Happiness and Alcohol consumption

A cura di Bolletta Oscar Maria, Galassi Leonardo e Sartini Alberto

Introduzione

Il progetto che stiamo per presentare riguarda la correlazione tra felicità e consumo di alcool. Infatti, il dataset, trovato in *Kaggle*, prende il nome di "Happiness and Alcohol Consumption" e ci è sembrato interessante capire se un aumento del consumo di bibite alcoliche portasse ad un aumento della felicità. Oltre a questo obiettivo, ci siamo prefissati anche di capire se ci sono altre variabili significative per una maggiore felicità, come si raggruppano gli Stati tra loro in base a delle loro caratteristiche comuni e, in generale, studiare per bene ogni particolare del dataset. Le tecniche e analisi che andremo ad utilizzare sono: l'analisi esplorativa, la regressione lineare, la cluster analysis, la PCA, la regressione logistica e la random forest.

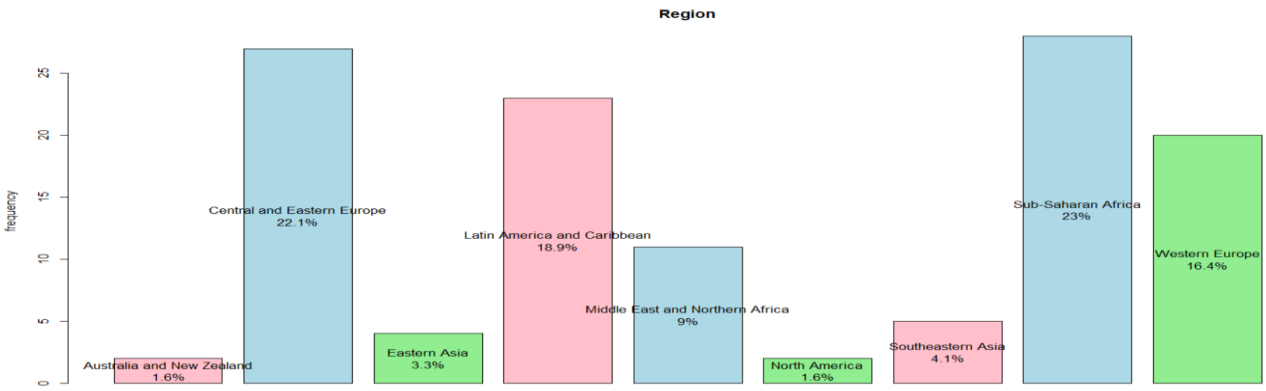
Analisi esplorativa

La prima analisi effettuata per studiare il nostro dataset è stata l'analisi esplorativa per indagare, comprendere e sintetizzare i dati forniti da esso. Il primo passo è stato quello di vedere quali fossero le variabili presenti e sono queste: Country (inteso come il nome dello Stato in questione), Region (la regione/il continente di appartenenza dello Stato in questione), Hemisphere (l'emisfero di appartenenza dello Stato in questione), HappinessScore (punteggio di felicità misurata con una scala che va da 0 a 10), HDI (Human Development Index ovvero l'Indice di Sviluppo Umano, misurato con una scala che va da 0 a 1000), GDP_PerCapita (il PIL pro capite), Beer_PerCapita (i litri, in media a persona, di consumo di birra in un anno), Spirit_PerCapita (i litri, in media a persona, di consumo di superalcolici in un anno) e Wine_PerCapita (i litri, in media a persona, di consumo di vino in un anno).
La prima vera operazione è stata quella di identificare eventuali NA e doppioni nel dataset ma, per nostra fortuna, non erano presenti e, di conseguenza, abbiamo proceduto subito con iniziare il nostro progetto.
Per capire com'è organizzato, ecco la "testa" e la "coda" del dataset per vedere graficamente come è composto.

	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

Come detto prima, c'è la variabile 'Region' che ci indica dove ogni Stato viene collocato (per esempio 'Western Europe', 'Eastern Asia', …) e ora vedremo con un grafico come sono distribuiti.

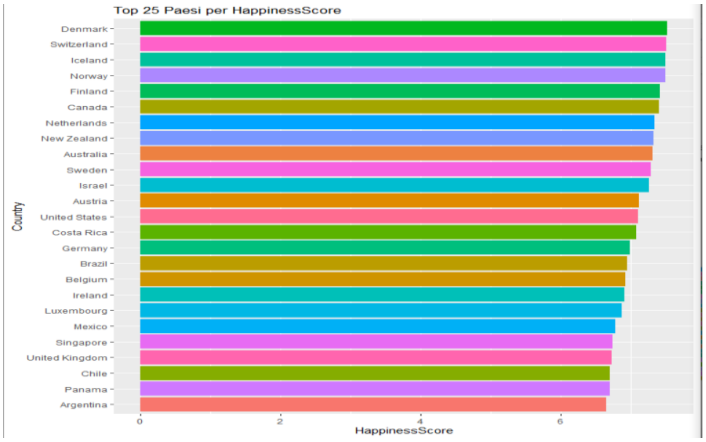


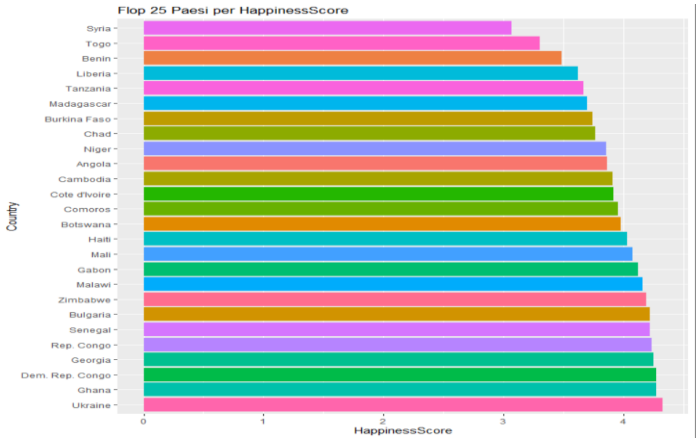
Possiamo notare come le aree geografiche siano classificate non dal continente di appartenenza generale ma da zone più specifiche di un determinato continente.
Ecco anche una visione più rigorosa della distribuzione dei Paesi.

	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

Il passo successivo è quello di studiare tutte le variabili quantitative, col fine di comprendere meglio cosa rappresentano e se ci sono dei valori e risultati anomali o improbabili. Per fare questo abbiamo deciso di fare dei grafici che ci mostrassero i top e i flop 25 Paesi per ognuna di queste variabili.

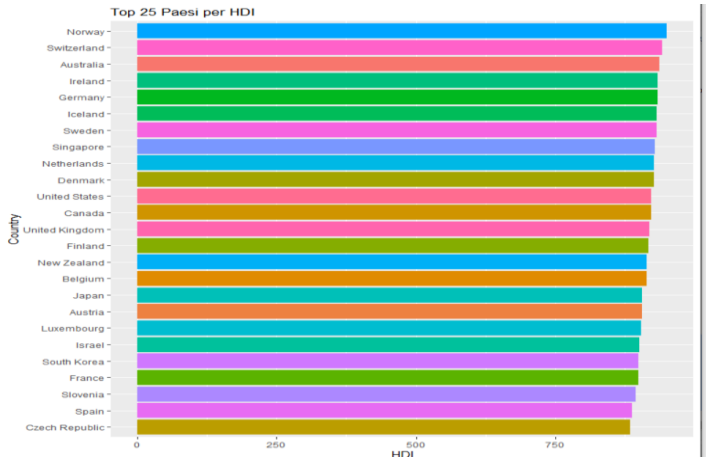
Top e flop 25 Stati per HappinessScore

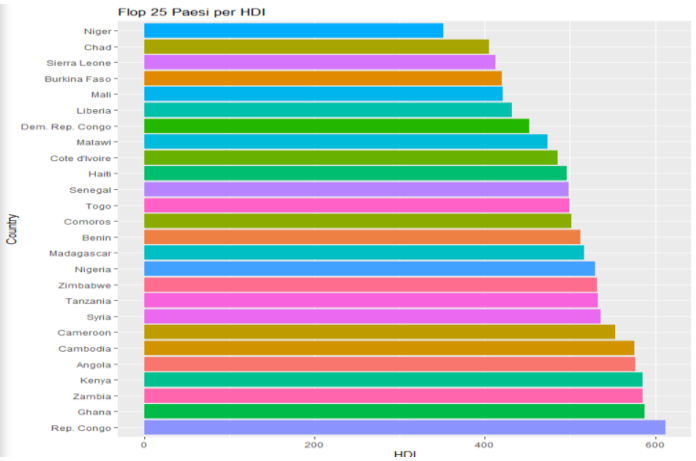




Possiamo subito notare come molti Stati americani ed europei si trovino nella top 25, mentre negli ultimi 25 troviamo quasi tutti Stati africani. Questo risultato è un qualcosa che ci aspettavamo, considerando le conoscenze base del mondo che ci circonda.

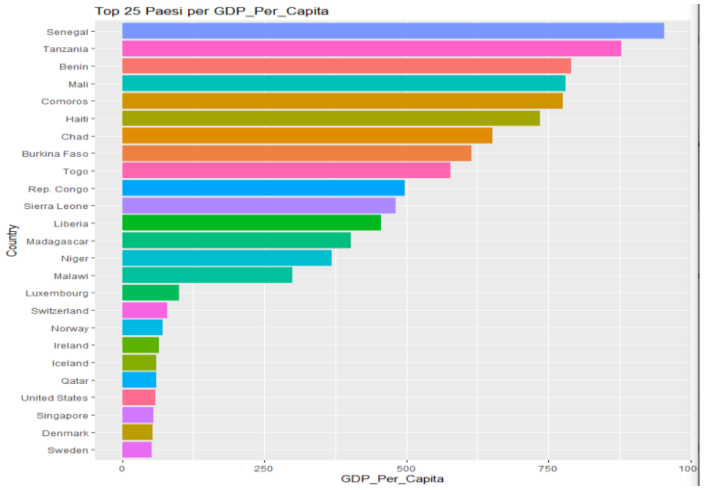
Top e bottom 25 Stati per HDI

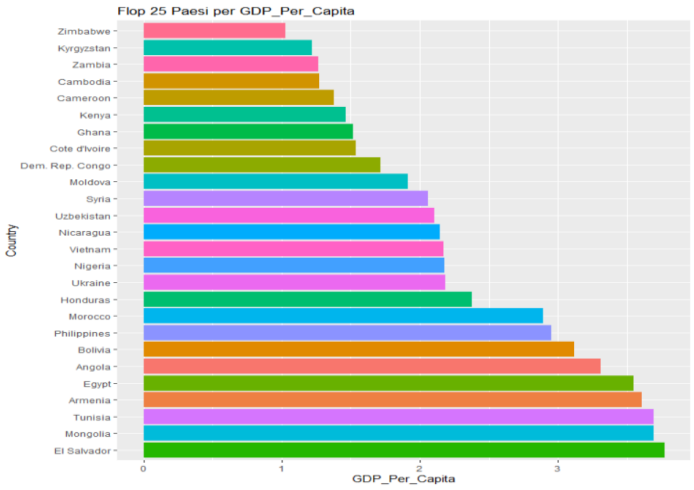




Come possiamo vedere, anche per la variabile HDI gli Stati presenti nelle due classifiche sono simili a quelli della variabile HappinessScore, con nei top gli Stati americani ed europei e nelle posizioni più basse quelli africani.

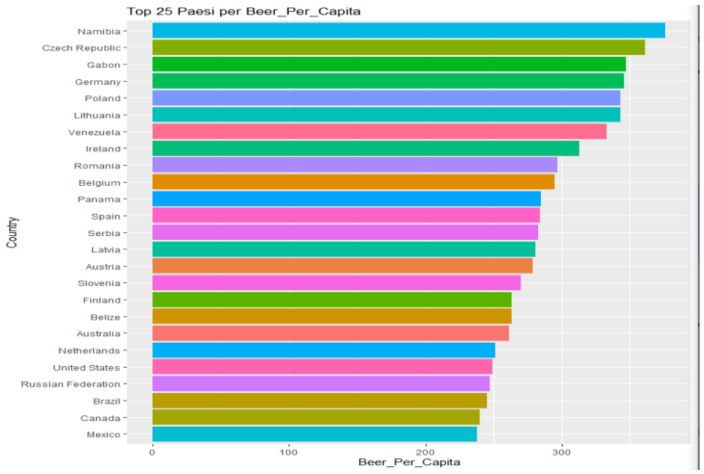
Top e bottom 25 Stati per GDP_PerCapita

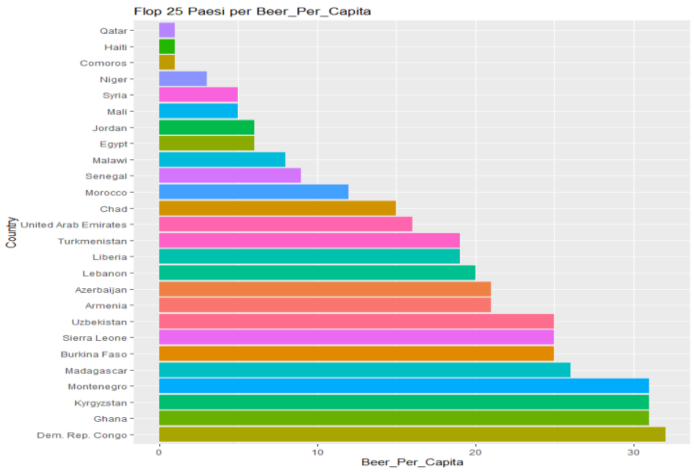




Nel caso della variabile GDP_PerCapita, sembra che ci sia qualcosa di sbagliato nei dati sul PIL di alcuni Paesi del dataset; infatti, con una rapida ricerca possiamo dire con certezza che Paesi come Senegal, Tanzania e altri non abbiano PIL pro capite così elevati. Pensavamo potesse essere un problema di valuta nazionale ma, con una verifica, abbiamo appurato che tutti i dati in questione sono espressi in dollari. Per questo motivo questa variabile è stata esclusa dalle successive analisi per evitare di avere risultati distorti. Inoltre, sono state fatte ulteriori analisi che ci hanno permesso di prendere questa decisione e le vedremo più in avanti.

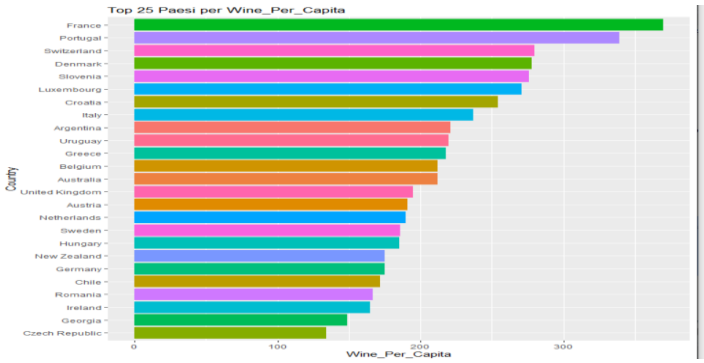
Top e bottom 25 Stati per Beer_PerCapita

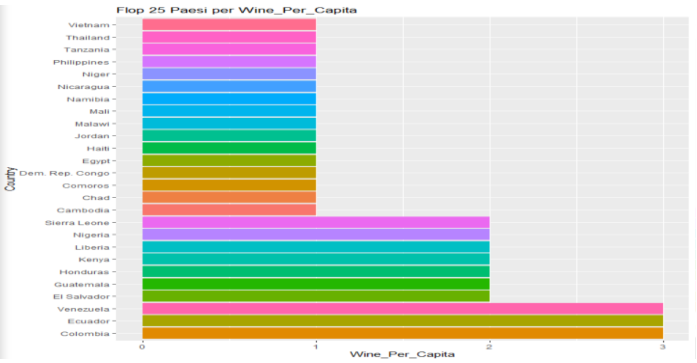




Per questa variabile, la distribuzione degli Stati è più casuale, infatti, per esempio, al primo posto di consumo di birra c'è uno Stato africano come la Namibia. Questo probabilmente ci può iniziare a far pensare che l'alcool non incide così positivamente sulla felicità, ma vedremo meglio con le restanti due variabili legate al consumo di alcool. Chiaramente negli ultimi posti troviamo nazioni come il Qatar dove il consumo è minimo considerando le restrizioni religiose che ci sono nel Paese.

Top e bottom 25 Stati per Wine_PerCapita

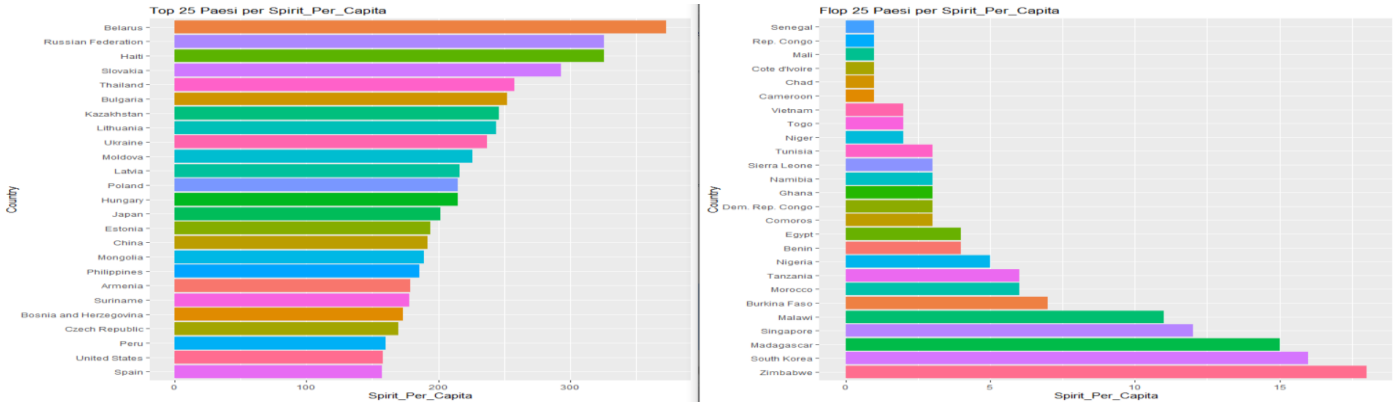




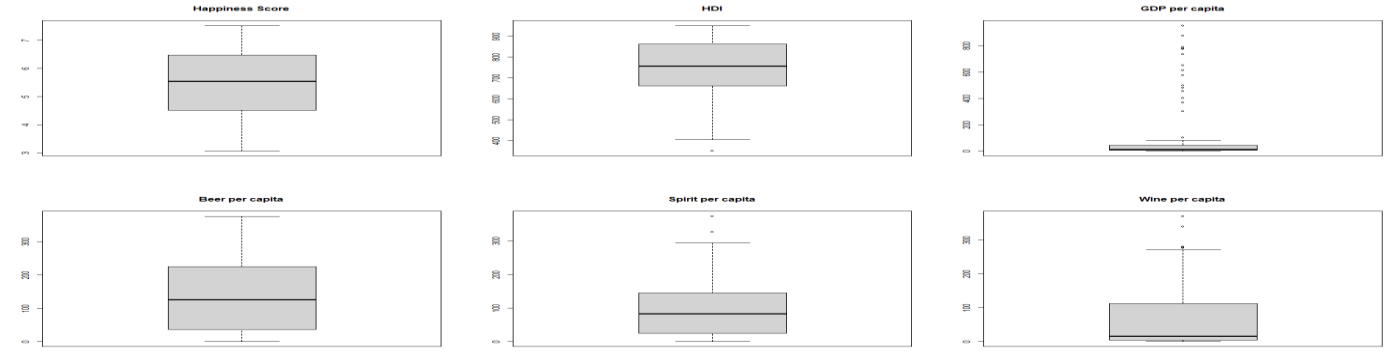
Con la variabile che indica il consumo medio a persona di vino, vale lo stesso discorso fatto per il consumo di birra. Però in questo caso notiamo subito che in testa ci sono Stati molto conosciuti per la produzione di vino, come Italia e Francia. In

successive analisi, come per esempio la PCA, troveremo ancora ciò che è stato appena detto. Per la natura della classifica degli ultimi Stati ci siamo posti delle domande riguardo all'effettiva precisione della raccolta dei dati, però abbiamo deciso di mantenere questa variabile perché, in base a ricerche aggiuntive, sappiamo che i risultati del ranking sono coerenti.

Top e bottom 25 Stati per Spirit_PerCapita

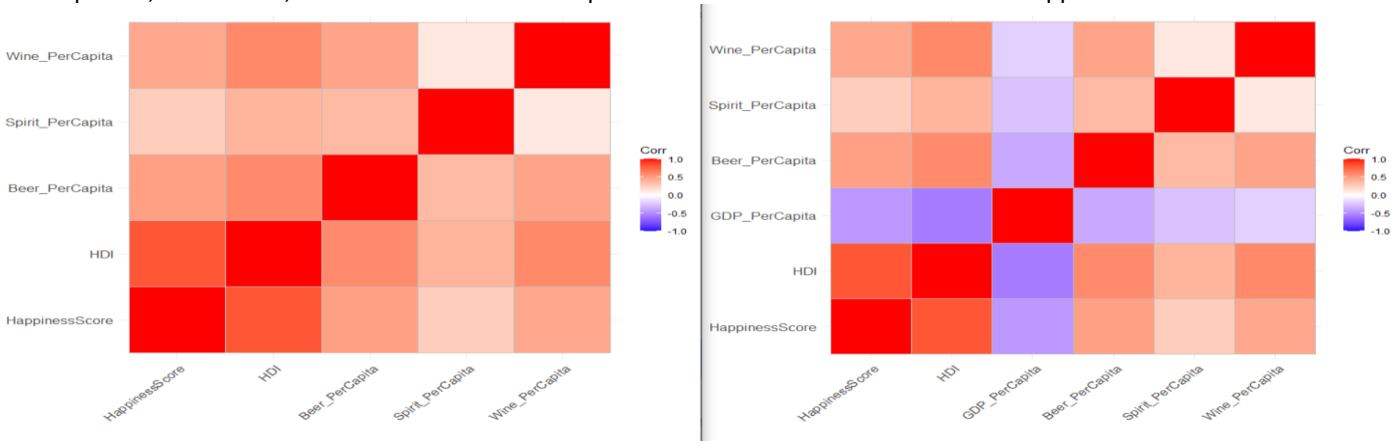


Qui vediamo Stati come Bielorussia e Russia abbastanza conosciuti per il consumo di superalcolici, come ad esempio la Vodka, che si trovano nelle prime posizioni della classifica. In particolare, per la Bielorussia vedremo come, nella PCA, si troverà in un punto dove potrebbe sembrare un outlier, probabilmente a causa del valore elevato che ha in questa variabile. Il Camerun si trova in fondo alla classifica e, guardando gli ultimi 25 Paesi, sembra che gli alcolici non siano molto popolari nell'Africa subsahariana. Questo perché molti Stati africani sono di religione musulmana, vi è disponibilità limitata e il costo che possono avere è elevato.
Continuiamo la nostra analisi esplorativa con l'aiuto dei boxplot, molto utili per vedere se sono presenti outliers nelle variabili quantitative.



Notiamo subito che, come detto prima, la variabile GDP_PerCapita presenta molti problemi e ci dà ulteriori motivi per escluderla dall'analisi. Invece le altre variabili non presentano, o quasi, nessun outlier e questo ci permette di continuare il lavoro senza troppi intoppi.
A parte la variabile GDP_PerCapita e, in maniera meno evidente, Wine_PerCapita, tutte le variabili sono distribuite normalmente.

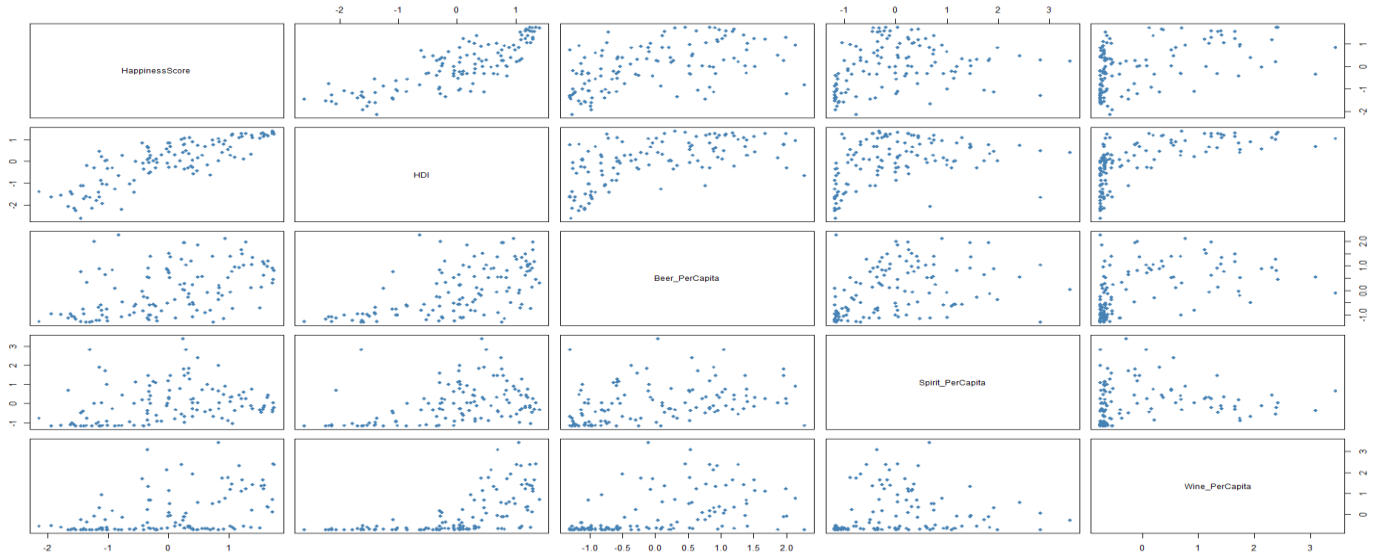
Come ultima procedura per l'analisi esplorativa, abbiamo effettuato due grafici di correlazione tra le variabili per identificare le relazioni tra esse. Abbiamo creato un grafico con la variabile GDP_PerCapita e uno senza. Vediamo a colpo d'occhio come solo la variabile PIL pro capite ha una correlazione negativa con le altre. Questa è l'ennesima dimostrazione di come questa variabile sia errata perché, idealmente, dovrebbe essere correlata positivamente con variabili come HDI e HappinessScore.



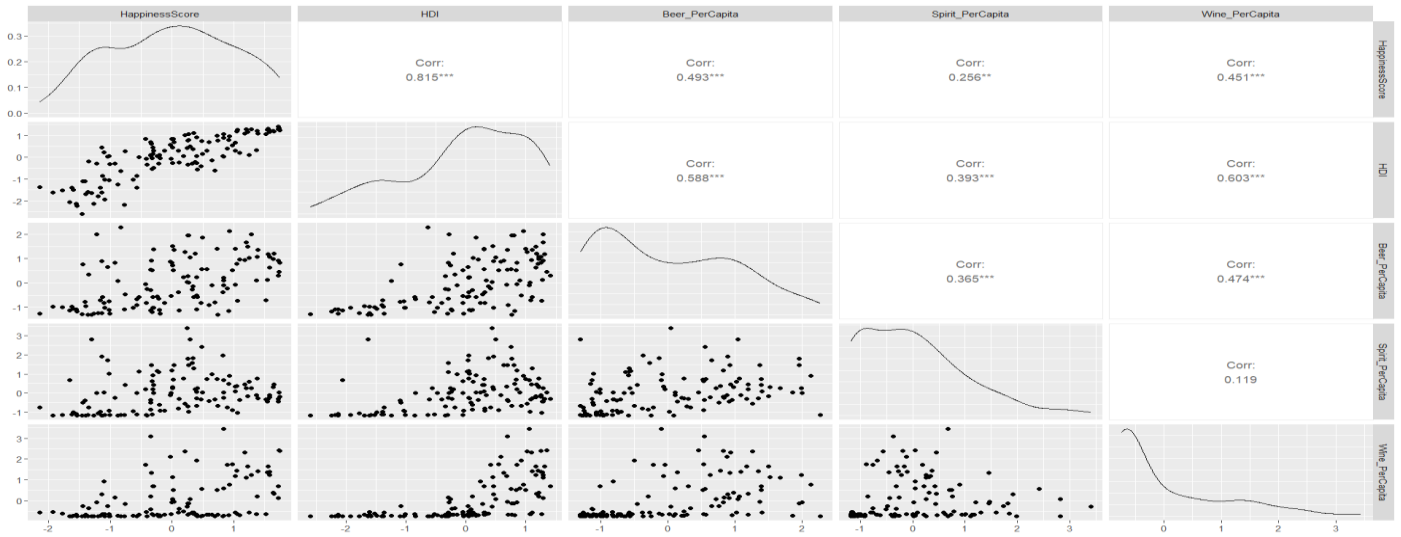
In sintesi, per concludere l'analisi esplorativa possiamo dire che il dataset utilizzato non presentava problemi a livello di valori NA o doppioni e le variabili sono tutte più o meno coerenti con i valori associati ai singoli Paesi e distribuite normalmente, eccetto per la variabile GDP_PerCapita che, per le ragioni elencate in precedenza, sarà esclusa dalle analisi che effettueremo.

Regressione lineare

Per studiare le variabili quantitative del dataset abbiamo usato la tecnica della regressione lineare. Questo passaggio è stato importante per conoscere quali variabili sono significative per il modello e fare stime sulla variabile dipendente. La variabile dipendente scelta è stata quella del 'HappinessScore' che ci è sembrata quella più rappresentativa da utilizzare in questa analisi poiché le altre variabili, ovvero quelle che identifichiamo come indipendenti, ci possono spiegare meglio come influiscono su quella dipendente, cioè, nel nostro caso, cosa influisce positivamente e in che modo la felicità di un popolo. Le variabili indipendenti usate sono tutte le altre quantitative viste in precedenza, meno GDP_PerCapita.
Per la regressione lineare abbiamo deciso di scalare il dataset poiché le misure delle variabili erano troppo diverse e quindi, di conseguenza, avrebbero potuto creare problemi visto il peso diverso che hanno tra loro. Prima di partire con la creazione del modello, vediamo come si relazionano le variabili tra loro con dei grafici a dispersione.



Possiamo subito notare come la relazione tra HappinessScore e HDI sembri essere lineare, mentre gli altri grafici si muovono in maniera più ambigua. Questo ci dice in particolare come all'aumentare dell'Indice di Sviluppo Umano aumenti anche la felicità mentre non è detto che all'aumentare dei litri consumati pro capite di birra, vino e superalcolici aumenti la felicità. Ecco un'altra serie di grafici che conferma ciò.



Ora, finalmente, andiamo ad effettuare la regressione lineare. Ecco quali sono i risultati.

```
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
```

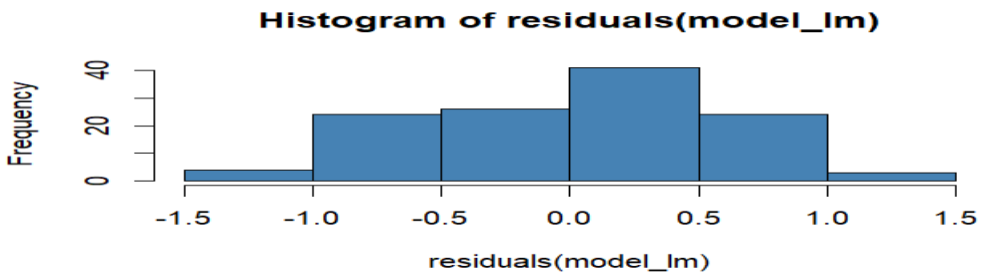
Come avevamo intuito prima, vediamo subito come l'HDI ha un forte effetto positivo sulla felicità, mentre le altre variabili non hanno un effetto significativo. Il valore di R-quadro e R-quadro adjusted è di circa il 0.67, il che significa che il modello spiega il 67% della varianza nella variabile indipendente e, quindi, possiamo dire che la bontà del modello è buona. La statistica F complessiva del modello è 60.85 e il corrispondente valore p è minore di 2.2e-16. Ciò indica che il modello complessivo è statisticamente significativo. In altre parole, il modello di regressione nel suo complesso è utile.
Per verificare che il risultato ottenuto sia esatto, abbiamo effettuato l'ANOVA test per mostrare i risultati dell'analisi della varianza per il modello di regressione lineare e ciò conferma il risultato ottenuto nella regressione lineare.

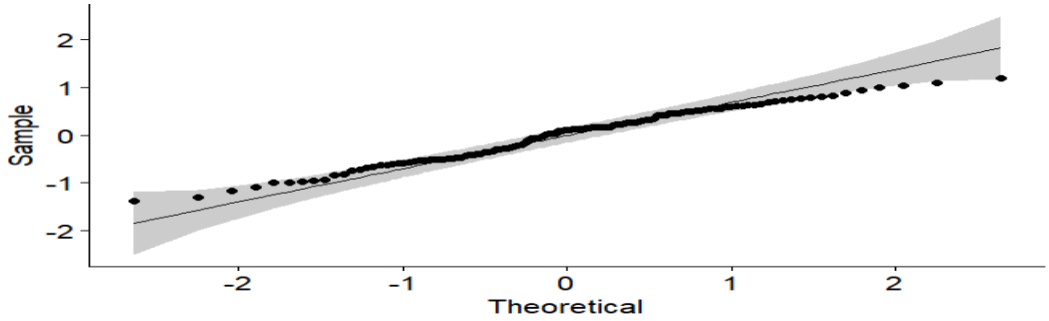
```
Analysis of Variance Table

Response: HappinessScore
                 Df Sum Sq Mean Sq  F value Pr(>F)
HDI               1 80.403  80.403 239.4928 <2e-16 ***
Beer_PerCapita    1  0.035   0.035   0.1039 0.7477
Spirit_PerCapita  1  0.666   0.666   1.9836 0.1617
Wine_PerCapita    1  0.616   0.616   1.8357 0.1781
Residuals       117 39.280   0.336
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Poi, per concludere con la regressione lineare in termini di valutazione dei coefficienti, abbiamo eseguito diversi test per valutare se ci fosse alta correlazione tra le variabili, i residui e, di conseguenza, l'omoschedasticità.
Parlando di correlazione tra variabili, grazie all'utilizzo del VIF test, cioè il fattore di inflazione della varianza, abbiamo visto come le variabili siano correlate moderatamente tra loro e questo non dovrebbe causare problemi significativi nell'analisi.

```
vif(model_lm)
              HDI   Beer_PerCapita Spirit_PerCapita
         2.119692         1.661492         1.274083
Wine_PerCapita
         1.694690
```

Per quanto riguarda i residui, possiamo affermare che essi seguono una distribuzione normale e sono lineari e quindi dire che il modello sia adeguato grazie a questi due grafici.





Sono stati fatti anche altri grafici per valutare la distribuzione dei residui, come ad esempio il density plot, e tutti confermano ciò che è stato affermato prima. Di conseguenza possiamo dire che c'è omoschedasticità e quindi, possiamo affermare nuovamente che il modello è buono. Per valutare l'omoschedasticità abbiamo utilizzato l'ncvTest e il bptest ed entrambi danno risultati coerenti.

```
Non-constant Variance Score Test          |        studentized Breusch-Pagan test
data:  model_lm                           |      Variance formula: ~ fitted.values
BP = 1.0061, df = 4, p-value = 0.9089     |  Chisquare = 0.007110081, Df = 1, p = 0.9328
```

In conclusione, possiamo dire che la regressione lineare ci conferma che il nostro modello è buono e i risultati ottenuti sono coerenti, anche grazie ai test che abbiamo svolto.

Cluster analysis

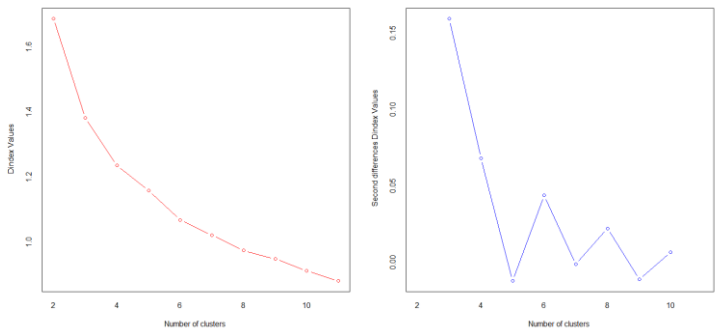
Analizzando il dataset da questo punto di vista ci siamo posti una domanda in particolare: quanto ogni paese effettivamente si discosta dagli altri in base alle variabili prese in considerazione? Ci siamo focalizzati principalmente su una analisi a cluster gerarchico e non gerarchico riscontrando alcune differenze significative.
Prima di tutto abbiamo analizzato il dataset e valutato se utilizzare la versione scalata o meno (prendendo in considerazione il fatto che GDP_PerCapita sia una variabile errata come riscontrato in analisi esplorativa). Possiamo comunque dire che anche in questo caso abbiamo scelto il dataset scalato come visto, anche qui, in regressione lineare e come vedremo qui di seguito nel calcolo delle distanze.
La prima cosa da sapere è che abbiamo utilizzato esclusivamente la distanza euclidea tra le variabili.

```
> summary(d1)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  4.154 185.567 269.110 281.000 362.809 707.848
```

Utilizzando il dataset non scalato, abbiamo un'altissima variabilità delle distanze e non ci sembrava un'ottima soluzione per poter procedere con questo tipo di approccio rendendo inutile o falsata la creazione di clusters.

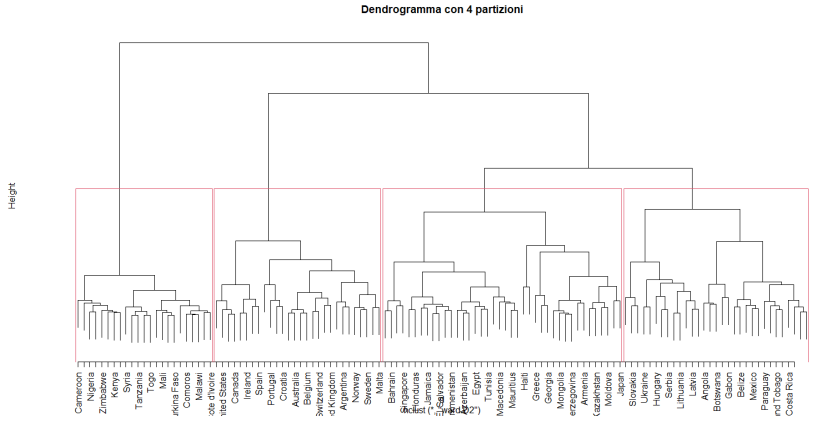
```
> summary(d)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.1721  2.0356  2.8670  2.9255  3.7429  6.4064
```

Per questo motivo abbiamo optato per il dataset scalato.
A questo punto abbiamo sfruttato due indici (Hubert Index e il D Index) che ci hanno suggerito quanti cluster scegliere.

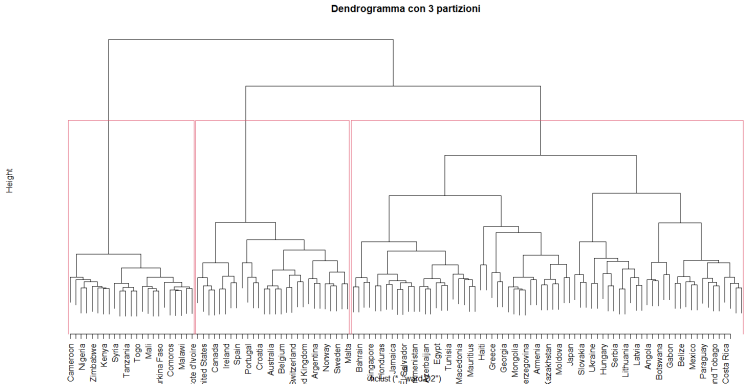


Come possiamo notare, i grafici, ma anche il codice stesso, ci suggeriscono che la miglior partizione in cluster è 3 (sia nel primo che nel secondo grafico vediamo una maggiore distanza tra il primo e il secondo punto e questo ci fa comprendere che, nel passaggio tra 3 e 4 cluster gli indici si abbassano notevolmente).

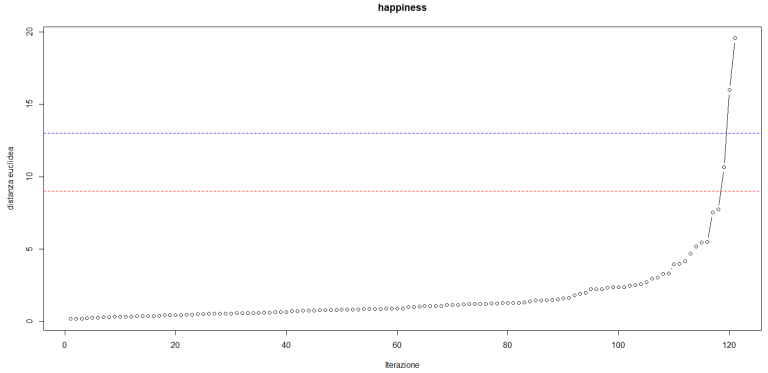
Successivamente abbiamo proceduto con la costruzione del dendrogramma. Abbiamo utilizzato il metodo Ward in quanto abbiamo ritenuto fosse il più adatto per massimizzare la devianza tra i gruppi (Between) e minimizzare quella all'interno dei gruppi (Within).



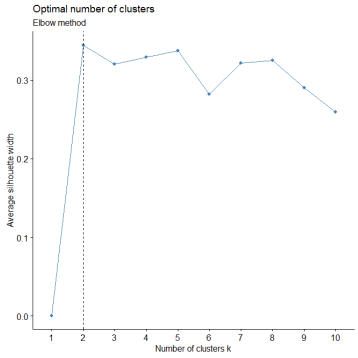
Come possiamo notare a prima vista i 4 gruppi formati sono pressocché omogenei. È possibile notare come gli Stati, in base a nostre conoscenze pregresse della natura politico-sociale e anche a ricerche conseguite successivamente, si distribuiscono in modo omogeneo in base al proprio continente di riferimento. Quindi possiamo vedere che i Paesi occidentali si trovano per una buona parte in uno stesso cluster. È possibile però che alcuni Paesi non si trovano nel proprio cluster di "riferimento". A questo punto abbiamo avanzato l'ipotesi di diversi tipi di approcci al consumo di alcol da parte dei cittadini, le diverse tipologie di regolamentazioni da parte dei singoli Stati o i limiti imposti dalle religioni. Qui sotto abbiamo analizzato il dendrogramma con una partizione di 3 gruppi.



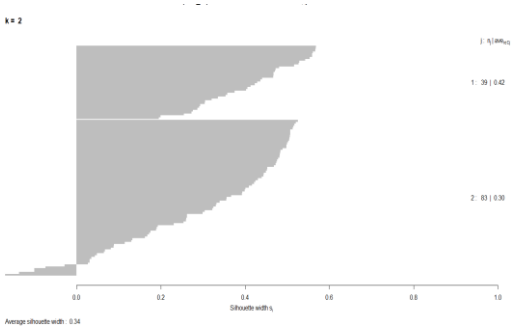
Al contrario qui notiamo una differenza principalmente per quanto riguarda l'ultima partizione, in quanto è estremamente più grande rispetto agli altri due.



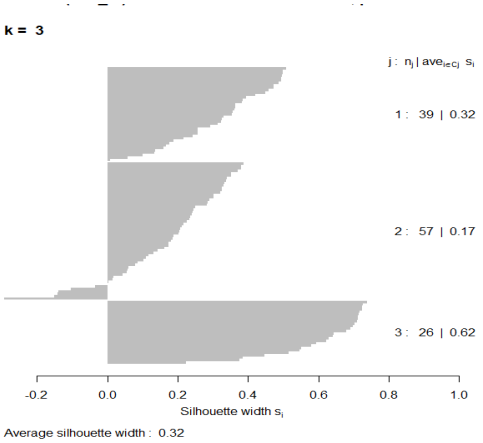
Come possiamo notare anche in questo grafico, la migliore partizione che ci suggerisce è proprio quella con 3 clusters, in quanto il salto maggiore che vediamo è proprio all'altezza di 13 (h=13), rispetto alla formazione di 4 clusters che comunque si presenta in una forma migliore proprio perché i gruppi sono più omogenei.
Passando al metodo del k-means, per prima cosa abbiamo analizzato quanti gruppi è preferibile scegliere.



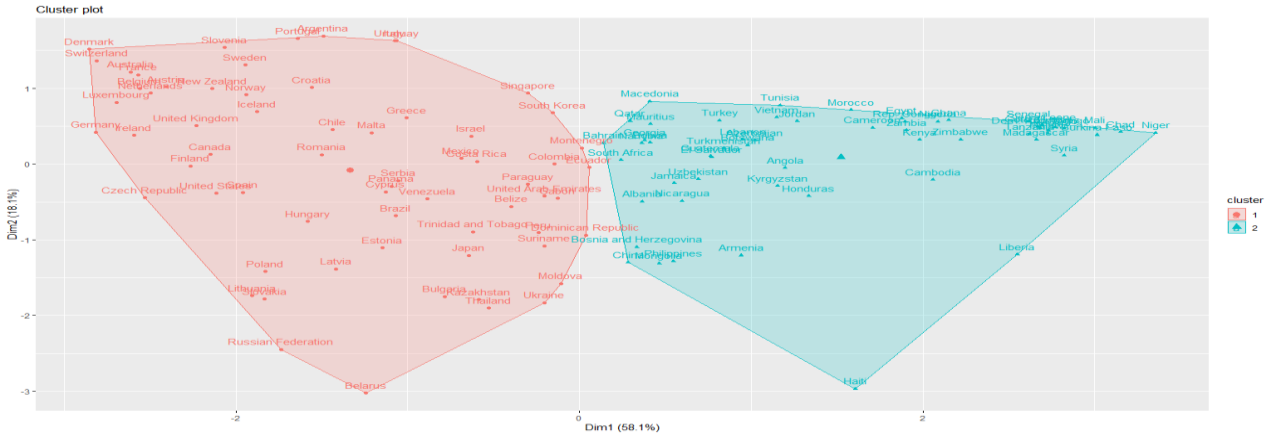
In questo caso possiamo notare che il grafico forma un gomito (elbow) in prossimità del numero di clusters ottimali (ovvero 2). Analizzando anche la Silhouette vedremo in che modo vengono distribuite le variabili all'interno dei gruppi e se ci saranno degli errori nella classificazione.
In questo caso abbiamo optato per una analisi di 2 e 3 formazioni di gruppi in quanto ci è sembrato più completo mostrare anche come si formano 3 gruppi e confrontarli con il metodo gerarchico.



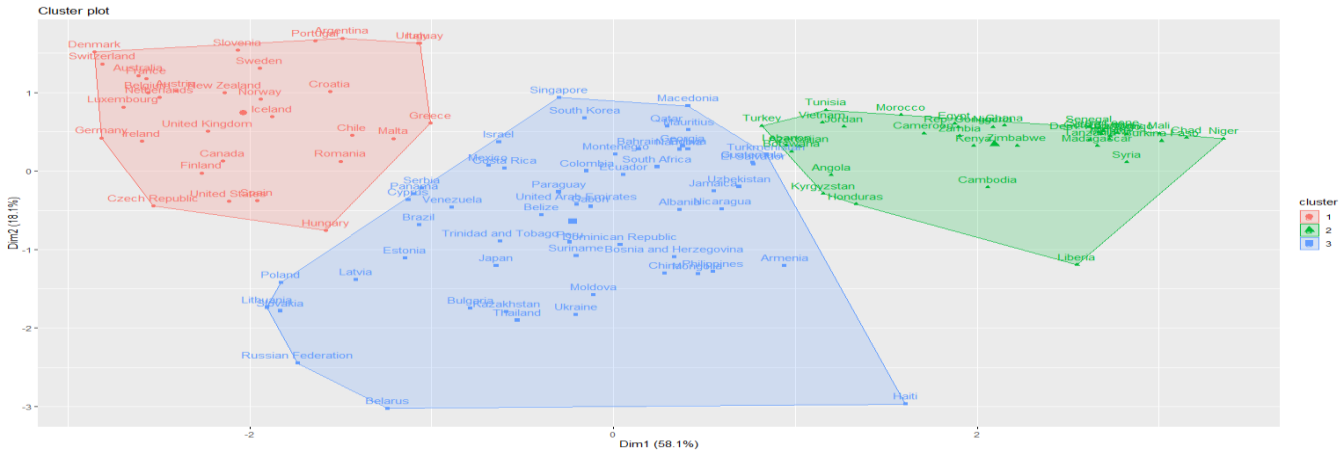
Nella figura sopra abbiamo composto la Silhouette per 2 gruppi. Come possiamo vedere la composizione dei 2 cluster è soddisfacente nonostante la piccola coda che si trova nel secondo gruppo. Possiamo dire quindi che questo tipo di formazione è più che buona.



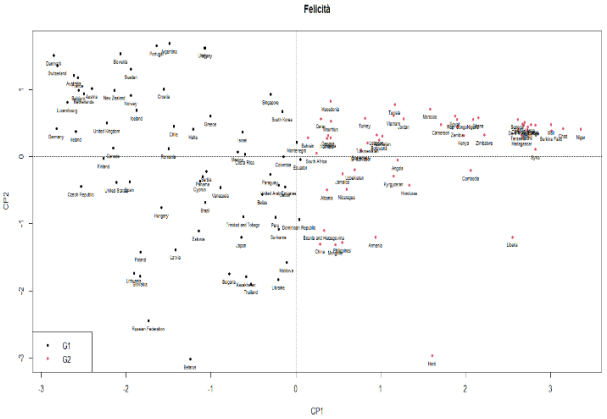
In questo caso abbiamo una buona formazione per il terzo gruppo. I primi due hanno una costituzione accettabile, a parte per la piccola coda presente nel secondo gruppo ed infatti anche qui le unità sono state inserite nei rispettivi gruppi pressocché correttamente. La sbagliata classificazione per entrambe le formazioni non è così male da rifiutarle completamente; quindi, procediamo con la creazione dei gruppi veri e propri.
In questo caso vedremo una divisione abbastanza netta ma allo stesso tempo la distanza tra i gruppi non è così grande come ci saremmo aspettati.

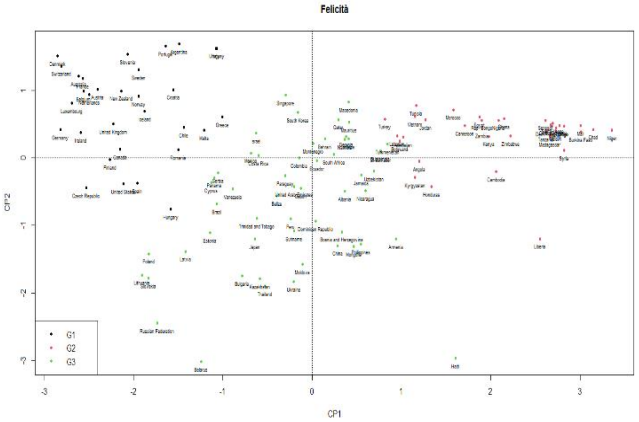


Qui possiamo vedere che effettivamente i due gruppi creati sono ben distinti e a primo impatto riconosciamo una netta distinzione. Nel primo gruppo (a sinistra) si trovano i Paesi cosiddetti sviluppati, che hanno un livello di sviluppo e "felicità" maggiori rispetto a Paesi del terzo mondo che si trovano a confrontarsi con gravi problematiche sociopolitiche e culturali; infatti, in questi ultimi Paesi si può notare un grande consumo di alcol (principalmente birra o super-alcolici) sia per mancanza di istituzioni che possano regolamentarne il consumo, sia per il degrado stesso di molte zone di questi Paesi. Allo stesso tempo, sempre a sinistra, troviamo Paesi che non sono quasi del tutto sviluppati però potrebbero avere un rapporto abbastanza "maturo" con l'alcol da non avere problematiche particolari di abuso. Per completezza, la composizione a 2 gruppi non presenta alcun elephant clusters (65 e 57 unità) e le distanze within sono 221.06 – 134.37 e la distanza between è 249.56.



In questo caso vediamo una maggiore distinzione tra i Paesi sviluppati (a sinistra), Paesi in via di sviluppo o del terzo mondo (a destra) e Paesi sviluppati o meno (al centro) che in un modo o nell'altro hanno un approccio all'alcol diverso da quelli "occidentali" (ovvero non hanno una cultura dell'alcol come in Europa, ad esempio, o hanno delle regolamentazioni diverse). In questo caso la grandezza dei gruppi è appena più sproporzionata rispetto alla classificazione di 2 clusters (31, 36 e 55 unità nei gruppi) e le distanze within sono 57.07 – 44.93 – 161.52 e la distanza between è 341.46.
Passando alla PCA abbiamo notato che comunque ci sono variabili che pesano più di altre. Prima di entrare in merito a questo tipo di analisi ci sembrava opportuno puntualizzare ancora una cosa in merito alla cluster analysis.

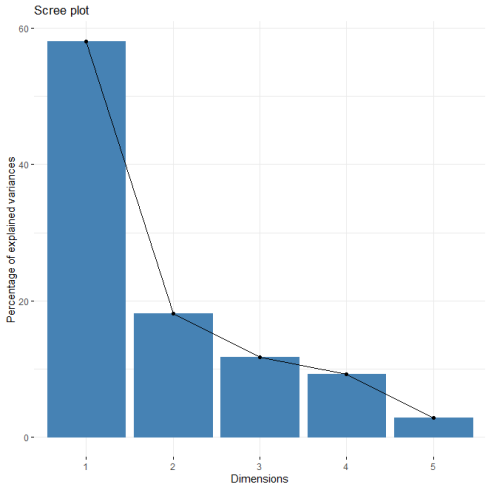


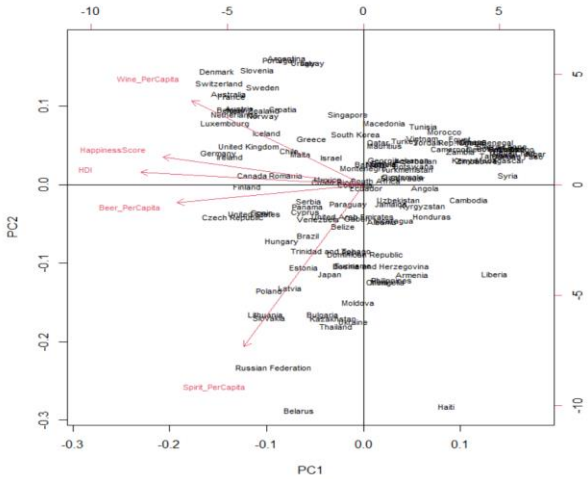


In questo caso possiamo notare una divisione tra 2 e 3 gruppi concordante con l'analisi del k-means. Le prime due componenti principali sono quindi sufficienti per spiegare anche la cluster analysis. In quanto alla interpretazione delle componenti 1 e 2 ci ritorneremo in seguito.

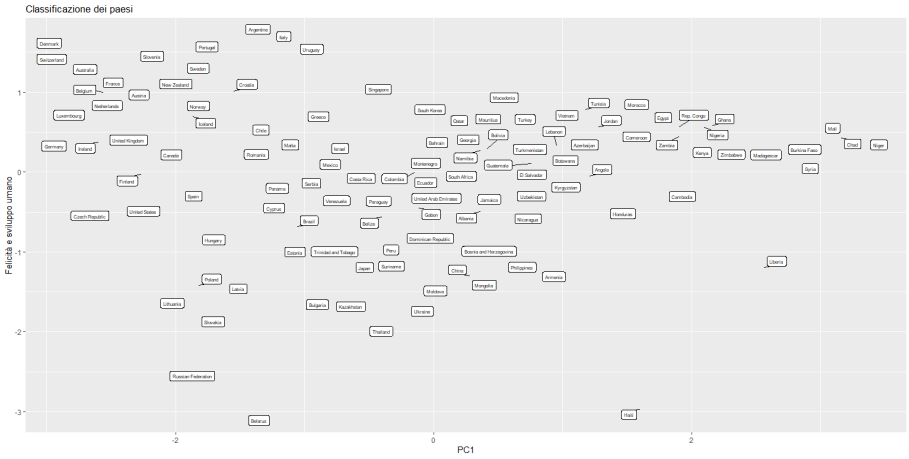
PCA

Abbiamo riscontrato una buona struttura per quanto riguarda la PCA. Prima di tutto abbiamo visto che le prime 2 componenti principali riescono a spiegare quasi l'80% della varianza. Questo si può notare anche nello scree plot di seguito, in cui la prima componente principale spiega di per sé il 58% della varianza.





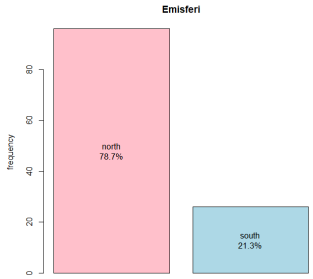
Successivamente abbiamo composto il biplot per capire come interpretare le prime 2 componenti principali.
Possiamo notare come per quanto riguarda la prima componente principale tutte le variabili influiscono negativamente, in primis HappinessScore e HDI. Non siamo riusciti purtroppo a capire per quale motivo abbiamo avuto questo tipo di risultato, ma è fortemente significativo il fatto che proprio le 2 variabili che ci saremmo aspettati che avessero influito positivamente, sono proprio quelle che hanno l'effetto contrario. Abbiamo pensato che questo tipo di risultato sia congenito al dataset stesso (infatti più di una volta, nel corso delle analisi di altri dataset, ci siamo ritrovati con problemi del genere, in cui variabili come HappinessScore, HDI o GDP hanno segni negativi nelle analisi delle componenti principali). Di conseguenza non siamo riusciti propriamente a dare una interpretazione vera e propria alla prima componente principale. Diverso è il discorso per la seconda componente, infatti notiamo subito come Happinessscore e HDI influiscono positivamente ma non tanto quanto la variabile Wine_PerCapita. In merito a questa ultima variabile abbiamo pensato che non fosse tanto un'influenza ma più una conseguenza. Infatti, abbiamo notato che i Paesi più sviluppati come Italia, Francia, Portogallo e altri Paesi europei tendono ad avere un maggiore grado di HappinessScore, HDI e Wine_PerCapita. Ci sembrava dunque che il vino non influisse tanto sulla felicità delle persone ma è più una conseguenza di quest'ultima in quanto in Paesi più ricchi si tende a consumare alcol più "pregiato", in più ovviamente abbiamo tenuto in considerazione anche la cultura legata al vino di ogni paese.



Infine, qui vediamo la classificazione di ogni paese attraverso le influenze delle prime 2 componenti principali. In questo caso possiamo dire che ovviamente non si discostano dalle analisi fatte in precedenza in quanto i Paesi si trovano nella stessa posizione (quindi le 2 componenti principali sono sufficienti per spiegare l'intero dataset).

Regressione logistica

Il nostro obiettivo è quello di trovare un modello di regressione logistica che permetta di prevedere l'emisfero di appartenenza delle Nazioni utilizzando le variabili che abbiamo a disposizione.
Andremo ad effettuare i vari test per vedere se i modelli proposti possono essere adatti a effettuare la nostra previsione. La prima difficoltà che riscontriamo è che il nostro dataset è particolarmente sbilanciato.



Andremo successivamente a valutare la curva Roc per trovare il threshold ottimale da applicare alla logistica.
Per prima cosa sostituiamo ai valori "south" e "north" della variabile "Hemisphere" rispettivamente "0" e "1" e dividiamo il dataset in training (75% delle osservazioni) e testing (25%).

Applichiamo il modello di regressione logistica contenente tutte le variabili (full_model) al dataset "train". Effettuiamo il *likelyhood ratio test* tra il full_model e diversi nested model, concludiamo che il modello ideale è quello che ha come regressori: Beer_PerCapita + Spirit_PerCapita + HDI; chiameremo questo modello logit2.
Il Wald test su logit2 ci suggerisce che abbiamo abbastanza evidenza empirica per poter rifiutare H0, di conseguenza i regressori sono statisticamente significativi per il nostro modello.
Lo pseudo-R2 di Mcfadden assume un valore di 0.20 circa, appena sopra la soglia di accettazione,
i valori dei vif sono sotto a 3 per ogni variabile, questo significa che le nostre variabili non sono correlate.

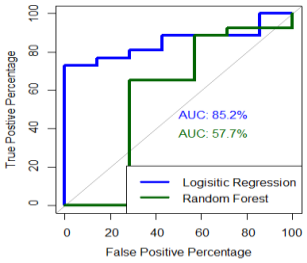
```
Wald test:
----------

Chi-squared test:
X2 = 8.9, df = 3, P(> X2) = 0.03
```

```
> pR2(logit2)["McFadden"]
fitting null model for pseudo-r2
 McFadden
0.2011446
> vif(logit2)
  Beer_PerCapita Spirit_PerCapita             HDI
        2.746580         1.807144        2.439175
```

Esaminiamo l'importanza delle variabili

```
> varImp(logit2)
                 Overall
Beer_PerCapita  2.265563
Spirit_PerCapita 2.341091
HDI             1.871687
```

Ora utilizziamo la funzione predicted e diamo in pasto al modello "logit2" il test set, successivamente calcoliamo la curva Roc relativa al modello logit2 e quella relativa a un modello Random forest, entrambe le Roc-curve sono calcolate sul test set.

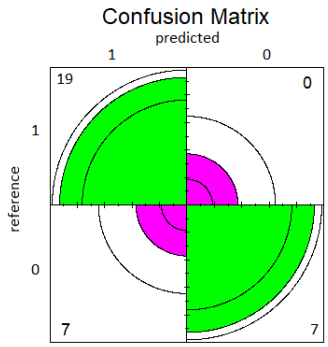


L'indice AUC "area under the curve" ci mostra che in questo caso la Random Forest non è un buon modello predittivo. Andiamo ora a valutare nella roc blu qual è il miglior threshold. Costruiamo un dataframe costituito da tre colonne True positive percentage (TPP), False Postitive Percentage (FPP) e threshold.

Il nostro obiettivo è massimizzare TPP e minimizzare FPP, osservando il grafico vediamo che questo punto si trova in un intorno di TPP: (65<TPP<75).

```
> roc.df[roc.df$tpp > 65 & roc.df$tpp < 76,]
        tpp      fpp thresholds
14 73.07692 14.28571  0.8547989
15 73.07692  0.00000  0.8638175
16 69.23077  0.00000  0.8682961
17 65.38462  0.00000  0.8802020
>
```

Trovato il nostro threshold ottimale (0.8638175) lo applichiamo alla nostra funzione predicted e andiamo a costruire la matrice di confusione.



tp	fn
fp	tn

Accuracy	0.789
Precision	0.731
Sensitivity	1.000
specificity	0.500

In conclusione, possiamo dire che il modello è di qualità discreta, commette alcuni errori nel classificare bene Paesi appartenenti all'emisfero Sud ma ce lo aspettavamo data la misera numerosità del campione e lo sbilanciamento del dataset.

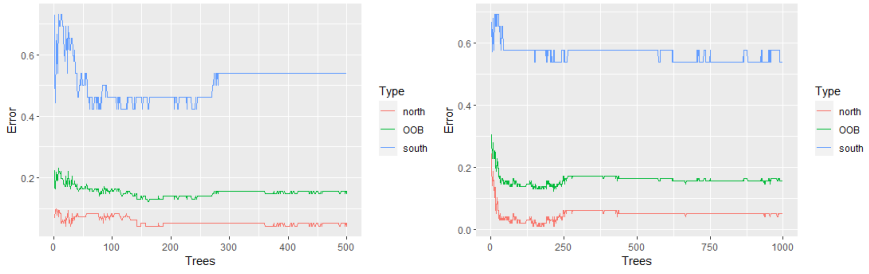
Random Forest

In questa fase dell'analisi utilizzeremo il metodo della Random Forest sull'intero dataset (122 Paesi) per predire se le Nazioni sono locate nell'emisfero boreale o australe (come nella logistica); di conseguenza la variabile Hemisphere sarà la nostra variabile dicotomica dipendente. Dal dataset originario elimineremo la variabile GDP (dopo aver verificato essere "fallata") e la variabile HappinessScore (poiché eccessivamente correlata con HDI) e utilizzeremo le restanti. Il nostro obiettivo è dimostrare che la numerosità del campione preso in partenza incide sulla qualità del modello.

Effettuiamo la prima RF con 500 alberi e mtry = 2, otteniamo un OOB (Out Of Bag) err.rate di 15,57%.
Effettuiamo la seconda RF con 1000 alberi, mtry = 2 ottenendo un OOB err.rate di 15,75% e una matrice di confusione identica.

```
        OOB estimate of  error rate: 15.57%
Confusion matrix:
      north south class.error
north    91     5  0.05208333
south    14    12  0.53846154
```

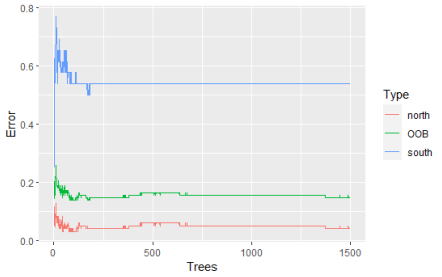
La linea blu indica l'error rate commesso per aver classificato "south" alcune nazioni che sono OOB.
La linea verde l'OOB error rate generico.
La linea rossa indica l'error rate commesso per aver classificato "north" alcune nazioni che sono OOB.



Non contenti dell'instabilità dei nostri errori, ci spingiamo oltre e cerchiamo di calcolare un Random Forest con 1500 alberi e mtry = 2. Otteniamo un OOB err.rate di 14,75% e una matrice di confusione migliorata

```
        OOB estimate of  error rate: 14.75%
Confusion matrix:
      north south class.error
north    92     4  0.04166667
south    14    12  0.53846154
```

Accuracy	0.852
Precision	0.867
Sensitivity	0.958
specificity	0.461



Notiamo che l'OOB error rate generico (linea verde) migliora leggermente ma LlOOB error rate relativo all'emisfero sud (linea blu) si assesta comunque ad un valore realtivamente alto (0.53846154).
Al contrario, la linea rossa, che rappresenta l'error rate che viene commesso nel classificare in "nord" i Paesi che fanno parte del'OOB dataset mantiene sempre un livello particolarmente basso.
Per ciò che concerne i parametri della matrice di confusione osserviamo un leggero miglioramento di Accuracy e Precision rispetto a ciò che abbiamo ottenuto dalla regressione logistica.

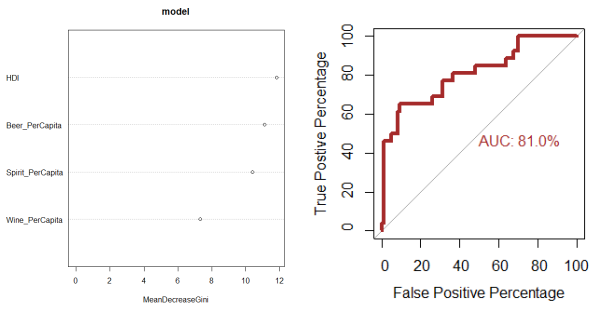
Ora andiamo a valutare il numero ottimale di mtry, cioè di variabili utilizzate per la costruzione degli alberi
Per fare ciò scegliamo quel numero di variabili che ci fa ottenre un OOB error rate generico minore.
L'mtry ottimale risulta pari ad 1, runniamo il nostro modello da 1500 alberi e mtry = 1 e otteniamo il seguente risultato:

```
        OOB estimate of  error rate: 14.75%
Confusion matrix:
      north south class.error
north    94     2  0.02083333
south    16    10  0.61538462
```

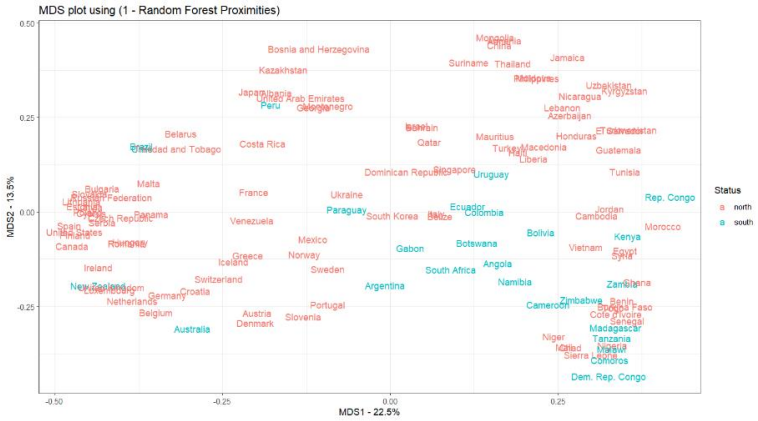
C'è un leggerissimo miglioramento dell'OOB error rate relativo all'emisfero nord a discapito di un aumento dell'OOB error rate relativo all'emisfero sud; decidiamo quindi di escludere quest'ultimo modello e di mantenere quello immediatamente precedente con un mtry = 2.

Utilizziamo la funzione VarImpPlot per valutare l'importanza delle nostre variabili e notiamo con sorpresa che la RF ritiene più importante la variabile HDI, al contrario della logistica in cui HDI era considerata meno importante.

Osserviamo ora il valore AUC della curva ROC e notiamo un netto miglioramento rispetto alla Random Forest implementata con soltanto il 25% delle osservazioni, questo significa che il modello è migliore nel classificare i samples.



Infine, andiamo a valutare la PCoA (Principal Coordinate Analysis) relativa al nostro modello di RF, per fare ciò covertiamo le distanze tra i sample in un grafico 2-D.
Notiamo che l'asse x e y del grafico sono poco rappresentativi della varianza, questo perché le variabili che abbiamo utilizzato sono poco correlate.
Si accenna una leggera dipendenza non lineare e c'è una polarizzazione dei Paesi dell'Emisfero Sud in basso a destra.
Ci accorgiamo della presenza di alcuni outliers come Brasile, Nuova Zelanda e Australia che si collocano nella nuvola di Paesi appartenenti all'Emisfero Nord. Questo può significare che, date le variabili prese in considerazione, il modello riconosce queste Nazioni più simili a quelle dell'emisfero nord.



Conclusione

Per concludere possiamo affermare che parametri come HDI, HappinessScore, consumo di alcol, birra e vino per capita possono fungere da discreti discriminanti nel suddividere e raggruppare le Nazioni nei rispettivi emisferi. Curiosa è la vicinanza di Paesi come Australia, Nuova Zelanda, UK e altri Paesi anglofoni che, pur appartenendo ad emisferi diversi vengono comunque aggregati; questo è intuitivamente legato a una somiglianza di alcuni Paesi nei parametri come il consumo di birra e HDI.

Tuttavia, sarebbe auspicabile avere a disposizione altre variabili in più in modo da ottenere una suddivisione migliore e una migliore caratterizzazione che distingue i Paesi dell'emisfero boreale da quello australe.