Ipotesi da testare

Le preferenze d'acquisto, di visione e di ascolto sono reciprocamente condizionate.

Gli utenti outlier hanno gusti più marcati e le loro preferenze sono più prevedibili.

Le variabili sociodemografiche influiscono nelle preferenze d'acquisto, visione e di ascolto.

Le piattaforme e-commerce, di streaming on-demand di musica e di serie-tv possono diminuire del 66% i costi di gestione e realizzazione di dati e algoritmi profilativi e suggeritivi perché possono attingere dallo stesso database.

Metodologia

Per indagare sui gusti musicali, d'acquisto e di visione di serie-tv, occorre sottoporre agli intervistati dei questionari che produrranno un database da cui estrarre le informazioni di cui si necessita. Come suggerito dal prof. Italiano, l'obiettivo era realizzare un questionario -e quindi un database- unico, con tutte le domande di cui necessitavamo conoscere la risposta.
L'obiettivo era evitare di avere tanti questionari e database diversi sia per agevolare coloro che avrebbero dovuto rispondere, sia per consentire all'analista di analizzare ulteriori variabili anche a cui prima non aveva pensato. Ho curato personalmente la realizzazione di un database unificato, coordinando tutti i gruppi che necessitavano di informazioni diverse.
Parallelamente, mi sono occupato della realizzazione di un questionario adeguato al tipo di informazioni di cui necessitassi.
La fase successiva era caratterizzata dalla trasformazione e pulizia dei dati, per disporli in un formato adeguato all'analisi della fase successiva.
In seguito, la fase di analisi vera e propria, per provare le ipotesi di partenza.

Design e coordinamento della fase di Data Collecting

Numerosi problemi sono sorti nel design della fase di data collection. Sono stati, tuttavia, risolti garantendo, comunque, massima flessibilità per l'analista, e comodità e anonimato per chi risponde.

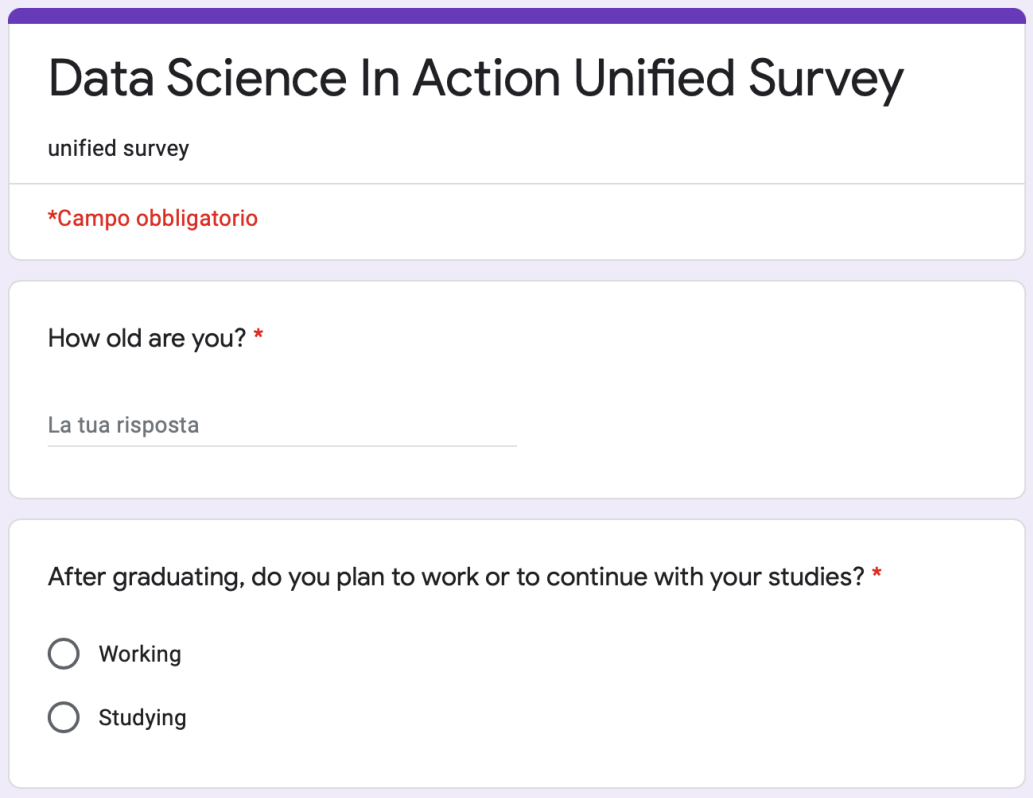
- Ho realizzato un dory su dory.app. Uno spazio dory è una bacheca online in cui chiunque può postare un testo o un'immagine, senza necessità di registrazione. Ogni post è un dory.
  In ogni momento, dal 14/12/2021 13:51 al 30/12/2021 23:59, agli analisti è stato possibile postare un dory per ciascuna domanda da sottoporre agli interrogati. Ogni dory si componeva:
  - Della domanda da sottoporre
  - Del tipo di risposta da aspettarsi [Open Answer] / [Multiple Answers] / + [Multiple Answers + Specify Other]
  - Del nome e cognome dell'analista, in modo che, qualora fosse sorto qualunque tipo di problema, avrei potuto mettermi in contatto con lui.

  Dato il gran numero di domande già presenti, agli analisti non è richiesto verificare se vi siano duplicati.
  Ho dato comunicazione dell'apertura del dory ai miei colleghi il 14/12/2021 13:51. Nello stesso momento ho scandito la timeline "dei lavori".

- Il 31/12/2021 alle 09:00 ho realizzato un Google Form con le domande estratte dallo spazio dory. Ho rimosso le domande esattamente duplicate (Quanti anni hai? – Quanti anni hai) e mantenuto quelle simili (Quanti anni hai? – In che anno sei nato?). All'analista spetterà l'onere di verificare eventuali collinearità. In caso di dubbi sul senso o chiarezza della domanda, ho contattato l'autore. Esempio:

  What do you think of Data Science and Management Master's Degree?
  In questo caso, l'autore non si aspettava che l'utente fornisse un riscontro sulla parte del corso già completata, ma intendeva ricevere un quanto l'interrogato si aspetti dal corso in futuro.

  - Le domande e le risposte sono mostrate in ordine casuale.
  - Le domande indagano principalmente su caratteristiche socio/demografiche.
  - Tutti i questionari, incluso questo, sono completamente anonimi.
  - È stato aggiunto un campo *keyword*, in cui l'utente dovrà inserire un nickname a sua scelta., che dovrà ripetere in ogni altro questionario gli sarà sottoposto. Qualora un gruppo intendesse condurre indagini ulteriori, in tempi o su piattaforme diverse, dal momento che tutti i questionari sono anonimi, non sarebbe possibile unificare i database di ciascun questionario: al posto di utilizzare come indice comune a tutti i questionari il proprio nome e cognome (o altra parola identificativa) è stata scelta la variabile *keyword* che funge da "nome utente anonimizzato". Infatti, è esplicitamente richiesto di non utilizzare parole che possano richiamare all'interrogato.
    Per evitare che due utenti utilizzassero la stessa keyword, ho imposto che una keyword sia composta da almeno 6 lettere e 2 cifre, tipo *aaaaaa00.* Con queste condizioni, ho trovato accettabile -e remota- la probabilità che due individui scegliessero la stessa keyword: *1,73611111111111E-23.* Ogni utente dovrà reimmettere la stessa keyword nei diversi questionari che gli saranno sottoposti.



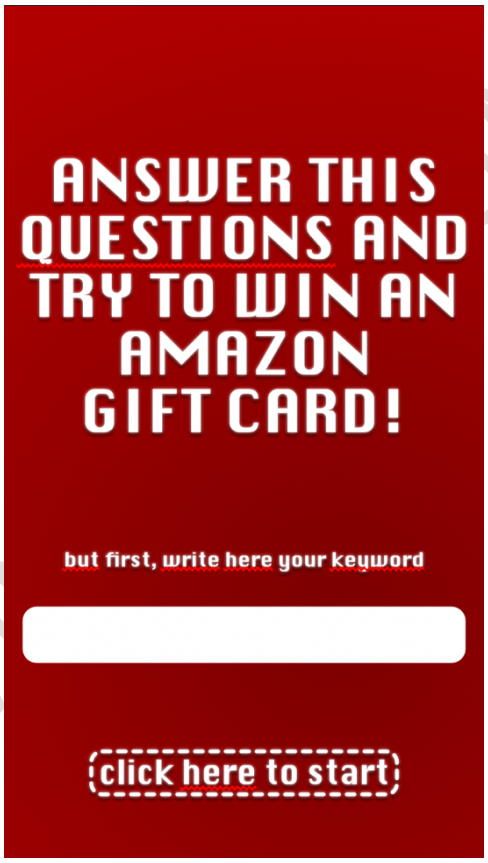
- Il 3 gennaio 2022 23:59, il Google Form è stato chiuso. Ne ho ricavato un file .xls di X osservazioni e Y variabili. Le osservazioni sono indicizzate per la variabile *keyword.*
- Il 4 gennaio 2022 ore 09:00 ho reso disponibile ai miei colleghi il database unificato.

Il database unificato risultante era così strutturato:

Keyword	Domanda 1	Domanda 2	Domanda 3	Domanda 4
forzanapoli1926	Risposta 1	Risposta 2	Risposta 3	Risposta 4
Vecchiaroma59	Risposta 1	Risposta 2	Risposta 3	Risposta 4

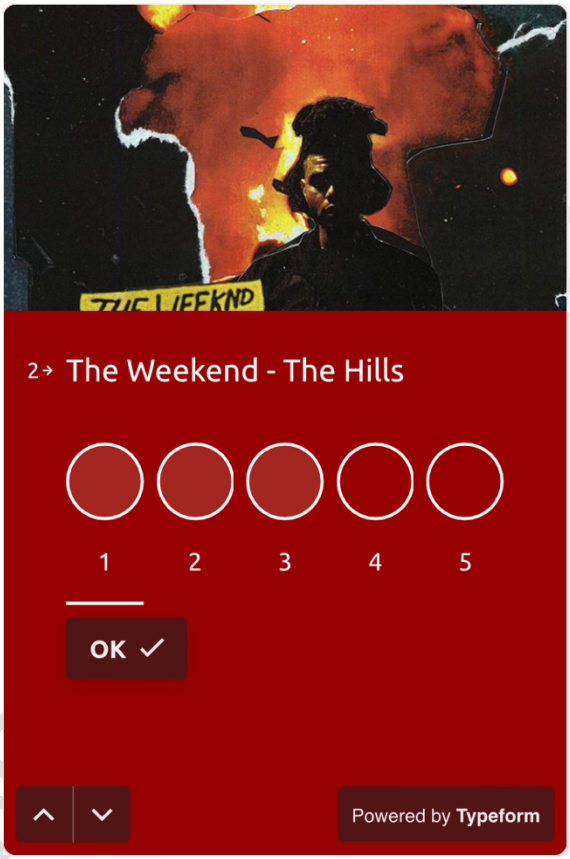
Parallelamente al lavoro di data collection "collettiva", io e il mio gruppo, ci siamo occupati della collezione di informazioni utili al nostro progetto.

- Ai fini della nostra analisi, occorreva conoscere le preferenze d'ascolto, di serie-tv e d'acquisto degli interrogati (chiamati di seguito *item*). L'interrogato avrebbe dovuto attribuire un punteggio da 1 (lo odio) a 5 (lo amo) o SKIP (nel caso in cui non conoscesse l'item mostratogli) per un certo numero di item, per ciascuna categoria. Visto il gran quantitativo di item necessari alla nostra analisi, abbiamo escluso la possibilità di condurre il questionario su Google Form, che risulta essere poco accattivante ed immediato. Consci del tempo richiesto per completare il form, abbiamo incentivato il completamento mettendo in palio due buoni Amazon da 15€. Ispirati dall'app di incontri *Tinder*, abbiamo realizzato un prototipo di come avrebbe dovuto essere l'interfaccia utente:



Necessitavamo di un questionario che producesse un dataframe dove, per ciascuna osservazione, è presente la variabile keyword e il punteggio da 1 a 5 o skip per ciascun item. Il sito web typeform.com ci ha consentito di realizzare un questionario web-app in poco tempo, che avesse un design accattivante e fosse facilmente utilizzabile dagli interrogati. (segue screenshot in basso). Il database prodotto da typeform presentava questa struttura:

Keyword	The Weekend – The Hills	La Casa di Carta	Massaggiatore per atleti	Lost
forzanapoli1926	5	SKIP	3	4
Vecchiaroma59	1	5	1	9
farina00	5	2	3	SKIP



Gli item da sono stati oggetto di un'accurata scelta.
In un esperimento ideale, avremmo dovuto far valutare all'utente tutte le canzoni disponibili nel catalogo Spotify, tutti i prodotti su Amazon e le serie-tv di Netflix. Non essendo possibile ciò, abbiamo dovuto bilanciare il rischio di non cogliere l'intera gamma delle preferenze dell'utente con la necessità di realizzare un form di un impiego temporale ragionevole. Siamo consci che è possibile vi sia una correlazione positiva tra la distanza dalla media del gusto musicale di un utente, con le sue preferenze d'acquisto o di visione. lecito supporre che un ventenne del 2021 che ascolti musica degli anni '30, sia anche poco avvezzo alla tecnologia, quindi agli acquisti online o agli acquisti "tech". Se così fosse, gli outlier potrebbero avere gusti più "marcati", e prevedibili.
Inoltre, includendo tutto il catalogo a disposizione per ciascuna piattaforma, avremmo rischiato di sottoporre agli utenti dei contenuti a loro sconosciuti.
Per la scelta degli item abbiamo proceduto così:

- Per i brani: abbiamo stabilito che 30 brani sarebbero stati sufficienti per delineare il "profilo di preferenze" di un utente.

        - Il 20% era composto da brani presi a caso tra i 100 più ascoltati del 2020. Questi hanno un peso residuale perché ci aspettiamo che, in media, essendo "hit", l'interrogato li valuti. È stato utile includerli per comparare i gusti musicali del 2020 con le preferenze di visione e d'ascolto dello stesso anno.
        - Il 30% da brani presi a caso tra i 300 brani più ascoltati di sempre. Il peso è maggiore, perché abbiamo ritenuto che i gusti musicali non cambino a seconda dell'anno.
        - Il 50% da brani presi a caso tra i primi 100 brani più ascoltati per ciascun genere musicale tra quelli proposti da Spotify. Abbiamo ritenuto che un gusto musicale sia maggiormente condizionato dal genere di un brano, più che periodo di tempo.

    - Per le serie tv: ancora, abbiamo stabilito che 30 serie-tv sarebbero state sufficienti per delineare i gusti di un utente. Le motivazioni della scelta degli item da valutare sono identiche per tutte i tipi di item.
        - Il 20% era composto da serie-tv prese a caso tra le 100 più viste nel 2020 dal catalogo di Netflix.
        - Il 30%, da serie-tv prese a caso tra le 100 più viste di sempre.
        - Il 50% da serie-tv prese a caso tra le 100 più viste per ciascun genere musicale, tra quelli proposti da Netflix.
    - Per gli acquisti online: abbiamo stabilito che 30 prodotti sarebbero stati sufficienti per delineare i gusti di un utente. Abbiamo forzatamente escluso, mascherine, DPI, alcol e altri dispositivi utili alla prevenzione contro il COVID19 che, a nostro avviso, non delineano i gusti personali di un utente, essendo spese "obbligate".
        - Il 20% era composto da prodotti acquistati presi a caso tra i 100 best-seller nel 2020.
        - Il 30%, da prodotti acquistati presi a caso tra i100 best-seller di sempre.
        - Il 50% da prodotti acquistati presi a caso tra i 100 best-seller per ciascuna categoria, tra quelle proposte da Amazon.

Data Collecting

Lo stesso metodo è applicato per le tre categorie, quindi per semplicità si prenderanno d'esempio le preferenze d'ascolto.
Per verificare l'ipotesi abbiamo profilato gli utenti in base ai loro gusti musicali: grazie ad un'analisi statistica, sulla base del punteggio dei brani, ad ogni utente è associata una lista di coefficienti che rappresentano il peso che l'utente attribuisce alle caratteristiche dei brani: se per questi sia più o meno importante la ballabilità, la strumentalità, il cantato, il tempo ecc. Quindi una lista di coefficienti rappresenta un certo gusto.
Abbiamo regredito tali coefficienti + altri predittori socio/demografici estratti dal database unificato, cercando di misurarne l'associazione con tutte le scelte d'acquisto (o di non acquisto) e di visione (e non visione).

1. Una volta scelte gli item secondo il criterio esposto nella sezione di data collection, abbiamo dovuto stabilire delle variabili che rappresentassero delle caratteristiche più o meno oggettive dell'item. Spotify mette a disposizione degli sviluppatori un'API contenente il catalogo dei brani dal 1922 al giugno 2021. Per ciascun brano sono disponibili le variabili autore, durata, album più alcune che Spotify utilizza per il suo algoritmo di profilazione e raccomandazione, tra cui ballabilità, tempo, strumentalità eccetera. Per ciascuna di queste ultime variabili associa un punteggio che rappresenta quanto tale brano sia o meno ballabile, strumentato o veloce. Abbiamo ritenuto tali variabili, unite a quelle sociodemografiche disponibili nel database unificato, sufficienti per la profilazione degli utenti sulla base del proprio gusto musicale.

Brano	Ballabilità	Strumentalità	Cantato	Tempo
Brano 1	1.789	1.09	1.09	12.98
Brano 2	0.3	2.92	0.2	1.4

2. All'utente è richiesto di esprimere un punteggio da 1 a 5 o SKIP nel caso in cui non conoscesse la canzone. Dopo l'attribuzione del punteggio, per ciascun utente il database sarà così formato:

Brano	Ballabilità	Strumentalità	[…]	**Punteggio**
Brano 1	1.789	1.09	[…]	**9**
Brano 2	0.3	2.92	[…]	**8**

Con i dati collezionati abbiamo attuato una *conjoint analysis*, un'analisi statistica simile ad una regressione che consente di stabilire il peso che un utente da a ciascuna caratteristica di un determinato item. Tale strumento è ampiamente utilizzato in ambito strategico e di marketing: si supponga di essere proprietario di una pizzeria. È possibile personalizzare la propria pizza in base al condimento alla dimensione e al prezzo. Sono disponibili tre condimenti, tre dimensioni e tre fasce di prezzo.
Il pizzaiolo interessato a predire quale sia la pizza più vendibile dovrebbe sottoporre a ciascun cliente l'assaggio di 27 pizze per un certo numero di clienti: **l'operazione di marketing più costosa della storia**.
Attraverso la *conjoint analysis* è possibile sottoporre ai clienti 3 pizze caratterizzate ciascuna da 3 modalità diverse rispetto alle altre.
L'utente sarà chiamato ad attribuire un punteggio a ciascuna pizza e grazie alla *conjoint analysis* sarà possibile stabilire quale sia il peso attribuito all'utente al condimento Marinara, Margherita, Quattro Formaggi, alla taglia Small, Medium eccetera.

Data Pre-Processing

3. Abbiamo importato il dataframe in formato .CSV in Python grazie alla libreria Pandas e attraverso il comando concat abbiamo concatenato orizzontalmente i dataframe provenienti da questionari diversi sulla base dell'indice keyword. Grazie ad un'operazione di slicing abbiamo aggiunto una colonna relativa all'anno di uscita ed eliminato quelle superflue, mantenendo solo il titolo della canzone più le variabili cui eravamo interessati.

4. Per ciascun brano abbiamo trasformato le variabili numeriche continue ballabilità, tempo, strumentalità in variabili discrete. Grazie al comando qcat di Pandas abbiamo attuato una trasformazione di tipo *quartile-data binning:* questa operazione sostituisce alla modalità numerica continua il quartile di appartenenza: se la distribuzione statistica di tutte le modalità *ballabilità* avesse un minimo di 0 e un massimo di 100 e il brano presentasse una ballabilità di 54 avremmo sostituito 54 con il quartile di appartenenza [50-75).
   Abbiamo ridotto a quattro le infinite modalità originali.

   (AGGIUNGERE IL TRATTAMENTO PER SKIP)

5. Abbiamo esportato il dataframe risultante su Excel per una manipolazione più immediata dei dati.

Brano	Bin_ballabilità	Bin_tempo	Bin_strumentalità	Bin_cantato	Punteggio
Brano 1	[50–75)	[75-100)	[75-100)	[50–75)	9
Brano 2	[1-25)	[50–75)	[25-50)	[75-100)	8

6. Abbiamo riformattato i dati in maniera diversa, grazie ad un semplice script VBA: necessitavamo di un dataframe sulle cui righe fossero presenti tutti brani e sulle cui colonne tutti i possibili quartili di ciascuna variabile. Le celle avrebbero contenuto 0 nel caso in cui il brano non ricadesse in un certo quartile e 1 se lo avesse fatto.

	Ballabilità				Tempo				[…]	
Brano	[1-25)	[25-50)	[50–75)	[75-100)	[1-25)	[25-50)	[50–75)	[75-100)	[…]	Punteggio
Brano 1	0	0	1	0	0	0	0	1	[…]	9
Brano 2	1	0	0	0	0	0	1	0	[…]	8

7. Abbiamo sostituto agli 0 e 1, il prodotto di riga tra ciascuno di questi e il punteggio del brano. Abbiamo rimosso la colonna punteggio che è "incorporata" negli ex 0 e 1.

	Ballabilità				Tempo				[…]
Brano	[1-25)	[25-50)	[50–75)	[75-100)	[1-25)	[25-50)	[50–75)	[75-100)	[…]
Brano 1	0	0	**9**	0	0	0	0	**9**	[…]
Brano 2	**8**	0	0	0	0	0	**8**	0	[…]

8. Abbiamo calcolato la Media di colonna per ciascuna colonna e la Media delle Medie per ciascuna variabile. Quindi abbiamo sottratto ciascuna Media per la Media minima per variabile, ottenendo i coefficienti utili alla profilazione del gusto dell'utente (Media sottratta o Part-Worth)

	Ballabilità				Tempo				[…]
Brano	[1-25)	[25-50)	[50–75)	[75-100)	[1-25)	[25-50)	[50–75)	[75-100)	[…]
Brano 1	0	0	9	0	0	0	0	9	[…]
Brano 2	8	0	0	0	0	0	8	0	[…]
Media	**5**	**4**	**6**	**8**	**9**	**8**	**10**	**3**	[…]
Media delle Medie	**5.75**				**7.5**				
Media min. per variabile	**4**				**3**				[…]
Media sottratta (Part-Worth)	**1**	**0**	**2**	**4**	**6**	**5**	**7**	**0**	[…]

Da un punto di vista interpretativo, la Media delle Medie misura quanto una variabile sia importante per la valutazione di un brano, rispetto alle altre. In particolare, l'utente preso in esame, valuta più positivamente del 30% i brani con un tempo più incalzante, rispetto ai brani più ballabili.
I coefficienti Media sottratta o Part-Worth misurano quanto pesi una modalità di una variabile, rispetto alle altre. In particolare, i brani con ballabilità più alta sono preferiti quattro volte di più rispetto ai brani con ballabilità più bassa, e il doppio rispetto ad una ballabilità poco superiore alla media.

I coefficienti part-worth sono i coefficienti di una *dummy variable regression:*

$$PUNTEGGIO\ CANZONE = 1 * x1 + 0 * x2 + 2 * x3 + 4 * x4 + 6 * x5 + 5 * x6 + 7 * x7 + 0 * x8$$

Dove le x sono gli 0 e 1 caratteristici di un brano. Per ciascun brano è possibile una sola x per categoria di coefficienti (infatti i coefficienti 1, 0, 2, 4 si riferiscono al livello di ballabilità e non è possibile avere due livelli di ballabilità per un solo brano). Il brano che massimizza il punteggio è quello che presenta 1 in corrispondenza dei coefficienti part-worth per ciascuna variabile.

I coefficienti part-worth sono espressione anche del brano ideale per un utente: il brano che rispecchia al massimo i suoi gusti sarà quello che in corrispondenza del coefficiente più alto per ciascuna variabile presenti un 1 (e he quindi abbia quella caratteristica). In particolare:

	Ballabilità				Tempo				[…]
	Part Worth 1	Part Worth 2	Part Worth 3	Part Worth 4	Part Worth 1	Part Worth 2	Part Worth 3	Part Worth 4	[…]
Part-Worth	1	0	2	4	6	5	7	0	[…]
Brano Ideale	0	0	0	1	0	0	1	0	[…]

Più gli score si avvicinano a 4 *1 + 7 * 1 = 11, maggiore un certo brano sarà affine ai gusti dell'utente.
Lo stesso ragionamento è stato applicato per le serie-tv e acquisti Amazon **(SPIEGARE I CRITERI DI SCELTA E QUALI SONO STATE LE VARIABILI OGGETTIVE)**

Alla fine, per ciascun utente avevamo un dataframe aggregato (che verrà chiamato d'ora in avanti *dag*):

	songs_ballabilità		songs_tempo		products_vita di tutti i giorni	
keyword	Part Worth 1	Parth Worth 2	Part Worth 1	Parth Worth 2	Parth Worth 2	Parth Worth 2
forzanapoli1926	0.78	0.23	1.34	1.67	1.78	10.1
Vecchiaroma59	1.67	0.78	0.79	1.34	1.56	0.11

	series_commedia		series_intrattenimento			
keyword	Part Worth 1	Parth Worth 2	Part Worth 1	Parth Worth 2	Genere	Età
forzanapoli1926	11.3	6.7	1.5	7.8	M	23
Vecchiaroma59	0.63	1.98	0.32	11.4	F	24

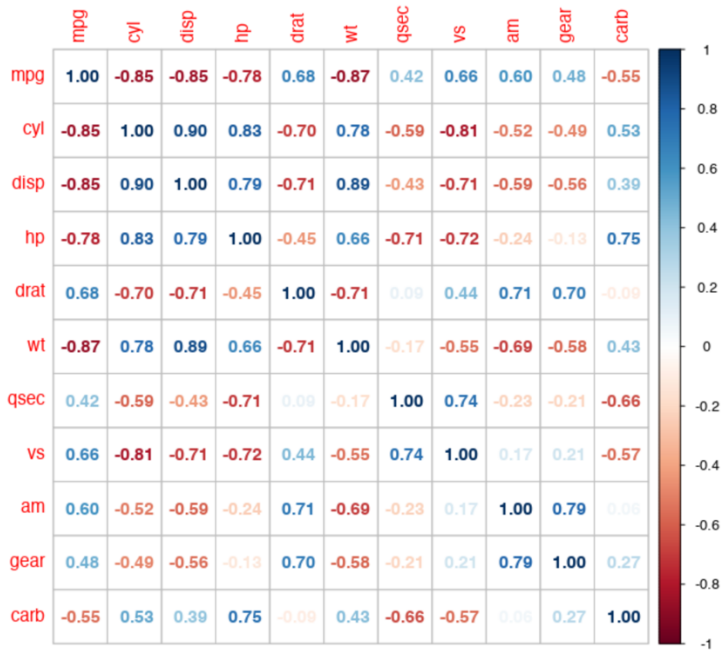
Abbiamo realizzato un ulteriore dataframe (chiamato *score*) che contenesse i punteggi assegnati da ciascun utente per ciascun item mostrato. È stato sintetizzato il punteggio da 1 a 5, in un variabile binaria che valesse 1 (mi piace), quando il punteggio era maggiore o uguale a 3; 0 (non mi piace) quando il punteggio era inferiore a 3. La modalità SKIP è stata sostituita dallo 0, perché considerata di scarso interesse per l'utente.

Non ancora processato						
keyword	Canzone 1	Canzone 2	Prodotto 1	Prodotto 2	Serie-tv 1	Serie-tv 2
forzanapoli1926	2	1	5	3	5	SKIP
Vecchiaroma59	5	5	1	2	3	2

Dopo la trasformazione (score)						
keyword	Canzone 1	Canzone 2	Prodotto 1	Prodotto 2	Serie-tv 1	Serie-tv 2
forzanapoli1926	0	0	1	1	0	0
Vecchiaroma59	1	1	0	0	1	0

Data Analysis (to be done; everything is displayed for reference)

Con l'aiuto del software statistico di R e il package corrplot abbiamo calcolato la matrice di correlazione per scovare pattern nascosti tra le variabili del dataframe *dag*.



I coefficienti più trasparenti (per esempio quello dato dall'intersezione tra qsec e drat), indicano che non vi è correlazione tra le variabili a questo associati. I coefficienti tendenti al blu, indicano che c'è correlazione positiva tra le due variabili (wt e disp). I coefficienti tendenti al rosso indicano una correlazione inversa tra le variabili (wt e mpg). In teoria, possiamo aspettarci una correlazione *blu* tra età giovanile e acquisti di prodotti tech, e una correlazione *rossa* tra sesso mashile e acquisto di assorbenti femminili.

(ANALIZZARE CORRELAZIONE TRA VARIABILI)

È altrettanto interessante conoscere quale sia l'associazione tra le variabili nel dataframe *dag* e ciascuno degli item contenuti nel dataframe *score.*
Prendo, ad esempio, l'associazione tra tutti i part-worth del gusto musicale + variabili demografiche e i part-worth delle preferenze d'acquisto. **Non è possibile regredire più X (gusto musicale + variabili demografiche) per un numero di Y maggiore di 1 (abbiamo infatti più part-worth che profilano le preferenze d'acquisto).** È questo il motivo per cui è stato creato il dataframe *score* che sintetizza le preferenze d'acquisto di un prodotto in una sola variabile binaria. Ripeteremo l'analisi di associazione tra le stesse X e tante Y quanti sono i prodotti di cui conosciamo le preferenze 0-1.

1. Selezioniamo le X da includere nell'analisi con il metodo basato sul punteggio AIC, forward selection.
2. Dopo la selezione, otteniamo un modello 60 osservazioni e circa 50 variabili, quindi ci troviamo davanti un problema di ultradimensionalità. Riduciamo il numero di variabili con la *Sure Independence Screening*, che rimuovere quelle variabili mostrano una correlazione con le X sotto una certa soglia. Il metodo di misurazione della correlazione sarà quello di **Kendall/Spearman/Pearson.** Selezioniamo una soglia del **X%**.
3. Vista la natura binaria della variabile Y, che in questo caso è il gradimento (1) o meno (0) del Prodotto 1, eseguiamo una regressione binomiale su r, includendo come predittori, i part-worth che delineano il gusto musicale e le variabili sociodemografiche età e genere.

```
Call:
glm(formula = Prodotto 1 ~ songs_part_worth1 + songs_part_worth2 + età + genere, family = binomial)

Residuals:
Min              1Q              Median          3Q              Max
-0.27238         -0.2428         -0.02762        0.16014         0.47238

Coefficients:
                 Estimate        Std. Error      t value  Pr(>|t|)
(Intercept)       64.9283        0.5084          127.71   <2e-16   ***
songs_part_worth1 45.9034        0.0214          29.66    <2e-16   ***
songs_part_worth2 1.94304        1.1234          5.69     0.067    .
età               23.3234        0.1239          12.23    0.01     *
genere            4.23445        1.4355          3.923    0.09
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.256 on 10 degrees of freedom
Multiple R-squared: 0.6789,       Adjusted R-squared: 0.6697
F-statistic:     880 on 1 and 10 DF, p-value: 4.428e-11
```

4. Le variabili con più asterischi sono quelle più significative, quindi rimuoviamo quelle mostrano un punto o uno spazio vuoto affianco la colonna Pr(>|t|)

(ECCETERA ECCETERA)