UNIVERSITÁ DEGLI STUDI DI MILANO-BICOCCA

Scuola di Economia e Statistica

Corso di laurea in

Scienze Statistiche ed Economiche

UN'ANALISI GEO-QUALITATIVA DELLA PRODUZIONE DELLA BIRRA NEL MONDO

Relatore: Prof. Christian Garavaglia

Tesi di Laurea di:

Davide Luperi

Matr. N. 826249

Anno Accademico 2019/2020

INDICE

CAPITOLO 1 – PRESENTAZIONE

INTRODUZIONE

Il presente elaborato ha come oggetto lo studio della produzione della birra nel mondo, analizzandone analogie e differenze da un punto di vista qualitativo tra i diversi Paesi del globo.

Lo stimolo per la scrittura di questo elaborato nasce da un interesse personale per il mondo della birra e, nello specifico, della birra artigianale. In particolare, mi premeva svolgere un lavoro che approfondisse delle mie curiosità circa alcuni aspetti della birra ignoti alla maggior parte dei consumatori e come questi aspetti si differenziassero geograficamente. L'obbiettivo che mi sono posto è stato quello di capire come varia la produzione di questa bevanda tra le diverse nazioni prese in esame. Per fare ciò mi sono posto delle domande cardine su cui ho incentrato la mia analisi, tra cui quanto influisce la cultura e la 'storia' sulla produzione odierna, quali sono gli stili più diffusi, in che relazione stanno tra di loro le variabili che caratterizzano il prodotto.

Per realizzare questo lavoro mi sono servito di due database contenenti i dati necessari per le mie analisi effettuate sul programma Rstudio. I processi statistici che ho affrontato sono stati quelli di visualizzazione ed esplorazione dei dati, di clusterizzazione e, in piccola parte, di modellizzazione. Per quanto riguarda la parte di programmazione ho implementato algoritmi utilizzando varie metodologie viste durante il corso di laurea.

È bene precisare che le analisi che andrò a implementare non tengono conto della quantità in cui è prodotta ogni singola birra all'interno dei dataset. Infatti, le birre industriali sono prodotte in quantità molto maggiori delle birre artigianali ma nei dataset, e dunque nel mio lavoro, sono considerate come singole unità e dunque con lo stesso peso. Questa considerazione non crea grandi problemi poiché lo scopo dell'elaborato è quello di andare a studiare la birra da un punto di vista qualitativo, quindi andando a prendere ogni birra come unità per analizzarne le caratteristiche.

Nel primo capitolo, oltre alla presente introduzione, sono inseriti un paragrafo che riassume brevemente la storia della birra e che permette al lettore di avere un'inquadratura

più ampia del lavoro oltre ad introdurlo nella scoperta della bevanda, ed un paragrafo in cui sono presentati i due dataset di cui mi sono servito.
Il secondo capitolo riguarda la visualizzazione dei dati, per comprendere che tipo di dati i dataset contengono. Vengono inoltre approfonditi alcuni argomenti che riguardano la birra più nel dettaglio.
Nel terzo capitolo si trova la parte dell'analisi in cui ho risposto alle domande centrali del lavoro per raggiungere il mio obiettivo.
Il quarto capitolo contiene i riferimenti bibliografici di cui mi sono servito per la stesura dell'elaborato.

STORIA DELLA BIRRA

La birra è una bevanda antichissima avente origine nel VI millennio avanti cristo, come testimonia una tavola rinvenuta durante gli scavi archeologici nella regione della Mesopotamia e datata intorno al 6000 a.c. sulla quale fu scritta una ricetta della bevanda. Inizialmente si trattava semplicemente di malto d'orzo, un cereale già conosciuto e coltivato nell'antichità, che fermentando naturalmente creava una bevanda alcolica.

Nel 3500 ca. nella città sumera di Uruk i cittadini già usavano la birra come materiale prezioso negli scambi commerciali.

Nel 3000 ca. la birra venne introdotta anche nel mondo egizio di cui divenne una bevanda privilegiata e apprezzata anche da faraoni. Da qui si diffuse poi in tutta Europa e quando i romani crearono l'impero che la storia ci ha insegnato, introducendo il vino come bevanda alcolica d'eccellenza, la birra rimase comunque prediletta nelle province più settentrionali e dai popoli germanici oltre il *limes*[1] romano.

Con la caduta dell'Impero Romano e la creazione dei regni romano-germanici il consumo della birra crebbe in tutta Europa sotto l'influenza delle popolazioni germaniche. In particolare, Carlo Magno fece costruire monasteri nell'Europa centro-settentrionale che divennero sostanzialmente gli unici punti di produzione di birra, prodotta per consumo personale, per essere offerta ai pellegrini all'inizio e per essere anche venduta successivamente. Questa cultura di monasteri si estese poi anche nelle isole britanniche, in Germania e in Scandinavia.

Un'importante rivoluzione nella produzione di questa bevanda fu l'introduzione dell'utilizzo del luppolo, che portò a migliorare il gusto della birra e a poterla preservare più a lungo, permettendo così di poter essere anche trasportata per lunghi tragitti.

Un'altra svolta importante si ebbe nel XII secolo con l'avvento della tecnica della 'bassa fermentazione'. Inizia qui la distinzione tra birre 'nuove' o lager (da cold lagering ovvero a fermentazione fredda) e birre 'vecchie' o ale (a fermentazione alta).

Dal XV secolo si ebbe un incremento della domanda di birra, domanda che non poteva più essere soddisfatta dai soli monasteri. La produzione si espanse a livello industriale e si dovettero creare leggi e regolamenti ad hoc, tra cui la più famosa è la 'legge di purezza'

[1] *Limes*, parola latina che significa letteralmente 'limite' o 'strada' e che indicava la linea di confine del territorio dell'impero romano

(‘*Reinheitsgebot’*) di Norimberga del 1487 nella quale si stabilì che la birra dovesse essere composta da soli 4 ingredienti: acqua, lievito, malto e luppolo.
Nel XVI secolo, con la colonizzazione delle Americhe, la birra venne esportata anche nel nuovo continente. La globalizzazione verso le Americhe e le Indie, portò sul mercato europeo diversi competitori della birra quali bevande a base di spezie (thè), cacao, caffè, oltre allo sviluppo di superalcolici quali gin, rum e whisky.
Nel XVIII e XIX scoperte circa proprietà del lievito e innovazioni tecniche portarono ad un maggiore controllo nella produzione e ad una diversificazione della birra prodotta. Lo stile predominante divenne quindi quello delle lager, che a differenza delle altre birre, era più limpida e ‘pulita’ grazie alla scoperta e all’utilizzo di due processi di fermentazioni separati.
Nel corso del XX secolo la crescita della produzione di birre continuò costante, ad eccezione della crisi economica a cavallo tra le due guerre. Gli anni del boom economico segnarono un importante aumento della produzione della bevanda tanto negli USA quanto in Europa e grazie ad innovazioni tecnologiche la birra divenne un bene a basso costo, disponibile per un’ampia fascia della popolazione.

A partire dagli anni ’80 nacquero in California i primi birrifici artigianali che affiancarono alle birre industriali, le birre artigianali. Questo movimento si diffuse poi negli anni successivi in America e in Europa andando a differenziare e innovare notevolmente il mercato della birra.
La prima definizione ufficiale di ‘birrificio artigianale’ fu fornita dalla *brewewr association*[2] la quale stabilì che un birrificio per essere definito artigianale dovesse essere ‘piccolo, indipendente e tradizionale’.
-Piccolo a indicare che la distribuzione non deve superare i 6 milioni di barili di birra all’ anno.
-Indipendente significa che non più del 25% della proprietà del birrificio deve essere posseduta da un membro dell’industria alcolica esterno al birrificio.
-Tradizionale, ovvero che la parte principale di produzione è costituita di birre i cui aromi derivano da ingredienti brassicoli tradizionali o innovativi e dalla loro fermentazione.

[2] La Brewer Association è un’associazione che promuove e protegge i birrifici artigianali americani

MATERIALE USATO (DATASET):

Per la stesura del mio lavoro mi sono servito di due database presenti nel sito Kaagle[3].
Il primo database, che per comodità chiamerò 'birre', è un dataset le cui unità statistiche (di numerosità 358873) corrispondono ad un'ampissima fetta delle birre prodotte in tutto il mondo. È bene precisare che non si tratta di un censimento, che sarebbe stato impossibile da effettuare per questo tipo di prodotto, ma bensì di un campione molto ampio che rappresenta esaustivamente la realtà.
Le variabili presenti nel dataset 'birre' sono:

"id"
"name"
"brewery_id"
"state"
"country"
"style"
"availability"
"abv"
"notes"
"retired"
"X11"
"X12"
"X13"
"X14"

Tra le quali ho tenuto in considerazione quelle a me utili, andando ad eliminare le altre.
Il dataset finale è così composto:

"id"
"name"
"state"
"country"
"style"
"availability"
"abv"

Dove:

'id' è un valore numerico che identifica unicamente un'unità statistica.

'name' nome della birra.

'country' nazione in cui la birra è prodotta.

[3] Kaagle è una piattaforma online in cui utenti e aziende condividono dataset, che possono essere usati per le analisi di tipo statistico

‘state’ stato degli USA in cui la birra è prodotta (quando ‘country’ non assume il valore USA questa variabile non ha valore).
‘style’ stile della birra in questione.
‘availability’ disponibilità in cui è presente la birra nel mercato.
‘abv’ indica il grado alcolico.

Il secondo dataset di cui mi sono servito, denominato ‘ricette’, contiene 73861 unità statistiche corrispondenti ad altrettante birre di cui sono stati misurati determinati parametri durante e alla fine del procedimento di preparazione.
Mi sono servito di questo secondo dataset poiché conteneva informazioni che mancavano nel primo dataset e che sarebbero state fondamentali per le analisi.
Le variabili sono:

"BeerID"
"Name"
"URL"
"Style"
"StyleID"
"Size.L."
"OG"
"FG"
"ABV"
"IBU"
"Color"
"BoilSize"
"BoilTime"
"BoilGravity”
"Efficiency"
"MashThickness"
"SugarScale"
"BrewMethod"
"PitchRate"
"PrimaryTemp"
"PrimingMethod"
"PrimingAmount"
"UserId"

Tra le quali ho tenuto in considerazione quelle a me utili andando ad eliminare le altre, il dataset finale è così composto:

"Name"
"Style"
"ABV"
"IBU"
"Color"

Dove:
‘style’ stile della birra.
‘abv’ grado alcolico.
‘ibu’ variabile quantitativa per calcolarne l’amarezza.
‘color’ variabile quantitativa a indicare il colore della birra.

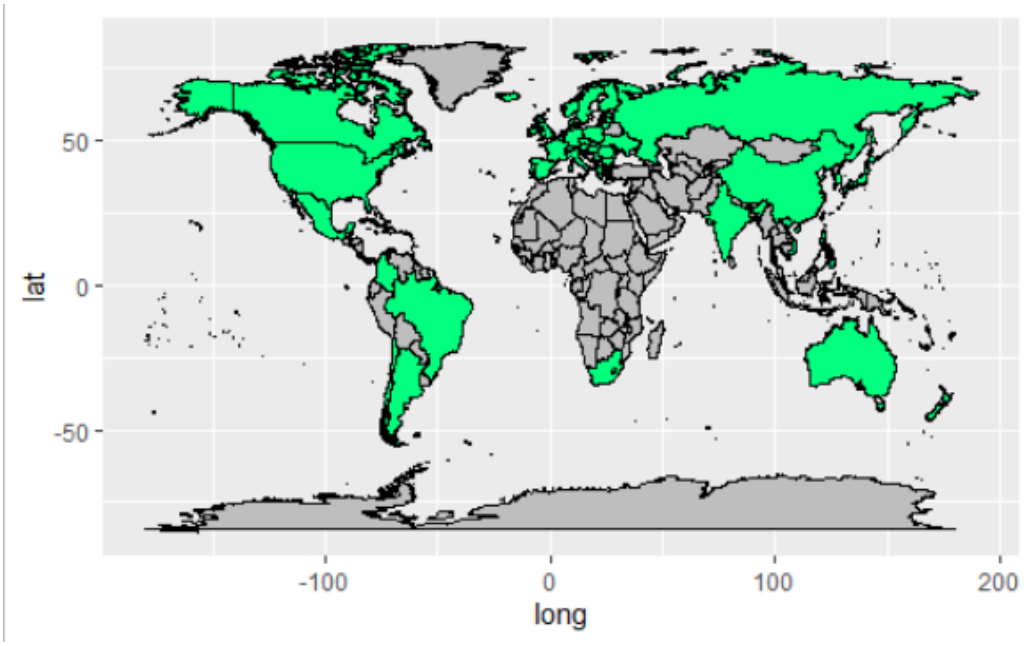
CAPITOLO 2 – VISUALIZZAZIONE DEI DATI

PRIMO DATASET

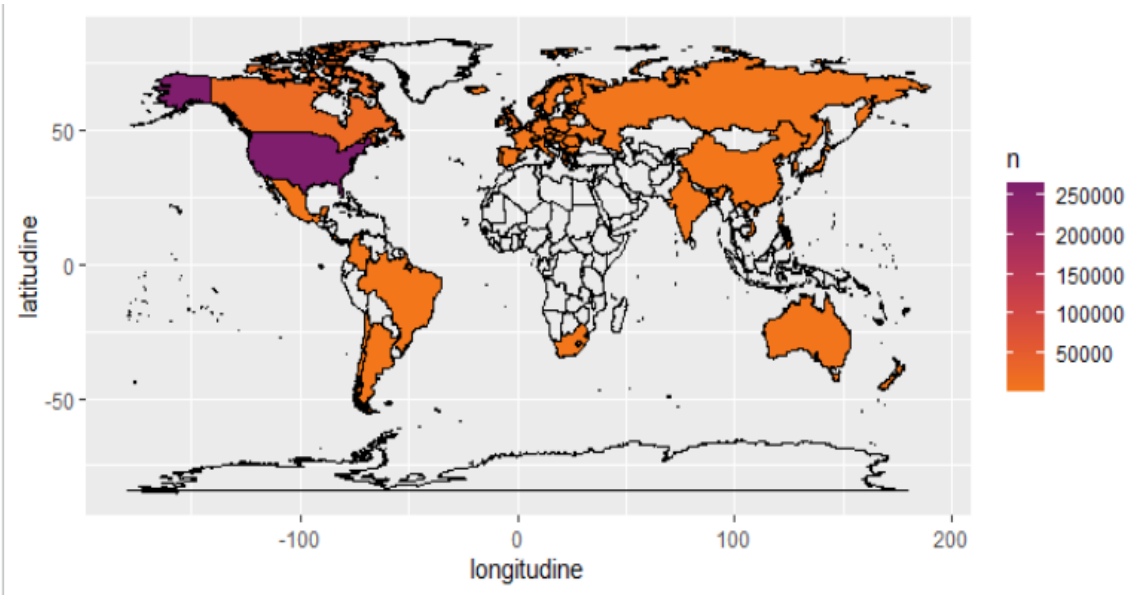
Inizio la visualizzazione dei dati osservando le variabili del primo dataset (‘birre’) interrogandomi sui valori che assumono.
Le variabili ‘*Id’* e ‘*Name’* servono esclusivamente a identificare in maniera univoca le mie osservazioni (ad ogni unità statistica corrisponde un numero id e un nome univoco), per cui non rientrano nell’analisi.

Country

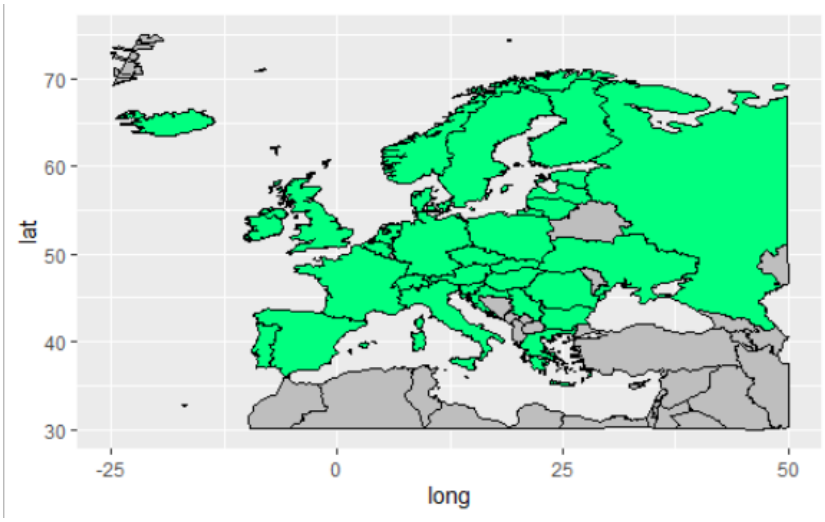
La prima variabile significativa che incontro è ‘country’, una variabile qualitativa che esprime la nazione di provenienza della birra. Il valore che può assumere è infatti l’insieme delle nazioni in cui viene prodotta almeno una birra.
Per rendere più chiara la mia analisi ho scelto di considerare solo quelle osservazioni la cui nazione (e quindi il valore della variabile country) si presentasse almeno cento volte nel dataset. In altre parole, ho raggruppato le osservazioni per nazione, ho sommato il numero delle volte che quella nazione si presentava nel dataset e ho escluso quelle nazioni con meno di cento osservazioni presenti.
Ho optato per questa scelta per confrontare le nazioni che hanno un peso abbastanza importante nella produzione di birra mondiale (ricordo dal punto di vista della diversità). Cento diverse birre inoltre significa avere un’adeguata diversificazione nella produzione per quanto riguarda stile e caratteristiche, mentre a numeri inferiori questo difficilmente avviene.
Considero dunque solo quelle nazioni che producono almeno cento tipi di birra e le visualizzo su una mappa:



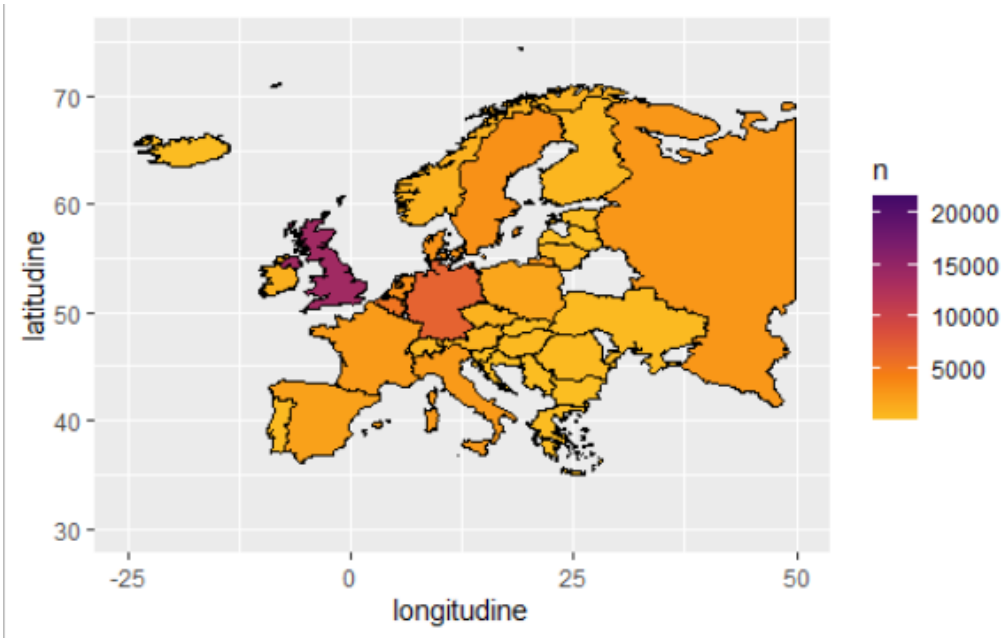
In questa successiva mappa viene mostrata la densità in base al numero di birre prodotte da ciascuna nazione:



Si vede subito che gli USA sono la nazione maggiormente rappresentata dal mio dataset (oltre il 60% delle osservazioni assume come valore della variabile 'country' 'USA') Poiché vedo che il continente maggiormente rappresentato è l'Europa, decido di fare un grafico (identico al primo) in cui focalizzo l'attenzione sul vecchio continente.



Anche in questo caso visualizzo il grafico di densità:



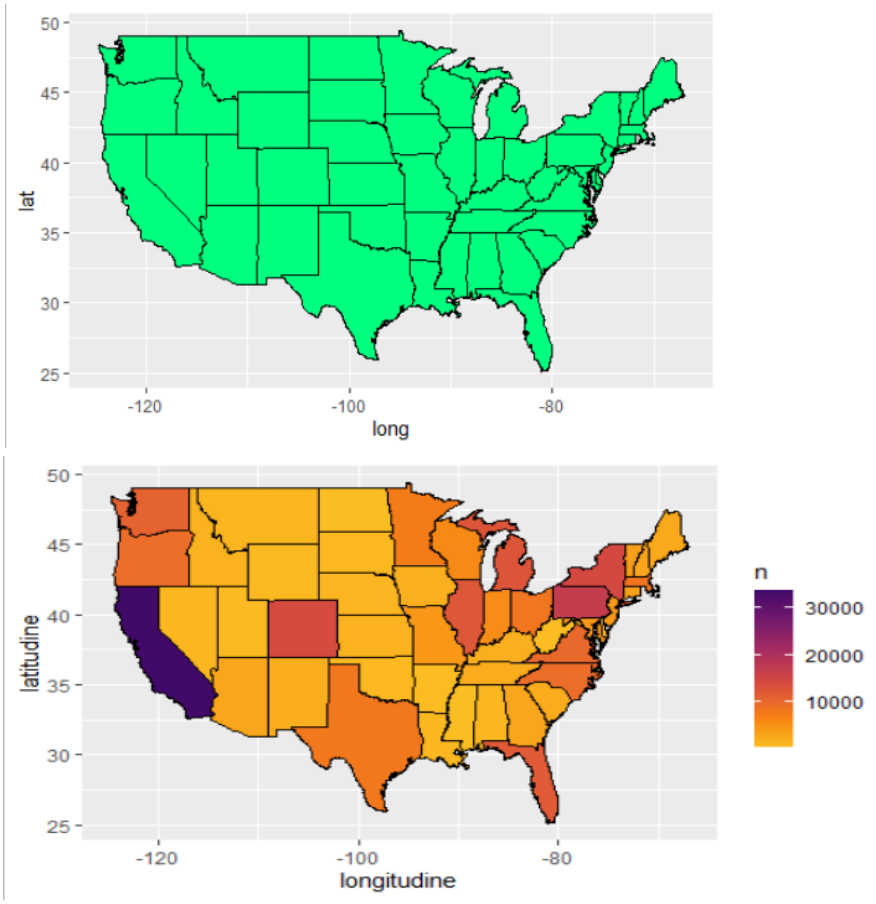
Poichè abbiamo visto che gli USA hanno un peso così influente, si possono andare a vedere i dati all'interno dei singoli stati andando a studiare la variabile ‘states’.

States

La variabile 'states' presente nel dataset è una variabile qualitativa il cui insieme di valori che può assumere corrisponde con gli stati degli Stati Uniti d'America.

Si deduce che dunque questa variabile ha senso solo per quelle osservazioni con valore 'country' = 'USA'.

Andiamo dunque a vedere su una mappa quali stati sono presenti nel dataset:



Piccola osservazione che devia leggermente dal focus dell'argomento:

È interessante vedere come la mappa della distribuzione numerica delle birre prodotte nei singoli stati della federazione, ricalchi, seppur con alcune evidenti differenze, quella della distribuzione della popolazione negli USA

Se guardiamo attentamente infatti gli stati colorati con colori più scuri (e quindi con 'n', a indicare numero di birre prodotte, più alto) quali California, New York, Illinois, Florida, Colorado, Washington, Pennsylvania, sono anche tra gli stati più popolosi[4].

[4] La California presenta valori così alti poiché, oltre ad essere lo stato più popoloso degli u.s. è anche quello in cui per primo è nato il concetto di birra artigianale e birrificio artigianale

Style

Arriviamo alla variabile 'Style'. Si tratta anche questa di una variabile qualitativa che riporta, per ogni birra, lo stile a cui essa appartiene.

Come abbiamo visto la birra è una bevanda che può sembrare semplice, con pochi ingredienti ed un unico processo di lavorazione. La realtà è però totalmente diversa

Come prima osservazione tutti i singoli ingredienti 'base' (quelli definiti dalla legge di Norimberga) quali acqua, lievito, malto e luppolo, hanno un'ampissima varietà che combinate determinano enormi differenze tra una birra e l'altra. Inoltre, spesso nelle birre si aggiungono elementi secondari come aromi, infusi, estratti o altri ingredienti come miele, frutti (o le rispettive scorze), cocco e molti altri che ne determinano un'ulteriore variabilità. Infine, anche il processo di produzione può presentare dei cambiamenti.

Birre con caratteristiche comuni vengono dunque classificate sotto uno stile. Per poter affermare che una birra appartiene ad un determinato stile deve avere delle caratteristiche che distinguono quello stile dagli altri.

Ci sono molte caratteristiche della birra che possono essere valutate ma le più importanti e identificative sono quelle ne determinano il gusto, il profumo e l'aspetto. Queste sono IBU, ABV, COLOR le quali rientrano nella mia analisi, oltre a FG e AG che però riguardano il processo di bollitura della birra e non sono valori finali, che quindi non prenderò in considerazione.

Come ho spiegato, l'assunzione da parte di una singola birra di valori differenti di queste variabili ne determina l'appartenenza ai vari stili di birra.

Vediamo allora nel concreto quali sono gli stili di birra.

Ci sono molti stili di birra, riconosciti ufficialmente da enti quali il *'Beer Judge Certification Program's'* che è periodicamente aggiornato poiché l'evoluzione e la creazione di nuovi stili avviene in maniera annuale. Il *Beer Judge Certification Program's*[5] completo può essere consultato sul sito ufficiale.

La suddivisione degli stili di birra è facilmente visualizzabile come una struttura ad albero Iniziamo con il dividere gli stili di birre in due grandi famiglie, da cui si ramificano poi tutti gli stili.

[5] Il BJCP è un'associazione che ha lo scopo di promuovere l'apprezzamento della birra e per riconoscere le capacità di valutazione e degustazione della birra

1. LAGER

Dall'introduzione abbiamo visto come queste birre siano birre 'nuove' e sono caratterizzate dall'utilizzo di un lievito che agisce a bassa fermentazione (il *Saccharomyces Pastoranus*). Questo lievito necessita appunto di basse temperature per avviare la fermentazione. Possiamo raggruppare le Lager in:

- *Pilsner*, nate nell'attuale Repubblica Ceca vicino alla città di Plzen, che vengono prodotte con un malto omonimo da cui prendono il nome

- German Lager, che sono appunto lager che hanno origine nell'attuale Germania, al cui interno troviamo stili come
    - *Vienna Lager* (originarie della capitale austriaca)
    - *Kölsch*
    - *Rauchbier* (birre con un gusto particolare di affumicato, da rauch che significa fumo)
    - *Bock*, stile molto diffuso in Germania, in cui rientrano birre dal colore ambrato scuro, con forti sentori di caramello ed un moderato grado alcolico
      Estensioni di questo stile sono le *Dopplebock* e le *icebock*

- American Lager, tra cui:
    - *American pale lager*
    - *American adjunct lager*

- *Dark/amber lager,* sono birre la cui particolarità risiede nel fatto di presentare un colore scuro o ambrato scuro dovuto all'utilizzo di malto tostato. Chiaramente ciò impatta notevolmente sul sapore della birra

2. ALES

In queste birre viene usato per la fermentazione un lievito (il *Saccharomyces cerevisiae*) che agisce a temperature più alte.
In questa categoria rientrano diversi stili, molto differenti tra di loro

- German ale. Anche nella categoria ale la Germania è patria di numerosi stili quali:
    - *Weissbier o weizen*, birre caratterizzate dalla torbidezza dovuta all'utilizzo del frumento, con tutti i suoi stili 'figli' (come *weizenbock*, *dunkelweizen*, *gose* ecc)

- Belgian/French ale. In questa categoria rientrano molti stili caratterizzati da una maltazione molto forte e da un grado alcolico decisamente elevato. Inoltre, troviamo tutti quegli stili classici delle famose birre d'abazia
    - Strong dark ale, tra cui le *dubble*
    - Belgian pale ale, tra cui le *tripel*
    - *Saison*
    - *Witbier*

- Lambic. Sulle birre lambic si deve aprire una piccola parentesi poiché hanno una caratteristica che le rende uniche in maniera decisamente particolare: sono birre a fermentazione spontanea. Ciò significa che nella produzione non viene aggiunto alcun lievito artificiale ma il lievito viene messo a contatto con lieviti 'selvaggi' autoctoni del luogo (queste birre sono principalmente prodotte in Vallonia, la regione meridionale del Belgio)
  Il risultato è una birra estremamente acida e sidrosa, molto diversa dal concetto comune di birra
  In questa categoria i principali stili di birra sono:
    - *Faro*
    - *Gueuze*

- *Pale ale*, stile nato in Inghilterra con la caratteristica di una moderata amarezza e profumo data dai luppoli. Il colore va dal chiaro all'ambrato a seconda dei malti utilizzati. Da questo stile ne discendono molti altri tra cui
    - *IPA (o Indian pale ale)* nata dalla necessità di trasportare birra ai coloni e soldati inglesi in India, così, affinché la bevanda venisse conservata per tutto il viaggio, venne aggiunto ulteriore luppolo durante la preparazione. Oggi è praticamente lo stile più diffuso al mondo. Da questo discendono altri stili come *Session IPA o Imperial IPA o Black IPA*
    - *Bitter* (dall'inglese 'amaro') sono birre molto simili alle ipa, sia per gusto sia per l'origine

- *Stout o Porter*. Come per le pale ale, anche le stout o porter (il nome è equivalente) sono uno stile da cui ne discendono molti altri. Le stout sono le classiche birre scure nate nelle isole britanniche, caratterizzate da sentori di fava di cioccolato e di caffè

  Da queste discendono
    - *Imperial stout* con stesse caratteristiche ma gusto molto più deciso e grado alcolico maggiore
    - *Sweet stout* che tendono più verso un gusto dolce
    - *Oatmeal stout* a cui viene aggiunto come cereale l'avena

Per quanto digressiva la mia è stata una panoramica molto veloce e sintetica del mondo degli stili della birra.
Ritornando al mio dataset, il numero di stili presenti all'interno del mio dataset è di 111.
Ho temporaneamente diviso il mio dataset in due: usa e resto del mondo e ho verificato che tutti i 111 stili siano presenti in entrambi i dataset (per vedere se ci fosse omogeneità nella distribuzione).

Essendo così alta la numerosità degli stili, risulta non prolifico andarne ad osservare le frequenze. Ci basti sapere che il dataset che stiamo analizzando presenta una diversificazione delle unità molto ampia che ci permetterà di svolgere al meglio le successive analisi.

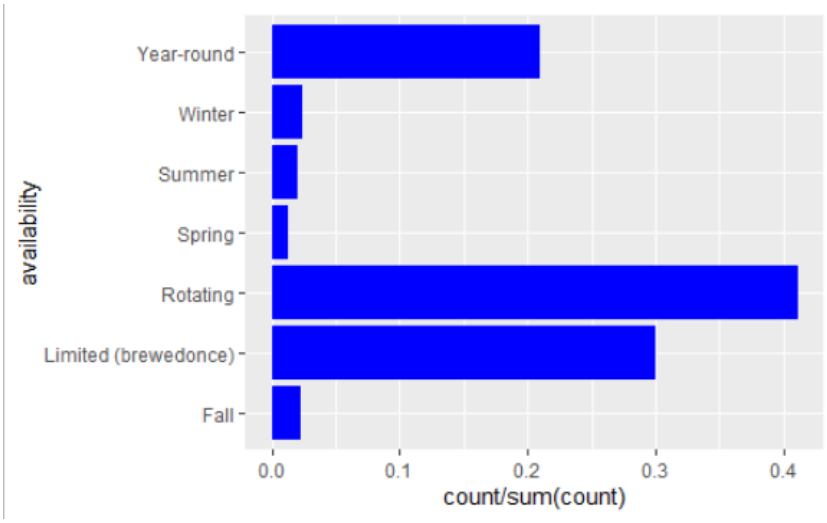
Availability

Il nome di questa variabile significa, traducendolo letteralmente, 'disponibilità' ed esprime dunque la disponibilità della birra sul mercato, nonché il suo periodo di produzione.

Come affermato in precedenza, molte birre presenti nel dataset provengono da birrifici artigianali i quali, a differenza di quelli industriali, producono una grande varietà di birre, seppur in quantità nettamente minore. Ciò porta alla conseguenza che alcune birre non siano sempre disponibili sul mercato.

La variabile *availability* spiega proprio questa caratteristica della birra in questione, ovvero quando è disponibile sul mercato.

È una variabile qualitativa che assume i seguenti valori (nel grafico si può vedere la distribuzione di frequenza che ogni livello assume)



Vediamo che circa il 90% delle osservazioni ha come valore 'year-round' (ovvero disponibili tutto l'anno), 'rotating' (prodotte a rotazione) o 'limited' (ovvero prodotte solo una volta esclusivamente).

Il 10% rimanente ha come valore i nomi delle stagioni indicando dunque che quella birra è prodotta solo in una determinata stagione.

Abv

La gradazione alcolica è la caratteristica forse più importante nelle birre, viene notata subito dal consumatore poiché descritta sull'etichetta ed è il primo parametro scelto dai mastri birrai per la creazione di una nuova birra.

La creazione dell'alcool nella birra avviene mediante un processo chimico in cui i lieviti reagiscono con gli zuccheri trasformandoli in alcool.

L'alcool contenuto in una birra viene espresso in termini di ABV (alcool by volume) che può variare da 0.1 fino a 12 e oltre.

Per calcolare il grado alcolico della birra ci sono diversi metodi tant'è che la "American Society of Brewing Chemists (ASBC)" lista sette differenti metodi.

Un metodo semplice per calcolare l'ABV durante la produzione della birra è quello di usare questa formula: (OG – FG) * 131, dove OG e FG sono rispettivamente la densità del mosto a inizio fermentazione e prima dell'imbottigliamento (quindi finita la fermentazione) a temperatura pari a 20º. Per misurare questi due valori si utilizza un semplice strumento, l'idrometro qui rappresentato, il cui utilizzo è molto intuitivo. Basta immergerlo in un liquido, nel nostro caso il mosto di birra, e leggerne la densità sulla scala posta sulla parte superiore.

Distribuzione della variabile ABV nel dataset

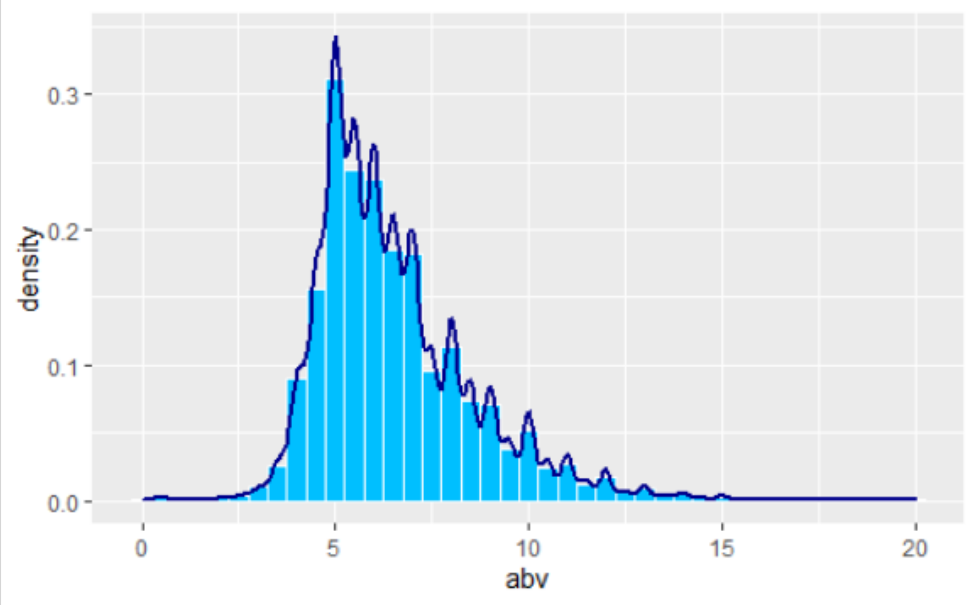


Tabella dei quantili:[6]

Ordine del quantile (α)	0.05	0.25	0.50	0.75	0.95
Valore del quantile (q)	4.1	5.0	6.0	7.5	10.5

È possibile osservare come il 50% delle osservazioni abbia un valore di *abv* compreso tra 5 e 7.5 e che il 90% l'abbia compreso tra i valori 4.1 e 10.5.

[6] Un quantile q di ordine α (con α compreso tra 0 e 1) è il valore di una distribuzione che la divide in due parti le cui frequenze relative sono α e 1-α.

SECONDO DATASET

Delle variabili appartenenti al secondo dataset non ne ho visualizzata la distribuzione poiché, come vedremo all'inizio del prossimo capitolo, questo dataset mi serve esclusivamente come input per crearne un altro ad hoc per la mia analisi. Mi limiterò quindi a spiegarne il significato, senza andare a visualizzarne le distribuzioni.

Name
Variabile che identifica la birra, non utile al fine dell'analisi.

Style
Variabile qualitativa che identifica lo stile della birra.
La descrizione è del tutto analoga a quella effettuata per la variabile omonima del dataset 'birre'.

Abv
La descrizione è del tutto analoga a quella effettuata per la variabile omonima del dataset 'birre'.

Ibu

Questa variabile quantitativa descrive il grado di amarezza delle birre. 'IBU' è infatti l'acronimo di International Bitterness Unit (unità di amarezza internazionale) ed è una scala usata per misurare l'amarezza nelle birre.
In realtà questa misura è una misura della quantità di iso-alpha acidi contenuti all'interno di una birra e non una misura della percezione dell'amarezza da parte di chi beve il prodotto. L'amarezza percepita può dunque risultare diversa dall'ibu indicato poiché dipende dalla composizione della birra stessa, nonostante la scala ibu dia un'indicazione importante.
Per la misura dell'ibu i mastri birrai utilizzano uno strumento chiamato spettrofotometro col quale, mediante un procedimento chimico, ottengono il valore di questo indicatore.

Color

È una variabile che rappresenta il colore della birra in questione. Come per l'amarezza, per quantificare questo aspetto nelle birre viene utilizzata una scala, SRM (standard reference method), con la quale è possibile associare ad ogni sfumatura di colore della birra un valore numerico.

Un'altra scala molto usata in Europa è l'EBC i cui valori sono uguali a quelli dell'SRM, raddoppiati.

Nel nostro dataset la variabile color è espressa in termini di SRM.

Entrando nel dettaglio, la sfumatura di colori che la birra può assumere ha un range che va da giallo molto chiaro (SRM tra 1 e 2) passando per ambrato, fino a quasi nero (SRM sui 40). In seguito, è riportata un'immagine indicativa della relazione tra valore del SRM e colore della birra.

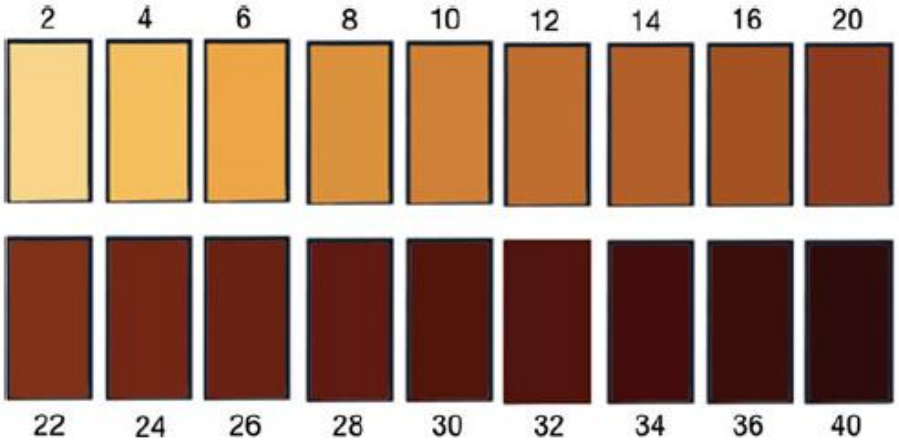


Fig. 2.3 SRM colors for beer and wort analysis

È bene notare come birre aventi lo stesso colore possano però appartenere a stili totalmente diversi e di conseguenza avere caratteristiche olfattive e gustative differenti.

CAPITOLO 3 - ANALISI DEI DATI

UNIONE DEI DATASET

Il mio obiettivo a questo punto è stato quello di far sì che gli stili del mio dataset principale trovassero una corrispondenza nel dataset 'ricette' da cui poterne reperire informazioni circa le caratteristiche.
Dal punto di vista statistico l'obiettivo era quello di creare una chiave che mi permettesse di collegare i due differenti dataset.

Mi sono dunque inizialmente concentrato sul dataset 'ricette' e da questo ho creato un dataset raggruppando le osservazioni (di numero 73861) in base allo stile e calcolandone la media delle tre caratteristiche (abv, ibu, color) per ognuno. Il nuovo dataset che ho creato ha tante osservazioni quanti sono gli stili di birra presenti nel dataset 'ricette' e ad ogni stile ho associato i corrispondenti valori medi delle variabili calcolati sempre sul dataset precedente.
Il dataset creato ha quindi le seguenti variabili:
"Style"
"n"
"abv"
"ibu"
"color"
In cui 'n' è il numero di osservazioni di un determinato stile del dataset 'ricette' di partenza e le restati tre variabili corrispondono, come detto in precedenza, alla media dei valori per stile.
Sarà questo dataset che userò per le mie analisi e lo chiamo 'stili'.

A questo punto sono andato a modificare manualmente gli stili nel primo dataset 'birre' (che ricordo essere di numerosità pari a 111) per far sì che questi trovassero una corrispondenza nel secondo 'stili'.
La prima operazione che ho fatto è stata quella di rimuovere le stopwords[7]: 'German ', 'American ', 'English ', 'Belgian ', 'Russian ', 'Irish ', 'Flanders '.

[7] Si tratta di un procedimento molto usato nel Text Mining in cui si crea un elenco di parole che si vogliono eliminare dal testo in questione

Trovando così 102 stili (alcuni che differivano per la stopwords iniziale sono stati uniti in uno solo es. 'american imperial stout' e 'russian imperial stout').

Ancora però non tutti trovavano una corrispondenza nel secondo dataset, a causa di scritture sbagliate, abbreviazioni, inversione di termini, ecc.

Sostanzialmente lo stile era presente nel secondo dataset ma era scritto in maniera diversa.

Ho quindi modificato manualmente i nomi degli stili per adattarli, arrivando così ad ottenere l'elenco riportato:

- Lager
- Altbier
- Red Ale
- International Amber Lager
- Baltic Porter
- Barleywine
- Berliner Weisse
- Null
- Ordinary Bitter
- Specialty IPA: Black IPA
- Blonde Ale
- Traditional Bock
- Bohemian Pilsener
- Braggot
- Brett Beer
- Brown Ale
- California Common Beer
- Spice Herb or Vegetable Beer
- Cream Ale
- Dark Mild
- Wheat or Rye Beer
- Doppelbock
- Dry Stout
- Dubbel
- Dunkelweizen
- Eisbock
- International Dark Lager
- Dortmunder Export
- International Pale Lager
- Premium Lager
- Extra Special/Strong Bitter (ESB)
- Lambic
- Sahti
- Foreign Extra Stout
- Bière de Garde
- Fruit Beer
- Fruit Lambic
- Gueuze
- Weizen/Weissbier
- Munich Helles
- Imperial IPA
- Pilsner (Pils)
- Imperial Stout
- IPA
- Kellerbier: Pale Kellerbier
- Kölsch
- Gose
- Light Lager
- Maibock/Helles Bock
- Oktoberfest/Märzen
- Munich Dunkel
- Oatmeal Stout
- Old Ale
- Oud Bruin
- Pale Ale
- Mild
- Wheat Beer
- Porter
- Tripel
- Rauchbier
- Robust Porter
- Roggenbier
- Saison
- Schwarzbier
- Strong Scotch Ale
  Scottish Export
- Classic Style Smoked Beer
- Stout
- Strong Ale
- Dark Strong Ale
- Sweet Stout
- Vienna Lager
  Weizenbock
- Wheatwine
- Wild Specialty Beer
- Holiday/Winter Special Spiced
- Witbier

Dopo le modifiche il totale degli stili è pari a 76 (più 1, 'null' in cui ho riunito tutti gli stili troppo particolari).
La riduzione è dovuta al fatto che alcuni stili che di partenza erano diversi sono stati uniti con gli aggiustamenti linguistici sotto un unico nome.
Tutti questi stili trovano una corrispondenza nel dataset creato all'inizio della analisi 'stili'.
Ho quindi ottenuto una chiave di collegamento tra i due dataset, ovvero lo stile delle birre, ognuno con le proprie caratteristiche espresse in termini numerici nelle variabili abv, color, ibu.
Ricapitolando, questi sono i dataset con cui mi trovo a lavorare adesso:
- 'birre'

Id	name	country	state	abv	availability	**STYLE**
..	..	..	..	..	..	..
..	..	..	..	..	..	..

Numero di osservazioni: 351699 (ho escluso dal dataset di partenza tutte le osservazioni con valore style pari a 'null').
Livelli della variabile style = 76.
- 'stili'

STYLE	n	ibu	abv	color
..	..	..	..	..
..	..	..	..	..

Numero di osservazioni: 76

CLUSTERIZZAZIONE

In questo paragrafo affronto la fase di clusterizzazione dei dati, in particolare, considerando il dataset 'stili' ho voluto raggruppare gli stili in 'macro stili' (o 'macro famiglie').
Per far ciò ho usato diverse tecniche di clustering che ho confrontato alla fine. Il Clustering è un procedimento statistico il cui scopo è quello di assegnare le unità statistiche del dataset a dei gruppi, chiamati appunto cluster.

Cluster gerarchico

Esistono due tipi di cluster gerarchico, quello agglomerativo (da me usato) e quello divisivo.

Il primo ha un approccio di tipo 'bottom-up', in cui inizialmente ogni osservazione viene vista come un unico cluster e ad ogni step dell'algoritmo vengono uniti i cluster più simili fino a quando tutte le osservazioni sono unite sotto un unico cluster. La formazione dei cluster può essere visualizzata attraverso un dendogramma[8], detto anche grafico ad albero.

Il metodo divisivo agisce in maniera opposta (approccio 'top down'), considerando inizialmente tutte le osservazioni come appartenenti ad un cluster e dividendole ad ogni step.

L'algoritmo di cluster agisce su una matrice di distanze o dissimilarità in cui ad ogni combinazione di due osservazioni viene associato un valore che ne indica la dissimilarità (più questo è alto, più le osservazioni saranno distanti e quindi dissimili tra di loro).

Per calcolare le distanze e quindi creare la matrice che si utilizzerà, si possono usare diverse metriche (una metrica è una misura di distanza tra un paio di osservazioni).

Nella mia analisi ho creato tre matrici di dissimilarità con i seguenti metodi (riportati con le formule) e ne ho confrontati i dendogrammi, tagliandoli ad un'altezza tale da avere quattro gruppi:

[8] Un dendogramma è un tipo di grafico che permette una visualizzazione rapida ed efficace del processo di raggruppamento di vari elementi

2.1 Euclidean Distance

Euclidean distance computes the root of square difference between co-ordinates of pair of objects.

$$Dist_{XY} = \sqrt{\sum_{k=1}^{m} (X_{ik} - X_{jk})^2}$$

2.2 Manhattan Distance

Manhattan distance computes the absolute differences between coordinates of pair of objects

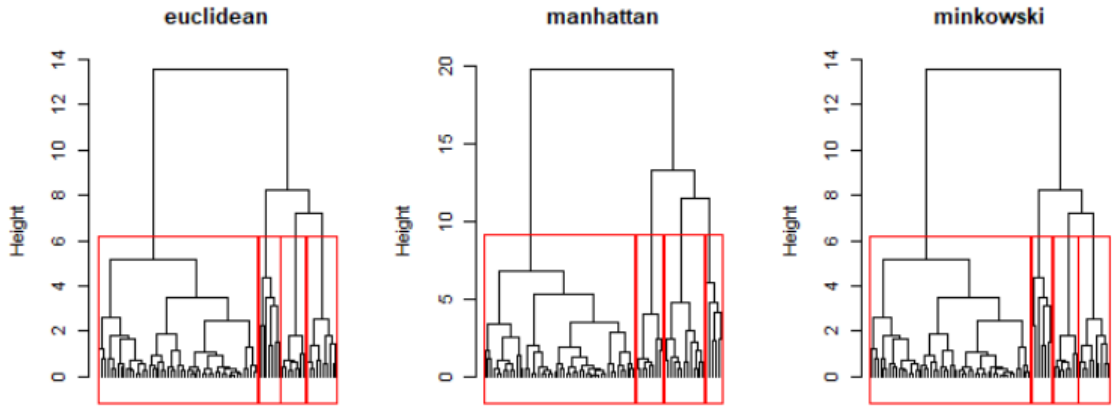
$$Dist_{XY} = |X_{ik} - X_{jk}|$$

2.4 Minkowski Distance

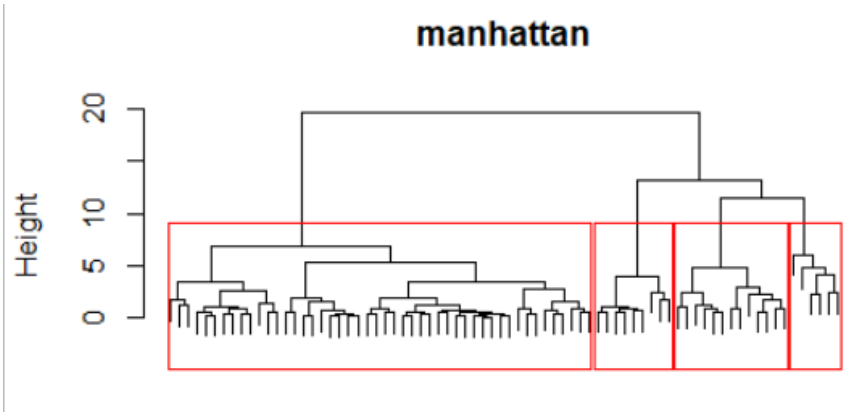
Minkowski Distance is the generalized metric distance.

$$Dist_{XY} = \left(\sum_{K=1}^{d} |X_{ik} - X_{jk}|^{\frac{1}{p}} \right)^{p}$$

Note that when p=2, the distance becomes the Euclidean distance. When p=1 it becomes city block distance. Chebyshev distance is a variant of Minkowski distance where p=∞ (taking a limit). This distance can be used for both ordinal and quantitative variables.



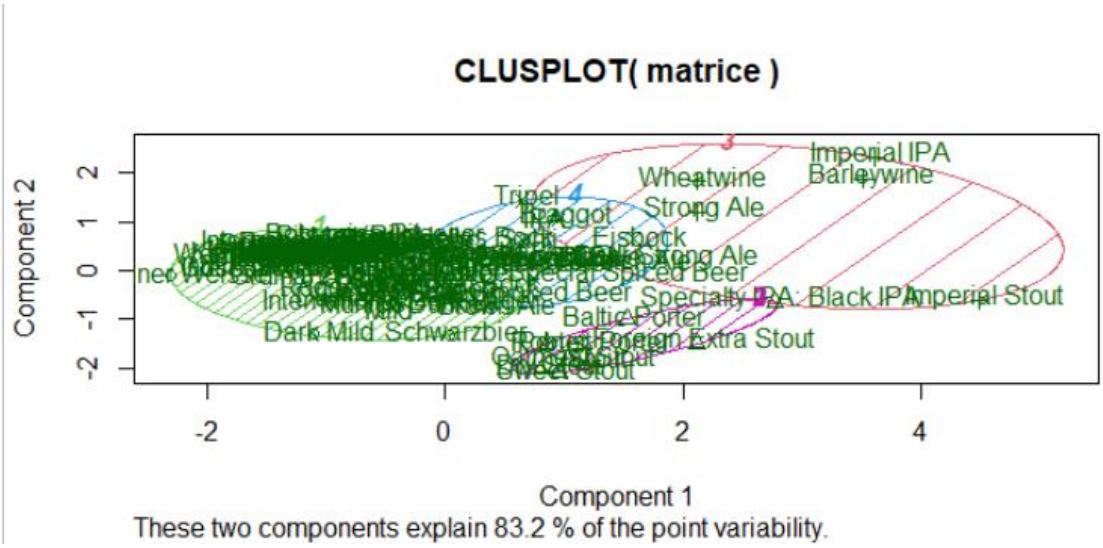
Opto la distanza di Manhattan poiché distingue meglio i 4 gruppi.



Ho poi calcolato, per ognuno dei quattro gruppi, le medie delle caratteristiche degli stili e ho creato una tabella per visualizzare i risultati.

Gruppo	abv	ibu	color	n
1	5.38	25.9	10.0	48
2	6.17	43.3	35.7	9
3	8.81	78.2	17.9	6
4	8.10	29.2	18.0	13

Riporto il relativo clusplot[9] (le due componenti principali spiegano l'83.2% della variabilità, il che è un risultato statistico molto buono).



[9] Il clusplot utilizza la Principal Component Analysis (PCA) per creare il grafico. Per spiegare i dati vengono utilizzate le due principali componenti. Se i dati hanno dimensionalità maggiore di due la percentuale della variabilità spiegata dalle due componenti principali è direttamente proporzionale alla correlazione tra le variabili

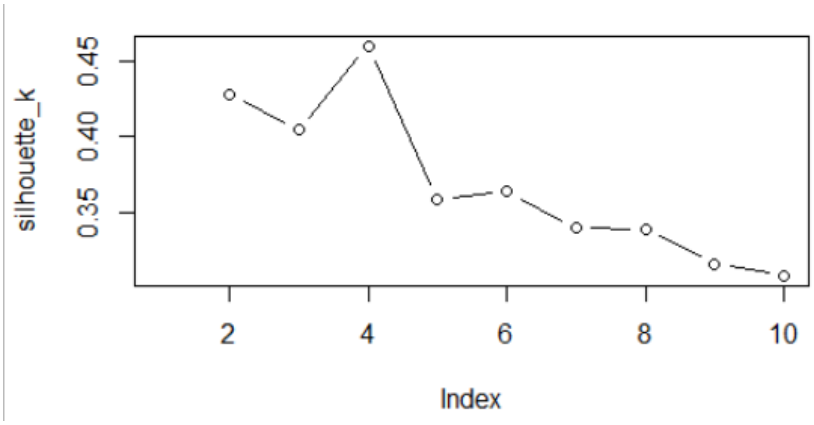
Cluster partizionale

L'algoritmo che viene usato nel clustering partizionale è quello detto delle k-medie ed è così strutturato:

1- Viene scelto il numero di cluster k a priori, e vengono scelti k *centroidi*[10] in maniera casuale.
2- Ogni osservazione viene assegnata al cluster con il cui *centroide* è più vicino.
3- Vengono ricalcolati i *centroidi* di ogni cluster come media delle osservazioni appartenenti al cluster allo step precedente.
4- Ogni osservazione viene riassegnata al cluster con il cui *centroide* è più vicino.

Gli step 3 e 4 vengono ripetuti fino alla convergenza, ovvero fino a quando solo un numero molto limitato di punti cambia cluster di appartenenza.

Per scegliere il miglior k ho cercato quello che massimizzasse la *silhouette*[11], uno strumento che ci permette di capire in quanti gruppi è più opportuno dividere il dataset, in un intervallo di k che va da 2 a 10.



Vediamo che il numero di gruppi migliore è 4, con valore della silhouette pari a 0.45997087.

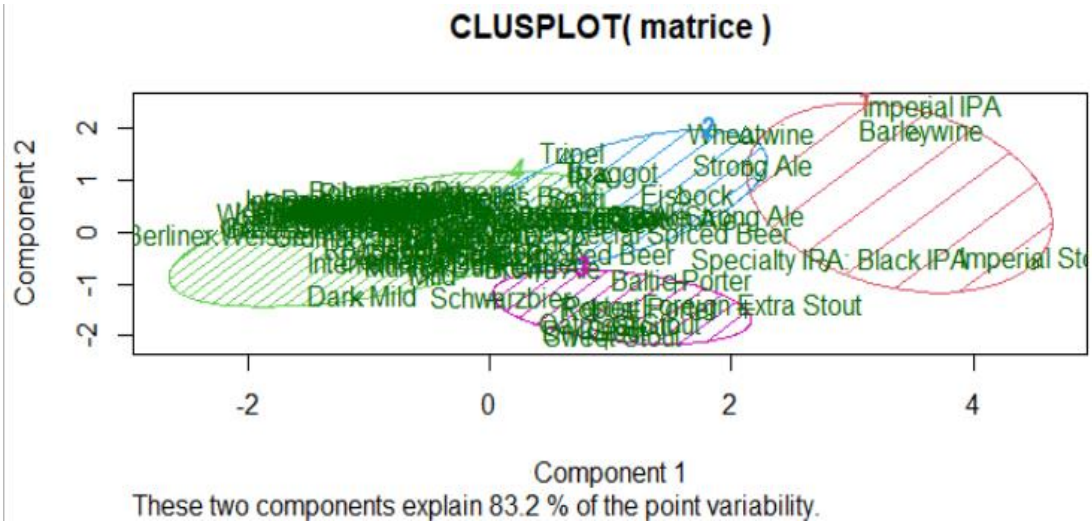
[10] Un centroide di un gruppo è un punto avente come valori delle variabili analizzate le medie dei valori assunti dalle variabili in quel gruppo

[11] Il valore della è una misura di quanto è simile un oggetto al proprio cluster di appartenenza rispetto agli altri cluster. Questo valore varia da -1 a +1

Tabella delle caratteristiche medie dei 4 gruppi:

Gruppo	**abv**	**ibu**	**color**	**n**
1	8.74	85.1	24.3	5
2	8.37	31.6	17.4	13
3	5.96	37.7	33.8	10
4	5.41	26.4	9.52	48

Con relativo clusplot:



Model based clustering

Il model based clustering è una tecnica di clusterizzazione fondata sull'ipotesi che la popolazione sia costituita da cluster (non osservati) in cui la legge di distribuzione varia, ovvero la popolazione si distribuisce come una mistura finita[12].

Questo procedimento consiste nell'utilizzare un modello ad hoc che assegni le unità statistiche ai diversi gruppi.

Le operazioni da fare sono una selezione del modello e la stima dei parametri del modello scelto.

Per scegliere il modello più adatto si vanno a confrontare diversi modelli, misture di normali, che differiscono per il numero k di cluster e per i vincoli. I vincoli influiscono sul numero di parametri, ovvero più sono i vincoli imposti al modello, minore saranno i parametri, e di conseguenza il modello risulterà più semplice.

Uno strumento che ci permette di ottenere il trade-off ottimale tra semplicità del modello e numero di cluster è l'ICL[13], che è una misura della bontà di clusterizzazione. Difatti ogni modello ha associato un valore per il parametro ICL e il modello migliore tra tutti sarà quello che ha il valore ICL maggiore.

Per la mia analisi ho utilizzato il model based clustering facendo variare il k (numero di gruppi) tra 2 e 10 così da confrontare i vari modelli e ottenere il modello migliore.

Riporto l'output da Rstudio:

'Mclust' model object: (VVI,4)

Dall'analisi risulta che il modello è del tipo VVI (in riferimento ai parametri dello stesso) e che il numero di gruppi ottimo è 4. Questo modello presenta un ICL pari a -545.1258.

[12] Si definisce mistura (finita) una variabile casuale avente una funzione di densità dipendente da η (vettore dei parametri) composta dalla sommatoria delle funzioni di densità delle singole componenti per la probabilità associata ad ognuna

[13] ICL è definito come la somma tra il BIC (una misura della bontà del fit del modello) e l'entropia (che misura la bontà dei cluster)

La tabella delle caratteristiche medie dei quattro gruppi è la seguente:

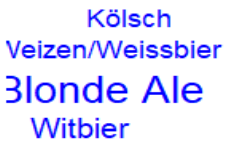
Gruppo	**abv**	**ibu**	**color**	**n**
1	5.60	26.3	13.1	37
2	8.45	50.6	20.5	18
3	5.22	24.9	4.78	15
4	5.71	37.3	36.0	6

Confronto e scelta

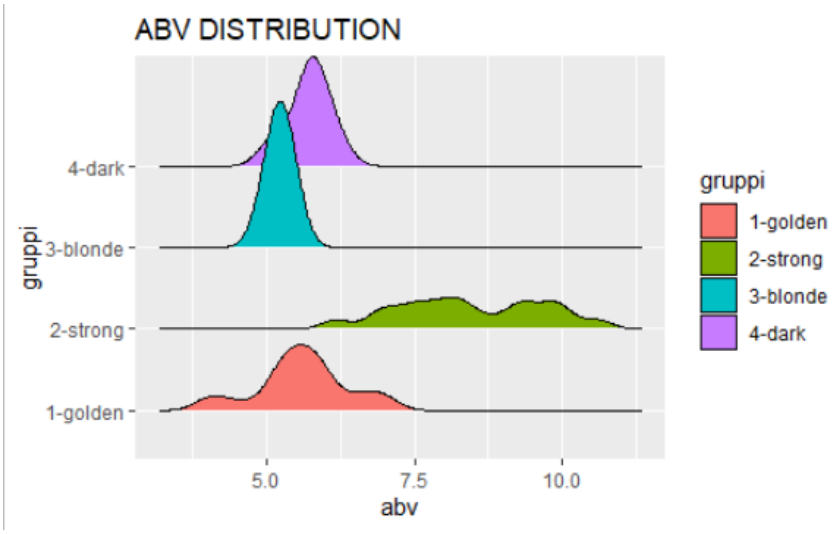
Dopo aver eseguito l'operazione di clustering mediante diverse tecniche vado a confrontarle per scegliere quella da utilizzare.
La grande differenza che esiste tra i primi due metodi e il terzo è che i primi sono tecniche euristiche, ovvero non sono poste ipotesi a priori sulla popolazione. Inoltre, sono algoritmi che possono essere sviluppati mediante diversi metodi (nella mia analisi ho utilizzato il k-means che è il più comune ma ne esistono altri come Wold o k-medians) e presentano soluzioni discordi. Infine, proprio perché non si può definire oggettivamente quale soluzione sia la migliore (ad esempio nel cluster gerarchico, a che livello tagliare il dendogramma), questa risulta essere spesso soggettiva.
Il model based clustering permette invece di confrontare tutti i modelli applicabili ad un problema di clusterizzazione (sotto ipotesi che la distribuzione della popolazione sia una mistura finita) e selezionare il migliore in maniera oggettiva. Nel calcolo dell'ICL si tiene conto infatti di diversi fattori quali la massima verosimiglianza, la dimensione del vettore dei parametri, la numerosità della popolazione e l'entropia.
Per i motivi sopra elencati utilizzerò come gruppi creati l'output del processo del model based clustering.

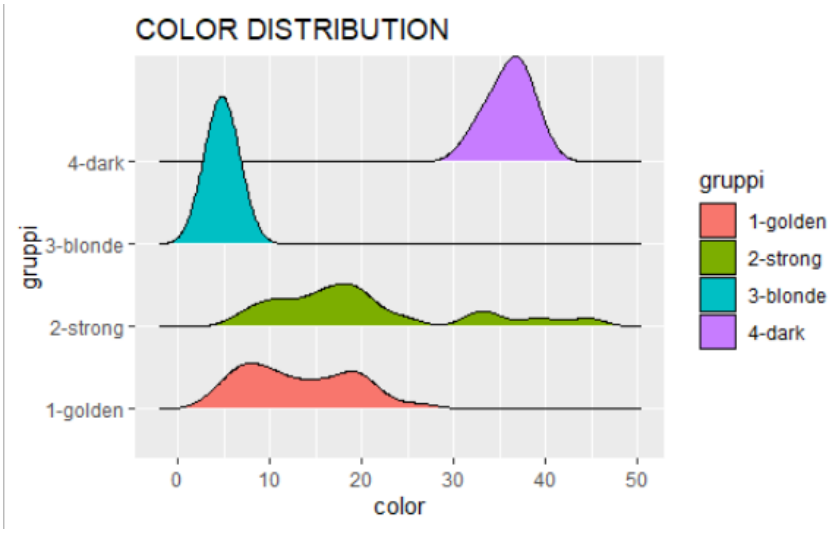
Ho rinominato dunque in maniera totalmente soggettiva le 4 classi in modo tale che abbiano un nome in base alle caratteristiche che presentano.
- i gruppi 1 e 3 essendo molto simili tra loro li ho distinti per la variabile 'color' (chiamo 'golden' per il primo gruppo, che ha un colore medio leggermente più scuro del terzo, rinominato 'blonde').
- per il secondo gruppo, poiché presenta tutte le caratteristiche con valori medi alti (quindi alto grado alcolico, alta amarezza e anche un colore ambrato) l'ho denominato 'strong'.
- per l'ultimo gruppo ho anche qui fatto risaltare la caratteristica che lo differenzia dagli altri, ovvero il colore (denominandolo dunque 'dark').

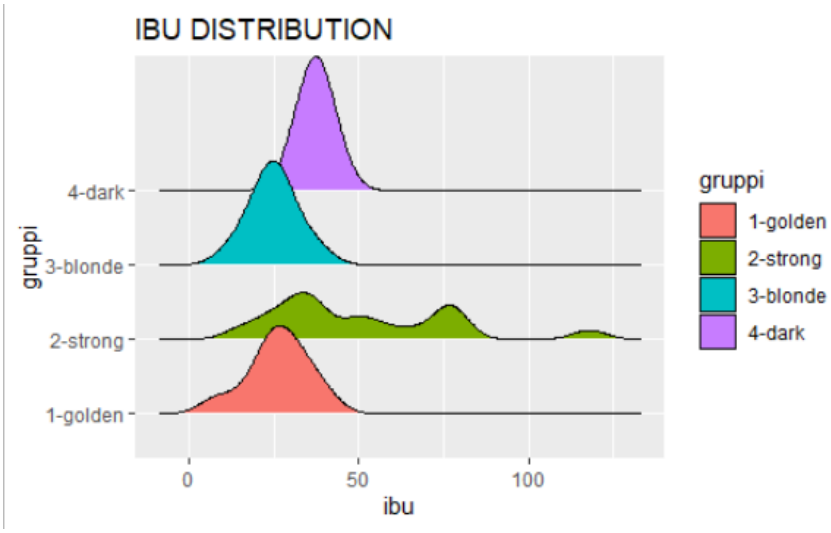
Visualizzo il corrispondente wordcloud (si tratta di uno strumento che ci permette di visualizzare le principali parole dei vari gruppi, più è alta la numerosità della parola, più grande sarà visualizzata la stessa all'interno del wordcloud).



Vediamo dunque la distribuzione delle variabili nei vari gruppi che ho creato.







Commenti:

- Partiamo dal quarto gruppo ‘dark’, questo raggruppa le birre che hanno esclusivamente valori di color molto alti ma valori di ibu e abv nella media.

- Il secondo gruppo strong presenta una variabilità alta in tutte e tre le variabili, come si può vedere dalla densità di frequenza. Ciò implica che il secondo gruppo include quegli stili che hanno o un abv alto o un ibu alto (o entrambi, la condizione è che ALMENO uno dei due lo sia) oppure un color alto E uno tra abv e ibu alti (o entrambi). Il solo color alto non basta infatti perché quegli stili rientrano nel quarto gruppo.

-Notiamo infine che la differenza tra il primo e il terzo gruppo (che presentano valori medi delle variabili molto simili) sta nel fatto che il terzo gruppo ‘blonde’ ha una

variabilità molto minore rispetto al primo ‘golden’. Quindi, prendendo gli stili con valori bassi di tutte le variabili, quegli stili che hanno tutti i valori strettamente vicini alle medie (di entrambi i gruppi poiché è molto simile) appartengono al terzo gruppo, mentre quegli stili con valori si bassi, ma più distanti rispetto alla media (anche solo in una variabile) appartengono al primo.

A questo punto aggiungo al mio dataset ‘stili’ una variabile, *gruppi*, che assumerà come valore uno dei quattro gruppi a seconda dell’assegnazione precedentemente effettuata con il model based clustering.
Il dataset ‘stili’ diventa dunque:

STYLE	n	ibu	abv	color	**GRUPPI**
..	..	..	..	..	..
..	..	..	..	..	..

E per completezza riporto il dataset dei gruppi con le caratteristiche medie di ognuno (la tabella completa con i valori è quella risultante dal model based clustering).

GRUPPI	Abv medio	Ibu medio	Color medio	n
..	..	..	..	..
..	..	..	..	..

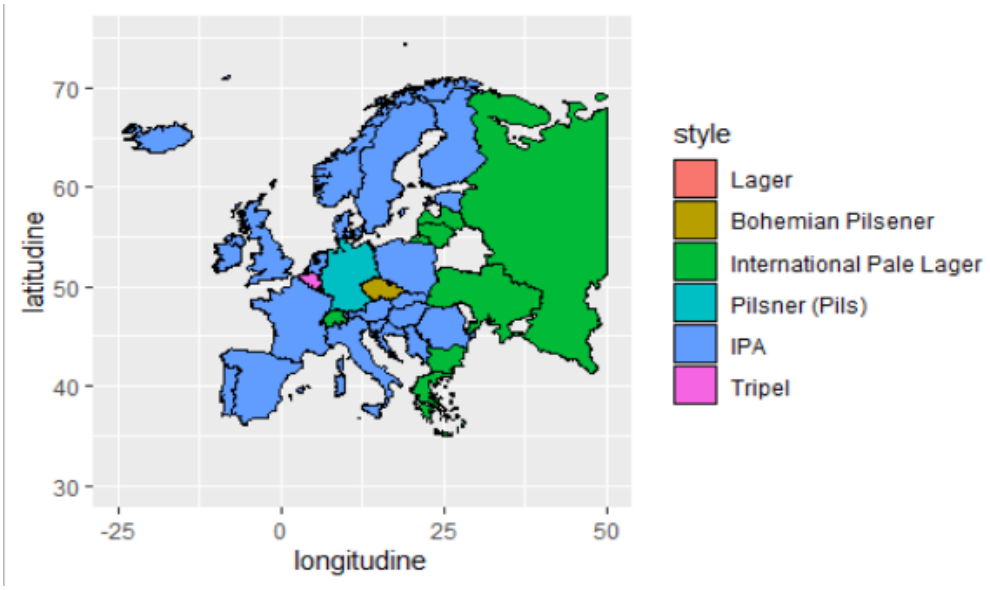
Dove la variabile *gruppi* svolge la funzione di chiave nel collegare i due dataset.
A questo punto tutti e tre i dataset sono connessi tra di loro.

ANALISI

Questa è la parte del mio lavoro in cui cerco di trovare le risposte alle domande poste inizialmente.
Dovrò quindi analizzare la distribuzione dei vari stili, delle marco-famiglie e delle variabili che caratterizzano gli stili per trovare analogie e differenze tra nazioni, e, cosa più importante, verificare se la distribuzione geografica delle caratteristiche è coerente con la distribuzione degli stili di birra.
Per fare ciò analizzo i singoli grafici che riguardano solo l'Europa, per diverse motivazioni.
La prima è che l'Europa è il continente più rappresentato e abbiamo già visto come le mappe globali presentino una qualità minore (per la presenza di ampi spazi bianchi, dovuta alle nazioni non analizzate le cui birre prodotte erano di numero inferiore a 100) che non quelle europee.
La seconda è che essendo l'Italia in Europa è molto più interessante concentrarsi su nazioni a noi vicine, le cui birre sono facilmente reperibili anche in Italia.
Si tratta comunque di una scelta logistica soggettiva.
Le analisi che farò in seguito possono anche essere fatte su scale globale oppure concentrandosi sui singoli stati degli USA (che ricordo coprono il 60% del dataset).

Conclusa questa premessa vado a rispondere alle domande con delle mappe apposite che poi a commenterò.

- In ogni stato, quale stile è il più rappresentato?



Vediamo che gli stili predominanti sono ‘ipa’ e ‘International Pale Lager’ che insieme ricoprono quasi la totalità degli stati considerati. È bene ricordare che ogni birra è considerata in maniera univoca e non si tengono conto delle quantità. La mappa non ci riporta difatti quale stile occupa la più ampia fetta di mercato, ma lo stile in cui sono prodotte più birre in ogni nazione. Lo stile ‘ipa’ è predominante in molti stati in quanto è molto diffuso nei birrifici artigianali, che producono dunque molto birre in questo stile, seppur in quantità limitate (non è detto che lo stile ‘ipa’ ricopra la più ampia fetta di mercato).

È interessante fare una constatazione su Germania, Belgio e Repubblica Ceca, in cui gli stili di birra maggiormente prodotti sono rispettivamente ‘pilsner’ ‘bohemian pilsner’ e ‘tripel’.

Questo dato non è assolutamente casuale ma rispecchia una precisa cultura e storia della birra nei tre paesi. Questo risultato, seppur quasi scontato, ci dà invece un’indicazione fondamentale nello studio della produzione della birra attuale. Il fatto che dopo secoli, nonostante la forte innovazione nei processi di produzione, determinate regioni continuino a produrre un determinato stile o tipo di birra, è indice di un forte legame tra il territorio e la cultura della birra.

La motivazione è determinata dal fatto che i nuovi birrifici hanno spesso preso spunto dai birrifici storici sia per quanto riguarda il prodotto finale sia per gli ingredienti e i processi di lavorazione.

- Relazione tra le variabili

  - Grafico correlazioni[14]:

In tutti e tre i casi troviamo una correlazione positiva tra due variabili, il che indica che mediamente all'aumentare di una delle tre variabili aumentano anche le altre due.

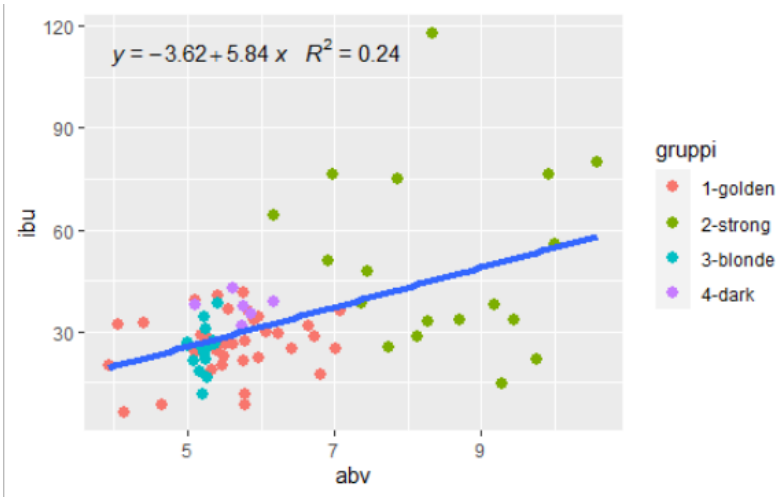
Andiamo a vedere nello specifico le rette calcolate mediante un modello di regressione lineare semplice, osservando i valori assunti dalle costanti.
Date *n* osservazioni che assumono *n* valori per le variabili X (v. indipendente) e Y (v. dipendente), un modello di regressione lineare semplice permette di stimare l'equazione della retta che meglio interpreta l'andamento tra le due variabili.
L'equazione della retta da stimare ha la seguente struttura: $Y = a + bX$.
Il metodo mediante cui vengono stimati i parametri *a* (valore assunto quando $X = 0$) e *b* (pendenza) è quello dei minimi quadrati (OLS)[15].

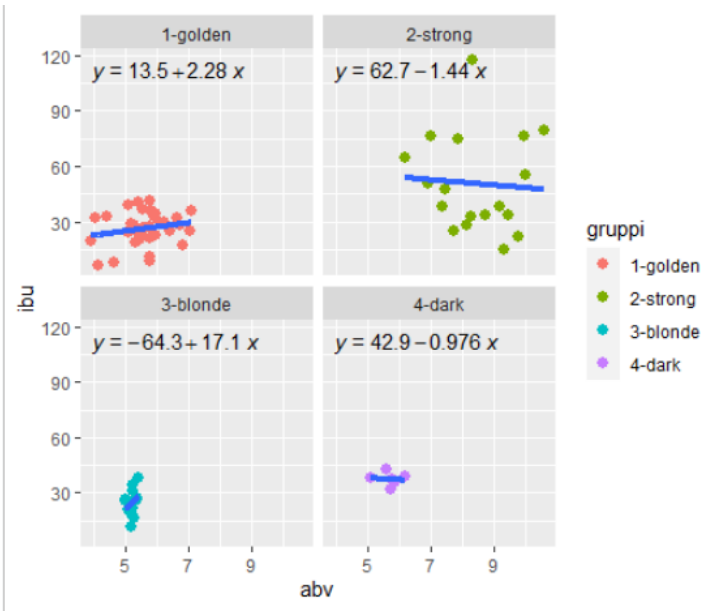
[14] L'Indice di correlazione lineare di Pearson esprime quanto le due variabili si muovono concordemente: oscilla fra -1 e +1, dove il segno indica l'andamento della relazione (positivo se le due variabili crescono o decrescono assieme; negativo se al crescere di una, l'altra decresce), e dove il valore 1 della correlazione (in valore assoluto) indica la correlazione perfetta, 0 la correlazione nulla.

[15] Il metodo dei minimi quadrati stima i parametri di regressione in modo tale da minimizzare l'errore $\sum_{i=1}^{n} e_i^2$ dove $e_i = Y_i - \widehat{Y_i}$, per ogni *i* unità statistica ($\hat{Y}_i$ è il valore che il modello prevede o fitta per Y_i)

Decido di riportare per ogni relazione, oltre i grafici con il totale delle osservazioni, anche i grafici per i gruppi distinti, per avere un'idea di come si comportano le variabili all'interno dei singoli gruppi. A causa del basso numero di osservazioni all'interno dei gruppi, per le analisi cosidererò i risultati numerici derivati dai grafici complessivi che contengono un numero sufficientemente alto di unità (76).

- Plot IBU ~ ABV (con distinzione per i gruppi)



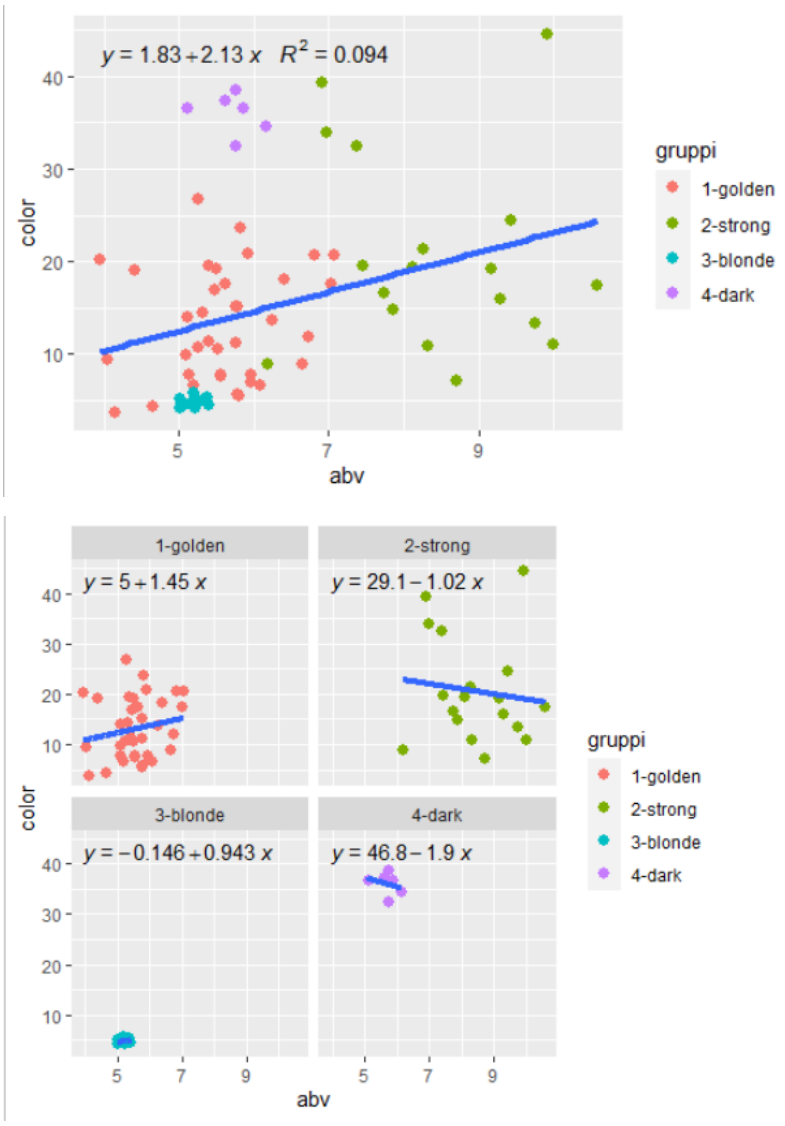


Come ci aspettavamo al crescere della variabile indipendente *abv*, troviamo una crescita della variabile indipendente *ibu*. Per correttezza statistica si dice che al crescere di un'unità di *abv*, *ibu* cresce mediamente

di 5.84 al netto delle altre variabili. Ovvero il modello ci informa che, lasciando invariati i valori delle altre variabili, se aumentiamo *abv* di un'unità l'*ibu* sale mediamente di 5.84 unità.

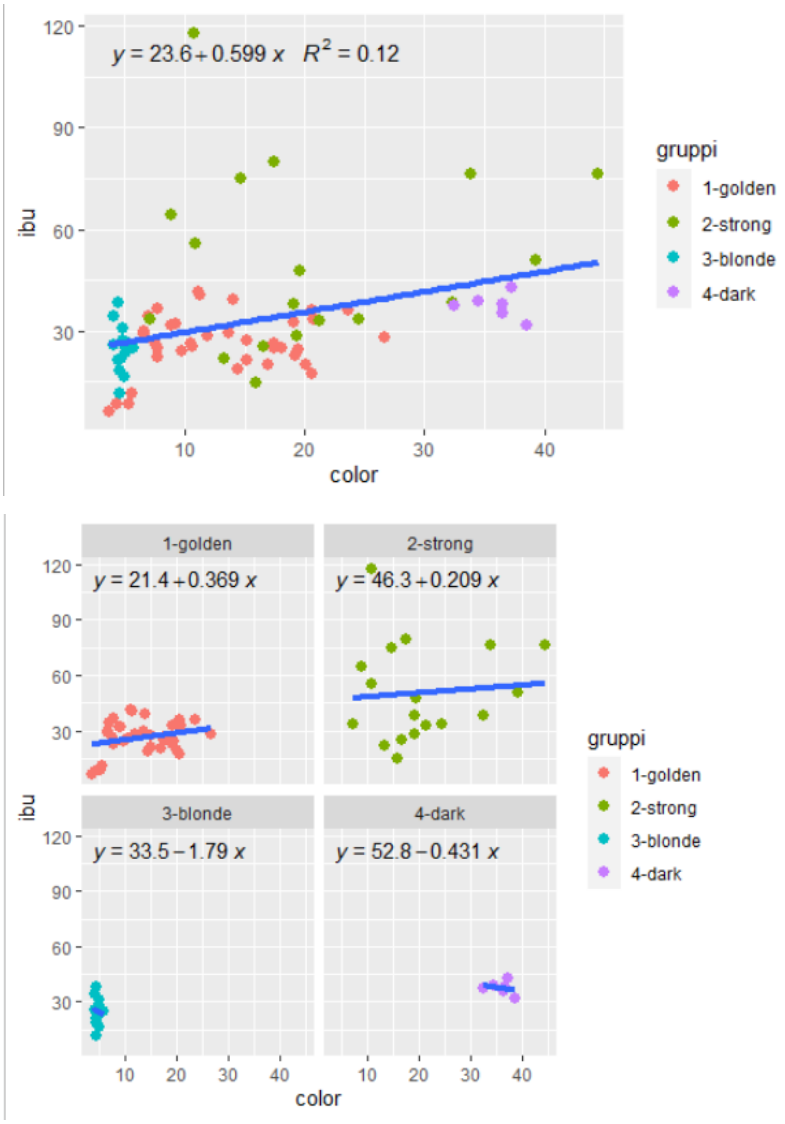
Dalla seconda immagine, si osserva come la retta abbia pendenza positiva nei gruppi 1 e 3, e negativa nei gruppi 2 e 4.

- Plot COLOR ~ ABV (con distinzione per i gruppi)



Al crescere di un'unità di *abv*, *color* cresce mediamente di 2.13 al netto delle altre variabili. Anche in questo caso, i gruppi 1 e 3 presentano una pendenza positiva mentre nei gruppi 2 e 4 è negativa.

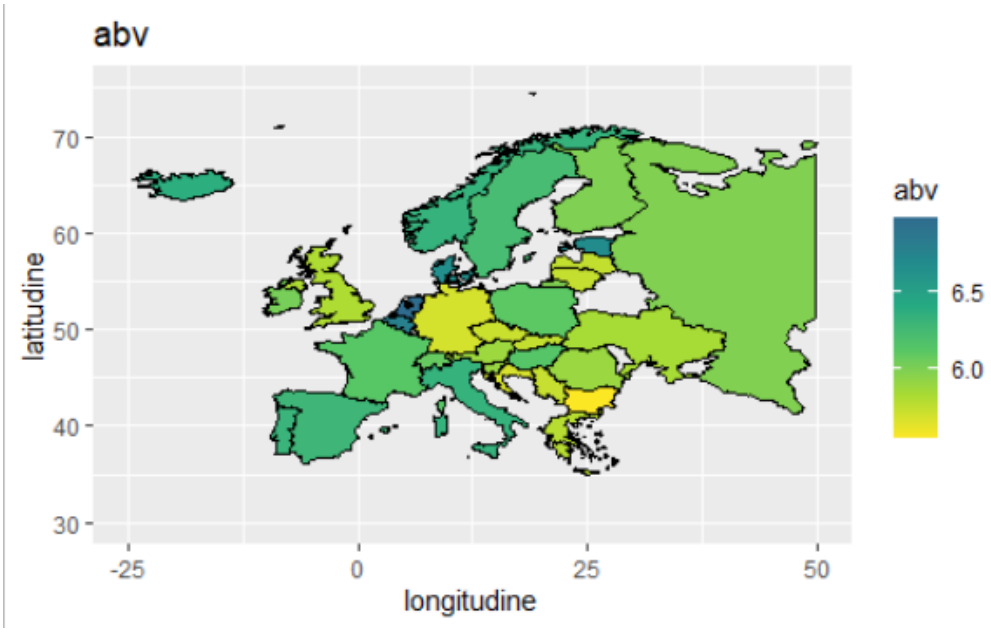
- Plot IBU ~ COLOR (con distinzione per i gruppi)



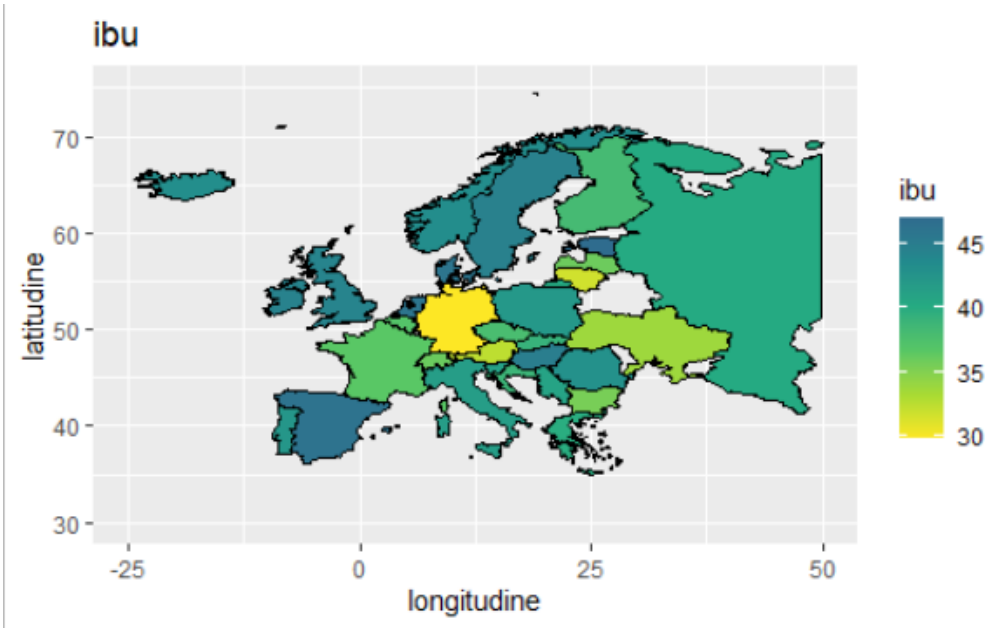
Al crescere di un'unità di *color*, *ibu* cresce mediamente di 0.599 al netto delle altre variabili. Come si può vedere dai grafici, i gruppi esprimono anche in questa situazione andamenti delle rette discordi.

Dalle considerazioni appena fatte ci aspettiamo che la distribuzione dei valori delle tre caratteristiche nei vari stati rispecchi l'andamento analizzato. Ovvero l'aspettativa è quella di avere una discreta continuità nei valori assunti dalle tre variabili dal punto di vista geografico.

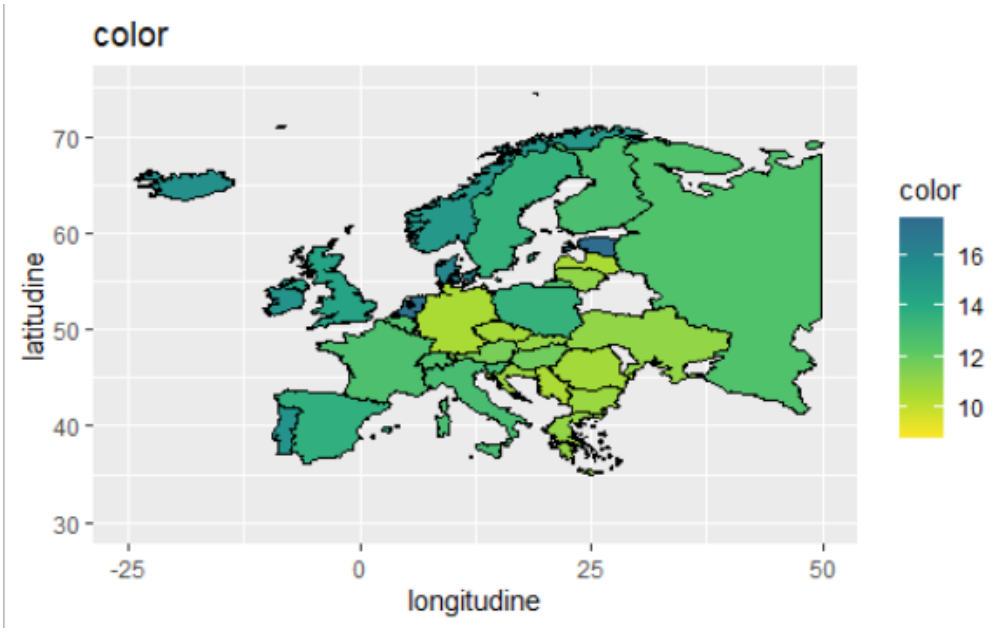
- Distribuzione ABV per nazione



- Distribuzione IBU per nazione



- Distribuzione COLOR per nazione



Andiamo ad osservare come in Germania i valori delle tre variabili siano evidentemente più bassi che non nei paesi confinanti. Notiamo come tutti i valori siano bassi, confermando le ipotesi precedenti.
L'esempio opposto può essere espresso dai Paesi Bassi in cui i valori di tutte e tre le variabili si presentano alti, confermando ancora una volta la dipendenza tra di esse.
Anche analizzando nazioni i cui valori delle variabili si presentano intermedi (come ad esempio l'Italia) notiamo la stessa situazione ovvero una costanza in tutte e tre le variabili nella posizione dei valori lungo la scala.

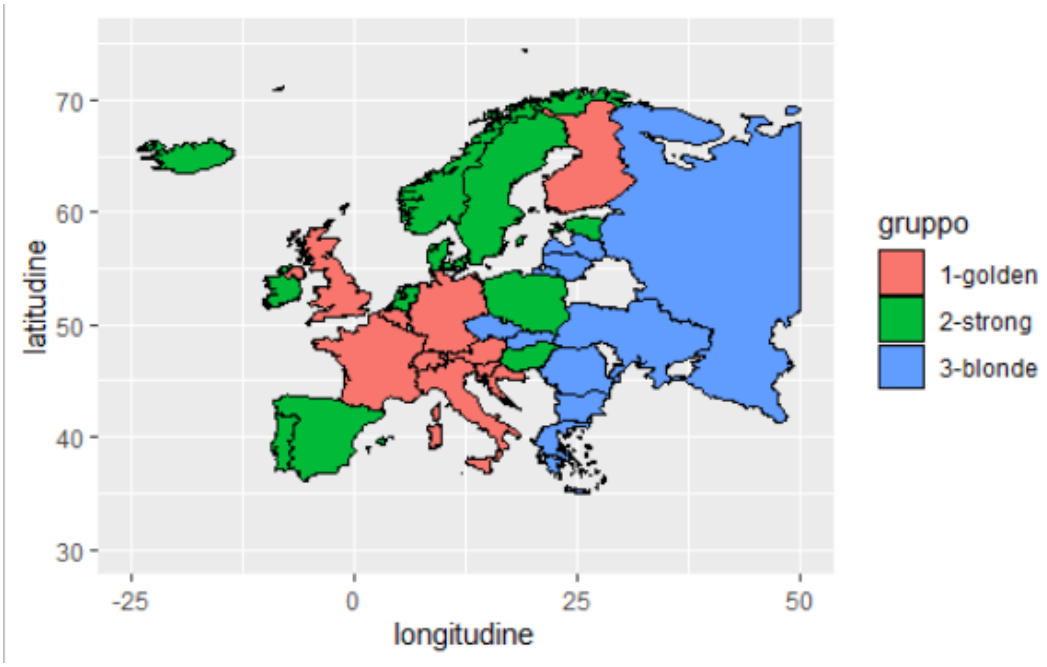
Per non creare interpretazioni errate non stiamo verificando che CERTAMENTE a valori di una variabile bassi (in una nazione) corrispondono valori bassi di altre variabili ma che MEDIAMENTE si conferma il trend e la correlazione delle variabili assunta in precedenza.

Questo risultato è molto importante perché oltre a confermare le informazioni assunte in precedenza circa la connessione tra le caratteristiche, ci permette di andare a vedere le differenze qualitative nella la produzione di birra nei vari paesi.

- In ogni stato, quale macro-famiglia è la più rappresentata?

  Riporto la tabella con i dati dei quattro gruppi calcolata in precedenza.

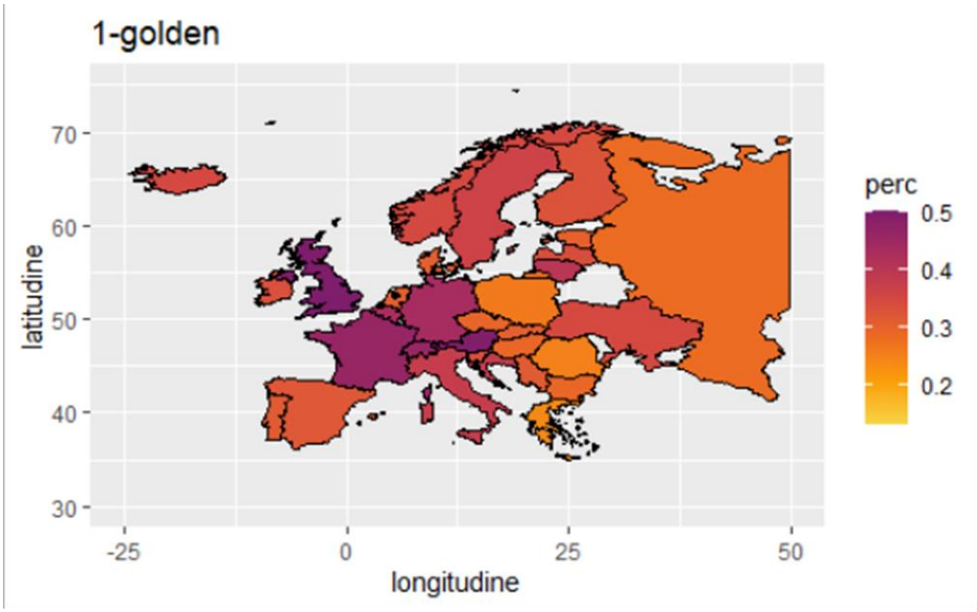
GRUPPI	abv	ibu	color	n
1-golden	5.60	26.3	13.1	37
2-strong	8.45	50.6	20.5	18
3-blonde	5.22	24.9	4.78	15
4-dark	5.71	37.3	36.0	6

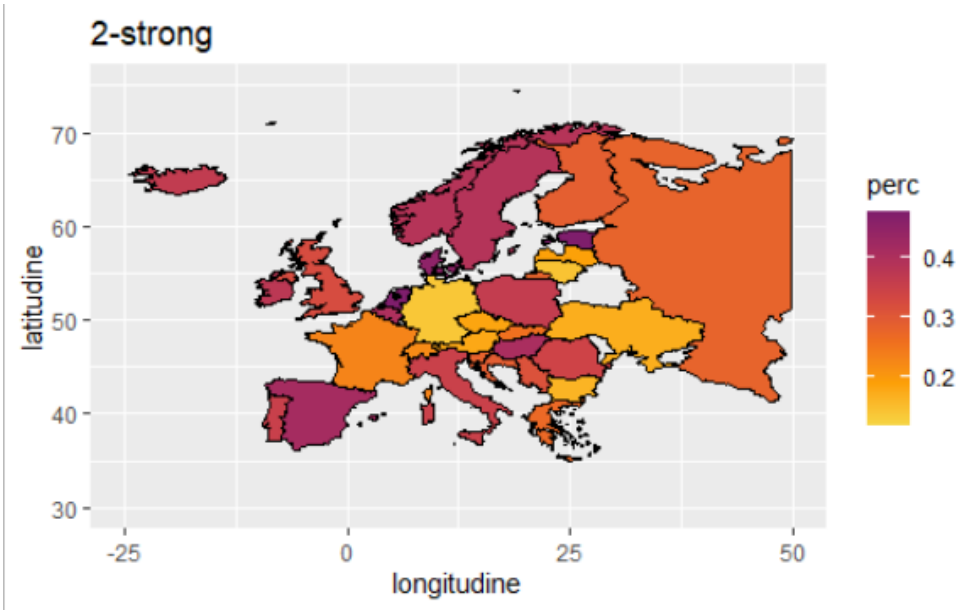


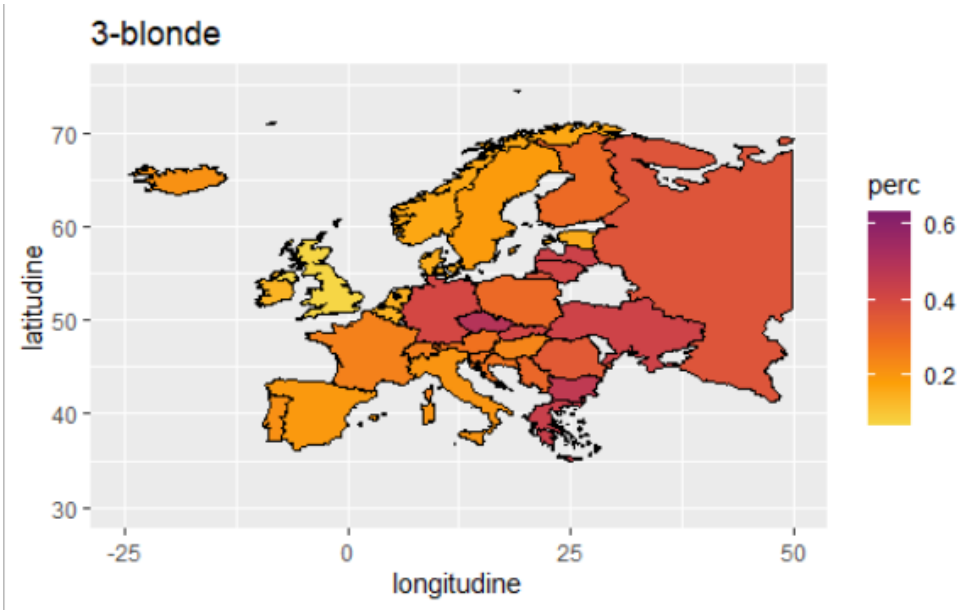
Notiamo subito dalla mappa che il quarto gruppo ‘dark’ non è predominante in nessuno stato.

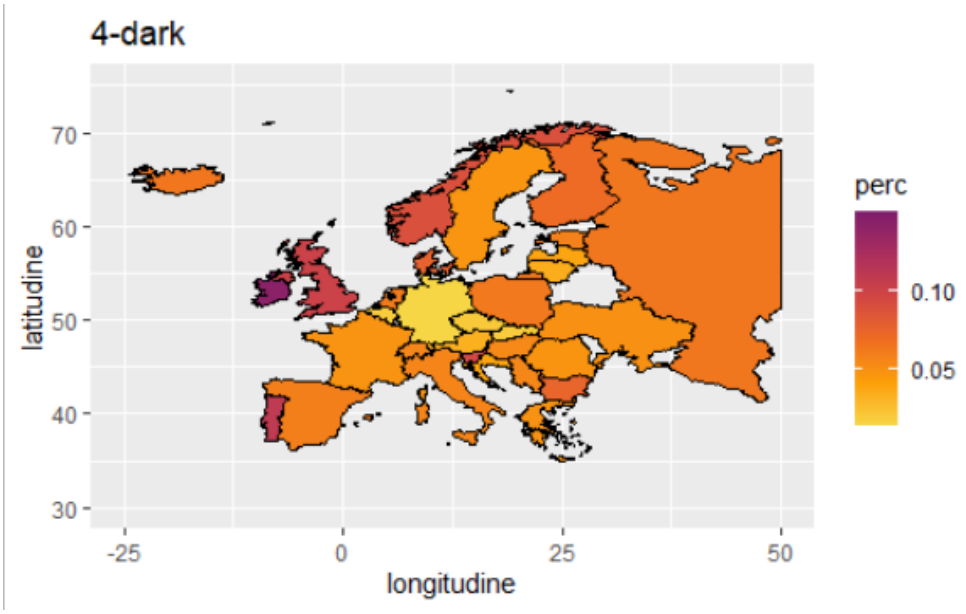
Questo è spiegato dal fatto che in quel gruppo sono riunite solo 6 stili di birra su 76 a differenza degli altri gruppi in cui convergono un numero decisamente maggiore di birre. È naturale dunque che il quarto gruppo sia percentualmente inferiore rispetto agli altri.

- Qual è la percentuale di ogni macro-famiglia in ogni stato?









Queste quattro mappe ci consentono di fare un'ulteriore e approfondita analisi:

Come si può vedere dalla tabella i gruppi creati hanno caratteristiche ben distinte. Poiché abbiamo le mappe della distribuzione in percentuale di ogni gruppo nelle nazioni, posso verificare se effettivamente esiste una relazione tra i valori medi delle caratteristiche assunti da una nazione e la distribuzione della macro-famiglia in quella.

In altre parole, posso andare a vedere se la percentuale della distribuzione delle macro-famiglie nelle varie nazioni trova una corrispondenza nei valori delle variabili che assumono le caratteristiche.

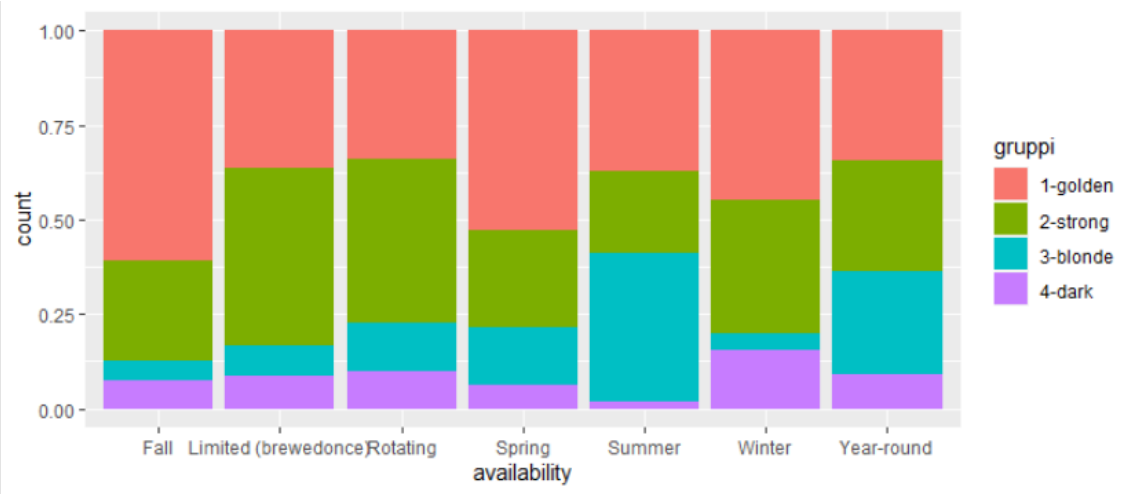
Andiamo a riprendere le nazioni usate in precedenza come esempio.

La Germania abbiamo visto avere valori delle tre variabili bassi, quindi ci aspettiamo che abbia un'alta percentuale di birre prodotte che rientrano nei gruppi 1 e 3 (che appunto presentano dei valori medi delle caratteristiche bassi).
Guardando la mappa notiamo che in effetti è proprio così e guardando i grafici dei gruppi 2 e 4 assume valori quasi minimi.

Nel caso opposto, in cui abbiamo preso i Paesi Bassi come riferimento di nazione avente valori delle tre variabili alti, ci aspettiamo che questi abbiano un'alta percentuale nei gruppi 2 e 4. Le mappe ci confermano questa tendenza, soprattutto per quanto riguarda il gruppo 2.

Un'altra analisi interessante può essere la seguente:
Abbiamo detto in precedenza che il quarto gruppo raggruppa quegli stili con valore del colore molto alto (quindi birre scure) ma con altre variabili basse.
Possiamo verificare come appunto ci sia una certa analogia tra il grafico del quarto gruppo e quello che riporta la distribuzione del colore (paesi come Portogallo, Paesi Bassi, Irlanda, UK, Estonia e Norvegia, che presentano un 'color' medio alto rispetto alla media, hanno anche un'alta percentuale nel grafico che riporta la distribuzione del quarto gruppo).

- Confronto Availability e Gruppi



Il grafico creato ci permette di visualizzare come si distribuiscono i vari gruppi in termini di percentuale, per ogni livello della variabile *availability*.

Con questo grafico possiamo analizzare se ci sono delle differenze, o analogie, circa la produzione degli stili di birra in base al periodo in cui vengono prodotte.
Mi soffermo in particolare sulle colonne 'Summer' e 'Winter'.
In estate vediamo che la quantità di birre 'dark' prodotte è quasi nullo mentre è molto più ampia della media il tratto della colonna azzurro, corrispondente alla tipologia di birre 'blonde'. In inverno notiamo invece il contrario ovvero un aumento notevole nella porzione di colonna viola ('dark') e una significativa diminuzione in quella corrispondente a 'blonde'.
Ciò significa che in estate abbiamo un deciso aumento della produzione di birre 'blonde' e una diminuzione delle birre 'dark', e viceversa accade in inverno.
Questo fatto non è assolutamente casuale.
In estate, con l'aumento delle temperature il consumo di birra viene indirizzato verso birre 'fresche' ovvero birre leggere, dal sapore non troppo forte, e molto beverine. Proprio per questo troviamo all'interno del gruppo 'blonde' birre di stile Weissbier, Pale lager, Pilsner ecc.
In inverno invece, quando le temperature sono più basse (nella maggior parte delle nazioni considerate) si prediligono birre 'calde' cioè birre consistenti dal gusto più deciso. Proprio per questo il gruppo dark è rappresentato da birre di stile stout e porter (con le

loro varianti). Queste birre sono nate in Irlanda e Inghilterra, e sono considerate le classiche birre da pub, da bere in un ambiente chiuso e quindi in contrapposizione totale con le birre ‘blonde’ che possono essere viste come birre da bere in contesti più miti all’aperto.

CONCLUSIONE

Il lavoro da me proposto ha come scopo quello di avvicinare il lettore ad un mondo tanto vicino quanto inesplorato, ovvero quello della birra, fornendone spunti per un consumo più consapevole.
Grazie all'analisi di due dataset, ho potuto rispondere alle domande poste all'inizio che mi hanno spinto a svolgere questo lavoro.
Tra le più importanti:
Quanto influisce la cultura o storia della birra sulla produzione odierna? Si è visto come la cultura della birra nei diversi paesi ha un impatto determinate nella produzione della birra ai nostri giorni, indice del fatto che si tratta di un prodotto fortemente tradizionale e legato alla cultura in senso lato di un popolo.
Come si relazionano le caratteristiche che identificano il prodotto? Le tre caratteristiche principali dei vari stili della birra, presentate nel secondo capitolo, hanno una correlazione positiva tra di loro e anche i modelli illustrati ci mostrano in quale modo variando il valore di una, cambiano in media i valori delle altre.
In che modo può essere utile studiare la birra da un punto di vista geografico? Le analisi geografiche ci hanno permesso da un lato di confermare le ipotesi di correlazione tra le caratteristiche, e da un lato di fornire mappe concettuali per osservare come si differenzia il prodotto nell'area geografica considerata.

Va considerato che il presente elaborato si tratta di un'analisi per lo più di tipo esplorativa e visualizzativa, che dà un'indicazione generale delle caratteristiche del prodotto nell'area considerata. Non si tratta dunque di un testo che approfondisce in maniera precisa l'idea di birra nei vari stati, ma fornisce delle informazioni sintetiche circa le caratteristiche analizzate.

Uno spunto che può dare questo elaborato è quello di un'analisi più approfondita concentrandosi su pochi stati andando a studiarli più nel dettaglio. Oppure è possibile utilizzarlo come punto di partenza per una possibile previsione della produzione di birra nel futuro.

CAPITOLO 4 - BIBLIOGRAFIA

Poelmans, Eline, and Johan FM Swinnen. "A brief economic history of beer." *The economics of beer* (2011): 3-28

Pascua, Michael, Scott Guenhe, and Mdsadaf Mondal. "History of craft beer." (2016).

<https://www.kaggle.com/>

<https://www.census.gov/data/tables/time-series/demo/popest/2010s-state-total.html>

Mosher, Michael, and Kenneth Trantham. "Beer Styles." *Brewing Science: A Multidisciplinary Approach*. Springer, Cham, 2017. 35-61.

beer Judge Certification Program's 2015 style guidelines

<https://www.scattidigusto.it/2014/05/31/12-stili-birra-artigianale/>

Mosher, Michael, and Kenneth Trantham. "Beer Styles." *Brewing Science: A Multidisciplinary Approach*. Springer, Cham, 2017. 35-61.

Sasirekha, K., and P. Baby. "Agglomerative hierarchical clustering algorithm-a." *International Journal of Scientific and Research Publications* 83 (2013): 83.

Singh, Archana, Avantika Yadav, and Ajay Rana. "K-means with Three different Distance Metrics." *International Journal of Computer Applications* 67.10 (2013).

Wagstaff, Kiri, et al. "Constrained k-means clustering with background knowledge." *Icml*. Vol. 1. 2001.

Thinsungnoena, Tippaya, et al. "The clustering validity with silhouette and sum of squared errors." *learning* 3.7 (2015).

Rossi, Germano. "Appunti sulla regressione lineare semplice e multipla." (2004).