UNIVERSITÁ DEGLI STUDI DI MILANO-BICOCCA

Dipartimento di Informatica, Sistemistica e Comunicazione

Corso di laurea in

Data Science

RELAZIONE FOUDATION of PROBABILITY and STATISTICS

Relatori:

Omar Gallo 808203

Eric Spinelli 799827

Alessandro Cerni 861474

Anno Accademico 2019/2020

Introduzione

Il dataset oggetto di studio si chiama “Happines and Alcohol Consumption”, la fonte è il sito Kaggle. Il dataset è composto da 122 osservazioni, per cui si hanno 9 variabili:

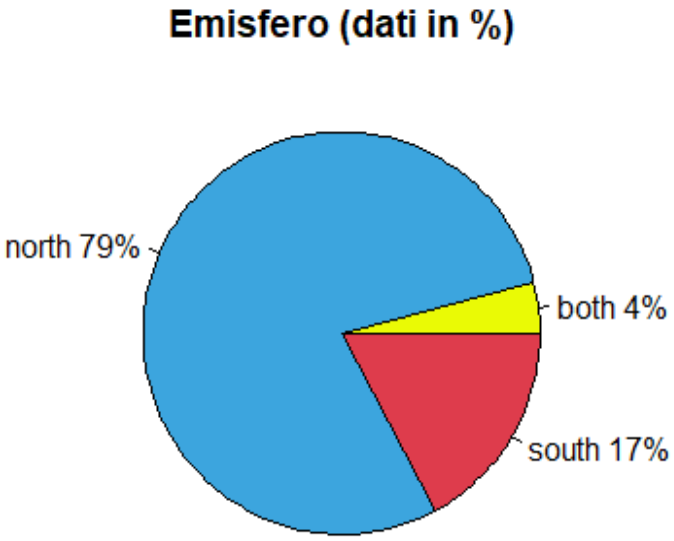
- *Country:* è l’informazione sulla nazione presa in considerazione è la chiave del dataset, univoca per ciascuna osservazione.
- *Region:* è una suddivisione in aree geografiche relativa alla nazione, è suddivisa in 9 modalità: *Australia and New Zealand, Central and Eastern Europe, Eastern Asia, Latin America and Caribbean, Middle East and Northern Africa, North America, Southestern Asia, Sub-Saharan Africa, Western Europe.*
- *Hemisphere:* è una variabile di 3 modalità in cui si indica l’emisfero di appartenenza di ciascuna nazione: *North, South, Both.* In particolare, la categoria *both* indica che la nazione si trova parzialmente in entrambi gli emisferi.
- *HappinessScore:* è una variabile quantitativa continua, è un indice di felicità relativo a ciascuna nazione
- *HDI:* è una variabile quantitativa continua con valori da 0 a 1, è anche detta Indice Di Sviluppo Umano e dà un’indicazione dello sviluppo di ciascuna nazione. L’HDI è calcolato rispetto a delle variabili indicative dello sviluppo di un Paese, ad esempio il PIL, la salute pubblica, l’istruzione.
- *GDP:* si intende il PIL pro capite riferito a ciascuna nazione, varabile quantitativa continua. Il prodotto interno lordo è un indicatore generalmente usato per esprimere il livello di ricchezza per abitante prodotto da un paese in un determinato periodo, e che consente di operare confronti tra aree di dimensione demografica differente.
- *Beer_PerCapita:* indica il consumo in litri medio pro capite di birra per ciascuna nazione, variabile quantitativa discreta.
- *Spirit_ProCapita:* indica il consumo in litri medio pro capite di liquore per ciascuna nazione, variabile quantitativa discreta.
- *Wine_PerCapita:* indica il consumo in litri medio pro capite di vino per ciascuna nazione, variabile quantitativa discreta.

I dati raccolti si riferiscono al 2016. Si vuole specificare che il World Happiness Report è una pubblicazione annuale delle Nazioni Unite, da cui derivano i dati raccolti riguardo alla felicità. Periodicamente la World Health Organization pubblica un report relativamente a delle stime di quanto alcol viene consumato in ogni paese in riferimento alle tre sostanze alcoliche sopra menzionate.

Il dataset di partenza palesava due errori che abbiamo corretto prima di svolgere le analisi: il GDP pro capite era relativo alla valuta di ciascun paese di riferimento, l’abbiamo quindi convertito rispetto al dollaro americano; l’HDI era erroneamente considerato da 0 a 1000, quindi l’abbiamo convertito ai suoi valori originali.

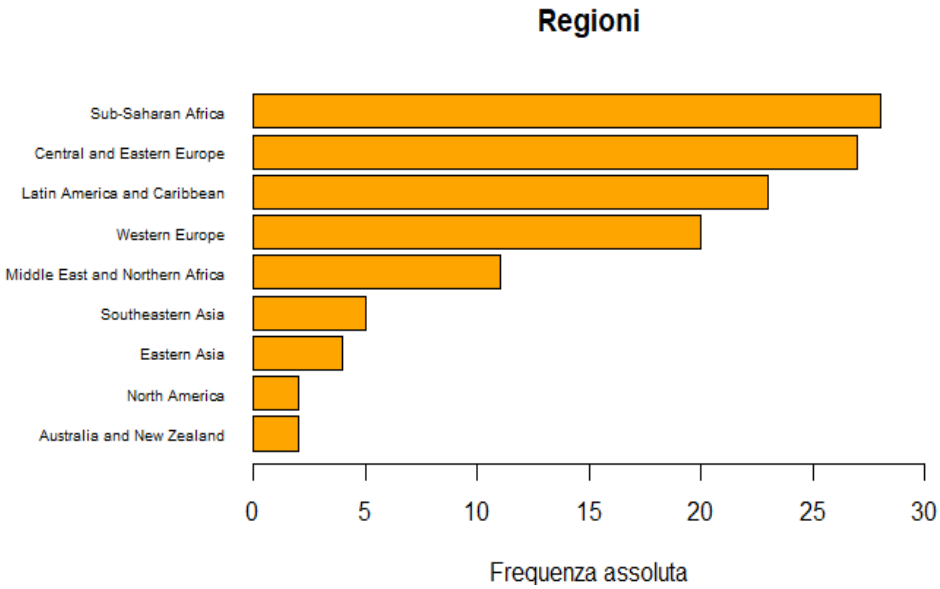
Variabili qualitative

La numerosità della popolazione per ciascuna delle classificazioni relative alla variabile *Hemisphere* evidenzia una sproporzione a favore delle nazioni collocate interamente nell'emisfero Boreale (North). Queste, infatti, rappresentano il 79% della popolazione e sono quindi la classe modale. A questa percentuale, si affianca il dato delle nazioni la cui estensione ricade sia l'emisfero Boreale che quello Australe (both), a cui appartiene il 4% delle nazioni in studio. La parte residuale, vale a dire il 17%, si colloca invece interamente nell'emisfero Australe (south).



In termini numerici, alla regione north appartengono 96 nazioni, 21 al south e 5 ad entrambe.

Affiancato al dato relativo all'emisfero, è presente un ulteriore informazione geografica di tipo qualitativo, nella fattispecie una suddivisione in 9 sub-aree geografiche o regioni. La classe modale è la Sub-Saharan Africa, con 28 record, con una frequenza relativa pari a 0.23.



Variabili quantitative

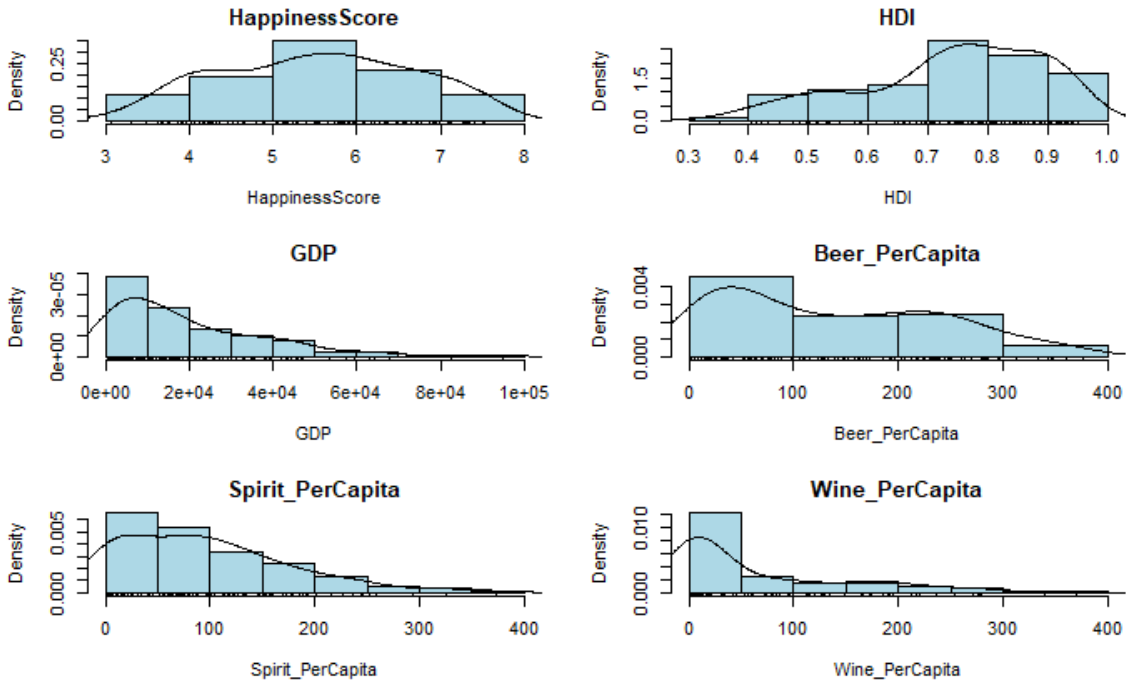
Passando alla trattazione delle variabili di tipo quantitativo, la prima riportata nel dataset rappresenta l'happiness score. I valori assunti da questa variabile sono racchiusi in un range tra 3.069 (Syria) e 7.526 (Denmark). Dividendo in classi la variabile *HappinessScore* secondo la suddivisione "Freedman-Diaconis" si nota che la frequenza più alta ricade nella classe il cui intervallo è compreso tra 5 e 6. Analizzando inoltre i valori della media e della mediana, rispettivamente pari 5.525 e 5.542, si nota un discostamento lieve dei due indici, rendendo la distribuzione lievemente positivamente asimmetrica ma tendente alla distribuzione Normale. In virtù di questo, le frequenze associati alle classi periferiche tendono invece a diminuire similmente in entrambe le code.

Per quanto riguarda l'HDI (Human Development Index), il range di valori è delimitato da un minimo corrispondente a 0.3510 associato al Niger ed un massimo di 0.9510 associato alla Norvegia. In questo caso, la distribuzione dei valori risulta essere negativamente asimmetrica, con una media dei valori pari a 0.7409 a fronte del valore assunto dalla mediana, pari a 0.7575. La classe modale nella suddivisione scelta è 0.7-0.8.

Analizzando il GDP, il cui range è compreso tra i valori limite di 444.4 (Dem. Rep. Congo) e 94920.99 (Luxembourg), si nota un comportamento estremamente asimmetrico, con alta frequenza associabile alla classe di valori compresi tra 0 e 10000$ (classe modale) e decrescente procedendo con le classi successive. La distribuzione risulta quindi positivamente asimmetrica, con valori di media e mediana discretamente differenti, il primo pari a 19824.6 ed il secondo pari a 14098.7.

Le ultime tre variabili fanno invece riferimento al consumo di alcolici l/anno per abitante. Per tutte e tre le variabili, vale a dire consumo di birra, liquori e vino, vi è una distribuzione di frequenze sempre positivamente asimmetrica, ma di differente natura. La variabile *Beer_PerCapita* assume maggiore frequenza nella classe 0-100 (classe modale), decrescendo all'aumentare della quantità. La variabile *Spirit_PerCapita* assume la classe modale in 0-50, decrescendo anch'essa. L'andamento della variabile *Wine_PerCapita* desta particolare attenzione concentrando le sue frequenze nella classe modale 0-50, in cui ha una frequenza relativa di 0.6.

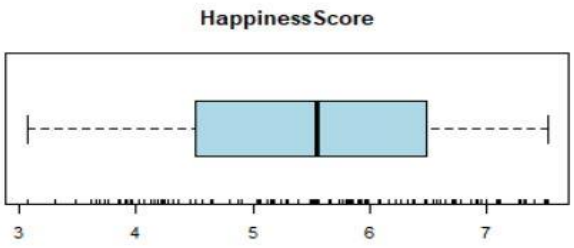
Andando poi a valutare i valori di limite, per tutte e tre le variabili il valore di minimo è pari a 1, mentre il valore massimo è simile tra le tre (rispettivamente 376 per la birra, 373 per i superalcolici e 370 per il vino). Analizzando invece le medie e le mediane, si nota ancora un maggior disallineamento per quanto riguarda il consumo del vino. Nello specifico, se per la birra e i liquori lo scostamento tra media e mediana è relativamente piccolo, essendo il consumo medio di birra pari a 137.57 e il valore mediano di 125.5 ed il consumo medio di liquori pari a 96.6 e il valore mediano di 82.5, per quanto riguarda il vino la differenza è assai marcata, essendo il consumo medio di 66.6 ed il valore mediano di 16.0.

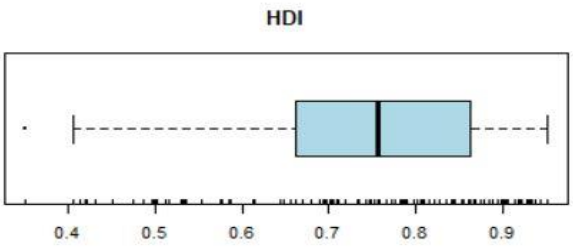


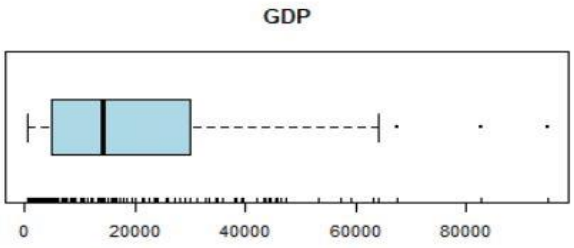
Fatte queste prime considerazioni sulle distribuzioni delle variabili si può mettere a confronto i coefficienti di variazione rispetto alle variabili. Il coefficiente di variazione CV permette di mettere a confronto la variabilità delle variabili. Si nota quindi una certa somiglianza tra la variabilità di *HappinessScore* e quella di *HDI,* rispettivamente di 0.2079 e 0.2023. Altre variabilità confrontabili sono quella di *GDP* (0.9474) e *Spirit_PerCapita* (0.843). La variabilità più estrema è quella di *Wine_PerCapita* di 1.323.

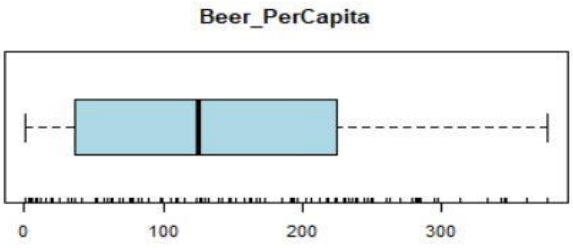
Osservando i boxplot di seguito rispetto a ciascuna variabile possiamo trovare conferma dei commenti fatti in precedenza. Inoltre, è utile notare gli outliers: rispetto ad *HDI* si ha un outlier oltre il baffo sinistro, che è il Niger; riguardo al *GDP* ci sono 3 regioni con un livello particolarmente altro che sono Lussemburgo, Singapore ed Emirati Arabi; Il consumo di liquore è particolarmente altro in Russia, Bielorussia e Haiti; in ultimo, il consumo di vino ha valori estremi in Danimarca, Svizzera, Francia, Slovenia, Polonia.

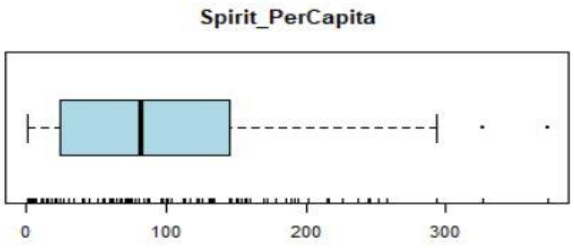
Inoltre, si vuole osservare come il range interquartile sia particolarmente ampio nella variabile *Beer_PerCapita* stando a significare che tra il 25% e il 75% dei consumatori si ha un vasto range di consumo a differenza delle altre due sostanze. Inoltre, la mediana di *Wine_Per Capita* è particolarmente prossima al primo quartile, comprovando la situazione descritta in precedenza di alta densità di consumo rispetto alla prima classe.

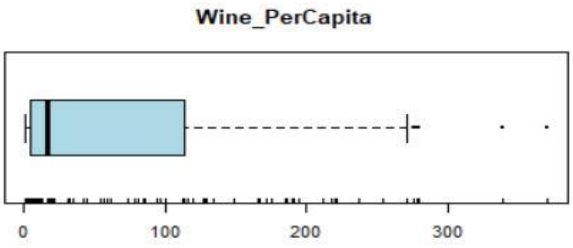












Abbiamo quindi scelto di suddividere le variabili quantitative in classi, per una maggiore comprensione dei fenomeni: *HappinessScore* è stata suddivisa in "Felicità bassa" (3-4.5), "Felicità parziale" (4.5-6.5), "Felicità piena" (6.5-8); *GDP* in "Molto basso" (0-5000), "Basso" (5000-10000), "Medio" (15000-40000), "Alto" (40000-100000); *HDI* in "Molto Arretrato" (0-0.554), "Arretrato" (0.554-0.699), "Avanzato" (0.699-0.799), "Molto Avanzato" (0.799-1); infine si è deciso di creare una nuova variabile "Sostanze Alcoliche" che racchiudesse la somma delle 3 bevande alcoliche per regione, suddividendola quindi nelle classi "Basso Consumo" (0-250), "Medio Consumo" (250-500), "Alto Consumo" (500-800).

Anova

Si vuole ora testare se esiste una differenza sostanziale tra le medie delle variabili sempre rispetto alle diverse regioni. H0: le medie sono tutte uguali. L'ipotesi nulla è rifiutata per le variabili *GDP, HDI* e *HappinessScore* con un pvalue prossimo allo 0. Ciò vuol dire che rispetto alle regioni, esiste almeno una media diversa dalle altre. Stesso esito si ha nel testare le medie di consumo per ogni sostanza alcolica rispetto a ciascuna regione, con il pvalue più alto dell'ordine di 10^{-9}, valore estremamente basso.

Gini

Si procede di seguito a presentare quanto emerso dall'analisi sull'indice di eterogeneità associato a ciascuna variabile raggruppando in base alla posizione geografica o regione.
Per quanto riguarda il GDP, è emerso un valore di Gini normalizzato pari a 0.9641, corrispondente ad un alto grado di eterogeneità in quanto, sebbene comparabili, vi sono marcate differenze sui GDP associati a ciascuna nazione del campione. Passando ad analizzare il valore dell'indice di eterogeneità di Gini normalizzato per la variabile Happiness Score, sempre tenendo conto del confronto su base regionale, si osserva un valore di 0.9966, vale a dire prossimo a 1, indice di una quasi completa eterogeneità del dato.

Stesso risultato si ha relativamente al consumo di sostanze alcoliche con un indice di Gini normalizzato che indica particolare eterogeneità nel consumo di alcol tra le varie regioni, con un valore di 0.968. L'indice di Gini normalizzato rispetto alle medie aritmetiche del consumo di ciascuna delle tre sostanze è però di 0.6439, indice del fatto che il consumo delle tre sostanze in media è sì eterogeneo, ma meno rispetto alle medie del consumo delle sostanze alcoliche per regione.

Si passa ora a calcolare l'eterogeneità riferita questa volta alle classi qualitative create in precedenza sulle variabili. La prima classe è quella relativa all'Happiness Score, divisa in tre classi: felicità bassa, felicità parziale e felicità piena. Il valore calcolato dell'Happiness Score normalizzato su queste tre classi risulta pari a 0.6167. Confrontandolo con l'indice normalizzato in base alla regione, si nota un valore decisamente inferiore, il che indicherebbe che l'introduzione delle classi così individuate introduce un indice di omogeneità superiore.

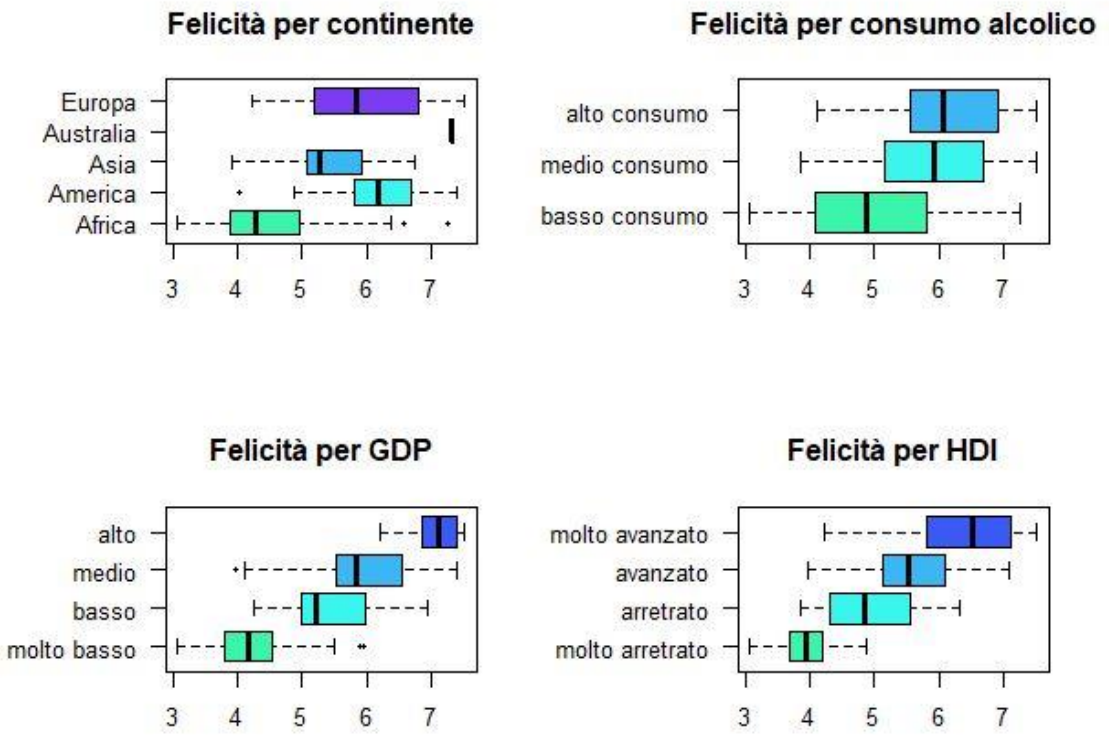
Analizzando invece l'indice sulle classi corrispondenti a differenti gradi di sviluppo umano (in questo caso 4: molto arretrato, arretrato, avanzato e molto avanzato), si ottiene un grado di eterogeneità normalizzato pari a 0.7197, associando quindi una discreta eterogeneità nella distribuzione delle frequenze.

Anche per la variabile GDP era stata effettuata una divisione in classi, suddividendo il campione in 4 fasce: molto basso, basso, medio e alto. Con un valore di Gini Normalizzato pari a 0.738, la divisione in classi introdotta ha comportato un abbassamento del grado di eterogeneità rispetto alla divisione per regioni, rimanendo tuttavia molto elevato e quindi ancora discretamente eterogeneo.

Connessione

Per studiare le relazioni tra variabili si è preferito raggruppare la variabile *Region* creando una nuova variabile "Continenti" con i livelli: Africa (includendo anche le regioni del Medio Oriente vicine al Nord Africa), America, Asia, Australia, Europa.

Consideriamo la variabile Continenti per il confronto tra boxplot. I seguenti boxplots sono stati condizionati per la variabile *HappinessScore* rispetto alle altre variabili suddivise in classi. Oltre che essere la variabile più significativa per i confronti in questo caso, si rivela anche la più adatta allo scopo in quanto distribuita normalmente, quindi sull'asse x si mantengono i valori originali della variabile non soggetta a trasformazioni.



Dal confronto per continente si nota subito che l'IQR (range interquartile) per l'Europa è il più ampio, quindi tra il primo e terzo quartile la popolazione assume livelli di felicità molto diversi tra loro, specialmente grazie al fatto che la numerosità delle città europee in questo dataset è molto alta. La situazione opposta accade ovviamente con l'Australia, la cui numerosità è di 2 paesi (Australia e Nuova Zelanda) il cui livello di felicità è molto simile (rispettivamente 7.334 e 7.313). Particolarmente utile al confronto grafico è la mediana, che condizionatamente al continente ci fa subito vedere che in Africa il valore mediano di felicità e molto basso (4.272), e che vi è molta disparità osservando i baffi e i due outliers che hanno valori ben al di sopra della mediana Europea. In particolare, Israele è l'outlier più estremo rispetto al continente africano e si paragona alla mediana del continente australiano avendo uno score di 7.267, contro il minimo score africano di 3.069, della Siria. Da notare anche che il terzo quartile dell'Africa neanche si avvicina alla media di HappinessScore di 5.525, mentre il primo quartile americano ne è evidentemente al di sopra.

Rispetto al grafico della felicità condizionata al consumo alcolico si nota una tendenza ad avere uno score maggiore all'aumentare del consumo. La tendenza di aumento della felicità legato all'aumento di GDP è molto simile a quella relativa all'aumento di HDI. Il primo caso però evidenzia degli outlier, in particolare Belize e Moldavia hanno rispettivamente uno score di 5.956 e 5.897 (quindi sopra la media totale) pur avendo un GDP pro capite molto basso. Per comprovare queste tendenze valutiamo il grado di connessione tra le variabili qualitative che abbiamo creato a partire dalle tabelle di contingenza, a patto che la connessione esista.

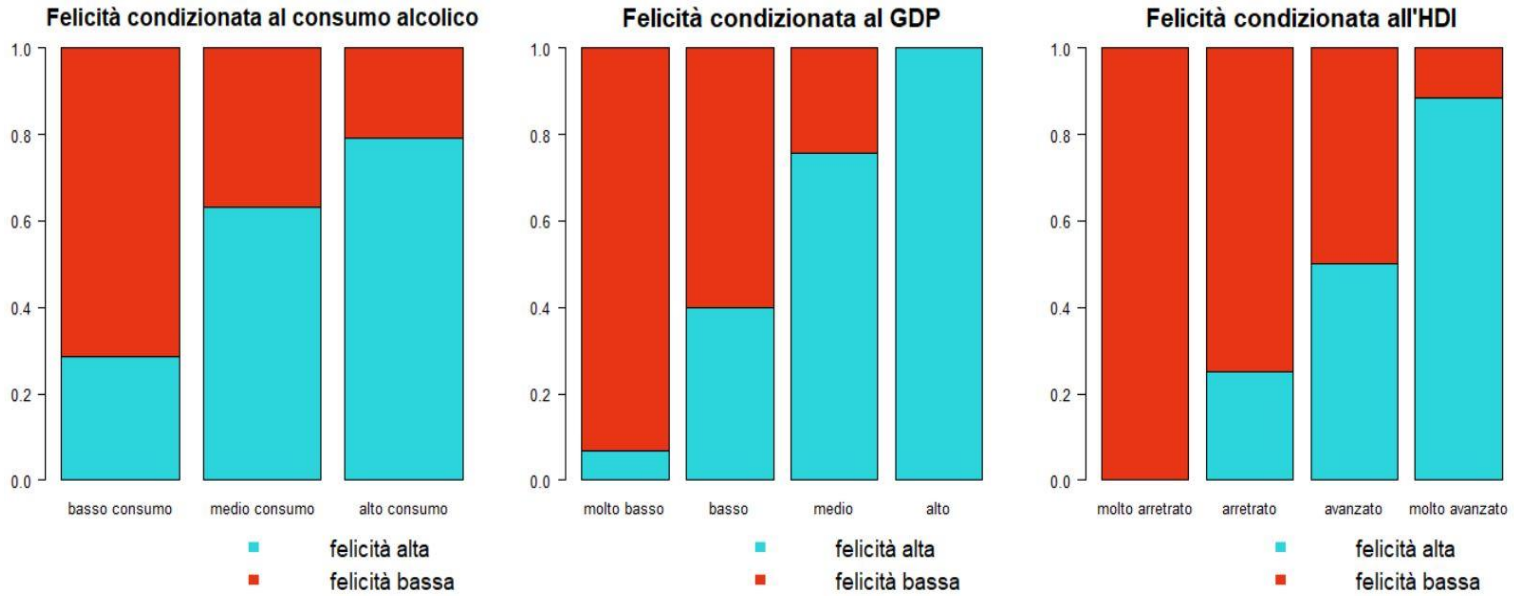
Le variabili messe a confronto convertite nelle loro qualitative sono: *HappinessScore* rispetto a *GDP, HDI, alcoholic_subtstances*; *GDP* rispetto a *alcoholic_substances* e *HDI*; *HDI* rispetto a *alcoholic_substances.*

Per prima cosa si osserva che tutte le variabili qualitative presentano un certo grado di connessione nelle rispettive tabelle di contingenza. Questo è dimostrato dal fatto che il pvalue del Test Chi Quadro per l'Indipendenza ha un valore molto vicino a 0 in tutti i casi, presentando il valore più alto nel confronto tra *acoholic_substances* e *HappinessScore* con pvalue 0.000085, rifiutando quindi sempre l'ipotesi nula ad un livello inferiore allo 0.0001, per cui si ha dipendenza tra le variabili $(H0: \chi^2 = 0)$. In particolare, le variabili che presentano maggiore connessione, in cui quindi il variare dell'una influisce maggiormente sul variare dell'altra, sono in ordine crescente: *HappinessScore* rispetto a *GDP,* con χ^2 Normalizzato di 0.4102; *GDP* rispetto a *HDI* con χ^2 Normalizzato di 0.3906; *HappinessScore* rispetto a *HDI*, con χ^2 Normalizzato di 0.3262. Secondo l'indice di connessione sembra esserci maggior effetto sul variare della felicità al variare del GDP.

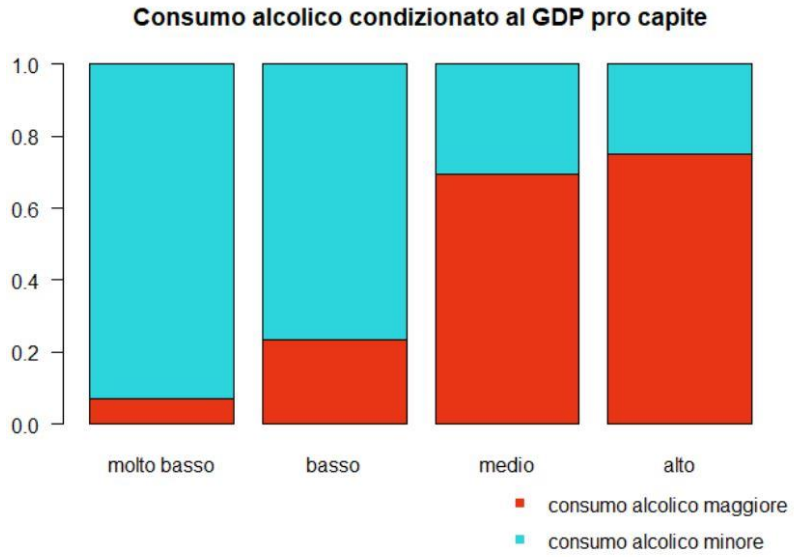
Barplots condizionati

Studiata la connessione, mettiamo a confronto diretto le variabili sfruttando i barplots condizionati. Abbiamo innanzitutto modificato le variabili qualitative rispetto a *HappinessScore* escludendo 11 elementi intermedi e nominando due classi "felicità alta" e "felicità bassa"; stesso procedimento nel caso della variabile qualitativa "Consumo Alcolico" per cui abbiamo individuato le classi "consumo alcolico maggiore" e consumo alcolico minore".

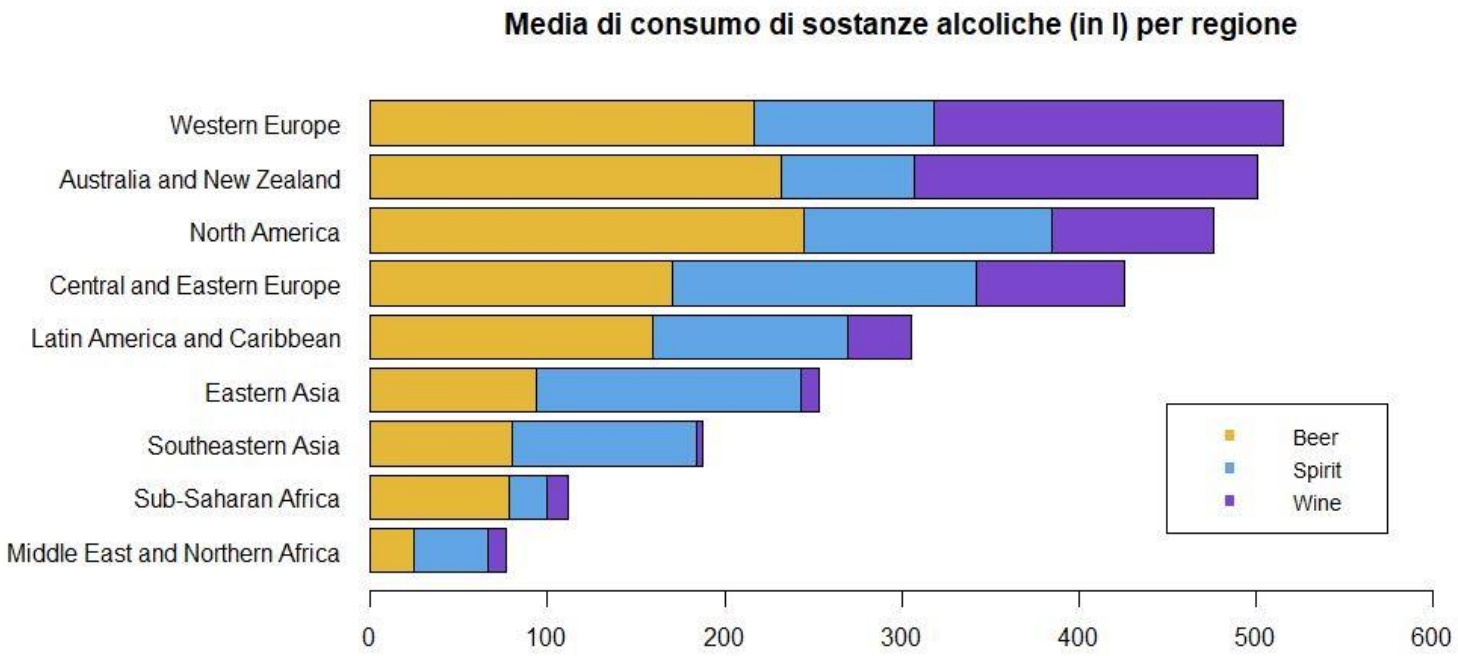
Le prime considerazioni si possono fare in base alla felicità rispetto alle altre variabili:



Dai grafici abbiamo un'idea ancora più chiara del fatto che nei paesi con maggiore consumo alcolico si abbia maggiore felicità. La stessa tendenza è dimostrata condizionatamente al GDP e all'HDI. Si evince inoltre che il GDP sembra avere un maggiore effetto sulla felicità rispetto all'HDI. Un'osservazione interessante è data dal seguente grafico, che dimostra una stretta relazione tra il GDP e il consumo alcolico: in paesi con medio e alto GDP si ha una tendenza a consumare più alcol rispetto ai paesi con GDP basso. L'uguaglianza della proporzione condizionata al GDP rispetto al consumo alcolico maggiore nei casi di GDP alto e medio è dimostrata da un test in cui H0: p1=p2 che accetta H0 con un pvalue di 0.6621. La proporzione di paesi con consumo alcolico maggiore e in cui il GDP pro capite è medio o alto è palesemente uguale.



Di seguito sono riportate in ordine di consumo alcolico le regioni dalla regione con consumo medio maggiore in alto a quella con consumo minore, suddividendo le barre in proporzione alla tipologia di sostanza alcolica. Prima di commentare il grafico si vuole specificare i risultati dei test sulla varianza delle variabili relative alle sostanze alcoliche con H0: var1=var2. In particolare, la variabile *Beer_PerCapita* ha una varianza sostanzialmente diversa da *Spirit_PerCapita* (pvalue inferiore a 10^{-5}); *Beer_PerCapita* a confronto con *Wine_PerCapita* rifiuta H0 per alfa 0.005 (pvalue 0.0037); *Wine_PerCapita* a confronto con *Spirit_PerCapita* accetta H0 per gli alfa convenzionali (pvalue 0.1926), quindi si può dire che le due variabili hanno stessa varianza.



Se sosteniamo l'ipotesi che al crescere del GDP i paesi tendono a consumare più alcol, al vertice superiore si hanno le regioni a GDP più elevato. La media totale di consumo di birra è pari a 137.565 litri pro capite ed è significativamente più alta di quella di liquore pari a 96.598, rifiutando l'ipotesi nulla di esserne inferiore con un pvalue inferiore a 10^{-7}. Stesse conclusioni si possono trarre riguardo alla media dello spirit che è significativamente più alta di quella del vino che è di 66.598, rifiutando l'ipotesi nulla di esserne inferiore con pvalue inferiore a 10^{-6}.

Osservando il grafico ci si potrebbe domandare se alcune medie di consumo alcolico relativamente alle tre sostanze siano uguali o diverse. Per togliere ogni dubbio si è eseguito un test di differenza tra medie in cui H0: media1-media2=0 rispetto alla popolazione (numero di Paesi) di ciascuna regione. In particolare, si sono eseguiti test per ogni tipologia di bevanda rispetto ad ogni regione e per stesse regioni test relativi alle tre diverse bevande. Qui si riporteranno soltanto i casi in cui il pvalue è al limite della soglia di accettazione o rifiuto.

Considerando la variabile *Beer* le differenze più interessanti tra medie sono tra le regioni: *Australia and New Zealand* rispetto a *Latin America and Caribbean* con un pvalue (0.03145) che rifiuta H0 con un'alfa di 0.05 ma accetta con 0.01; *Australia and New Zealand* rispetto a *Central and Eastern Europe* (pvalue 0.08837) che accetta con alfa 0.05; *Western Europe* rispetto a *Central and Eastern Europe* (pvalue 0.8324) che accetta con alfa 0.05; *Western Europe* rispetto a *North America* (pvalue 0.0804) che accetta con alfa 0.05.

Considerando invece la variabile *Spirit* le differenze più interessanti tra le medie si verificano tra le regioni: *Middle East and Northern Africa* rispetto a *Eastern Asia* (pvalue 0.0189) che rifiuta con alfa 0.05 ma accetta con alfa 0.01; *Sub-Saharian Africa* rispetto a *Southastern Asia* (pvalue 0.0992) che rifiuta con alfa 0.1 ma accetta con 0.05; *Middle East and Northern Africa* rispetto a *Eastern Asia* (pvalue 0.0189) che rifiuta con alfa 0.05 ma accetta con 0.01; *Australia and New Zealand* rispetto a *Latin America and Caribbean* (pvalue 0.0104) che rifiuta con 0.05 ma accetta con 0.01; *Western Europe* rispetto a *North America* (pvalue 0.0489) che rifiuta con alfa 0.05 ma accetta con 0.01.

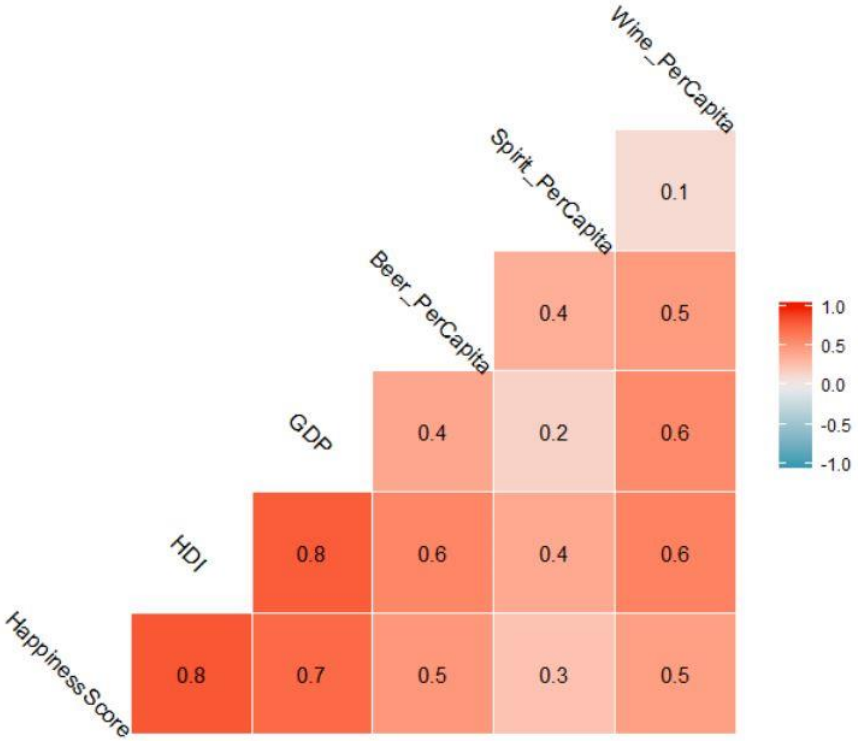
Considerando la variabile *Wine* i casi interessanti sono: *Middle East and Northern Africa* rispetto a *Southestern Asia* (pvalue 0.0277) che accetta H0 con alfa 0.05 ma rifiuta con 0.01; *Middle East and Northern Africa* rispetto a *Latin America and Caribbean* (pvalue 0.082) che rifiuta con 0.05 ma accetta con 0.1; *Sub-Saharan Africa* rispetto a *Southestern Asia* (pvalue 0.0287) che accetta con 0.05 ma rifiuta con 0.01; *Southestern Asia* rispetto a *Latin America and Caribbean* (pvalue 0.0238) che accetta con 0.05 ma rifiuta con 0.01; *Eastern Asia* rispetto a *Central and Eastern Europe* (pvalue 0.0782) che accetta con 0.1 ma rifiuta con 0.05; *Latin America and Caribbean* rispetto a *Central and Eastern Europe* (pvalue 0.0177) che accetta con alfa 0.05 ma rifiuta con 0.01.

Adesso valutiamo le differenze più interessanti tra le medie delle sostanze per ciascuna regione, tenendo come H0 il caso in cui la differenza tra le medie sia uguale a 0: in *Middle East and North Africa* la differenza tra le medie di *Beer* e *Wine* (pvalue 0.0474) è significativa per alfa 0.05, mentre non lo è per alfa 0.01; in *Southestern Asia* la differenza tra le medie di *Spirit* e *Wine* (pvalue 0.044) è significativa per alfa 0.05 mentre non lo è per alfa 0.01; in *Latin America and Caribbean* la differenza tra medie di *Beer* e *Spirit* è significativa per alfa 0.05 mentre non lo è per alfa 0.01; in *North America* la differenza tra le medie di *Spirit* e *Wine* (pvalue 0.0148) è significativa con alfa 0,05 mentre non lo è per alfa 0.01.

Ad eccezione dei casi sopra menzionati, le differenze tra le medie di consumo di sostanze alcoliche per regione sono palesemente uguali o palesemente diverse tra loro a seconda dei casi presi in considerazione.

Correlazione

Passiamo ora allo studio della correlazione tra le variabili.



Si sono eseguiti i test per valutare se le correlazioni siano significativamente diverse da zero. Il risultato è che tutte rifiutano l'ipotesi nulla H0: $\rho = 0$ tranne: *HappinessScore* in correlazione con *Spirit_PerCapita* che ha $\rho = 0.2564$ e il cui test restituisce un pvalue di 0.00436 che quindi rifiuta H0 se si considera un'alfa di 0.001 ma accetta H0 se alfa è di 0.05; *GDP* correlato a *Spirit_PerCapita* che ha $\rho = 0.1606$ e il cui test restituisce un pvalue di 0.0771 che rifiuta H0 per alfa 0.05; *Spirit_PerCapita* correlato a *Wine_PerCapita* che ha a $\rho = 0.1187$ e il cui test restituisce un pvalue di 0.1928 che rifiuta H0 per tutti gli alfa noti.

Dati gli esiti dei test si può dire che tutte le variabili hanno una correlazione positiva tra loro, quindi all'aumentare dei valori dell'una aumentano i valori dell'altra, fatta eccezione dei seguenti casi: la felicità che non aumenta all'aumentare del consumo di superalcolici; Il consumo di superalcolici non aumenta all'aumentare del GDP pro capite di un paese, considerato un certo livello di accettazione; non si tende a consumare più vino all'aumentare del consumo di superalcolici e viceversa.

Regressione multipla:

Come primo modello di regressione multipla si effettua la regressione della variabile HappinessScore (dipendente) sulle variabili GDP_PerCapita, HDI e Sostanze alcoliche:

HappinessScore $\sim$ log (GDP_PerCapita) + HDI + Sostanze_alcoliche

Si vuole quindi capire quanto la felicità della popolazione dei paesi sia influenzata dal Prodotto interno lordo, dall'indice di sviluppo umano del paese, e dalla quantità di sostanze alcoliche presenti nel paese. Attraverso lo studio della bontà di adattamento si ricava un coefficiente di determinazione multiplo (R^2) pari a 0,6656, quindi le variabili esplicative spiegano in modo notevole la variabile dipendente, inoltre le variabili GDP_PerCapita e HDI risultano significative.
Si procede poi allo studio della multicollinearità attraverso il Variance Inflation Factor (VIF):
VIF (GDP_PerCapita) = 9,659 ; VIF (HDI) = 11,24 ; VIF (Sostanze _alcoliche) = 5,066.
Come si evince dai risultati elevati le variabili esplicative sono correlate fra loro, siamo quindi in presenza di una forte situazione di multicollinearità.

Si passa ora ad un secondo modello di regressione multipla dove la variabile dipendente rimane HappinessScore, mentre le variabili esplicative sono GDP_PerCapita e HDI:

HappinessScore $\sim$ log (GDP_PerCapita) + HDI

In questo caso l'obbiettivo è capire quanto la felicità della popolazione dei paesi sia influenzata solo dal Prodotto interno lordo e dall'indice di sviluppo umano, senza quindi tener conto della quantità di sostanze alcoliche presenti nei vari paesi. Il coefficiente di determinazione lineare multiplo in questo caso risulta leggermente diverso, con un valore pari a 0,6665. Anche in questo caso le variabili esplicative sono significative e spiegano bene la variabile dipendente, quindi si può dedurre che la variabile Sostanze_alcoliche è poco influente nel modello. Si passa allo studio della multicollinearità, i valori del Variance Inflation Factor in questo caso sono sono:
VIF (GDP_PerCapita) = 9,584; VIF (HDI) = 9,584.
Come si vede anche in questo caso siamo in presenza di collinearità fra le variabili esplicative.

Si propone ora un altro modello di regressione multipla utilizzando come variabile dipendente sempre la felicità, quindi la variabile HappinessScore, mentre come variabili esplicative il prodotto interno lordo GDP_PerCapita e la quantità di sostanze alcoliche, tralasciando cosi l'indice di sviluppo HDI:

HappinessScore ~ log (GDP_PerCapita) + Sostanze_alcoliche

Si vuole quindi determinare l'influenza del prodotto interno lordo e delle sostanze alcoliche sul livello di felicità registrato nei paesi oggetto di studio. Il coefficiente di determinazione multiplo del modello assume un valore pari a 0,6639, quindi anche questo modello ha un fitting discreto. Lo studio della multicollinearità tramite Variance Inflation Factor restituisce i seguenti risultati: VIF (GDP_PerCapita) = 1,6712; VIF (Soostanze_alcoliche) = 1,6712.
Tale modello a differenza dei precedenti presenta una collinearità molto inferiore tra le variabili esplicative, non siamo quindi in presenza di una situazione di multicollinearità.

Ultimo modello di regressione multipla che si vuole studiare è quello che regredisce HappinessScore sulle variabili HDI e Sostanze alcoliche.

HappinessScore ~ HDI + Sostanze_alcoliche

Si vuole quindi determinare l'influenza dell'indice di sviluppo umano del paese e della quantità di alcolici sul livello di felicità registrato in quel determinato paese. Il coefficiente di determinazione del modello risulta pari a 0,6671, un valore elevato, si può quindi concludere che il modello spiega bene la variabile dipendente, inoltre la variabile HDI risulta significativa. Si procede allo studio della multicollinearità tramite Variance inflation factor (VIF), i risultati sono i seguenti:
VIF (HDI) = 2,05; VIF (Sostanze_alcoliche) = 2,05. I risultati del VIF non elevati ci fanno concludere che non siamo in presenza di una grave situazione di multicollinearità.

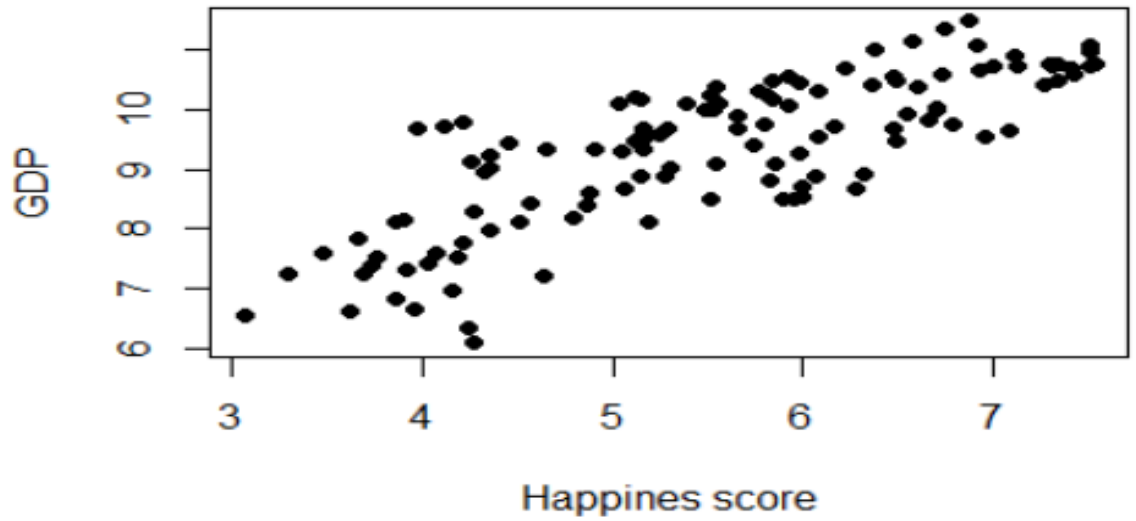
Regressione lineare semplice:

Si passa ora alla regressione lineare semplice. Come primo modello si prende quello della regressione di HappinessScore sulla variabile GDP_PerCapita:

HappinessScore ~ log (GDP_PerCapita)

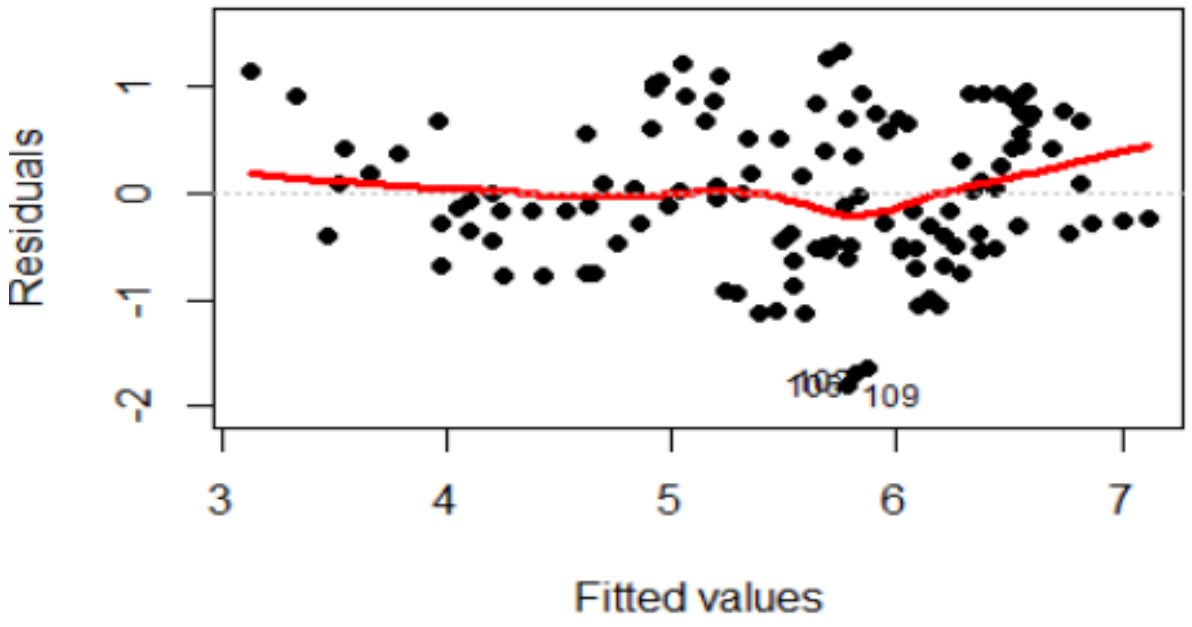
Quindi si vuole studiare il nesso lineare tra il livello di felicità registrato nei paesi e il prodotto interno lordo del paese. Altro obbiettivo del modello è capire se e come la variabile relativa alla quantità di alcolici, omessa in questo modello, influenzi la variabile HappinessScore, ovvero vogliamo capire se il livello di felicità registrato nei vari paesi sia effettivamente anche influenzato dalla quantità di alcolici o sia dovuta più ad un Prodotto interno lordo più o meno elevato. Si vuole quindi ottenere un confronto tra questo modello e il modello di regressione multipla con variabile dipendente HappinessScore e variabili esplicative GDP_PerCapita e Sostanze_Alcoliche.
Come prima cosa si studia la bontà di adattamento del modello ai dati, attraverso l'indice di determinazione (R^2=0,6391) si può dire che il fitting del modello è discreto e che la variabile esplicativa GDP_PerCapita è significativa, quindi il Prodotto interno lordo del paese ha una certa influenza nella felicità registrata sugli abitanti di quel paese. Si vuole ora studiare la correlazione tra le due variabili dal punto di vista grafico attraverso un diagramma a dispersione:

SCATTER PLOT



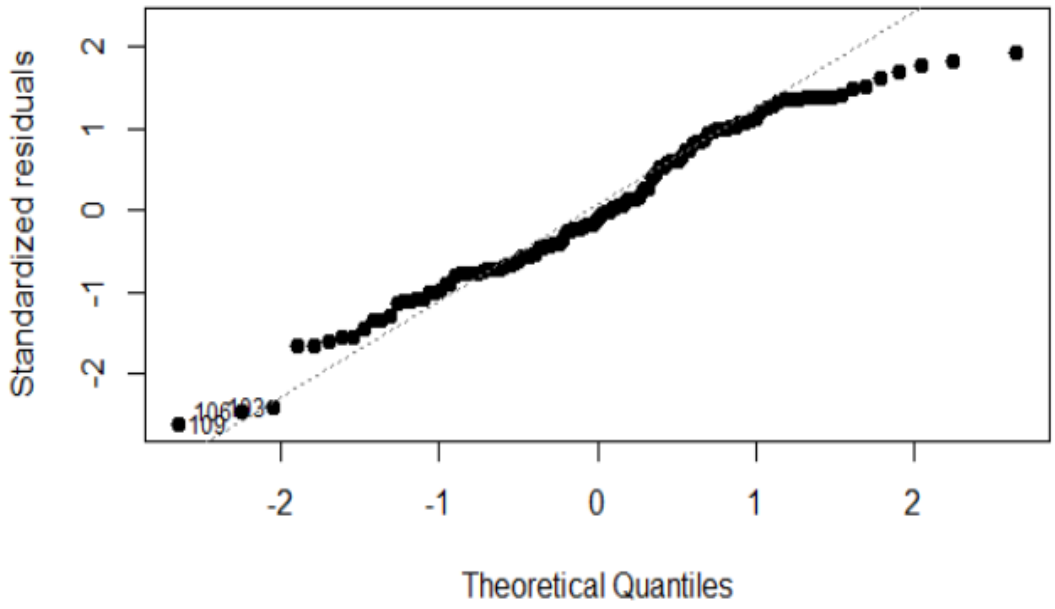
Dal diagramma a dispersione si vede come i punti abbiano un discreto andamento lineare, quindi siamo in presenza di correlazione lineare positiva tra le due variabili oggetto di studio. SI passa ora all'analisi del grafico Residuals vs Fitted. Come si vede dal confronto tra valori stimati e residui anche in questo caso l'ipotesi di linearità non è violata, la curva rossa che interpola i dati ha un andamento approssimativamente lineare, si nota però la presenza di qualche outlier.

RESIDUALS VS FITTED



SI procede ora allo studio della normalità. A tal proposito si utilizza la rappresentazione grafica del Q-Q plot per confrontare i valori dei residui standardizzati con i quantili teorici. Osservando il grafico si vede che i punti giacciono sulla retta, ad eccezione che nelle due code, ciò ci porta a concludere che l'ipotesi di normalità non va rifiutata.

NORMAL Q-Q PLOT



Riguardo a tale modello lineare si può concludere che sia un buon modello per rappresentare la relazione esistente tra la variabile HappinessScore e GDP_PerCapita, confrontando inoltre tale modello d regressione multipla che regrediva HappinessScore su GDP_PerCapita e Sostanze_ alcoliche, si può dire che i modelli abbiano un fitting quasi uguale, rispettivamente un R^2 pari a 0,6391 e 0,6639. Quindi rimuovendo la variabile relativa alle sostanze alcoliche il fitting cambia di poco, si può quindi concludere che la felicità sia influenzata quasi esclusivamente dal Prodotto interno lordo del paese oggetto di studio, piuttosto che dalla quantità di sostanze alcoliche presenti nel paese, quest'ultima infatti spiega solo una piccolissima parte del modello.

Si esegue ora un'altra regressione lineare semplice, questa volta si regredisce la variabile HappinessScore sulla variabile HDI, per capire cosi quanto incide l'indice di sviluppo umano sulla felicità dei cittadini:

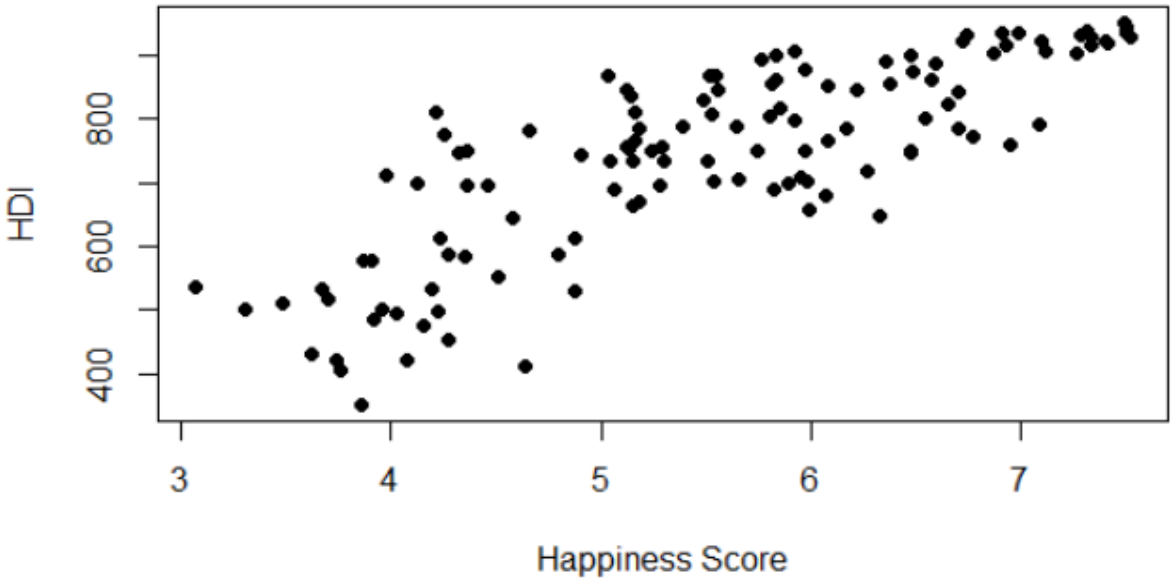
HappinessScore ~ HDI

Altro obbiettivo del modello è capire l'importanza della variabile Sostanze_alcoliche, omessa in tale modello a differenza del modello di regressione multipla studiato in precedenza avente come variabile dipendente HappinessScore e come variabili esplicative HDI e Sostanze alcoliche. Tale modello quindi ci permetterà di capire se la felicità dei cittadini dei vari paesi ha uno stretto collegamento con la quantità di sostanze alcoliche oppure la felicità dei cittadini è più legata all'indice di sviluppo umano.

Si procede subito con lo studio della bontà di adattamento, l'indice di determinazione del modello risulta pari a 0,6645, si può quindi concludere che il modello ha un fitting elevato e che la variabile esplicativa HDI è significativa. La felicità dei cittadini dei paesi oggetto di studio ha quindi un certo legame con l'indice di sviluppo umano (HDI). Si procede ora allo studio della correlazione lineare tra la variabile HappinessScore e HDI attraverso alcune rappresentazioni grafiche. Come prima rappresentazione grafica si procede con il diagramma a dispersione.

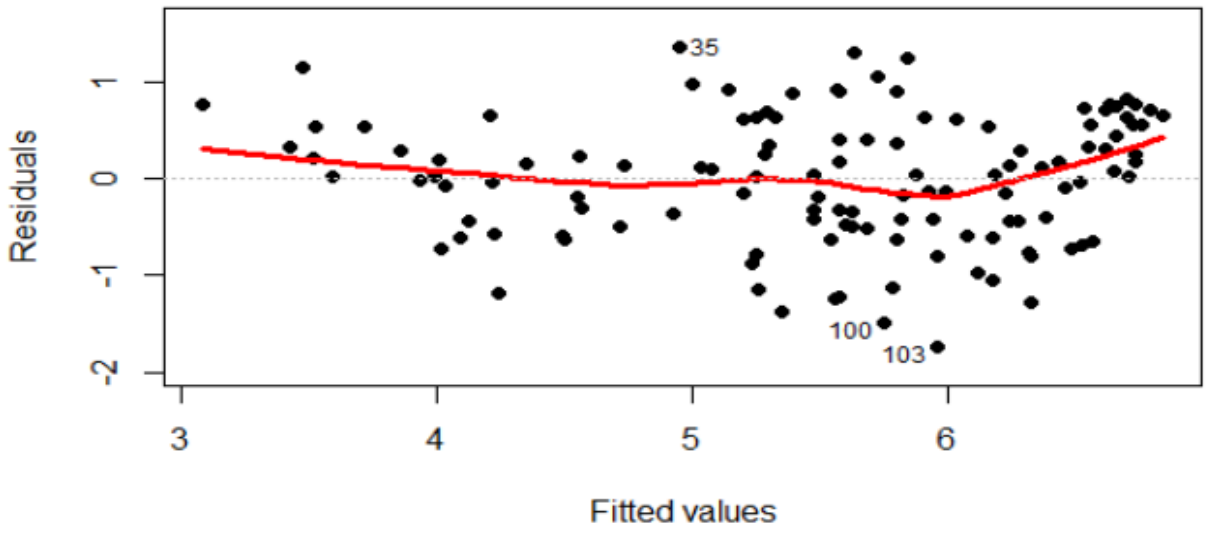
Osservando i punti sullo scatter plot si può notare come questi abbiano un andamento lineare, si può quindi concludere che le due variabili, HappinessScore e HDI, siano correlate positivamente.

SCATTER PLOT



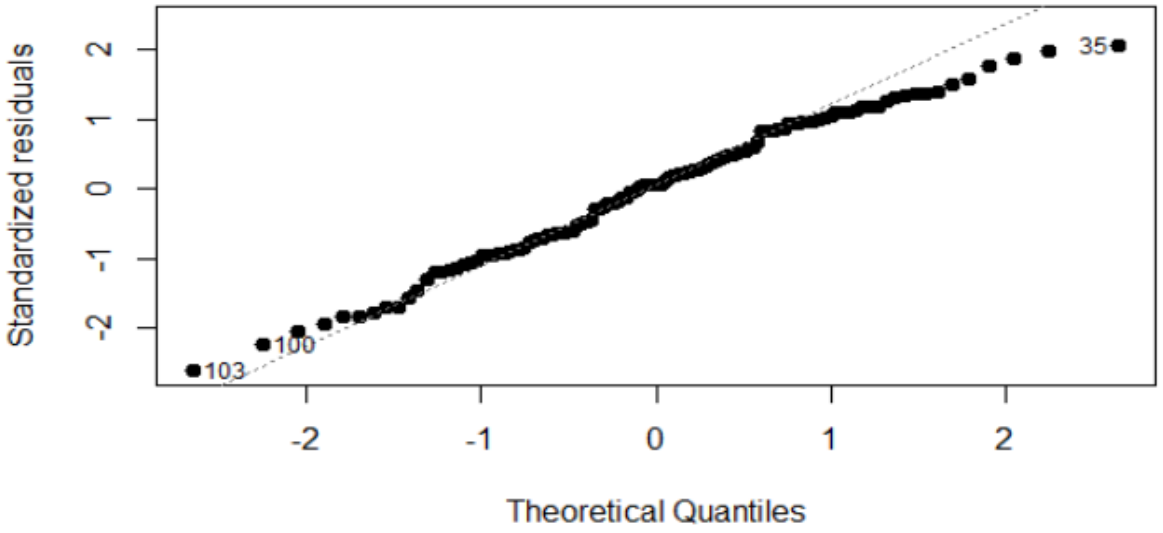
Si passa ora ad un'altra rappresentazione grafica, quella dei Residual vs Fitted. Anche dal confronto tra valori stimati e residui si può stabilire che l'ipotesi di linearità non va rifiutata, osservando infatti la curva che interpola i punti si nota che essa ha un buon andamento lineare, anche se si nota la presenza di qualche valore anomalo.

RESIDUALS VS FITTED



Si presenta ora il grafico Q-Q plot per la verifica della normalità. Osservando il grafico si nota anche in questo caso che la distribuzione dei punti segue la linea retta, fatta ancora eccezione per le code, quindi anche in questo caso l'ipotesi di normalità non è da scartare.

NORMAL Q-Q PLOT



Si può infine concludere relativamente al modello studiato che sia un discreto modello per rappresentare la relazione tra le due variabili HappinessScore e HDI. Confrontando il modello con il modello di regressione multipla HappinessScore ~ HDI + Sostanze_alcoliche si può notare come i due indici di determinazione si discostino di poco, 0,6645 e 0,6671. Ciò porta a concludere che la variabile relativa alla quantità alcolica non influenza in maniera significativa la felicità, ma che sia più l'indice di sviluppo umano del paese il fattore che fa alzare il livello di felicità della popolazione, e la variabile Sostanze alcoliche spiega solo una piccola parte del modello.