satisfaction. Our Shapiro-Wilk normality test found student life satisfaction to not be normally distributed within the data. When using ANOVA, we concluded that the nationality rank is more related to student life satisfaction than nationality score is. Finally, our Kruskal-Wallis test showed that the averages for nationality score and student life satisfaction were not the same and they had a significant difference.

We recognize the various shortcomings of our project. First, the respondent’s countries in our survey dataset only represent a small portion of countries around the world. As we saw in our first pie chart, the data is strongly overpowered by just a few countries. If we wanted to gear our findings more towards the United States happiness reports, we would need to gather more responses from that area to balance out the data. Also, having data on more majors with potentially specific concentrations would improve the accuracy and complexity of our results. In order to analyze the majors within our current dataset, we had to group majors into a few broad groups. We could analyze the same data without grouping to attain more specific and accurate results. Finally, when we conducted a Shapiro-Wilk normality test on student life satisfaction and the resulting p-value was very low. However, after we built the histogram, the distribution appeared as “normal-looking”. We think maybe this had to do with rescaling the data to match the data sets.

**Works Cited**

American Addiction Centers. “Coping with College Stress; Exploring Perceived Stress and Coping Methods in College.” Resources for Addiction Treatment & Recovery, American Addiction Centers, <https://americanaddictioncenters.org/learn/college-coping-mechanisms/>.

1. <https://www.kaggle.com/shivamb/ideal-student-life-survey?select=survey_questions_meta.csv> [↑](#footnote-ref-4377)
2. <https://www.topuniversities.com/university-rankings-articles/university-subject-rankings/which-subjects-have-happiest-students> [↑](#footnote-ref-12957)
3. <https://en.wikipedia.org/wiki/World_Happiness_Report> [↑](#footnote-ref-17395)