

# Report Data visualization

Matteo Gaverini, Federico Manenti, Paolo Mariani

Settembre 2019

## Indice

<b>1 Raccolta dati</b>	<b>2</b>
<b>2 Protocollo Think-aloud con valutazioni euristiche</b>	<b>3</b>
2.1 Prima infografica . . . . .	3
2.2 Seconda infografica . . . . .	3
2.3 Terza infografica . . . . .	4
2.4 Quarta infografica . . . . .	4
<b>3 Descrizione Task e visualizzazione dei risultati relativi</b>	<b>4</b>
3.1 Prima infografica . . . . .	5
3.1.1 Primo Task . . . . .	5
3.1.2 Secondo Task . . . . .	6
3.1.3 Terzo Task . . . . .	7
3.1.4 Quarto Task . . . . .	8
3.2 Seconda infografica . . . . .	9
3.2.1 Quinto Task . . . . .	9
3.2.2 Sesto Task . . . . .	10
3.2.3 Settimo Task . . . . .	12
3.3 Terza Infografica . . . . .	13
3.3.1 Ottavo task . . . . .	13
3.4 Quarta infografica . . . . .	14
3.4.1 Nono Task . . . . .	14
3.4.2 Decimo Task . . . . .	15
<b>4 Questionario Psicometrico</b>	<b>17</b>
4.1 Prima infografica . . . . .	17
4.2 Seconda infografica . . . . .	18
4.3 Terza infografica . . . . .	19
4.4 Quarta infografica . . . . .	20

# 1 Raccolta dati

I dati raccolti provengono da quattro diverse fonti:

- [AllSportsAPI](#) per le statistiche live delle partite
- [Twitter](#) per i tweet in tempo reale
- [Simplemaps](#) per i dati geografici (città, provincia e nazione)
- [Basketball Reference](#) per le statistiche dei singoli giocatori

Per l'analisi dei dati si sono utilizzati diversi script *Python* (si veda il report per un approfondimento) mentre per la visualizzazione si è deciso di utilizzare il tool Tableau.

Le infografiche risultanti sono le seguenti:

- [Prima infografica](#): NBA Finals 2019: Golden State Warriors (GSW) vs Toronto Raptors (TOR)
- [Seconda infografica](#): Posizione dei tweet durante le NBA Finals 2019
- [Terza infografica](#): Distribuzione hashtag utilizzati dalle tifoserie nelle NBA Finals 2019
- [Quarta infografica](#): Statistiche giocatori nella Regular Season 2018-19 & NBA Finals 2019

La prima infografica contiene tre tipologie di plot: linechart, barchart e un bubble chart. Il linechart mostra l'andamento del sentiment medio delle due tifoserie in ogni partita in funzione del punteggio del match (aggiornato ogni 30 secondi), il barchart permette di confrontare le statistiche complessive delle due squadre ed il bubble chart mostra i tweet che hanno avuto il maggior numero di retweet.

La seconda infografica invece contiene due tipi di plot: dot map e barchart. La dot map mostra la posizione da cui provengono gli utenti che pubblicano e condividono i tweet per ogni partita, i barchart invece presentano due tipi di informazioni diverse: i primi due collocati in basso a sinistra indicano rispettivamente la distribuzione dei tifosi nel mondo (sinistra) e nelle città (accanto a destra), mentre quello situato a destra e colorato di azzurro mostra i dispositivi usati dagli utenti per interagire con il match.

La terza infografica contiene un barchart che per ogni hashtag utilizzato durante le NBA Finals ne mostra la sua distribuzione nelle tifoserie.

Infine la quarta infografica mostra la media (con intervallo di confidenza al 95%) di 4 statistiche relative ai 6 giocatori più rappresentativi per franchigia durante le Finals 2019 e la Regular Season 2018-19, al fine di suggerire un confronto tra diversi giocatori e tra le due fasi di campionato dello stesso giocatore.

## 2 Protocollo Think-aloud con valutazioni euristiche

Per valutare ogni infografica creata è stato chiesto a sette utenti di eseguire una valutazione tramite protocollo think-aloud per sottolineare (attraverso la loro interazione con ogni infografica e l'espressione di pensieri e commenti su di esse) eventuali criticità legate alla visualizzazione dei dati. Il fine ultimo è quello di intervenire, in funzione dei commenti riportati, sulle problematiche da esse sottolineate per rendere la comprensione e visualizzazione di ogni infografica più semplice e conforme alle euristiche di dominio. Ogni singola valutazione è avvenuta durante un'interazione di tipo olistica, in cui l'utente è lasciato libero di agire durante la fase di utilizzo, osservazione, ragionamento e comprensione dei contenuti; ogni interazione ha avuto una durata qualche minuto (circa 3-4 minuti ad utente per ogni infografica). Per ogni valutazione verranno sintetizzati i commenti da essi riportati.

Prima di procedere con l'analisi dei risultati si sottolinea che tutte le infografiche inizialmente erano commentate e composte da testo in inglese per rendere la comprensione facile anche a utenti stranieri: durante la fase embrionale delle infografiche alcuni soggetti interessati che hanno osservato il progetto crescere hanno sottolineato una difficile comprensione delle descrizioni e dei contenuti delle infografiche, è stato deciso di convertire il contenuto in linguaggio più comprensibile possibile all'utente (italiano).

### 2.1 Prima infografica

Gli utenti fraintendono il dato espresso sull'asse x (Score della partita GSW - TOR). Avviene nonostante sia evidenziato sopra (automaticamente da Tableau).

**Violazione di Design:** il titolo dell'asse delle ascisse (score, punteggio) secondo alcuni utenti dovrebbe essere esplicitata maggiormente in quanto quella assegnata automaticamente da Tableau risulta essere poco visibile.

*Per rimediare alla problematica abbiamo posto un ulteriore casella di testo al di sotto del grafico in quanto Tableau non concede di modificare la posizione del label che rimane fissa nella parte superiore del grafico.*

### 2.2 Seconda infografica

Un utente ha sollevato una problematica rispetto all'utilizzo del colore grigio per indicare la tifoseria neutro all'interno della Mappa di utilizzo degli hashtag: secondo la sua considerazione risulta essere poco distinguibile rispetto allo sfondo della mappa. Richiede che sia cambiato con un grigio più tenue.

**Violazione di Design:** per l'utente il contesto e l'oggetto di visualizzazione sono poco distinguibili, si potrebbe suggerire di usare una palette di colori diversa.

## 2.3 Terza infografica

Un utente si lamenta della difficoltà nel distinguere il label associato all'asse y del grafico "distribuzione hashtag" rispetto ai valori che l'asse y assume; a sua richiesta dovrebbe risultare più marcato (es: grassetto e aumento della dimensione del carattere) o posto in una posizione diversa.

**Violazione di Design:** per l'utente non risulta distinguibile il label della variabile "hashtag" visualizzato sopra a tutti i valori che la variabile assume.

*Per rimediare alla problematica è stata incrementata la dimensione del font del label ed inoltre è stato applicato alla scritta il grassetto per evidenziare la differenza; Tableau non permette di posizionare il nome "hashtag" del label in verticale lungo l'asse delle ordinate.*

## 2.4 Quarta infografica

Quattro utenti si lamentano per la difficile comprensione del grafico "Statistiche giocatori nella Regular Season 2018-19 & NBA Finals 2019":

- Per due di loro non risulta intuitivo effettuare un confronto delle medie attraverso gli intervalli di confidenza al 95%.
- I due rimanenti ammettono di non aver letto il testo esplicativo relativo al funzionamento del confronto delle medie con gli intervalli di confidenza

I quattro utenti che hanno sollevato le precedenti critiche hanno anche partecipato all'esecuzione dei task: tutti e quattro hanno sbagliato le risposte dell'ultimo task (relativo all'infografica interessata) evidenziando la mancata comprensione del meccanismo di confronto.

**Mancata prevenzione dell'errore:** probabilmente il testo è poco esplicativo o il grafico dovrebbe essere ideato tramite una soluzione diversa che non prevede l'utilizzo degli intervalli di confidenza (di difficile comprensione per gli utenti che non li conoscono già di loro).

# 3 Descrizione Task e visualizzazione dei risultati relativi

Di seguito vengono elencate le spiegazioni dei task raggruppati per infografiche e mostrati i relativi risultati tramite uno stacked bar chart e un violin plot.

Nei violin plot la linea tratteggiata indica la media dei tempi di esecuzione dei creatori delle infografiche, mentre la regione grigia è calcolata come  $t_{medio} \pm \sigma$ ; la linea retta su cui è posizionata la media dei tempi registrati dagli utenti (evidenziata con un triangolo rovesciato) corrisponde all'intervallo di confidenza al 95% calcolato attorno alla media stessa. Per quanto riguarda gli stacked bar char invece la linea tratteggiata indica la tolleranza di errore del 5%.

### 3.1 Prima infografica

#### 3.1.1 Primo Task

Nel primo task si richiede di identificare il risultato corretto in Gara 1 attraverso l'interazione con il tabellino "Punteggio Finale" ed il filtro relativo alla gara selezionata. Il risultati sono i seguenti:

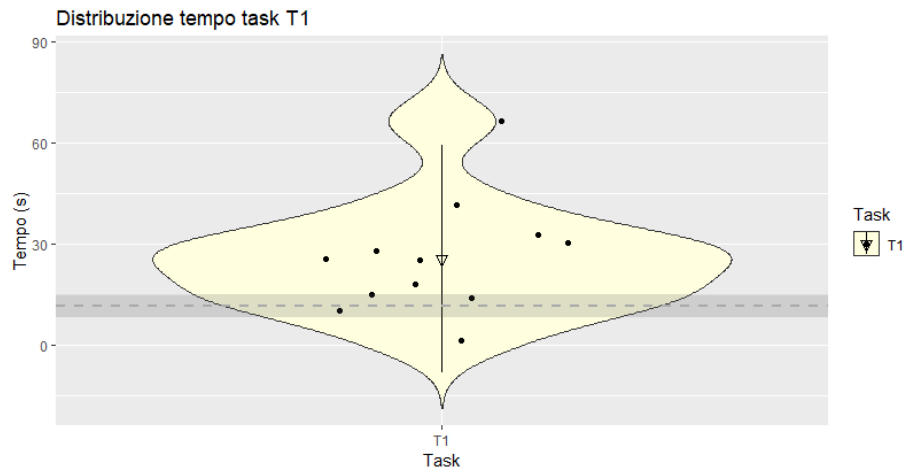


Figura 1: Tempi di esecuzione primo task

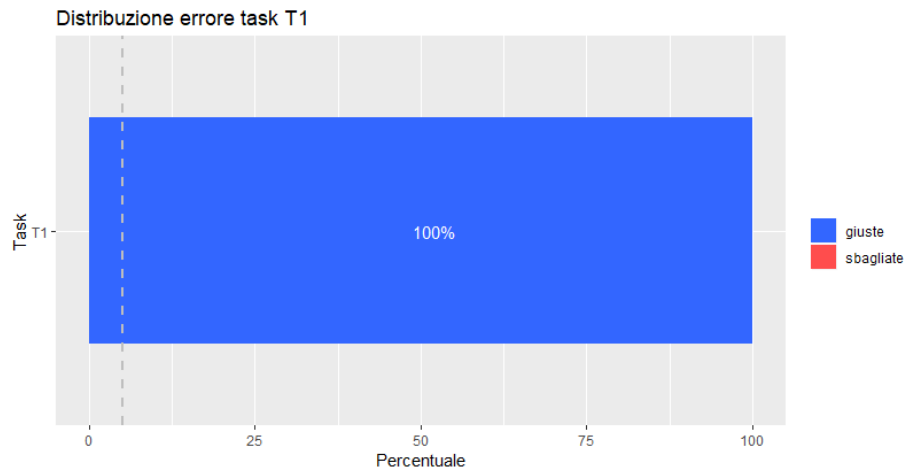


Figura 2: Percentuali di errore primo task

Nel primo task non si riporta nessun errore, per quanto riguarda i relativi tempi invece solo un soggetto si è distinto rispetto agli altri per via della lontananza rispetto alla media.

### 3.1.2 Secondo Task

Nel secondo task si richiede di individuare il sentiment medio più alto registrato durante gara 2 per la tifoseria di Golden State Warriors: richiede che gli utenti filtrino correttamente la gara con il filtro apposito e osservino il picco più alto di colore blu registrato nel grafico "Sentiment medio vs Punteggio di gara 1".

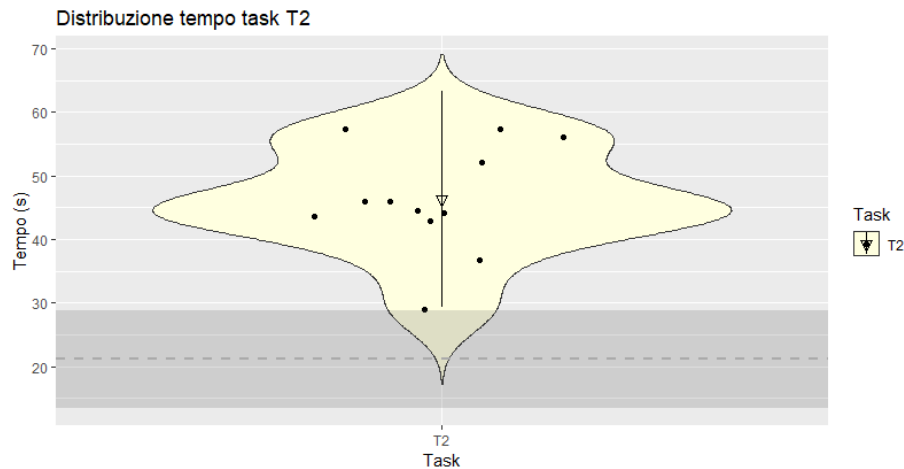


Figura 3: Tempi di esecuzione secondo task

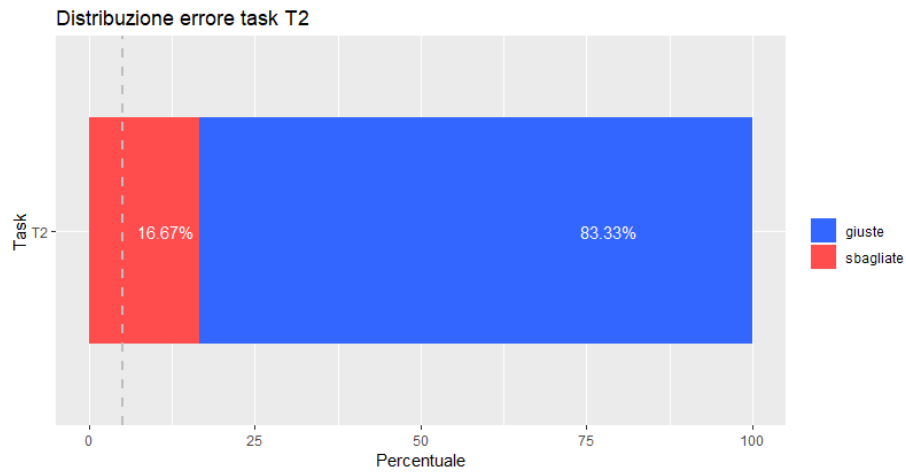


Figura 4: Percentuali di errore secondo task

Come si può osservare il secondo task riporta una errore pari a 16,67% e, osservando lo stacked bar chart, emerge una netta differenza tra i tempi degli utenti e quelli del gruppo di creatori.

### 3.1.3 Terzo Task

Il terzo task chiede di trovare il numero esatto di retweet associati al tweet più retwittato alla fine di gara 3 dai tifosi di Toronto Raptors; l'utente deve osservare il grafico "Tweet più ritwittati" e selezionare la "bolla" del bubble chart più grande tra tutte di colore rosso.

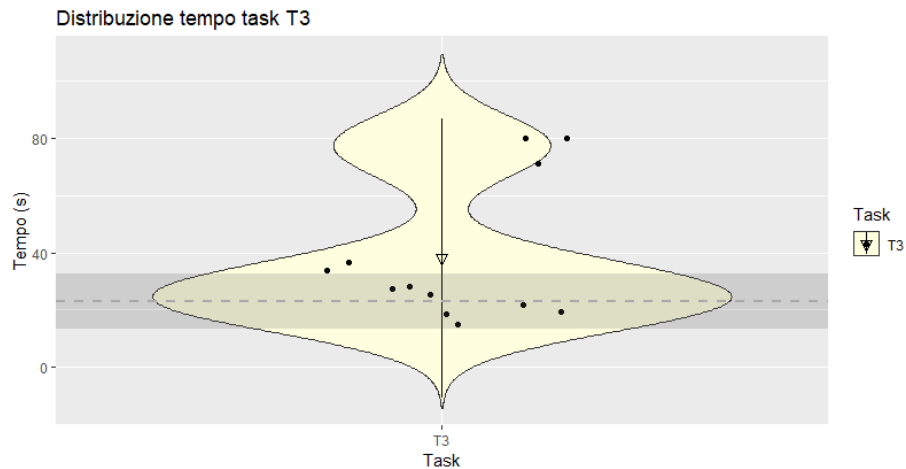


Figura 5: Tempi di esecuzione terzo task

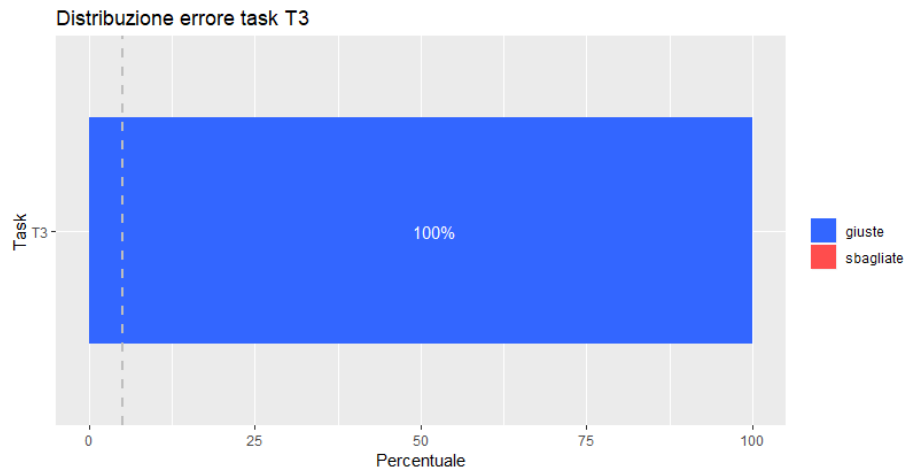


Figura 6: Percentuali di errore terzo task

In questo task tutti gli utenti hanno risposto correttamente e i tempi di esecuzione (escluse tre misurazioni) sono in linea rispetto ai tempi del gruppo.

### 3.1.4 Quarto Task

Nel quarto task si deve individuare la squadra che in gara 4 ha ottenuto la percentuale di tiri da 3 (3P%) più alta: ci si aspetta che l'utente rivolga l'attenzione al grafico "Statistiche delle squadre" per osservare nella parte destra la media della percentuale di tiri da 3 punti (3P%).

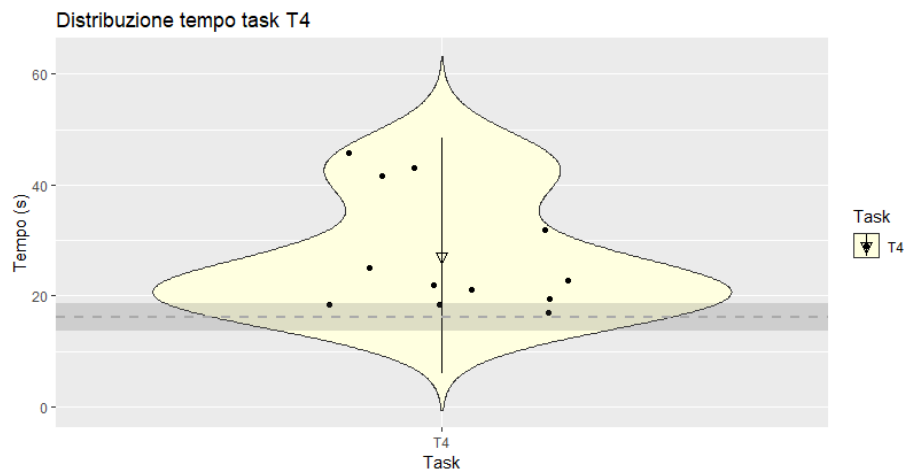


Figura 7: Tempi di esecuzione quarto task



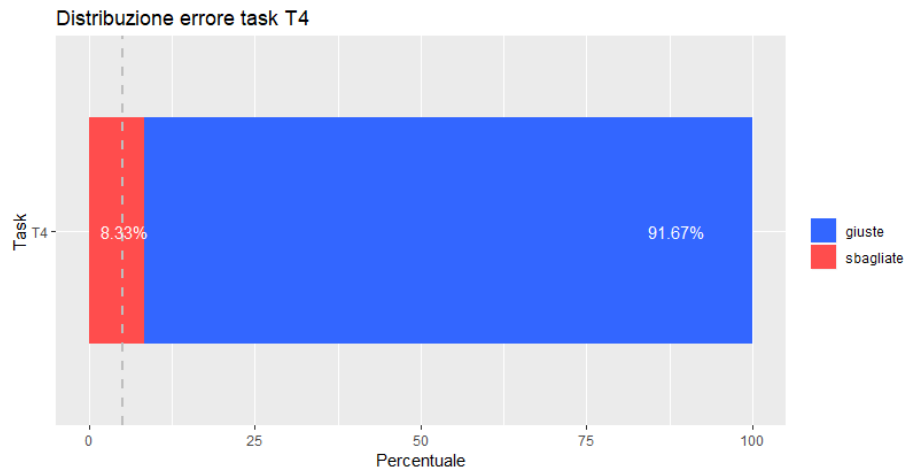


Figura 8: Percentuali di errore quarto task

Nell'ultimo task della prima infografica si registra un errore del 8,34% , la media dei tempi di esecuzione si discosta fortemente da quella dei tempi del gruppo di creatori.

## 3.2 Seconda infografica

In questa infografica sono richiesti tre task specifici (quinto, sesto e settimo dei dieci totali).

### 3.2.1 Quinto Task

Il quinto task consiste nella ricerca della partita con più alto numero di tweet (con posizione nota): l'utente dovrebbe utilizzare il filtro "Gara" per confrontare tra le diverse gare il valore che compare scritto all'interno del grafico "Numero di tweet (che possiedono una posizione nota)" e selezionare la sesta gara come risposta corretta.

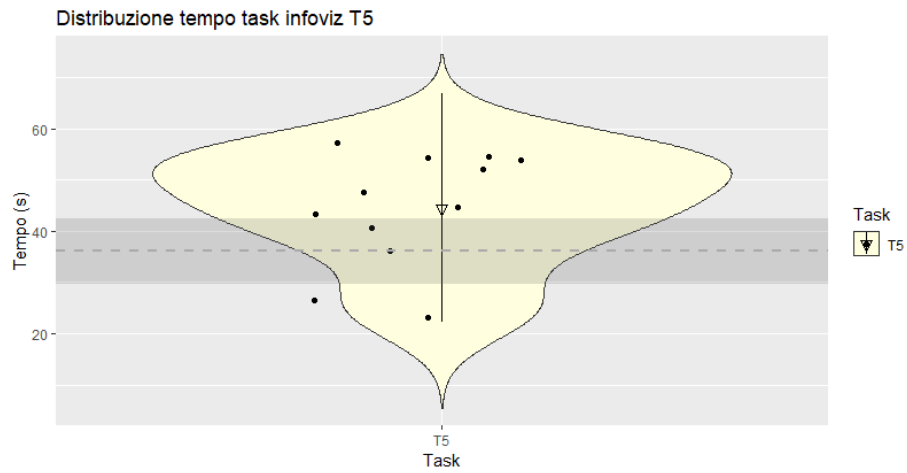


Figura 9: Tempi quinto task

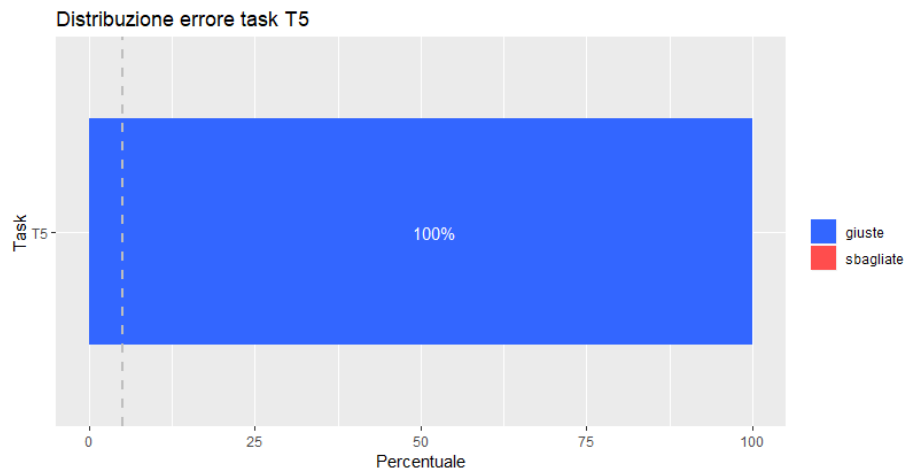


Figura 10: Errori quinto task

Osservando gli stacked bar chart non si sono registrati errori, per quanto riguarda i tempi invece ci sono valori sia inferiori che superiori alla media dei tempi dei creatori.

### 3.2.2 Sesto Task

Nel sesto task si richiede all'utente di individuare il dispositivo più utilizzato per twittare in Italia durante gara 4: ci si aspetta che utilizzi il filtro "Nazione" per selezionare il paese Italia, il filtro "Gara" per selezionare la quarta partita e

osservare il grafico "Dispositivi usati per twittare", selezionando come risposta corretta iOS.

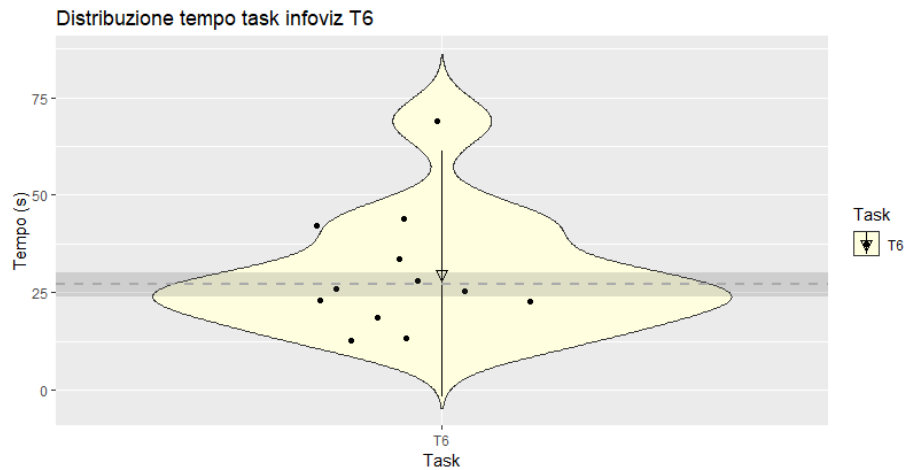


Figura 11: Tempi di esecuzione sesto task

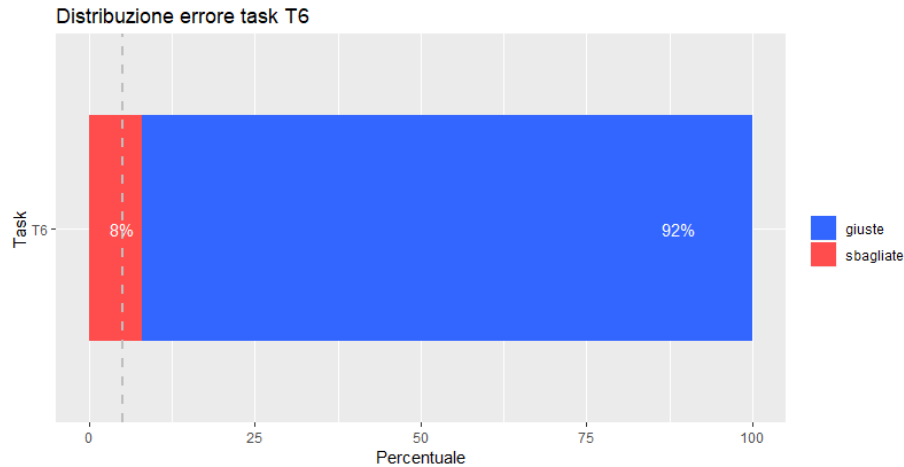


Figura 12: Percentuali di errore sesto task

Osservando i due grafici emerge che l'errore percentuale è pari a 8%, il tempo impiegato dagli utenti è vario: alcuni utilizzano il tempo ideale raggiunto dal gruppo di creatori, alcuni addirittura al di sotto, mentre un'osservazione risulta essere fuori dall'intervallo di confidenza; la media di conseguenza ricade nell'intervallo ideale.

### 3.2.3 Settimo Task

Nel settimo task si richiede all'utente di identificare la città brasiliana in cui si è registrato il numero più alto di tweet durante Gara 3: ci si aspetta che l'utente selezioni il Brasile tramite il filtro "Nazione", che selezioni con il filtro "Gara" la quarta partita e che osservi il grafico "Distribuzione dei tifosi nelle città presenti in \*\*\*" per indicare come valore corretto Rio de Janeiro.

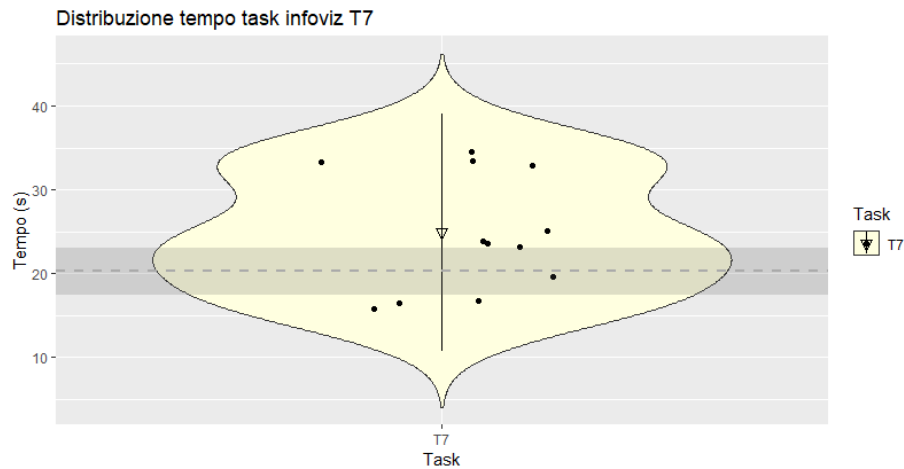


Figura 13: Tempi di esecuzione settimo task

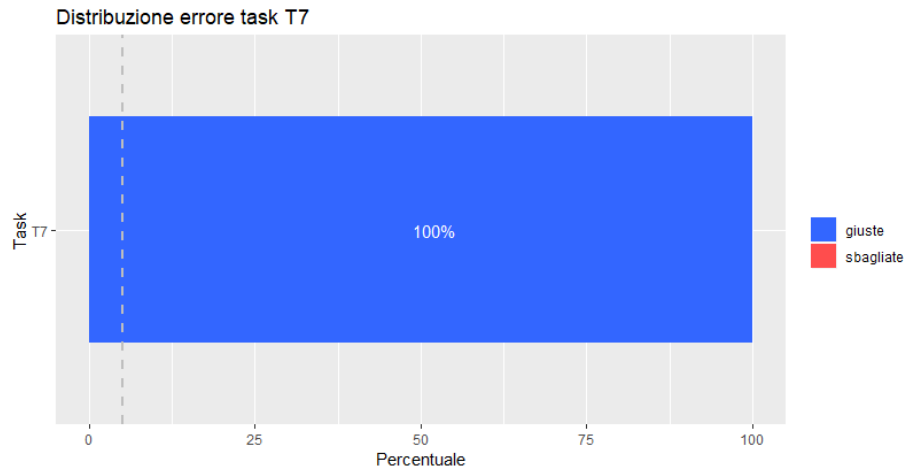


Figura 14: Percentuali di errore settimo task

Considerando i due grafici emerge che tutti gli utenti hanno risposto correttamente al quesito, e che la media dei tempi di esecuzione è superiore rispetto

a quello del gruppo di creatori, con alcuni valori registrati addirittura inferiori a quelli ideali.

### 3.3 Terza Infografica

#### 3.3.1 Ottavo task

In questa infografica viene richiesto un unico task (ottavo) che consiste nel trovare l'hashtag più utilizzato dai tifosi di Golden State: l'utente dovrebbe selezionare tramite il filtro "Tifosi" il valore Golden State Warriors e osservare nell'infografica aggiornata i valori più elevati, desumendo che sia "dubnation" la risposta corretta.

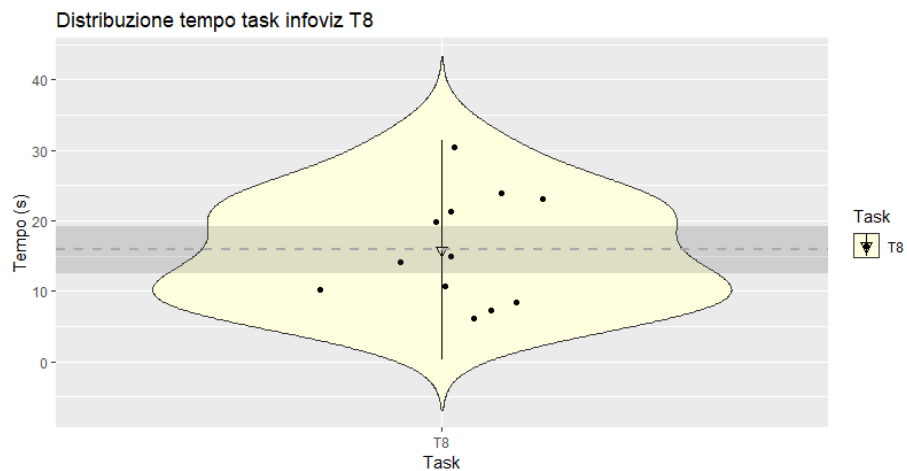


Figura 15: Tempi di esecuzione ottavo task

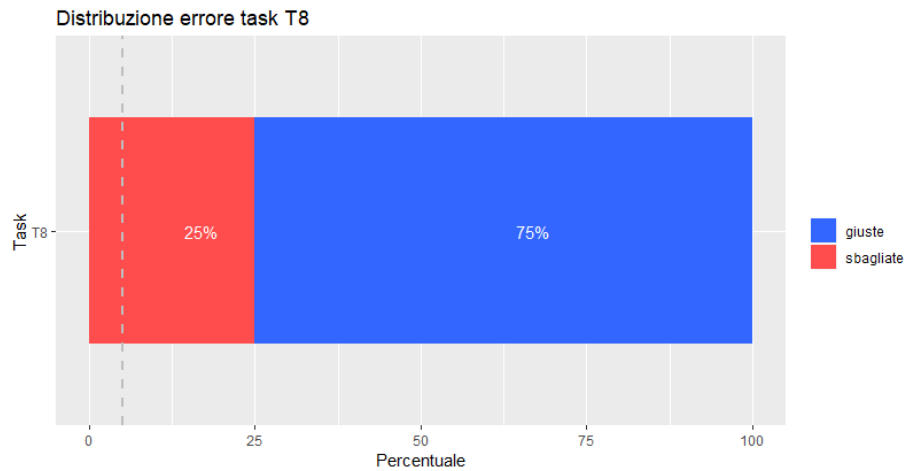


Figura 16: Percentuali di errore ottavo task

Il 25% degli utenti ha commesso un errore in questo task, i tempi di esecuzione degli utenti generano una media pressochè identica a quella registrata del gruppo di creatori.

### 3.4 Quarta infografica

Nell'ultima infografica, relativa alle statistiche dei giocatori, si richiede l'esecuzione di due task, che rispetto ai dieci totali che vengono sottoposti risultano il nono e il decimo.

#### 3.4.1 Nono Task

Il nono task richiede all'utente di comprendere in quale fase della stagione il giocatore *Cousins* ha avuto un rendimento migliore nella statistica punti (PTS): anche senza utilizzare il filtro "Giocatore" per selezionarlo, la sua statistica (PTS) risulta essere disponibile nell'infografica e la risposta al quesito (ottenuta con il confronto tramite le media e gli intervalli di confidenza) risulta essere "Regular Season".

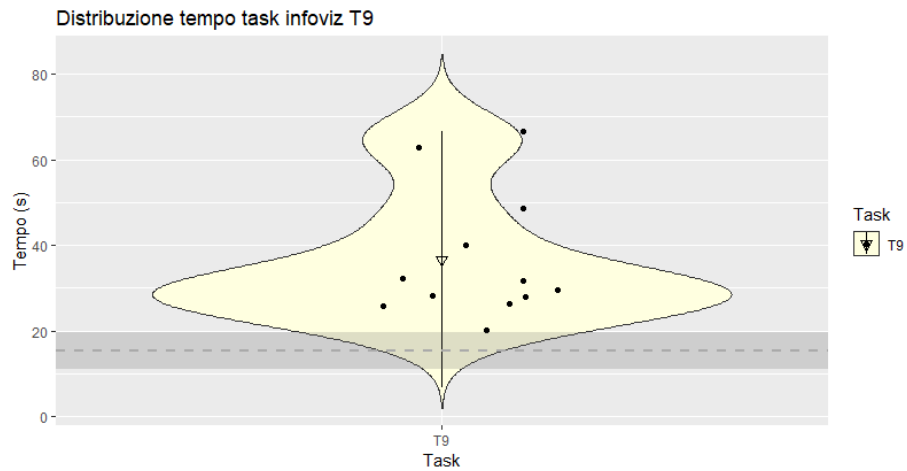


Figura 17: Tempi di esecuzione nono task

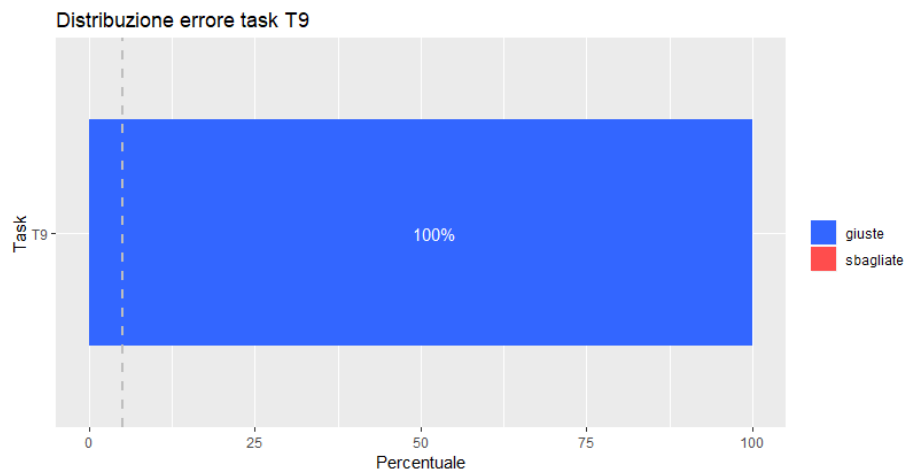


Figura 18: Percentuali di errore nono task

Nel penultimo task non si rilevano problemi relativi ad errori di risposta ai quesiti, ma la media dei tempi degli utenti risulta essere più elevata rispetto a quelli del gruppo.

### 3.4.2 Decimo Task

Nell'ultimo task si chiede di capire in quale fase della stagione il giocatore *Cousins* ha avuto un rendimento migliore nella statistica assist (AST): ci si aspetta dall'utente che utilizzi il filtro "Statistiche" per selezionare "Assist (AST)" e selezioni come risposta corretta "Non è possibile rispondere".

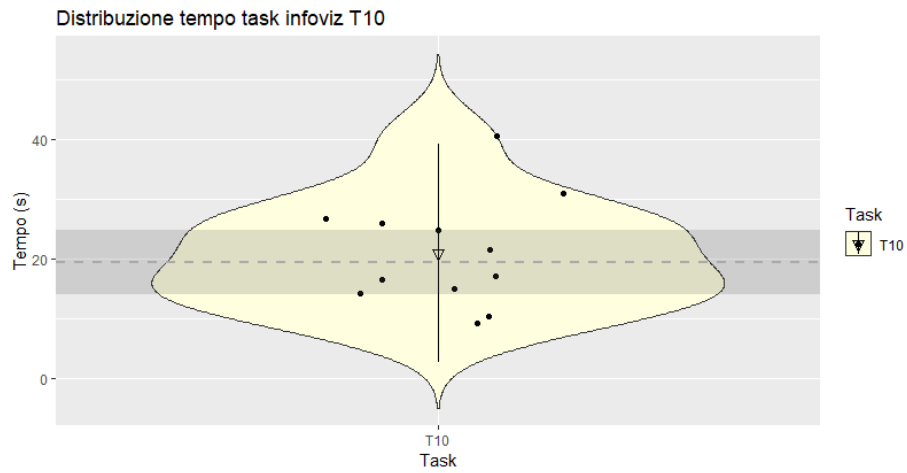


Figura 19: Tempi decimo task

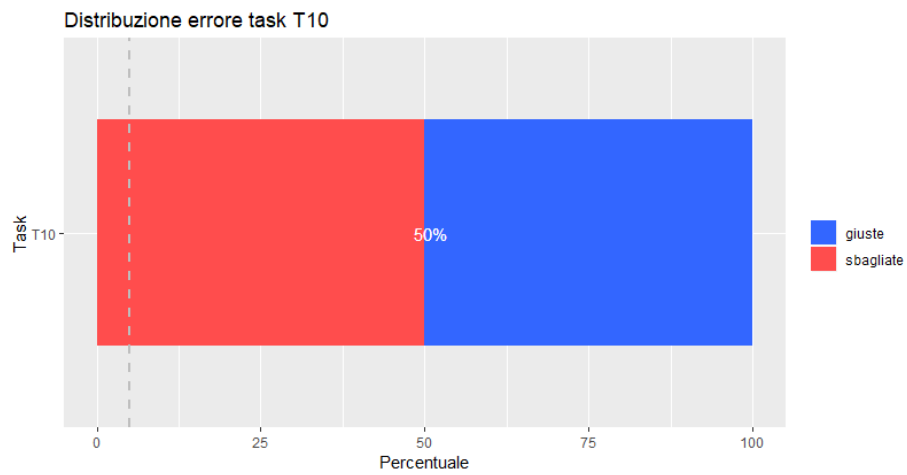


Figura 20: Errori decimo task

Nell'ultimo task si riporta la percentuale di errore più elevata tra tutte le altre infografiche: 50%, mentre i tempi degli utenti, rispetto al precedente task, risultano ridotti. Probabilmente gli utenti credendo di aver capito il funzionamento degli intervalli di confidenza (non leggendo per completo la descrizione della procedura) hanno prestato meno attenzione alle risposte.



## 4 Questionario Psicometrico

L'ultima fase di valutazione delle infografiche consiste nel sottoporre agli utenti un questionario psicometrico in modo da ottenere feedback quantificabili. Dopo una breve sezione di profilazione viene chiesto ad ogni utente la propria opinione, quantificata da 1 a 6, su cinque campi definiti dalla scala Cabitza-Locoro (utilità, intuitività, chiarezza, informatività e bellezza), infine è richiesta anche una valutazione generale dell'infografica. Verranno presentati i grafici delle valutazioni per ogni infografica raggruppando il punteggio con il seguente schema: 1-2 con valore basso, 3-4 con valore medio, 5-6 con valore alto. Inoltre si presenta il corrplot per valutare, attraverso i punteggi assegnati dagli utenti, la correlazione tra le variabili.

### 4.1 Prima infografica

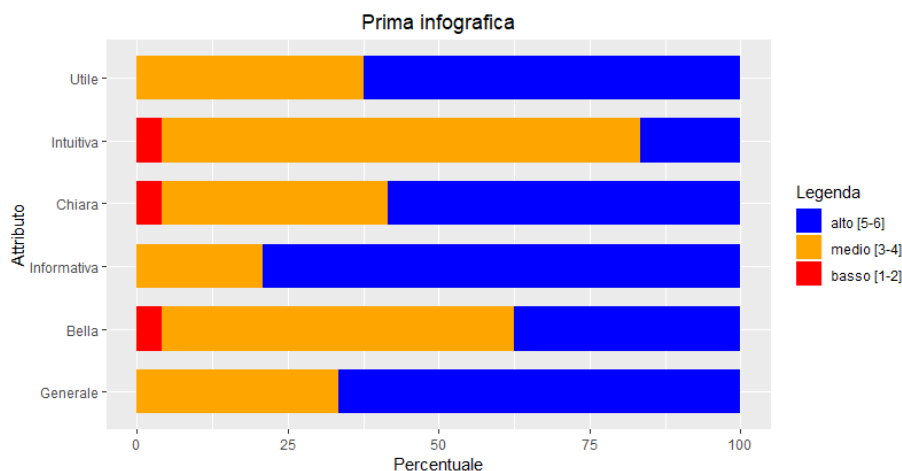


Figura 21: Valutazione prima infografica

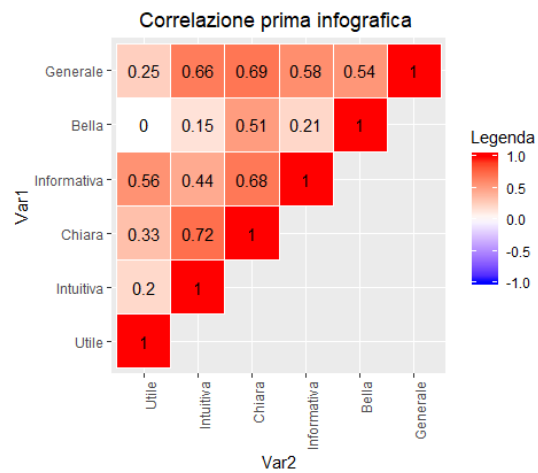


Figura 22: Corrplot prima infografica

Gli utenti criticano maggiormente l'intuitività, la chiarezza e la bellezza. Si nota inoltre che tra intuitività e chiarezza è presente una correlazione del 72%.

## 4.2 Seconda infografica

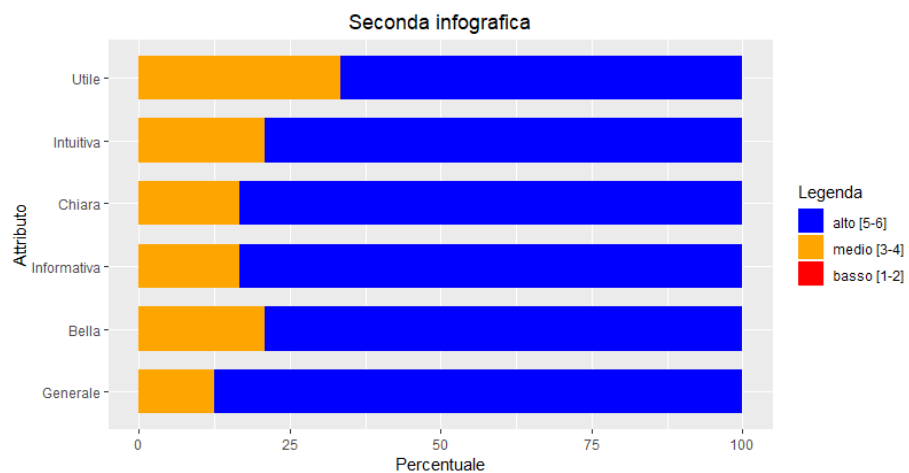


Figura 23: Valutazione seconda infografica

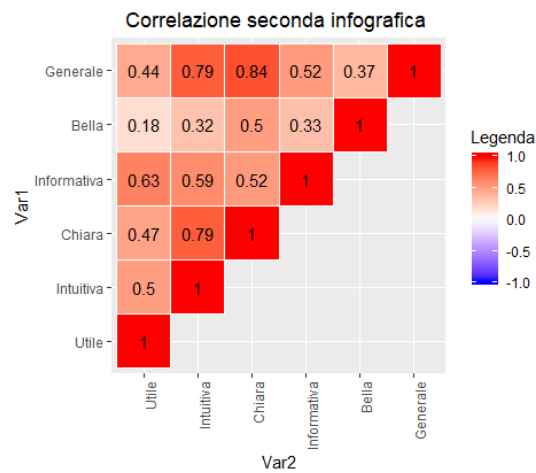


Figura 24: Corrplot seconda infografica

La seconda infografica è stata apprezzata maggiormente dagli utenti raggiungendo un valore alto in tutte le caratteristiche, nella matrice di correlazione si nota ancora una volta un legame tra chiarezza e intuitività.

### 4.3 Terza infografica

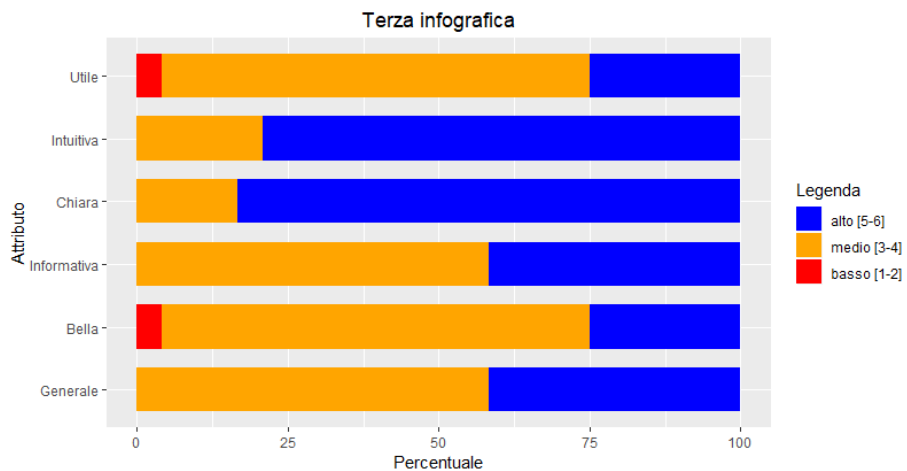


Figura 25: Valutazione terza infografica

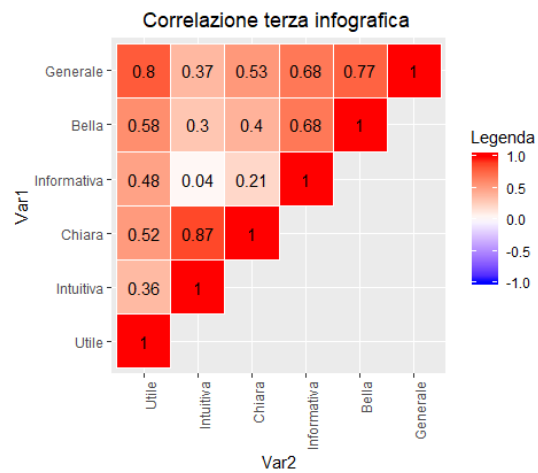


Figura 26: Corplot terza infografica

In questa infografica alcuni utenti hanno valutato negativamente i campi utilità e bellezza, ma nel complesso l'infografica è valutata positivamente. Si nota ancora una volta la correlazione tra chiarezza e intuitività.

#### 4.4 Quarta infografica

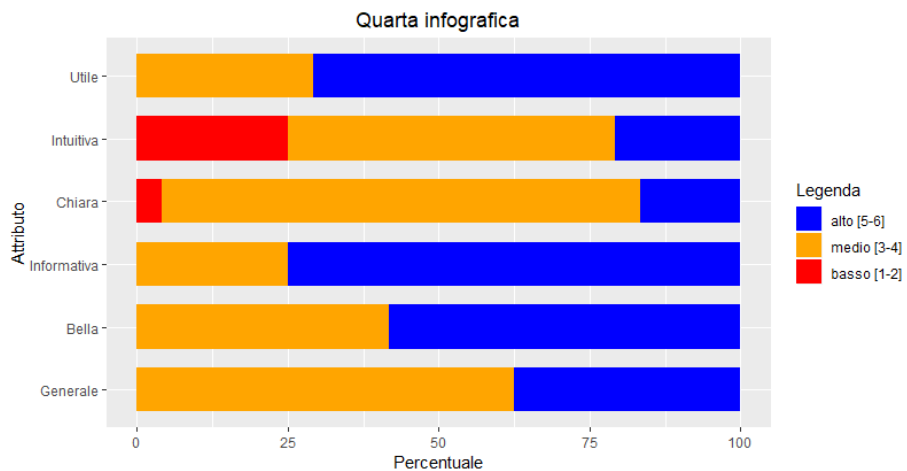


Figura 27: Valutazione quarta infografica

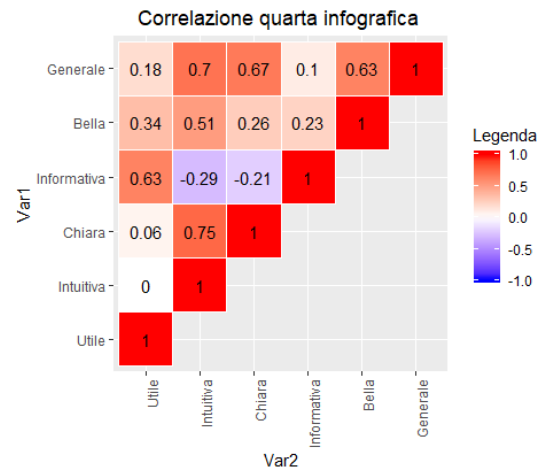


Figura 28: Corrplot quarta infografica

L'ultima infografica, come ci si aspettava, è quella che presenta maggiori criticità: l'intuitività è valutata negativamente da un quarto degli utenti. Nella matrice di correlazione emerge che gli attributi più correlati risultano essere chiarezza-intuitività e informatività-utilità.