The Relationship between Hobbies and Well-Being

By,
Omkar Prashant Juvatkar
and
Pritesh Das

DASC 5301-001: Data Science
Dr Marnim Galib

05 December 2025

Abstract:

In this project, based on the given dataset, we try to answer various scientific research questions along with the statistical models that are to be used. Firstly, we try to do Linear regression by checking the relationship between variables such as the active hobbies of students (consisting of hobbies like arts, music, sports) and their GPAs. Under linear regression, we check whether or not students with active hobbies have better GPAs than those who don’t have any active hobbies. Secondly, we make an attempt to predict academic performance (GPA) based on the combination of hobby type, study habits, and personality types by making use of Multiple Linear Regression. Additionally, we check for the existence of significant patterns or clusters among students’ preferences (such as beverage/cuisine, media/hobbies, along with sleep patterns) related to academic and social outcomes. Furthermore, we assess whether students identifying as extroverts are more likely to report multiple hobbies as compared to introverts.

Introduction

We found out that the given dataset contains various details of students (such as age, height, weight, study habits, number of hours studied, cuisine and beverage preferences, and number of hours slept). There are a total of 68 variables or columns. The purpose of the project is to determine if relationships, patterns, correlations, and trends exist in the dataset that can be used for a wide variety of applications, such as marketing, counseling, market research analysis, and policy making, among others. As always, with all large datasets, they have to be first cleaned, after which exploratory data analysis is performed so that we can apply various algorithms to analyze our data and make the required inferences.

Problem Statement

To analyze the given dataset while applying statistical models and come up with interpretations of these models to better understand the inferences derived from these models and help out relevant stakeholders with their operations.

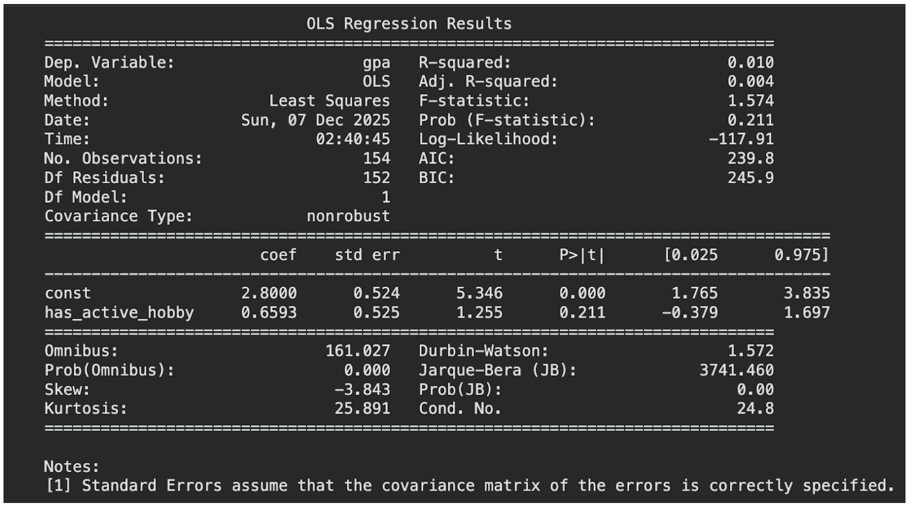
Proposed Methodology:

We make use of various statistical models to answer the following questions:

1. **Linear Regression**
2. **Multiple Linear Regression**
3. **K-Means Clustering**:
4. **Logistic Regression Classification**

Analysis and Results

1. Linear Regression:



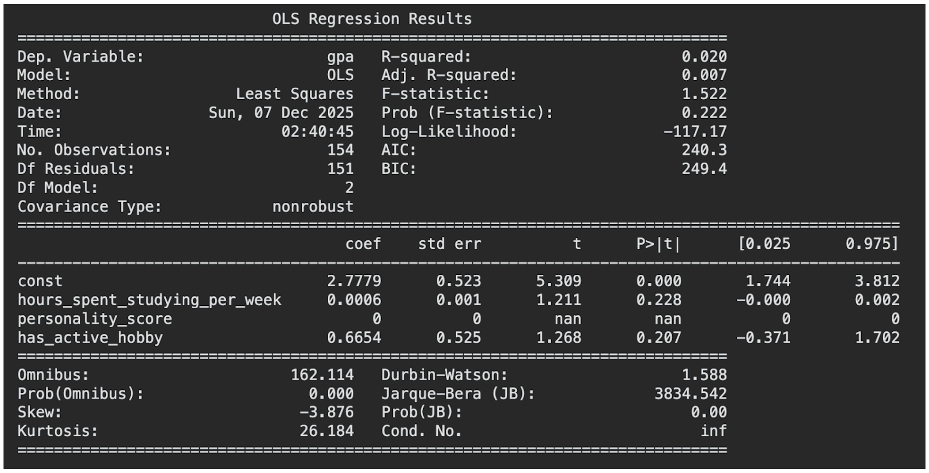
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    gpa   R-squared:                       0.010
Model:                            OLS   Adj. R-squared:                  0.004
Method:                 Least Squares   F-statistic:                     1.574
Date:                Sun, 07 Dec 2025   Prob (F-statistic):              0.211
Time:                        02:40:45   Log-Likelihood:                -117.91
No. Observations:                 154   AIC:                             239.8
Df Residuals:                     152   BIC:                             245.9
Df Model:                           1
Covariance Type:            nonrobust
====================================================================================
                       coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------
const                2.8000      0.524      5.346      0.000       1.765       3.835
has_active_hobby     0.6593      0.525      1.255      0.211      -0.379       1.697
==============================================================================
Omnibus:                      161.027   Durbin-Watson:                   1.572
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             3741.460
Skew:                          -3.843   Prob(JB):                         0.00
Kurtosis:                      25.891   Cond. No.                         24.8
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Image 1: Linear Regression Results

- **The relationship between having active hobbies and GPA is positive but not statistically significant in our dataset.**
- Although the model shows that students with hobbies tend to have higher GPAs, this effect is not strong enough to be considered reliable.
- Therefore, we cannot conclude that hobbies significantly influence academic performance in this sample.

2. Multiple Linear Regression:



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    gpa   R-squared:                       0.020
Model:                            OLS   Adj. R-squared:                  0.007
Method:                 Least Squares   F-statistic:                     1.522
Date:                Sun, 07 Dec 2025   Prob (F-statistic):              0.222
Time:                        02:40:45   Log-Likelihood:                -117.17
No. Observations:                 154   AIC:                             240.3
Df Residuals:                     151   BIC:                             249.4
Df Model:                           2
Covariance Type:            nonrobust
=================================================================================================
                                    coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------------
const                             2.7779      0.523      5.309      0.000       1.744       3.812
hours_spent_studying_per_week     0.0006      0.001      1.211      0.228      -0.000       0.002
personality_score                      0          0        nan        nan           0           0
has_active_hobby                  0.6654      0.525      1.268      0.207      -0.371       1.702
==============================================================================
Omnibus:                      162.114   Durbin-Watson:                   1.588
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             3834.542
Skew:                          -3.876   Prob(JB):                         0.00
Kurtosis:                      26.184   Cond. No.                          inf
==============================================================================
```

Image 2: Multiple Linear Regression

- None of the included lifestyle or personality factors: *study hours, hobbies, or personality*; show a statistically significant effect on GPA in this dataset.
- The model does not explain GPA variation, suggesting these variables are not strong predictors of academic performance based on the available data. Active Hobby parameter suggests that the relationship between GPA and these factors are not statistically significant.

3. K-Means Clustering

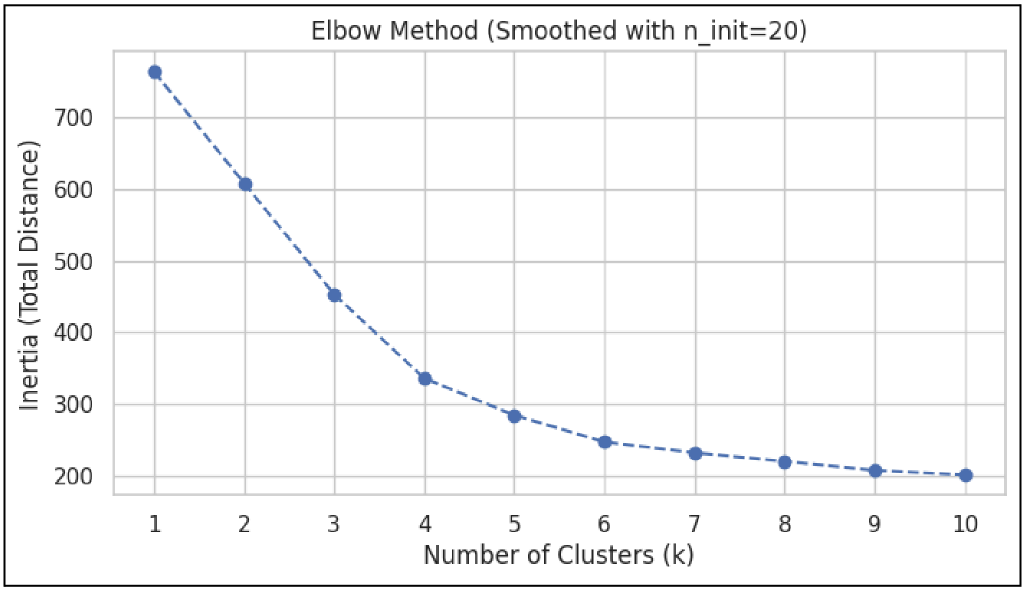


Image 3: Elbow chart

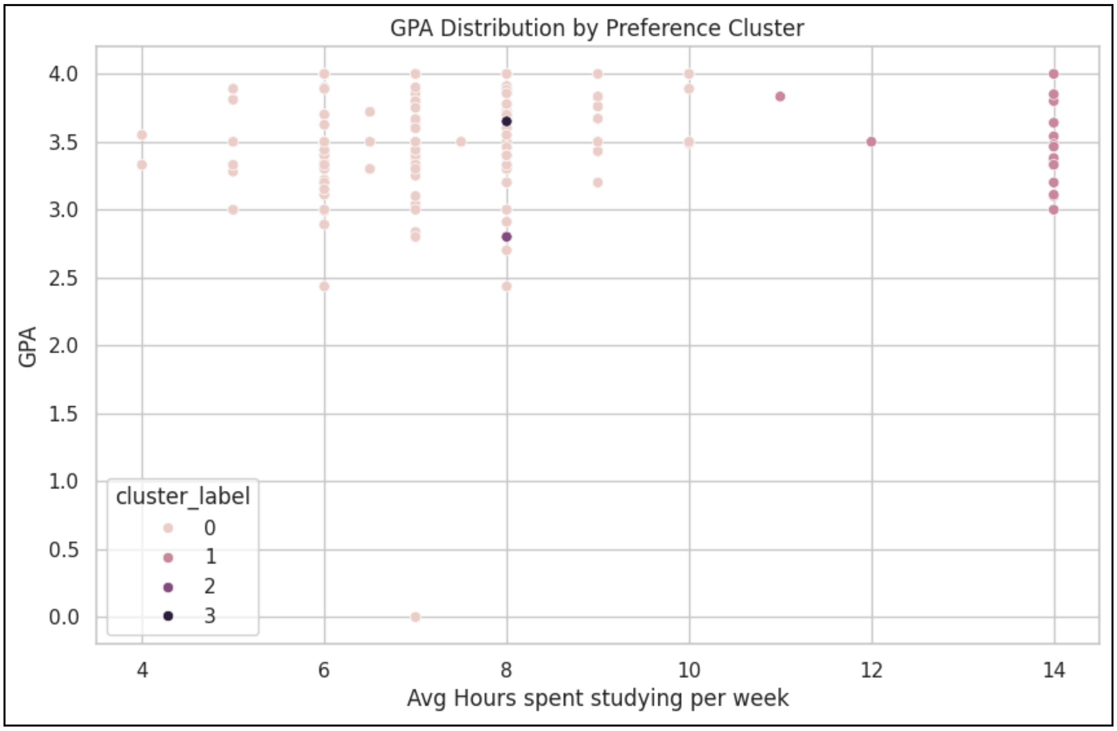


Image 4: GPA v Avg hours of sleep per night

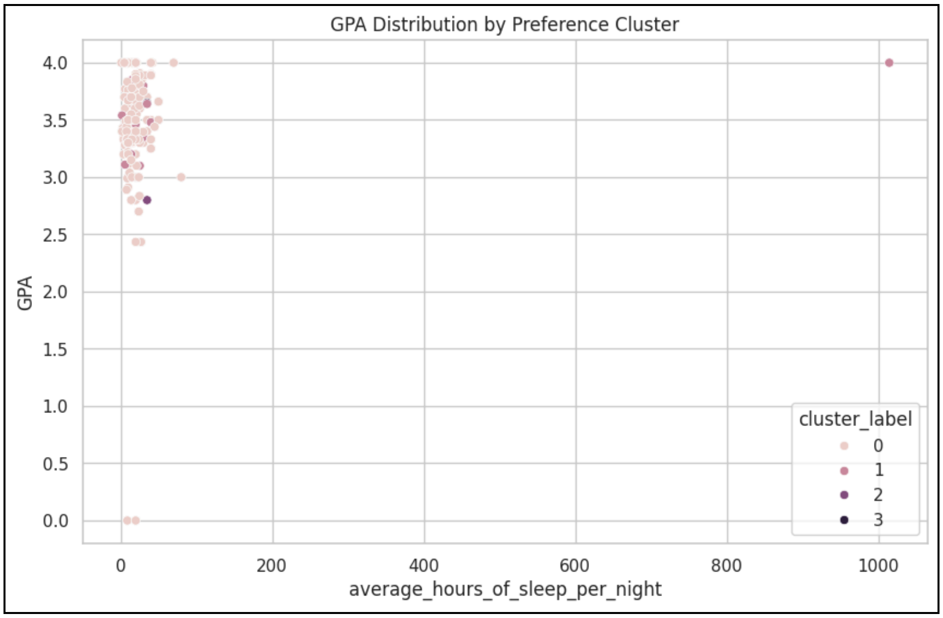
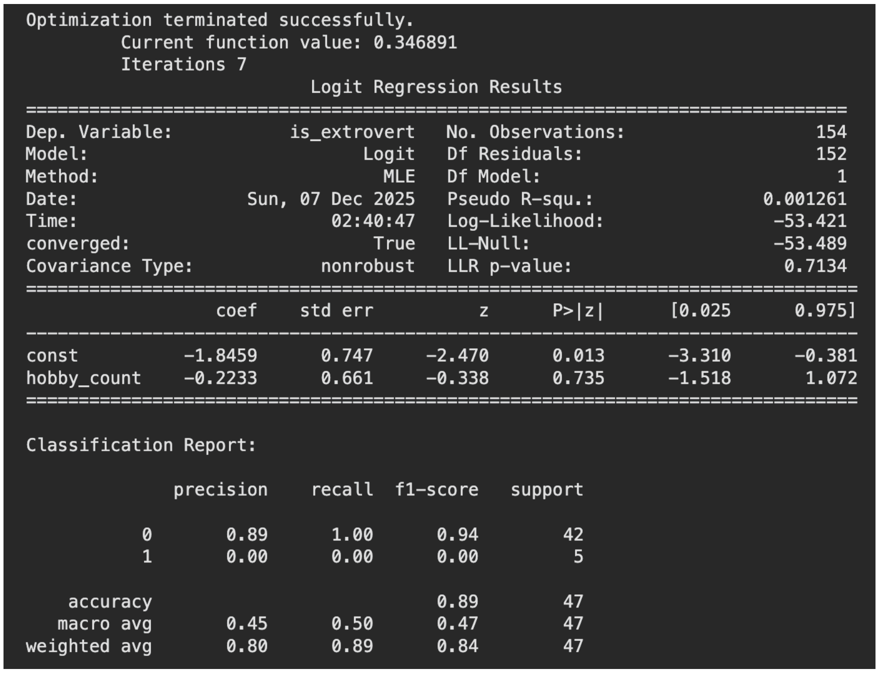


Image 5: GPA VS Average Hours of Sleep Per Night

- Academic success (GPA 3.65) is linked to a balance of study and sleep, while excessive studying (81.7 hours/week) does not increase GPA. The lowest-performing students (GPA 2.8) have similar study/sleep habits to high performers, suggesting non-lifestyle factors significantly impact academic outcomes.

4. Logistic Regression Classification:

Based on the Logistic Regression output, the results indicate **no significant relationship** between the number of hobbies a student has and whether they identify as an extrovert. The trend suggests that ambiverts have more hobbies than extroverts and introverts



```
Optimization terminated successfully.
         Current function value: 0.346891
         Iterations 7
                           Logit Regression Results
==============================================================================
Dep. Variable:            is_extrovert   No. Observations:                  154
Model:                           Logit   Df Residuals:                      152
Method:                            MLE   Df Model:                            1
Date:                 Sun, 07 Dec 2025   Pseudo R-squ.:                0.001261
Time:                         02:40:47   Log-Likelihood:                -53.421
converged:                        True   LL-Null:                       -53.489
Covariance Type:             nonrobust   LLR p-value:                    0.7134
===============================================================================
                  coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------
const          -1.8459      0.747     -2.470      0.013      -3.310      -0.381
hobby_count    -0.2233      0.661     -0.338      0.735      -1.518       1.072
===============================================================================

Classification Report:

              precision    recall  f1-score   support

           0       0.89      1.00      0.94        42
           1       0.00      0.00      0.00         5

    accuracy                           0.89        47
   macro avg       0.45      0.50      0.47        47
weighted avg       0.80      0.89      0.84        47
```

Image 6: Logistic regression results

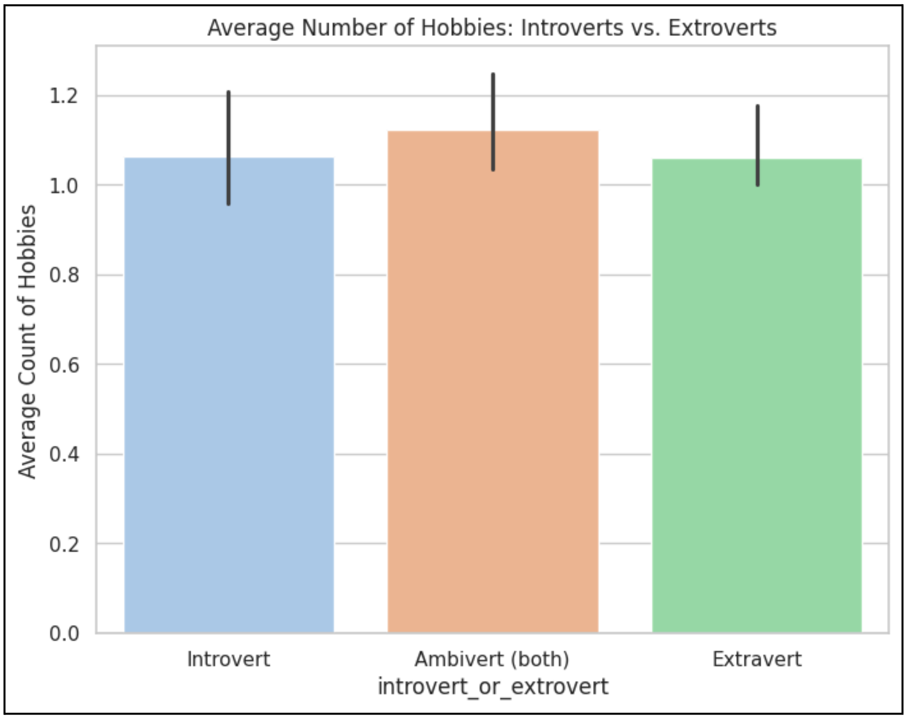


Image 7: Hobbies vs Personality

Conclusion Summary

This project aimed to determine if student lifestyle choices—specifically hobbies, sleep patterns, and personality traits—could reliably predict academic performance (GPA). After applying Linear Regression, Multiple Linear Regression, K-Means Clustering, and Logistic Regression to the dataset, we arrived at the following conclusions:

1. Hobbies are not a primary driver of GPA: We found no statistically significant evidence that having an active hobby influences GPA in this sample. While students with hobbies had marginally higher GPAs, the relationship is too weak to be considered a reliable predictor.
2. Lifestyle factors alone are poor predictors: The Multiple Linear Regression model failed to find significance among study hours, personality scores, or hobbies. This suggests that GPA is a complex metric influenced heavily by factors outside the scope of this survey (non-lifestyle factors).
3. Balance correlates with success: The strongest insight came from unsupervised learning (Clustering), which identified that high-performing students (GPA ~3.65) tend to maintain a balance between sleep and study, whereas excessive study hours do not correlate with higher grades.
4. Personality Trends: Finally, we concluded that personality type (Introvert/Extrovert) does not dictate how many hobbies a student pursues, with Ambiverts actually showing the highest engagement in diverse activities.

In summary, while we successfully cleaned the data and deployed multiple algorithms, the results suggest that academic success cannot be accurately predicted solely by the volume of studying or the presence of hobbies. Future work should focus on acquiring features related to academic background and classroom engagement to improve model accuracy.

Lessons Learned

Through the execution of this project, we encountered several key insights regarding the data science lifecycle and the complexity of modeling human behavior:

- <u>Statistical Significance vs. Intuition:</u> Our initial hypothesis suggested that active hobbies would positively correlate with GPA due to stress relief. However, the Linear Regression yielded a p-value of 0.211, which is well above the standard alpha of 0.05. This taught us that intuitive assumptions must always be validated by statistical rigor; in this dataset, the presence of a hobby was not a statistically significant predictor of academic success.
- <u>Low Explanatory Power (R-Squared) Challenges:</u> Both our Simple and multiple linear regression models resulted in extremely low R-squared values (0.010 and 0.020, respectively). This indicates that the variables we selected (study hours, hobbies, personality scores) explain less than 2% of the variance in GPA. The lesson learned here is that academic performance is likely driven by latent variables not present in this dataset (e.g., prior academic foundation, class attendance, or mental health metrics) rather than the lifestyle choices we analyzed.
- <u>The Nuance of "More is Better":</u> The K-Means Clustering analysis provided a critical lesson in non-linear relationships. While studying is essential, Cluster 2 revealed that students studying excessively (~81.7 hours/week) did not achieve higher GPAs than those with more balanced schedules. This demonstrates the concept of diminishing returns and highlights that data analysis can reveal optimal "sweet spots" (balance) that simple linear models might miss.

- Complexity of Personality Categorization: Our Logistic Regression showed that "Hobby Count" is not a strong predictor for Extroversion. Interestingly, the data revealed that Ambiverts (those identifying as both/neither) actually held the highest average number of hobbies. This taught us that binary classifications (Introvert vs. Extrovert) may oversimplify behavioral patterns, and distinct groups (like Ambiverts) often require separate analysis.

Bibliography:

- https://www.ramsayhealth.co.uk/treatments/weight-loss-surgery/bmi/bmi-formula
- https://scikit-learn.org/stable/supervised_learning.html
- https://scikit-learn.org/stable/modules/clustering.html

Appendix

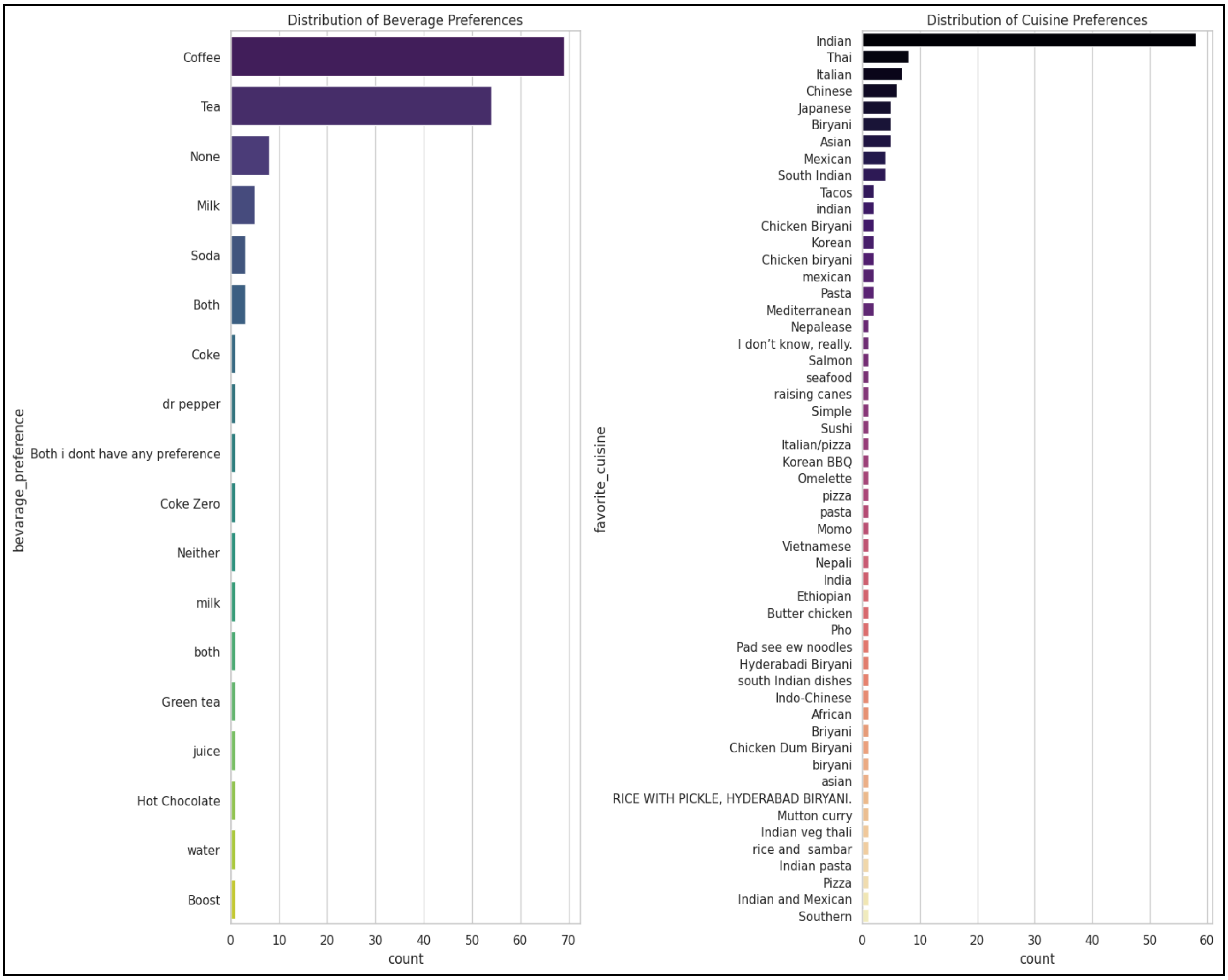


Image 8: Distribution of Beverage Preferences and Distribution of Cuisine Preferences

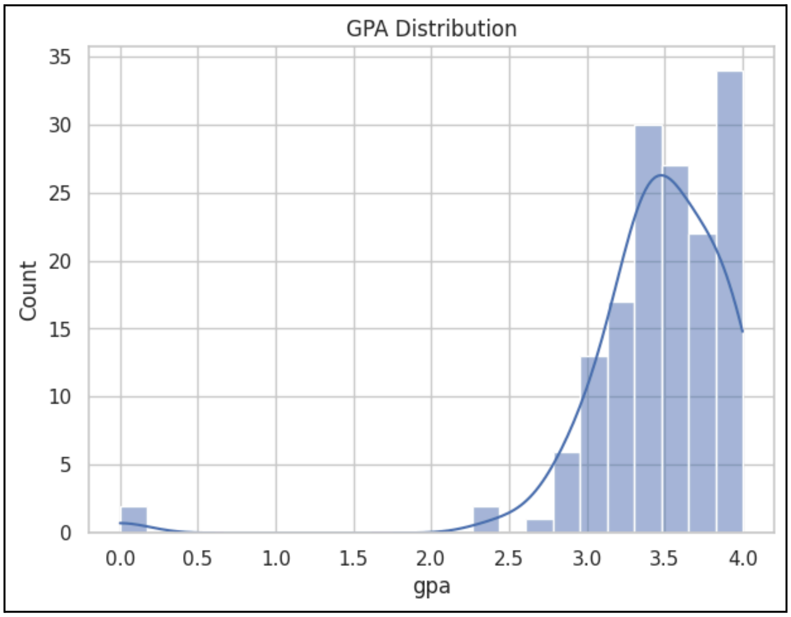


Image 9: GPA Distribution

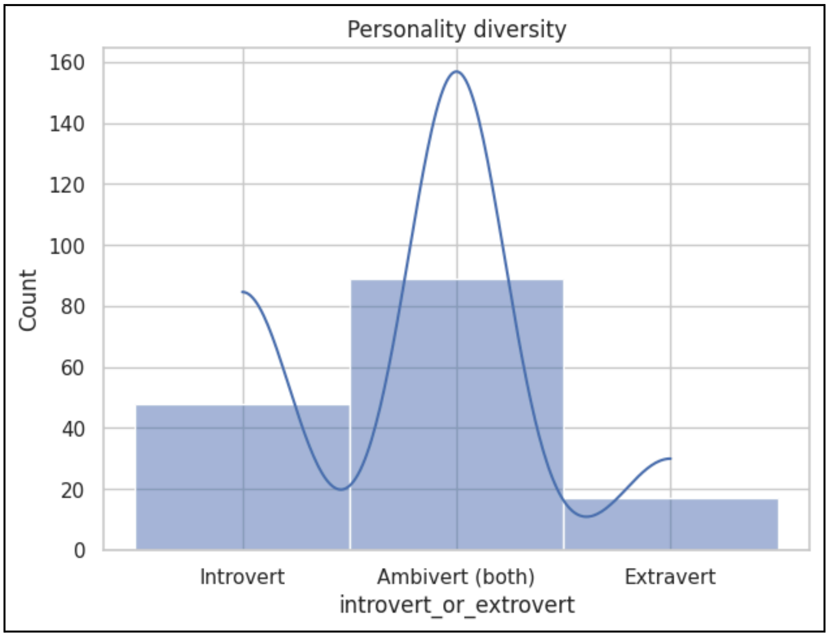


Image 10: Personality Diversity

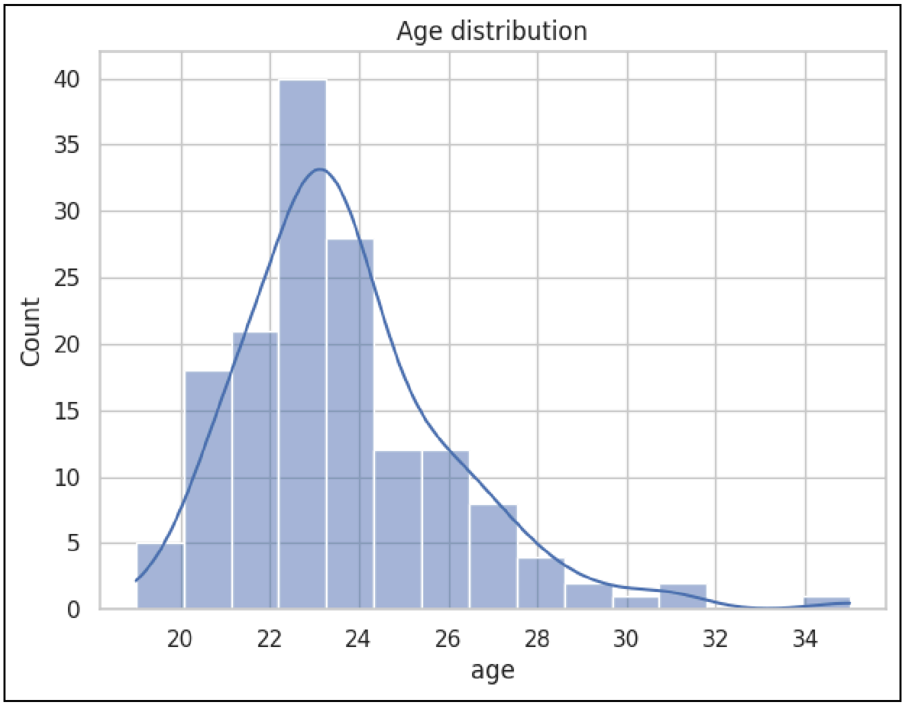


Image 11:

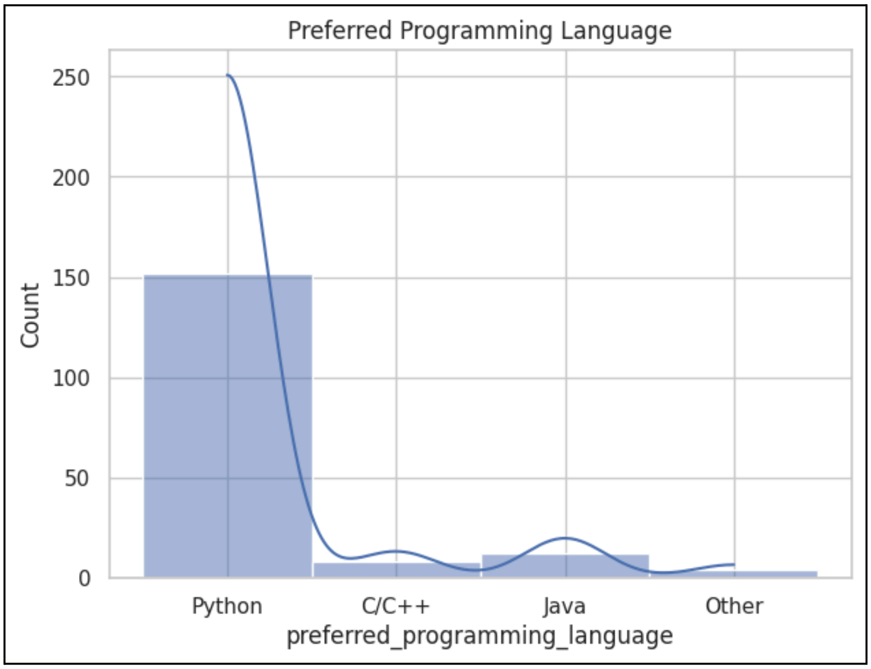


Image 12: Preferred Programming Language

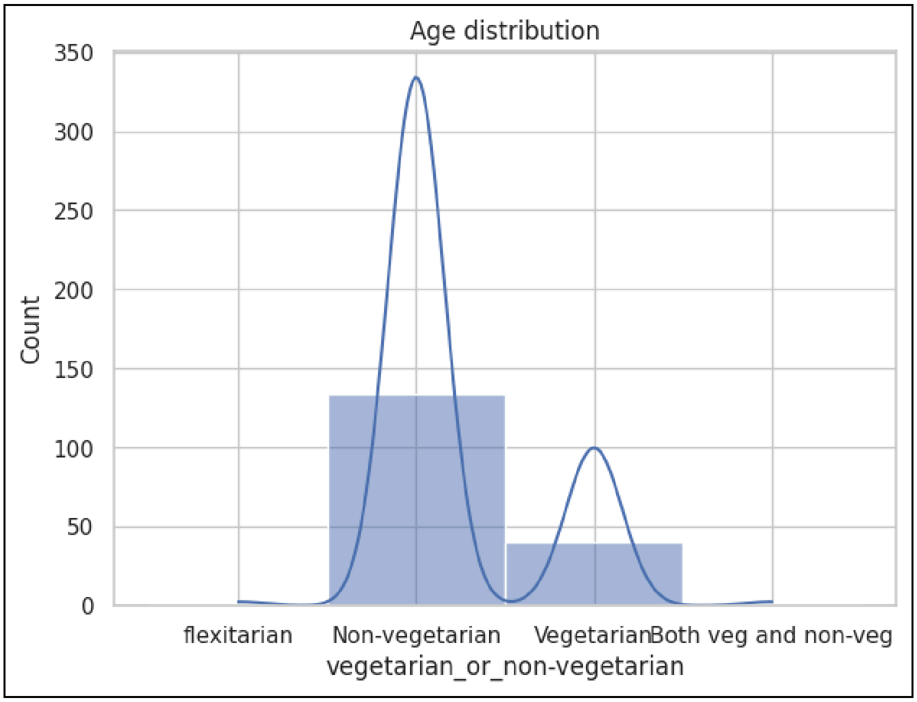


Image 13: Age Distribution

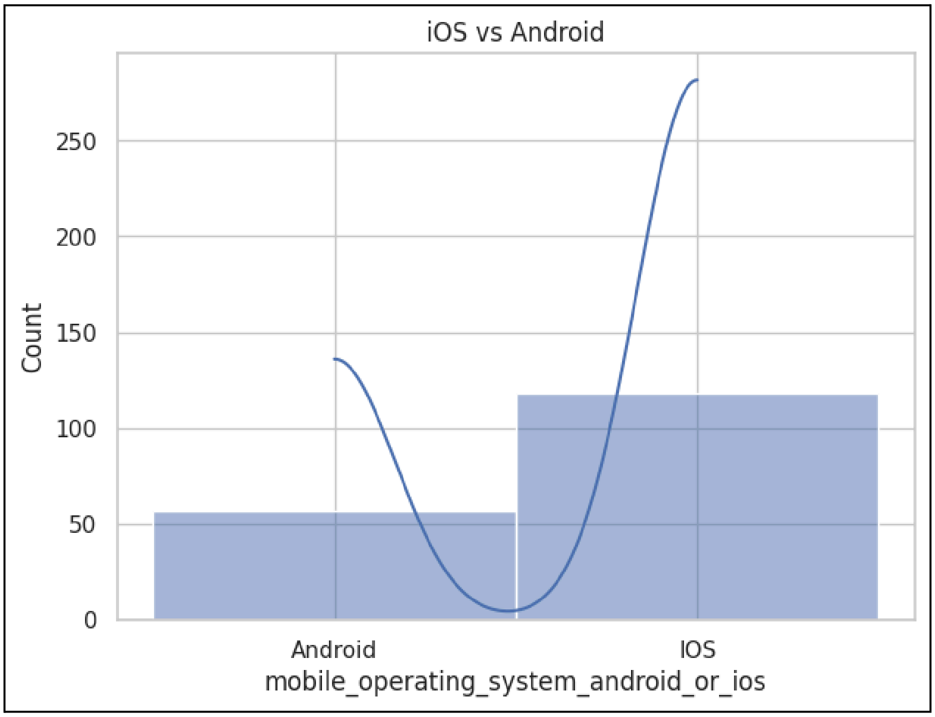


Image 14: iOS vs Android

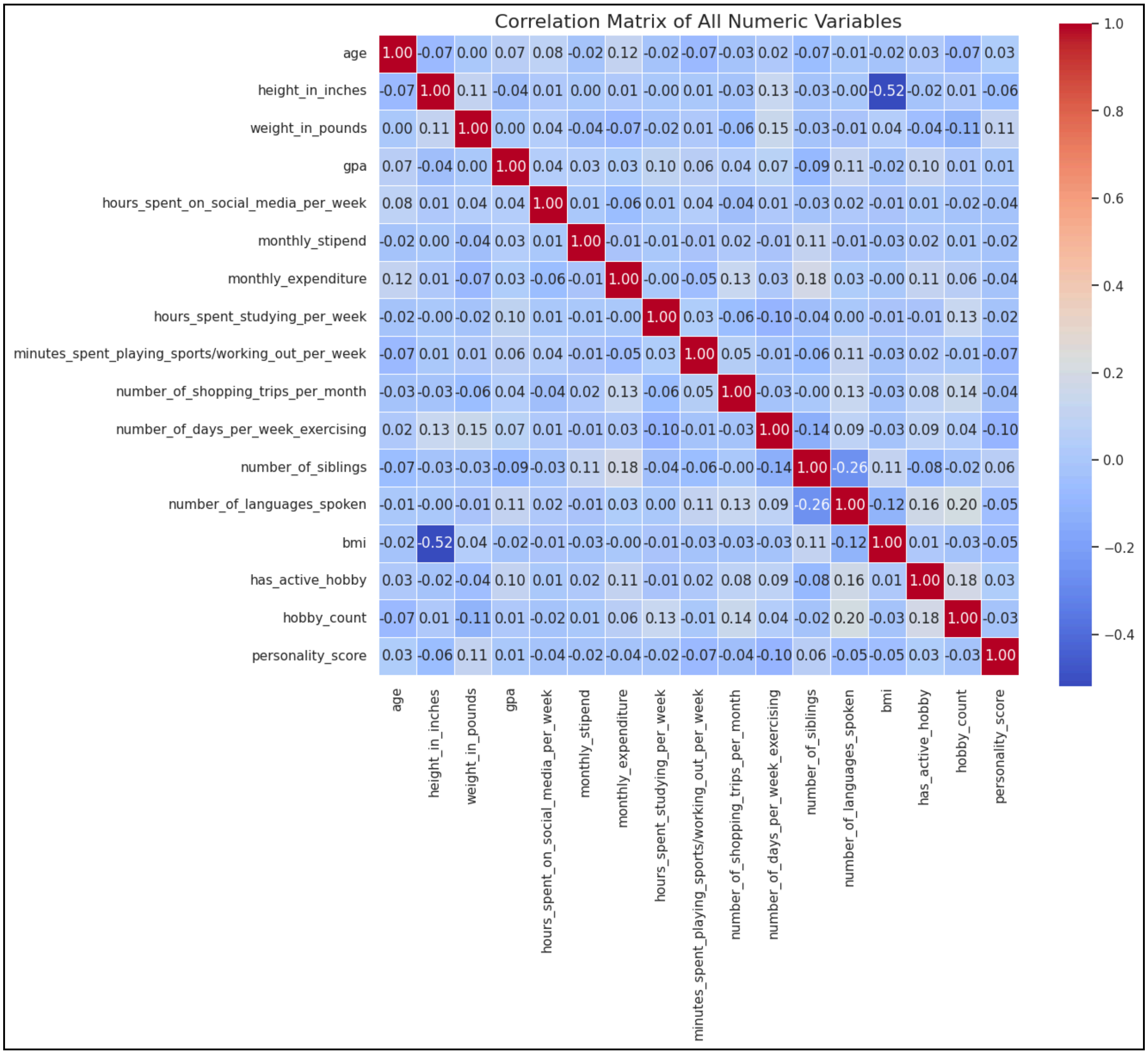


Image 15: Correlation matrix