**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.

Immagine che contiene testo, schermata, numero, Carattere

Descrizione generata automaticamente

Immagine che contiene testo, schermata, numero, Carattere

Descrizione generata automaticamente

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.

Immagine che contiene grafico

Descrizione generata automaticamente

Immagine che contiene tavolo

Descrizione generata automaticamente

**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.

Immagine che contiene grafico

Descrizione generata automaticamente

**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.

Immagine che contiene testo

Descrizione generata automaticamente

**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:**

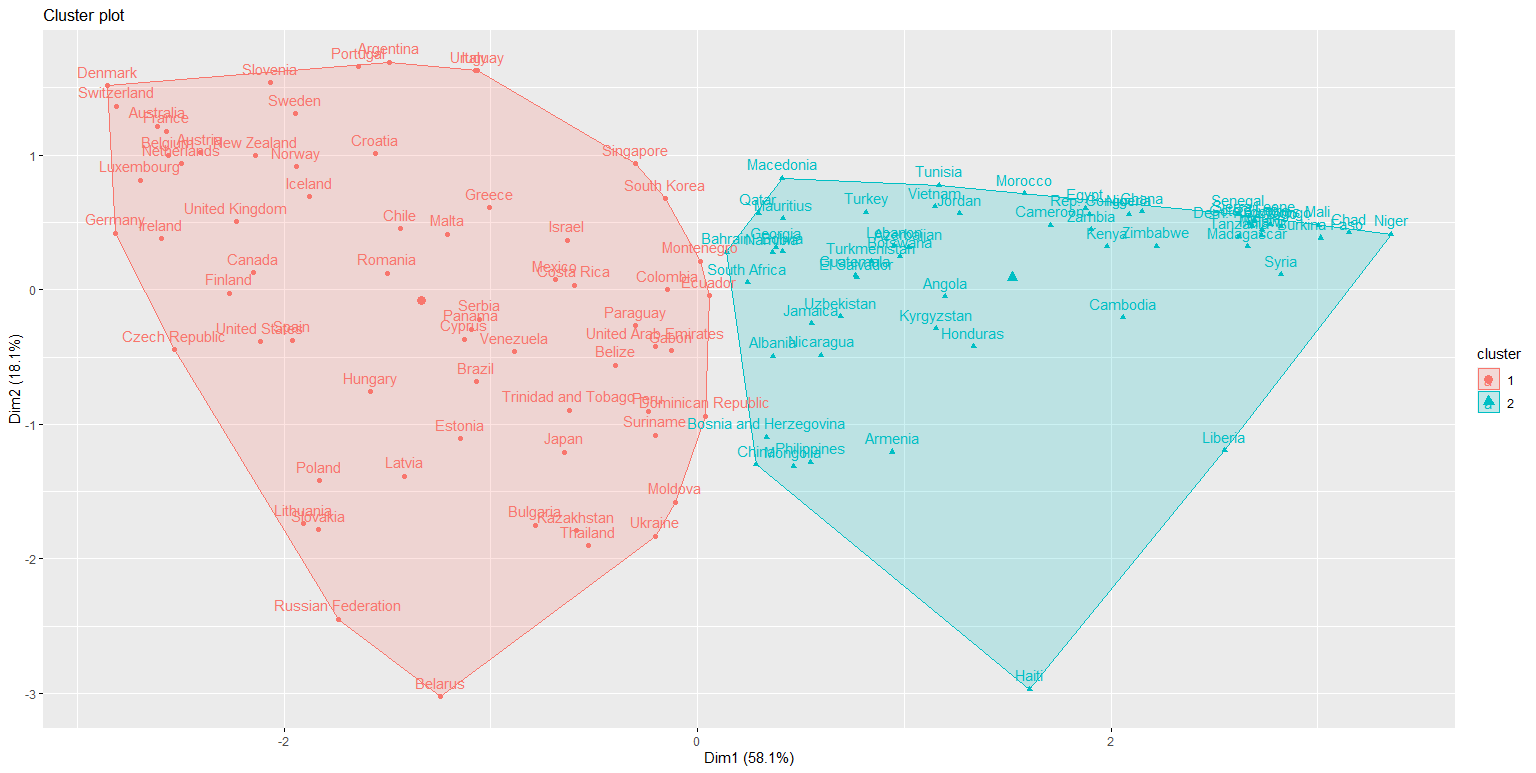
* We determined 3 clusters as optimal using the Hubert and D indices.
* The Ward method was used to maximize between-group variance and minimize within-group variance.
* Countries grouped mostly by continent, although some differences arose due to alcohol consumption, regulations, or religious influences.

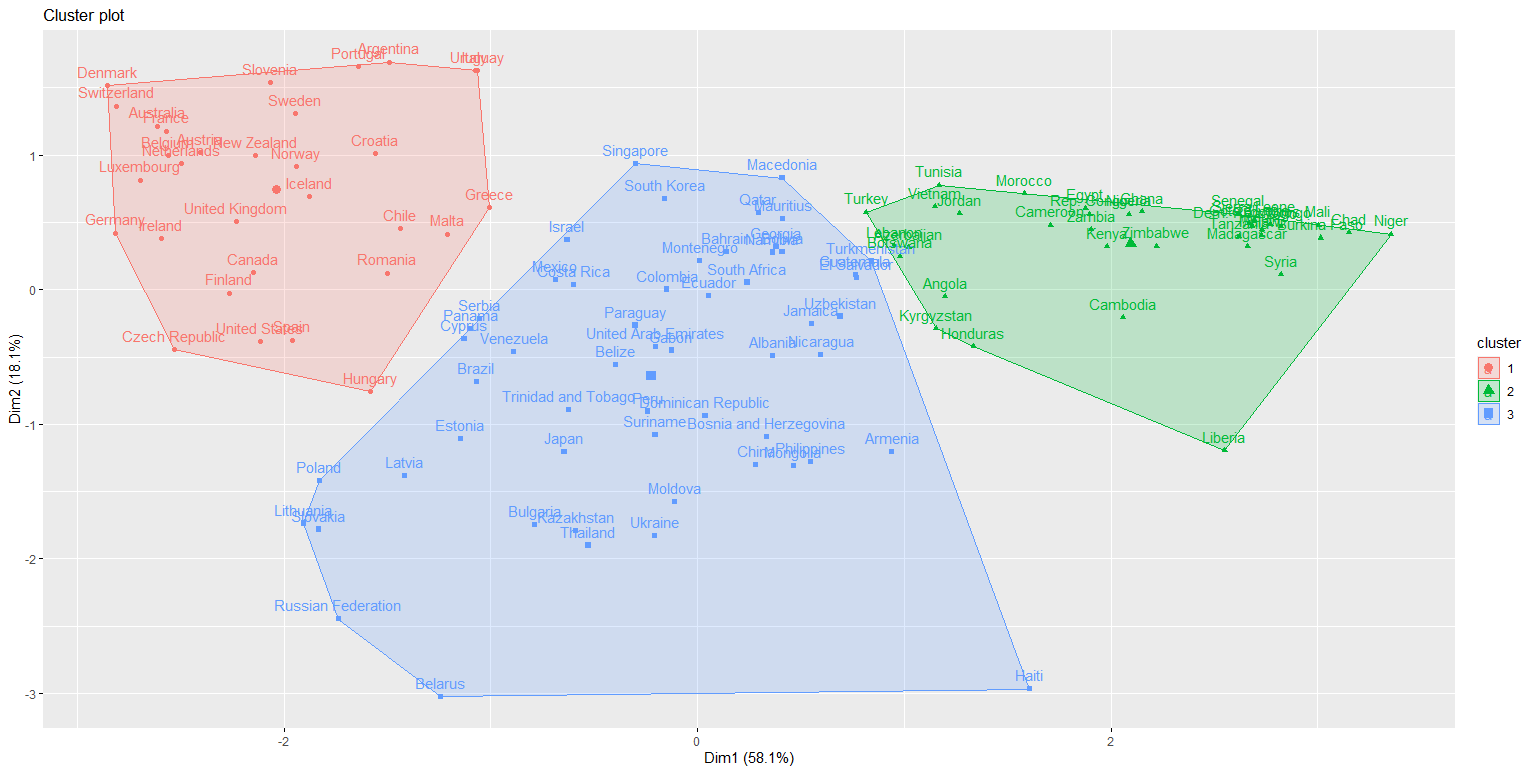
Immagine che contiene grafico

Descrizione generata automaticamente

**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.

Immagine che contiene grafico

Descrizione generata automaticamente

**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.