Happiness and Alcohol Consumption

By Bolletta Oscar Maria, Galassi Leonardo, and Sartini Alberto

Introduction

The project we are presenting explores the correlation between happiness and alcohol consumption. The dataset, sourced from Kaggle, is titled "Happiness and Alcohol Consumption." We were intrigued by the idea of investigating whether an increase in alcohol consumption correlates with higher levels of happiness. Additionally, we aimed to identify other variables that significantly influence happiness, analyze how countries cluster based on common characteristics, and thoroughly examine every detail of the dataset. The techniques and analyses we used include exploratory analysis, linear regression, cluster analysis, PCA, logistic regression, and random forest.

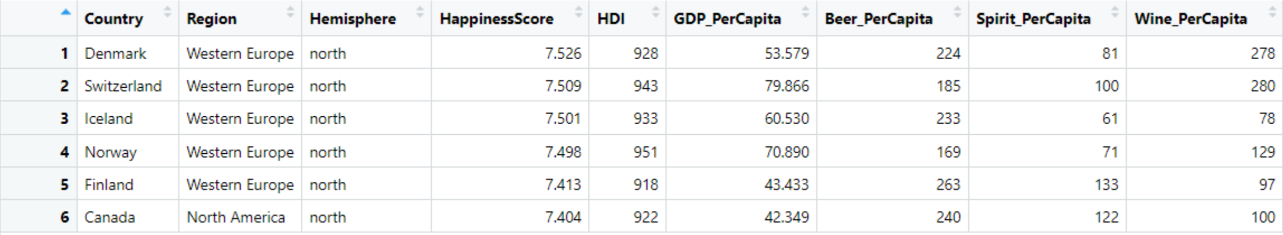
Exploratory Analysis

The first step in our analysis was to conduct an exploratory data analysis (EDA) to investigate, understand, and summarize the dataset. The dataset contains the following variables:

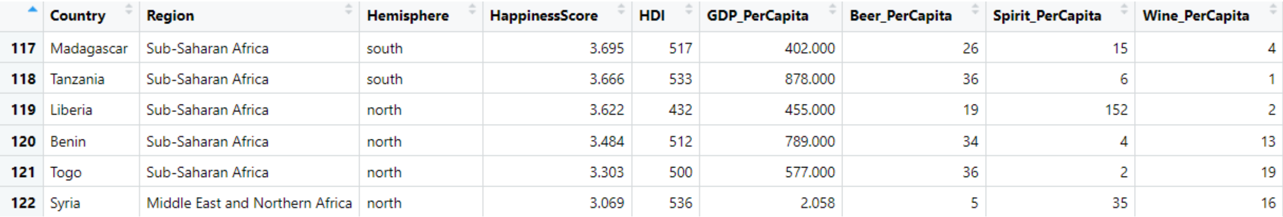
- **Country**: The name of the country.
- **Region**: The region/continent where the country is located.
- **Hemisphere**: The hemisphere the country belongs to.
- **HappinessScore**: A happiness index scored on a scale from 0 to 10.
- **HDI**: Human Development Index, measured on a scale from 0 to 1000.
- **GDP_PerCapita**: GDP per capita.
- **Beer_PerCapita**: Average annual beer consumption (liters per person).
- **Spirit_PerCapita**: Average annual spirit consumption (liters per person).
- **Wine_PerCapita**: Average annual wine consumption (liters per person).

We first checked for any NA values or duplicates in the dataset. Fortunately, there were none, allowing us to proceed with our analysis immediately.

We then examined the dataset's structure by viewing its head and tail to understand its composition.

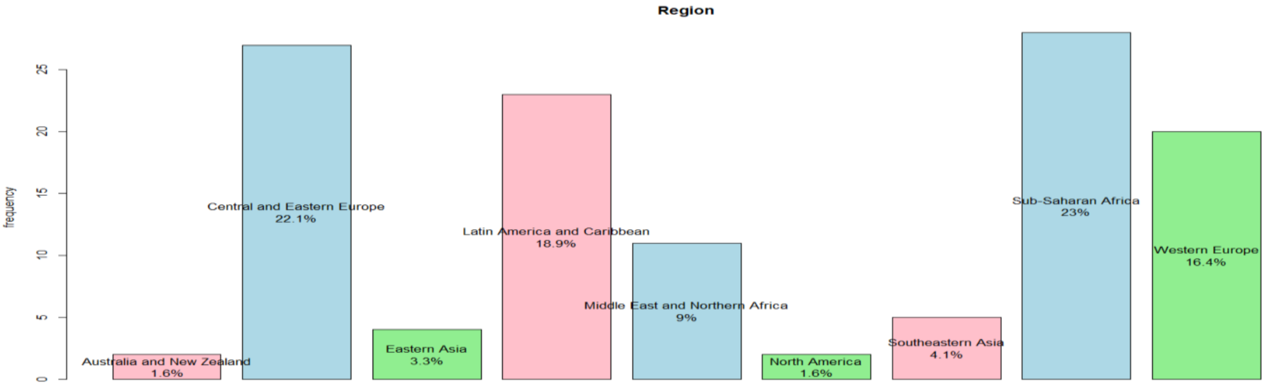


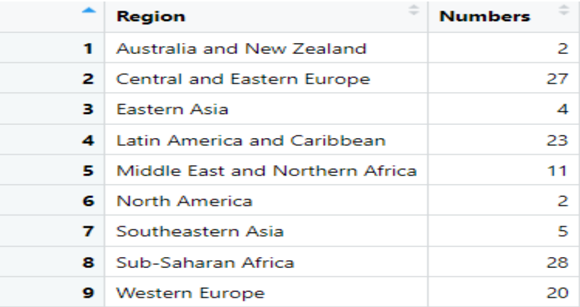
	Country	Region	Hemisphere	HappinessScore	HDI	GDP_PerCapita	Beer_PerCapita	Spirit_PerCapita	Wine_PerCapita
1	Denmark	Western Europe	north	7.526	928	53.579	224	81	278
2	Switzerland	Western Europe	north	7.509	943	79.866	185	100	280
3	Iceland	Western Europe	north	7.501	933	60.530	233	61	78
4	Norway	Western Europe	north	7.498	951	70.890	169	71	129
5	Finland	Western Europe	north	7.413	918	43.433	263	133	97
6	Canada	North America	north	7.404	922	42.349	240	122	100



	Country	Region	Hemisphere	HappinessScore	HDI	GDP_PerCapita	Beer_PerCapita	Spirit_PerCapita	Wine_PerCapita
117	Madagascar	Sub-Saharan Africa	south	3.695	517	402.000	26	15	4
118	Tanzania	Sub-Saharan Africa	south	3.666	533	878.000	36	6	1
119	Liberia	Sub-Saharan Africa	north	3.622	432	455.000	19	152	2
120	Benin	Sub-Saharan Africa	north	3.484	512	789.000	34	4	13
121	Togo	Sub-Saharan Africa	north	3.303	500	577.000	36	2	19
122	Syria	Middle East and Northern Africa	north	3.069	536	2.058	5	35	16

Additionally, we explored the distribution of countries by region through graphs, which revealed a more granular classification of geographic areas (e.g., “Western Europe,” “Eastern Asia,” etc.) rather than broad continental divisions.





	Region	Numbers
1	Australia and New Zealand	2
2	Central and Eastern Europe	27
3	Eastern Asia	4
4	Latin America and Caribbean	23
5	Middle East and Northern Africa	11
6	North America	2
7	Southeastern Asia	5
8	Sub-Saharan Africa	28
9	Western Europe	20

Top and Bottom 25 Countries by Variables

We analyzed the top and bottom 25 countries for each variable to understand their distribution and detect anomalies:

- **HappinessScore**:
  Many American and European countries appeared in the top 25, while African nations dominated the bottom 25. This distribution aligns with basic global knowledge.

- **HDI**:
  Similar to HappinessScore, higher HDI values were found in American and European countries, with African nations at the lower end.

- **GDP_PerCapita**:
  Some GDP data appeared incorrect (e.g., Senegal and Tanzania showed unrealistically high GDP per capita). We verified that all values were in USD but chose to exclude this variable from subsequent analyses due to its unreliability.

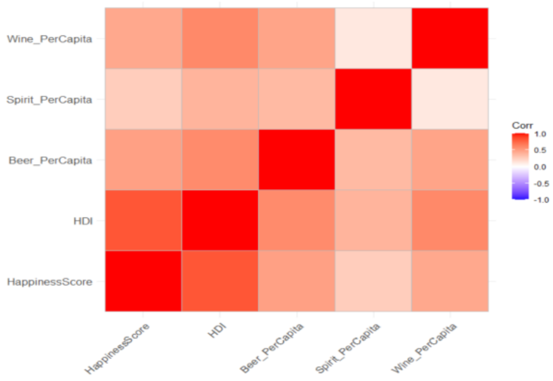
- **Beer_PerCapita**:
  The distribution was less consistent. For instance, Namibia topped the list, while countries like Qatar appeared at the bottom due to religious restrictions on alcohol consumption.

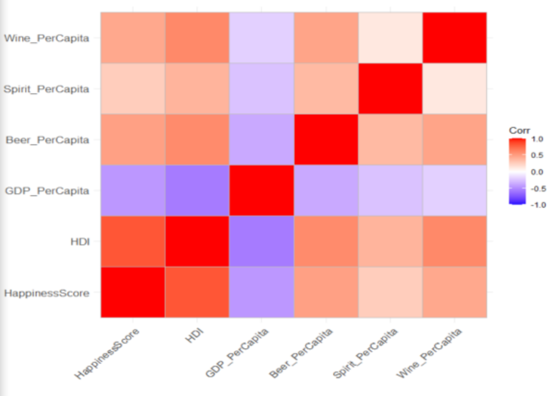
- **Wine_PerCapita**:
  Countries known for wine production, like Italy and France, dominated the top rankings. Despite some doubts about the accuracy of the bottom rankings, further research supported the data's validity.

- **Spirit_PerCapita**:
  Belarus and Russia, known for their vodka consumption, led the rankings. Conversely, sub-Saharan African countries appeared at the bottom due to religious, availability, or affordability reasons.

Using boxplots, we identified outliers and confirmed that most variables, except GDP_PerCapita and Wine_PerCapita, were normally distributed.

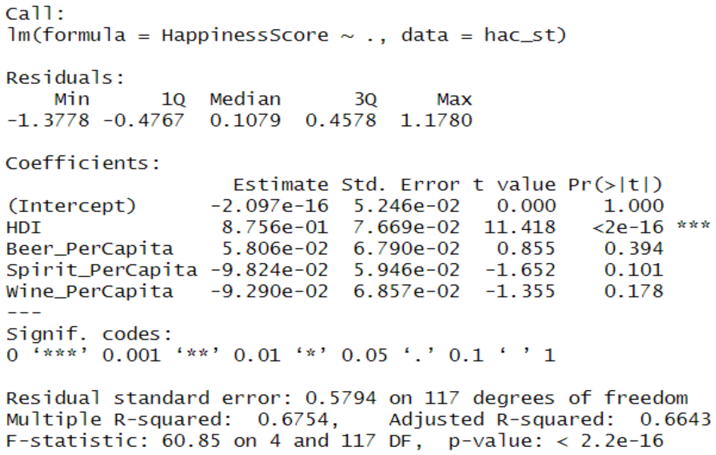
We also constructed correlation matrices (with and without GDP_PerCapita). GDP_PerCapita showed an anomalous negative correlation with other variables, further justifying its exclusion.





Linear Regression

We used linear regression to study the dataset's quantitative variables. HappinessScore was chosen as the dependent variable, with other quantitative variables (excluding GDP_PerCapita) as independent variables.



```
Call:
lm(formula = HappinessScore ~ ., data = hac_st)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.3778 -0.4767  0.1079  0.4578  1.1780 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)      -2.097e-16  5.246e-02   0.000    1.000    
HDI               8.756e-01  7.669e-02  11.418   <2e-16 ***
Beer_PerCapita    5.806e-02  6.790e-02   0.855    0.394    
Spirit_PerCapita -9.824e-02  5.946e-02  -1.652    0.101    
Wine_PerCapita   -9.290e-02  6.857e-02  -1.355    0.178    
---
Signif. codes:  
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5794 on 117 degrees of freedom
Multiple R-squared:  0.6754,	Adjusted R-squared:  0.6643 
F-statistic: 60.85 on 4 and 117 DF,  p-value: < 2.2e-16
```

Findings:

- HDI had a strong positive effect on HappinessScore.
- Beer_PerCapita, Spirit_PerCapita, and Wine_PerCapita had no significant impact.
- R-Squared and Adjusted R-Squared values were ~0.67, indicating that the model explained 67% of the variance in HappinessScore.
- ANOVA tests confirmed the validity of the regression model.

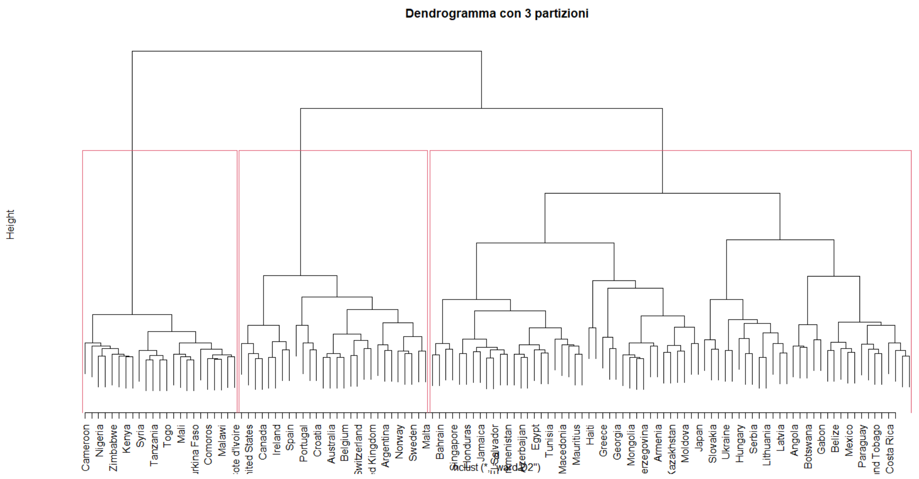
Residuals analysis showed normal distribution and linearity, confirming the model's adequacy. Additionally, VIF tests indicated no significant multicollinearity among variables.

Cluster Analysis

We analyzed how countries grouped based on the dataset’s variables using hierarchical and non-hierarchical clustering.

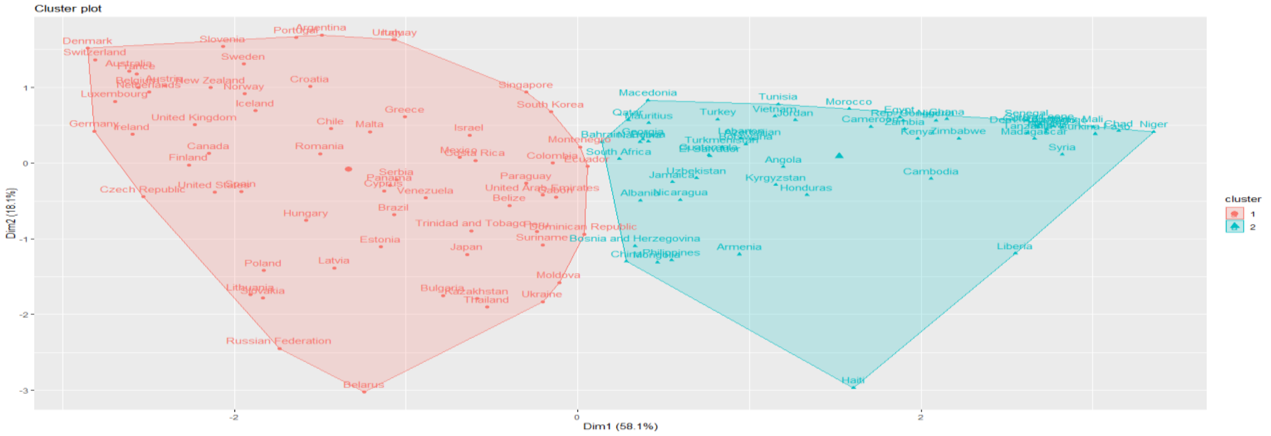
Hierarchical Clustering:

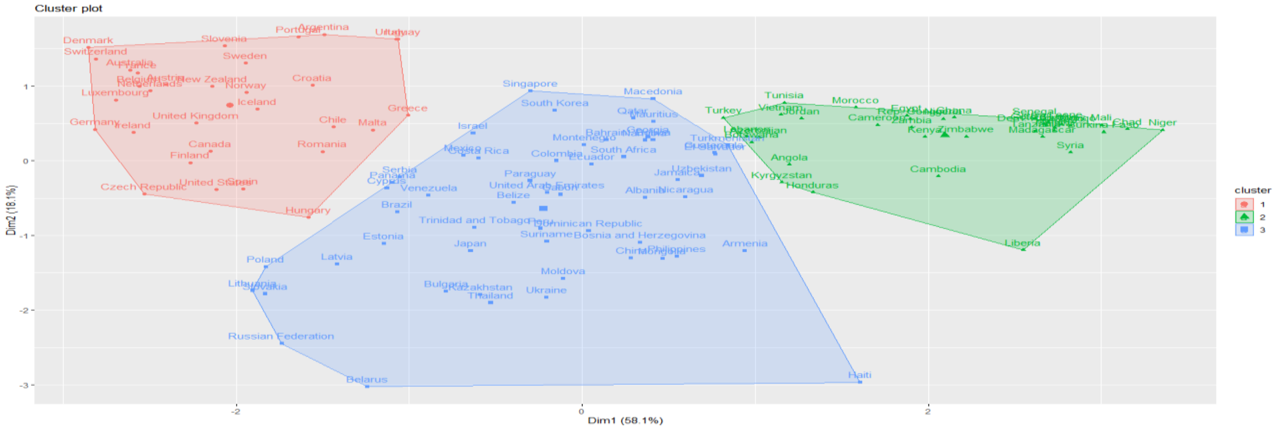
- We determined 3 clusters as optimal using the Hubert and D indices.
- The Ward method was used to maximize between-group variance and minimize within-group variance.
- Countries grouped mostly by continent, although some differences arose due to alcohol consumption, regulations, or religious influences.



K-Means Clustering:

- An elbow plot and silhouette analysis supported a 2- or 3-cluster solution.
- Clusters generally grouped countries into developed, developing, and culturally distinct categories.





PCA

Principal Component Analysis (PCA) revealed that the first two components explained nearly 80% of the variance.

- **First Principal Component (PC1)**: Influenced negatively by all variables, especially HappinessScore and HDI.
- **Second Principal Component (PC2)**: Positively influenced by HappinessScore, HDI, and Wine_PerCapita.

PCA supported previous clustering insights, showing a consistent country grouping pattern.

Logistic Regression

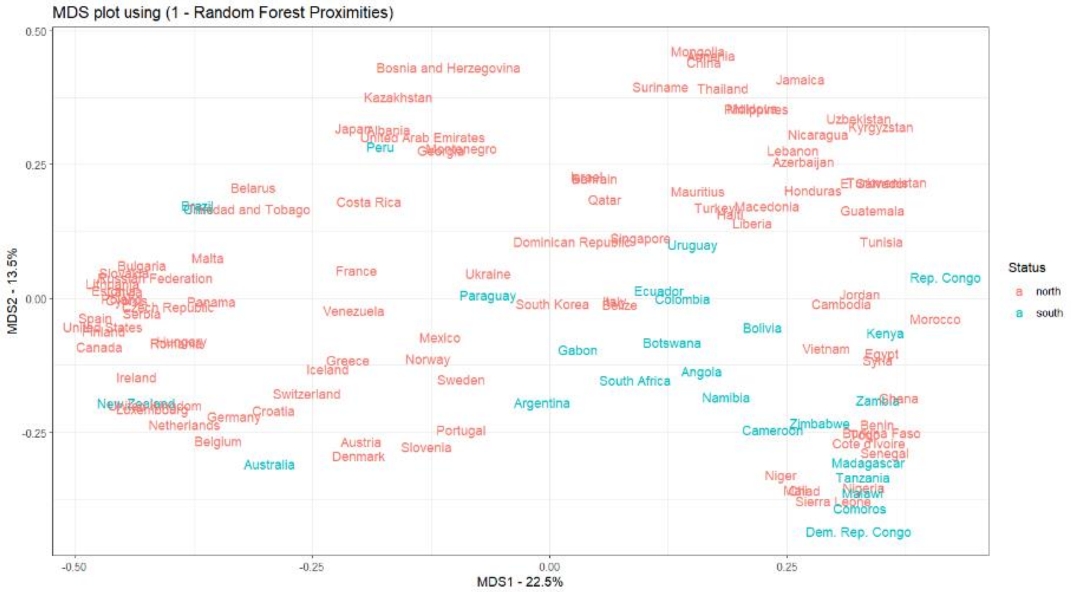
Our goal was to predict a country’s hemisphere based on the dataset's variables. After splitting the dataset into training (75%) and testing (25%) sets, we identified the optimal model using HDI, Beer_PerCapita, and Spirit_PerCapita as regressors.

The model achieved an accuracy of ~78.9% with good precision but struggled with specificity due to class imbalance.

Random Forest

We applied Random Forest to predict hemisphere classification, focusing on improving model performance:

- Models with 500, 1000, and 1500 trees showed gradual improvement, with the optimal model achieving ~85.2% accuracy.
- HDI emerged as the most important variable, followed by Beer_PerCapita and Spirit_PerCapita.



Conclusion

HDI, HappinessScore, and alcohol consumption variables moderately discriminate and group countries by hemisphere. Similarities among certain nations (e.g., Australia, New Zealand, and the UK) highlight shared cultural or socioeconomic characteristics. Future analyses would benefit from additional variables to enhance the model's accuracy and characterization.