



---

**Università degli Studi di Milano-Bicocca**  
**PROGETTO DATA MANAGEMENT &**  
**DATA VISUALIZATION**

---

Giorgia Antonicelli 872784\*(\*\*)  
Lorenzo Lorgna 829776\*(\*\*)  
Alessandro Maccario 865682\*(\*\*)

**\*CdLM Data Science**

Università degli Studi di Milano Bicocca  
Dipartimento di Informatica, Sistemistica e Comunicazione  
DISCo

28 Giugno 2021

(\*\*): *tutti i componenti del gruppo hanno partecipato attivamente e senza divisione di compiti al processo di creazione dell'intero progetto. Le uniche mansioni divise sono state le creazioni delle visualizzazioni ivi presentate, ovvero:*

- **Heatmap: Lorenzo Lorgna;**
- **Lollipop, Alluvial Diagram: Alessandro Maccario;**
- **Scatterplot, Barplot: Giorgia Antonicelli.**

*Tale approccio si è reso necessario per poter ottenere differenti prodotti visuali che fossero frutto dell'ingegno e creatività di ciascun componente e per sottoporre alla valutazione degli utenti quelle ritenute più valide da mantenere nel progetto finale.*

## Abstract

L'obiettivo di tale progetto è stato quello di indagare l'evento creato e promosso dalla British Broadcasting Corporation (BBC) denominato #BBC100Women, che si sostanzia con la pubblicazione annuale di una lista delle 100 donne più ispiratrici ed influenti, provenienti da ogni angolo del mondo.

Tale evento ha il fine di sensibilizzare la comunità mondiale intera, tanto maschile che femminile, al tema dell'importanza sempre crescente del ruolo della donna nelle società odierne che viene a scollarsi dai cliché del genere femminile come inferiore alla sua controparte maschile.

Il progetto in questione si propone di studiare la lista delle donne da diverse sfaccettature, analizzandone l'impatto mediatico, la composizione in termini di tipologia di donne premiate e gli Stati di provenienza delle stesse.

Per fare ciò, si sono reperiti dati da diverse fonti: il Social Network Twitter, i diversi siti della BBC in merito all'evento, gli indirizzi web delle maggiori istituzioni internazionali come l'*International Monetary Fund*.

Per lo svolgimento del progetto sono stati implementati script in Python 3.9 sfruttando la libreria Twint e il software Tableau, che ha permesso di creare infografiche che evidenziassero insights dettagliati e di semplice comprensione dei risultati.

**Keywords:** *Twitter, Twint, Python, Kafka, Nifi, MongoDB, Tableau, DataManagement, Data Visualization.*

# Indice

<b>1</b>	<b>INTRODUZIONE</b>	<b>4</b>
<b>2</b>	<b>DATA MANAGEMENT</b>	<b>5</b>
2.1	Obiettivo del progetto . . . . .	5
2.2	Approccio metodologico . . . . .	5
2.3	Velocità . . . . .	6
2.4	Varietà . . . . .	10
<b>3</b>	<b>DATA VISUALIZATION</b>	<b>13</b>
3.1	Progettazione, implementazione e valutazione della qualità delle visualizzazioni . . . . .	13
3.1.1	Visualizzazione 1 - Heatmap . . . . .	14
3.1.2	Visualizzazione 2 - Lollipop . . . . .	16
3.1.3	Visualizzazione 3 - Alluvial . . . . .	19
3.1.4	Visualizzazione 4 - Scatterplot . . . . .	21
3.1.5	Visualizzazione 5 - Boxplot/Barplot - Barplot e Donut chart . . . . .	22
3.2	Questionario Psicometrico . . . . .	24
3.3	User test . . . . .	26

# Elenco delle figure

2.1	Kafka Producer . . . . .	8
2.2	Workflow di NiFi . . . . .	8
2.3	Consumer NiFi . . . . .	9
2.4	Nodo di MongoDB . . . . .	9
2.5	Collezione in Mongo . . . . .	10
3.1	Quanto se ne parla su Twitter? . . . . .	15
3.2	Nel 2015 su Twitter quanto si parla dell'evento 100Women? . . . . .	15
3.3	Nel 2016 su Twitter quanto si parla dell'evento 100Women? . . . . .	16
3.4	Nel 2017 su Twitter quanto si parla dell'evento 100Women? . . . . .	16
3.5	Nel 2018 su Twitter quanto si parla dell'evento 100Women? . . . . .	17
3.6	Nel 2019 su Twitter quanto si parla dell'evento 100Women? . . . . .	17
3.7	Nel 2020 su Twitter quanto si parla dell'evento 100Women? . . . . .	18
3.8	Quali sono le categorie di riferimento? Qual è l'età mediana? . . . . .	18
3.9	A quali categorie appartengono le donne in lista? Da quali regioni mondiali provengono? . . . . .	20
3.10	Da quali Stati provengono le donne nominate dalla BBC? . . . . .	21
3.11	Come cambia la popolarità delle donne in lista? . . . . .	22
3.12	Come cambia la popolarità delle donne in lista? . . . . .	23
3.13	Correlogramma del questionario psicometrico . . . . .	24
3.14	Likert chart - Heatmap . . . . .	25
3.15	Likert chart - Lollipop . . . . .	26
3.16	Likert chart - Alluvial diagram . . . . .	27
3.17	Likert chart - Scatterplot . . . . .	27
3.18	Likert chart - Boxplot-Barplot . . . . .	28
3.19	User test - Efficacia . . . . .	29
3.20	Violin plot User test - Heatmap . . . . .	29
3.21	Violin plot User test - Lollipop . . . . .	30
3.22	Violin plot User test - Alluvial diagram . . . . .	30
3.23	Violin plot User test - Scatterplot . . . . .	31
3.24	Violin plot User test - Boxplot-Barplot . . . . .	31

# 1 INTRODUZIONE

È noto che, fino a pochi secoli fa, la donna venisse considerata come naturalmente inferiore all'uomo, alla quale venivano attribuite capacità e ruoli limitati alla procreazione e alla cura della prole e della famiglia. Dalla Rivoluzione francese in avanti, le donne si sono battute con sempre maggiore veemenza per l'acquisizione dei diritti di emancipazione e per il riconoscimento della piena dignità e uguaglianza rispetto alla categoria maschile (si pensi al ruolo delle suffragette e alle lotte per l'ottenimento del diritto di voto). Anni di storia hanno portato a grandi cambiamenti in questo senso e, seppur tale percorso non sia ancora concluso, in molti Paesi del mondo il genere femminile ha ottenuto le stesse opportunità degli uomini in ambito lavorativo, sociale e politico. Basti pensare a donne come Ursula von der Leyen, Presidente della Commissione europea, o Kamala Harris, la prima donna a diventare vicepresidente negli Stati Uniti d'America; mentre in molte altre culture, come ad esempio quelle islamiche, la donna ha ruoli, diritti e doveri considerati inferiori rispetto all'uomo. Nonostante ciò, il femminismo è un fenomeno diffuso a livello globale e *"non riguarda il rafforzamento delle donne. Le donne sono già forti, si tratta di cambiare il modo in cui il mondo percepisce quella forza"*. (G.D. Anderson)

Le parole di G.D Anderson, pseudonimo di Geena Dunne, fondatrice del Cova project, rispecchiano l'ideologia del mondo femminista nonché quella dell'emittente radiofonica inglese BBC. Ogni anno, infatti, la British Broadcasting Corporation pubblica una lista delle 100 donne, conosciute e meno conosciute, più ispiratrici ed influenti, provenienti da ogni angolo del mondo. Il loro impegno è incentrato nella costruzione di un futuro che porti con sé maggiori opportunità in tutti i campi della conoscenza e del vivere civile come i diritti umani, la medicina, la difesa dell'ambiente e l'ambito della cultura in senso lato.

La serie BBC 100 Women, da cui deriva tale lista, nasce con l'obiettivo di dare volto e lustro a queste personalità, non mettendo in evidenza un particolare ambito, ma inglobandovi all'interno qualsiasi esperienza in grado di cambiare, in positivo, il ruolo della donna nella società.

Tale serie nasce nel 2013 a seguito dello stupro di gruppo avvenuto a Delhi l'anno precedente, conclusosi con la morte della ventitreenne Jyoti Singh (2012 Delhi gang rape and murder case). Il seguito di proteste e di forte esposizione mediatica, portò Fiona Crack, editrice della BBC insieme ad altri giornalisti, a creare una serie incentrata sulle questioni e le conquiste delle donne nel mondo odierno.

Da allora la serie si ripete con cadenza annuale: dapprima vi è l'uscita della lista, in cui vengono premiate le 100 donne dell'anno; successivamente inizia la cosiddetta *Stagione delle donne della BBC*, della durata di tre settimane, che comprende una serie di eventi sul mondo femminile. In quest'occasione, le donne di tutto il mondo sono incoraggiate a partecipare tramite il canale social Twitter, commentando la lista, nonché seguendo le interviste e i dibattiti successivi alla sua pubblicazione.

## 2 DATA MANAGEMENT

### 2.1 Obiettivo del progetto

Il progetto ha avuto origine dall'importanza e dall'interesse suscitato dai temi sopracitati. In particolare, ci si è posti i seguenti quesiti:

- Quanta risonanza hanno su Twitter la pubblicazione della lista e la stagione delle donne della BBC?
- Qual è l'identikit delle donne nominate? La loro popolarità aumenta dopo esser state inserite nella lista?
- Nel corso degli anni, l'evento è diventato più popolare? È cambiata la tipologia di soggetti in lista?

A partire dalla volontà di trovare risposta a queste domande, si è proceduto come viene di seguito presentato.

Sono stati dapprima scaricati da Twitter i tweets con gli hashtags ufficiali dell'evento, sono stati altresì scaricati quelli in cui venivano menzionate le donne nominate nella lista, di cui si dirà più dettagliatamente nel seguito. Parallelamente sono state raccolte varie informazioni sugli Stati di provenienza delle donne elette: in particolare si sono scelti degli indici ritenuti significativi e rappresentativi dello sviluppo degli Stati, non solo inteso come sviluppo economico, ma anche sviluppo umano e disparità di genere. Durante lo svolgimento del progetto, è emersa anche la volontà e l'interesse nell'evoluzione tanto dell'evento BBC, quanto della risonanza su Twitter, che dell'evoluzione degli Stati dal punto di vista degli indici raccolti. Per questo motivo, si sono presi in considerazione due anni in particolare per svolgere le analisi di cui si discuterà nel seguito:

- Il 2015 perché di sufficiente distanza temporale dalla prima edizione della BBC 100 Women, svoltasi nel 2013, ma al contempo abbastanza ricco di informazioni a differenza dei due anni precedenti (maggiori dettagli relativi alle donne inserite nella lista, come per esempio età, occupazione, motivo della nomina);
- Il 2019 anche se inizialmente si era scelto il 2020. Da una prima analisi, ci si è accorti dell'enorme influenza del particolare anno della pandemia di COVID-19 sugli indicatori degli Stati e sul tipo di donne nominate, con una evidente distorsione dei loro valori principalmente motivata da tale evento globale. Si è quindi preferito trattare l'anno precedente in quanto lo scopo dello studio è stato quello di analizzare una situazione più generale e non così particolare come quella in cui si trova il mondo a seguito dell'epidemia. Inoltre, al momento della raccolta dati, molte informazioni (indicatori) relative agli Stati non erano ancora reperibili dalle fonti ufficiali.

### 2.2 Approccio metodologico

#### Big Data e 3V

In aggiunta agli obiettivi sopracitati, lo scopo del progetto è stato quello di essere in grado di gestire alcune delle dimensioni che caratterizzano oggi i Big Data. Alle 3V (volume, velocità e varietà) associate ai Big Data negli anni 2000, oggi si sono aggiunte veridicità, valore, viscosità e volatilità.

E' fondamentale al giorno d'oggi saper gestire, manipolare ed analizzare grandi quantità di dati, in modo

da poter trarre valore dal loro utilizzo. Nella realizzazione del progetto l'intento è stato quello di affrontare almeno due delle 3V quali la **velocità** e la **varietà**. La velocità indica la rapidità con cui i nuovi dati sono resi disponibili e si accumulano. La vera sfida è quella di raccogliere queste grandi quantità di dati per analizzarle in tempo reale.

Quando si parla di varietà, invece, è bene considerare che, per essere nella condizione di avere un'informazione completa e di valore, bisogna sapere gestire diverse tipologie di informazioni provenienti da fonti diverse, in formati spesso differenti.

Relativamente al progetto in questione, per quanto riguarda la velocità è stata simulata la raccolta di dati, in questo caso tweets, provenienti da una fonte di tipo streaming. Per quanto riguarda invece la varietà i dati raccolti dalla fonte streaming sono stati integrati con ulteriori dati provenienti da diverse sorgenti.

## 2.3 Velocità

### Limiti dell'API di Twitter

Per lo scaricamento dei tweets si è inizialmente valutato di utilizzare le API di Twitter, tramite la libreria in Python Tweepy che permette di fare richieste in maniera molto semplice. Dopo alcuni test, tuttavia, sono stati evidenziati alcuni limiti, fra i quali:

- È possibile effettuare non più di 450 richieste ogni 15 minuti (*rate limit*). Dunque, il numero delle richieste per intervallo temporale risulta essere limitato;
- Il numero massimo di tweets accessibili coincide con gli ultimi 3200 tweets disponibili;
- Non possono essere eseguite più di 500.000 richieste al mese. Perciò si ha anche un limite nel numero massimo di richieste al mese che è possibile effettuare;
- Con le API, versione standard, è possibile accedere solo ai tweets degli ultimi 7 giorni;
- La lunghezza della query è limitata a 512 caratteri.

Preso atto di tali limitazioni si è deciso di optare per *Twint*, una libreria Python che offre una notevole flessibilità e che permette di superare con grande semplicità i vincoli appena presentati.

### Twint

Tramite Twint è possibile dunque scaricare tweets facendo a meno dell'API ufficiale di Twitter. I principali vantaggi ottenuti utilizzando tale libreria sono i seguenti:

- È possibile scaricare tutti i tweets, anche passati, senza alcun limite temporale;
- Non sono richieste versioni premium;
- È possibile operare in modo anonimo, senza dover aver un account Twitter;
- La configurazione è molto rapida.

Nella definizione della query da eseguire per lo scaricamento di tweets è necessario impostare alcuni parametri, come la data di inizio e fine ricerca, gli hashtags e le eventuali menzioni da considerare. L'estrazione dei tweets viene effettuata utilizzando gli operatori di ricerca di Twitter.

In un primo momento, considerando singolarmente ogni anno dal 2015 al 2020, sono stati scaricati tutti i tweets, da gennaio a dicembre, contenenti gli hashtags ufficiale dell'evento *#BBC100Women* (e sue variazioni lessicali) oppure la menzione *@BBC100Women*, eliminando eventuali duplicati. Questi tweets sono stati estrapolati con il fine di comprendere e delineare in maniera più chiara la potenza mediatica dell'evento su Twitter.



La struttura della query è stata la seguente:

```
(@BBC100Women OR #BBC100Women OR #BBC100women OR #bbc100women OR #Bbc100Women
OR #bbc100WOMEN OR #bBc100women OR #BBC100WOMEN)
```

I tweets ottenuti dalla ricerca effettuata con Twint sono stati restituiti corredati delle seguenti informazioni: *date, place, tweet, language, hashtags, user\_id, username, name, nlikes, nreplies, nretweets, geo*.

Successivamente, sempre sfruttando Twint, prendendo in considerazione le singole donne nominate nelle liste degli anni 2015 e 2019, sono stati scaricati tutti i tweets che in qualche modo potessero essere associati alle donne in questione.

Per strutturare la query è stato necessario ricavare gli username delle donne, qualora queste ne disponessero, attraverso una ricerca manuale su Twitter. Congiuntamente sono stati anche cercati hashtags personali delle donne in base alle attività in cui si impegnano, in base al lavoro che svolgono e ad altri criteri. Una volta ottenute queste informazioni è stata definita la query di ricerca come segue.

Struttura query:

```
( USERNAME_TWITTER_DONNA OR ( NOME_DONNA AND ( HASHTAGS_CONSIDERATI
) ) )
```

Dove gli hashtags considerati sono sia quelli personali della donna, che quelli della BBC menzionati precedentemente. Per coloro che non dispongono di un account su Twitter la query è stata effettuata considerando il loro nome e gli hashtags della BBC. Questa strategia ha permesso di ottenere risultati coerenti evitando problemi quali di omonimia, univocamente individuando la persona di riferimento.

I tweets ottenuti dalla ricerca sono stati restituiti con le seguenti informazioni: *date, tweet, language, hashtags, user\_id, username, name, nlikes, nreplies, nretweets*.

Come periodo di ricerca dei tweets è stato deciso di considerare due intervalli temporali simmetrici rispetto all'uscita della lista della BBC, di circa 60 giorni l'uno.

Per ogni donna sono stati scaricati i tweets nei due periodi ottenendo il numero di tweets complessivo per il periodo pre e post pubblicazione della lista, mantenendone unicamente il conteggio totale.

In questo caso l'intento è stato quello di capire se per le donne, essere nominate nella lista, potesse avere un qualche impatto positivo sulla loro popolarità su Twitter. Per fare ciò, per ogni donna, è stata calcolata la differenza tra il numero di tweets successivi la pubblicazione della lista e quelli precedenti, con l'obiettivo di valutare se e quanto la nomina influisse sulla loro popolarità sul Social.

## Kafka

Una volta che i tweets sono stati scaricati, sia per i singoli anni sia per le singole donne presenti nelle liste nei due anni di interesse per la ricerca (2015 e 2019), è stata simulata l'acquisizione di tali dati in tempo reale. Per rispondere a tale esigenza è stato considerata **Apache Kafka**, una tecnologia middleware che si pone tra chi produce i dati e chi li consuma, realizzando una comunicazione asincrona. Tramite l'API di Kafka per Python è stato configurato un *Producer*, come mostrato in Figura 2.1 che, con l'intento di simulare una fonte dati streaming, legge i tweets contenuti in un file csv e li spedisce ad un Topic specifico, implementato su Apache Nifi. Tanto per l'implementazione del Producer quanto per quella del Consumer ci si è avvalsi di macchine *Azure* fornite dall'Università, con il fine di avere maggiori risorse computazionali a disposizione da poter sfruttare per il progetto.

Per simulare al meglio lo streaming in tempo reale di dati è stato inserito anche un `time.sleep` tra l'invio di un tweet e l'altro.

```

: # !pip3 install git+https://github.com/dpkp/kafka-python
from kafka import KafkaProducer
import json
import time

# producer creation
weets_2020weets_2020weets_2020weets_2020weets_2020producer = KafkaProducer(
    bootstrap_servers = ["kafka:9092"],
    value_serializer = lambda v: json.dumps(v).encode("UTF-8"))

# lettura file json tweets scaricati
tweets_final = json.load(open("tweets_2020.json"))

# tweets -> topic -> mongodb
for tweet in tweets_final:
    producer.send(topic='tweets', value=tweet)
    #time.sleep(0.005)

```

Figura 2.1: Kafka Producer

## NiFi

Su **Apache NiFi**, invece, è stato implementato un Consumer. *NiFi* è un applicativo Java di modellazione di flussi di dati che favorisce l'automazione e lo spostamento degli stessi tra sistemi diversi. Nell'ambito dei Big Data utilizzare NiFi diventa rilevante quando si hanno grandi quantità di informazioni da importare e si vuole evitare il collasso delle risorse: tutto ciò viene garantito dalla gestione del carico. NiFi permette di lavorare ad alto livello, con un certo grado di astrazione. E' stato dunque sviluppato il consumer che, una volta ricevuti i messaggi inviati dal Producer sviluppato in Kafka, li consuma. Per fare ciò, è stato implementato un workflow (Figura 2.2 che rappresenta il workflow finale) di NiFi contenente i seguenti nodi o processori (un processore è l'unità base di un flusso di dati e ognuno di essi ha una funzionalità diversa):

- **ConsumeKafkaRecord** (Figura 2.3): il messaggio ricevuto da Kafka viene deserializzato tramite il componente *Record Reader*. A seguire viene immesso nel FlowFile, dopo essere stato processato dal *Record Writer*;
- **PutMongoRecord** (Figura 2.4): permette di inserire dati all'interno di un database Mongo DB.

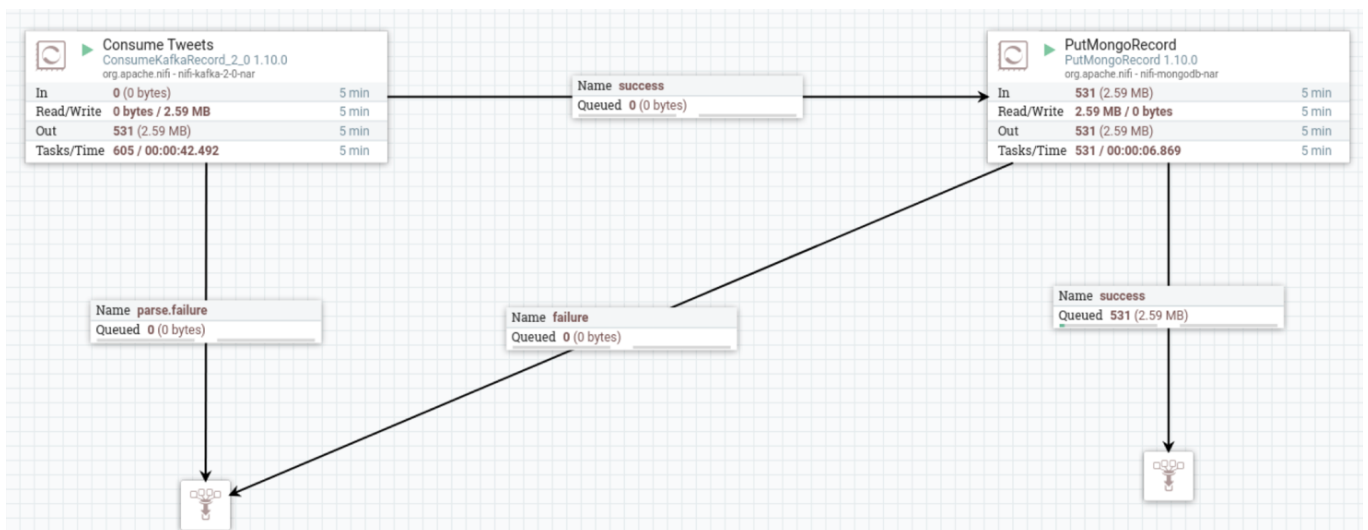




Figura 2.2: Workflow di NiFi

## MongoDB

Tramite l'interfaccia di NiFi e l'utilizzo del processore PutMongoRecord, è stato predisposto il salvataggio dei tweets consumati dal Consumer in un database NoSql *Mongo DB*. La scelta di utilizzare questa

	<b>Consume Tweets</b> ConsumeKafkaRecord_2_0 1.10.0 org.apache.nifi - nifi-kafka-2-0-nar	 1
In	0 (0 bytes)	5 min
Read/Write	0 bytes / 34.85 KB	5 min
Out	68 (34.85 KB)	5 min
Tasks/Time	185 / 00:00:05.707	5 min

**Configure Processor**

Stopped

SETTINGS SCHEDULING PROPERTIES COMMENTS

Required field +

Property	Value
Kafka Brokers	kafka:9092
Topic Name(s)	tweets
Topic Name Format	names
Record Reader	JsonTreeReader
Record Writer	JsonRecordSetWriter
Honor Transactions	true
Security Protocol	PLAINTEXT
SASL Mechanism	GSSAPI
Kerberos Credentials Service	No value set
Kerberos Service Name	No value set
Kerberos Principal	No value set
Kerberos Keytab	No value set

CANCEL APPLY

Figura 2.3: Consumer NiFi

	<b>PutMongoRecord</b> PutMongoRecord 1.10.0 org.apache.nifi - nifi-mongodb-nar	
In	68 (34.85 KB)	5 min
Read/Write	34.85 KB / 0 bytes	5 min
Out	68 (34.85 KB)	5 min
Tasks/Time	68 / 00:00:01.895	5 min

**Configure Processor**

Stopped

SETTINGS SCHEDULING PROPERTIES COMMENTS

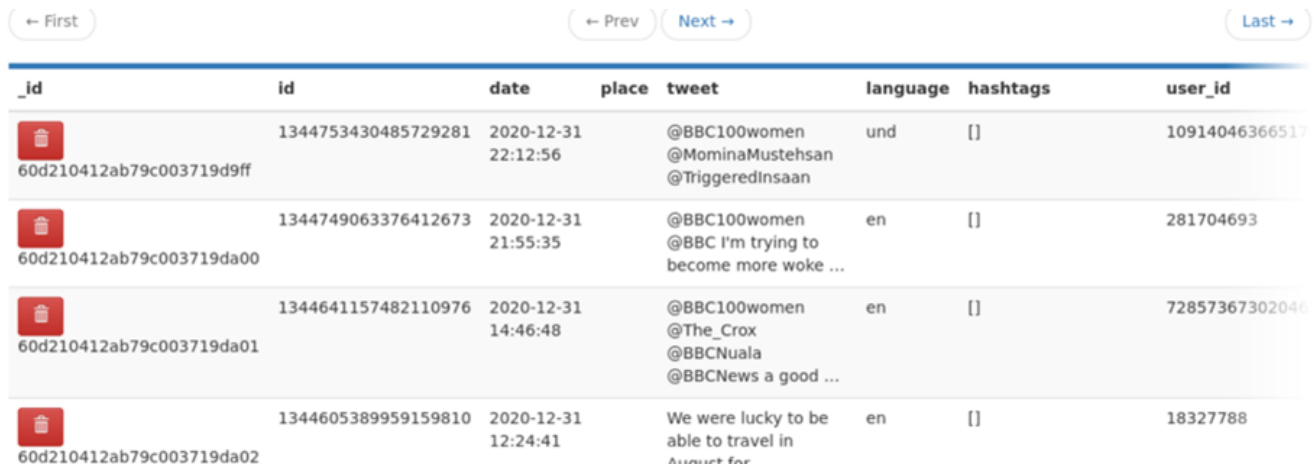
Required field +

Property	Value
Client Service	MongoDBControllerService
Mongo URI	No value set
Mongo Database Name	twitter
Mongo Collection Name	tweets_2020
SSL Context Service	No value set
Client Auth	REQUIRED
Write Concern	ACKNOWLEDGED
Record Reader	JsonTreeReader
Insert Batch Size	100

CANCEL APPLY

Figura 2.4: Nodo di MongoDB

tipologia di database è dovuta al fatto di poter strutturare i dati in collezioni e documenti JSON, superando la rigidità dei database relazionali. Il fatto di avere documenti garantisce la possibilità di memorizzare un volume maggiore di dati, anche tramite approccio *embedding*. Inoltre, l'utilizzo di un database come Mongo DB ha permesso di manipolare grandi quantità di dati eterogenei e senza uno schema prefissato, garantendo al tempo stesso alta disponibilità e scalabilità (Figura 2.5).



_id	id	date	place	tweet	language	hashtags	user_id
60d210412ab79c003719d9ff	1344753430485729281	2020-12-31 22:12:56		@BBC100women @MominaMustehsan @TriggeredInsaan	und	[]	10914046366517
60d210412ab79c003719da00	1344749063376412673	2020-12-31 21:55:35		@BBC100women @BBC I'm trying to become more woke ...	en	[]	281704693
60d210412ab79c003719da01	1344641157482110976	2020-12-31 14:46:48		@BBC100women @The_Crox @BBCNuala @BBCNews a good ...	en	[]	72857367302040
60d210412ab79c003719da02	1344605389959159810	2020-12-31 12:24:41		We were lucky to be able to travel in Assam for	en	[]	18327788

Figura 2.5: Collezione in Mongo

## 2.4 Varietà

Una volta raccolti i tweets sia relativi all'intero evento, sia relativi alle singole donne per il 2015 e il 2019, e completata la parte di simulazione della velocità, si è passati a considerare la seconda delle due V da trattare: la **varietà**. Come accennato inizialmente, l'obiettivo dell'intera analisi è stato quello di analizzare e studiare l'evento della BBC e ciò che è ad esso collegato per fornire una visione complessiva del fenomeno; per farlo è stato necessario arricchire i dati. Considerando il singolo anno dunque, si è provveduto ad integrare quelli delle donne rientranti nella lista con alcune altre informazioni quali l'età, lo username Twitter, la presenza o meno di un account verificato <sup>1</sup>, con quelli relativi agli Stati di appartenenza. In questa fase di integrazione alcuni dati come, ad esempio, l'età di alcune donne o il loro username Twitter, sono stati recuperati ed inseriti manualmente nella base dati. Altra operazione effettuata prima di passare all'integrazione vera e propria riguardava il trattamento dei tweets scaricati: l'intento dell'intero progetto non è stato quello di analizzare il *sentiment* (sebbene ne possa essere uno sviluppo futuro) dei vari tweets, ma di capirne la portata, in termini di numerosità. Considerando ciò, per i tweets scaricati relativi all'intero evento si è svolto un conteggio giornaliero del loro numero (per ogni anno), mentre per i tweets inerenti le singole donne si è sintetizzata l'informazione calcolandone il numero prima e dopo la pubblicazione della lista. Per quanto concerne invece i dati relativi agli Stati si sono cercate innanzitutto fonti attendibili ed autorevoli a livello globale. In seguito si è svolta una prima fase di *data preparation* per uniformare i nomi degli Stati in quanto, essendo le informazioni provenienti da fonti differenti, presentavano diversità nella sintassi. Infine, si è svolta la vera e propria integrazione tramite l'esecuzione di appositi script in linguaggio Python. I dati utilizzati nelle analisi relative agli Stati sono rappresentati da indicatori quali:

- **Pil e Pil pro capite**, i cui dati sono resi disponibili dall'International Monetary Fund;
- **Global gender gap**, i cui dati sono resi disponibili dal World Economic Forum;
- **Indice di sviluppo umano**, i cui dati sono resi disponibili dallo United Nations Development Programme.

<sup>1</sup>a seguito di una ricerca manuale, si è inserito un attributo e per ogni donna si è specificato se questa avesse o meno un account verificato su twitter (badge con spunta blu). Questo metodo per valutare se una donna è conosciuta o meno è stato preferito ad altri in quanto oggettivo e indipendente dalla percezione personale.

Inizialmente erano stati considerati anche altri indicatori come ad esempio il *Labor Percentage*, la percentuale di donne ministre e quella delle parlamentari. Il primo indicatore, reso disponibile dalla *World Bank Group*, rappresenta il rapporto tra tasso di partecipazione femminile e maschile alla forza lavoro; mentre gli ultimi due sono resi disponibili dall'*Organization for Economic Co-operation and Development* (OECD). Tuttavia, molti Stati presentavano valori mancanti per questi indici e, di conseguenza, si è preferito utilizzare quelli sopra riportati in quanto considerati significativi e utili a delineare le caratteristiche di interesse dei Paesi in analisi.

Inizialmente si era pensato di aggregare i vari dati in un unico file json ma, vista la limitata capacità del software Tableau nel gestire tale formate di dimensioni superiori a 128 Kb, si è preferito far ricorso a files in formato csv, più semplicemente gestiti dal software per l'implementazione delle visualizzazioni. La struttura finale delle tabelle utilizzate nelle visualizzazioni è la seguente:

**NUMERO TWEETS:** tabella per conservare il conteggio dei tweets che, giornalmente, si riferiscono all'evento BBC 100 Women (raccolti considerando gli hashtags e il tag):

- **Data:** dal 01-01-2015 al 31-12-2020;
- **Numero tweets:** conteggio del numero dei tweets giornalieri;

**IDENTIKIT DONNE:** tabella in cui sono contenute tutte le informazioni delle donne inserite nella lista della BBC:

- **Id:** identificatore generato in modo randomico per l'identificazione univoca della donna;
- **Anno:** anno in cui la donna è stata inserita nella lista. Assume valore 2015 o 2019;
- **Nome:** nome della donna, così come riportato nella lista della BBC;
- **Età:** età della donna (quando non reso disponibile dalla BBC, è stato cercato e inserito manualmente);
- **Categoria:** categoria in cui la donna si è distinta;
- **Stato:** Stato di provenienza della donna;
- **Regione:** zona geografica del mondo in cui rientra lo Stato. Assume valori quali: Africa, Asia, Australia, Europa, Nord America e Sud America. Tale divisione in *macroregioni* si è basata sulla ripartizione creata dalle Nazioni Unite (*Geoschema delle Nazioni Unite*) con una distinzione ulteriore fra Nord America e Sud America per sottolineare anche la differenza sostanziale per le donne nominate nei due Paesi;
- **Username:** username della donna su Twitter (per alcune donne l'informazione è mancante in quanto non hanno un account sul social network considerato);
- **Hashtags:** hashtags associati alla donna, inseriti a seguito di ricerca manuale;
- **Verificato:** account verificato o meno a seconda che il profilo della donna presenti o meno il badge con spunta blu;
- **Numero tweets pre pubblicazione:** numero dei tweets scaricati, relativi alla donna, nel periodo precedente all'uscita della lista BBC;
- **Numero tweets post pubblicazione:** numero dei tweets scaricati, relativi alla donna, nel periodo successivo all'uscita della lista BBC;

**INDICATORI STATI:** tabella in cui sono presenti gli Stati (non solo quelli da cui provengono le donne, ma tutti quelli per cui sono disponibili indicatori) e i relativi indicatori:

- **Stato:** Paese del mondo (scritto in inglese);

- **Anno:** anno a cui fanno riferimento gli indicatori. Assume valore 2015 o 2019;
- **Pil:** prodotto interno lordo dello Stato nell'anno corrispondente;
- **Pil pro capite:** prodotto interno lordo pro capite dello Stato nell'anno corrispondente;
- **Global Gender Gap:** indice di disparità di genere dello Stato nell'anno corrispondente;
- **HDI:** Indice di sviluppo umano dello Stato nell'anno corrispondente;
- **Id donna:** identificativo della donna, presente solo per gli Stati che hanno una donna eletta nell'anno corrispondente (se ci sono più donne in uno Stato per un dato anno, ci sono più righe nella tabella);
- **Nome donna:** nome della donna associata all'id.

## 3 DATA VISUALIZATION

### 3.1 Progettazione, implementazione e valutazione della qualità delle visualizzazioni

L'intenzione alla base del progetto è stata quella di analizzare l'evento della BBC *#BBC100Women* prendendo in considerazione differenti punti di vista.

Uno di questi ha riguardato l'impatto mediatico dell'evento per la cui analisi si è considerato un arco temporale dal 2015 al 2020. Per altri aspetti che verranno ripresi successivamente, l'obiettivo è stato principalmente quello di analizzarne l'evoluzione confrontando due anni in particolare: il 2015 e 2019. Tale scelta ha permesso di considerare un sufficiente periodo temporale di quattro anni (e non cinque a causa del particolare anno del COVID) per fornire un'immagine dei cambiamenti riguardanti i diversi elementi dell'evento.

Concluso il lavoro di raccolta, pulizia, elaborazione e analisi dei dati, le infografiche che vengono di seguito presentate ne sono il risultato, ottenute tramite l'utilizzo del software *Tableau* (una piattaforma di *business intelligence* focalizzata sull'analisi dei dati il cui *core* è indirizzato alla visualizzazione grafica delle informazioni, piuttosto che alla presentazione delle stesse in formato tabellare). Conclusa la fase di progettazione e creazione delle prime visualizzazioni, quest'ultime sono state sottoposte, tramite la *valutazione euristica*, a 6 utenti per raccogliere il loro giudizio: tale procedimento ha riguardato tutte le visualizzazioni realizzate.

Infine, una volta modificate in base ai suggerimenti raccolti dagli utenti, sono stati somministrati a 30 utenti i *questionari psicometrici* (nello specifico, è stato utilizzato il Cabitza-Locoro) per la valutazione delle grandezze relative all'*utilità*, alla *chiarezza*, all'*informatività*, alla *bellezza*, all'*intuitività* e al *valore complessivo* assegnato alla visualizzazione. Gli user test, che hanno coinvolto 12 utenti, hanno invece avuto lo scopo di valutare tramite la somministrazione di una domanda per ciascuna visualizzazione, la reattività e la capacità di comprensione del prodotto visivo da parte dell'utente e la correttezza o meno delle risposte fornite.

Riprendendo le domande di ricerca presentate di seguito:

1. Quanta eco ha sul social network Twitter la pubblicazione della lista e la stagione delle donne della BBC? È vero che se ne discute attivamente portando in primo piano il tema delle donne e il loro ruolo nella società?
2. Quali sono le caratteristiche delle donne maggiormente nominate? Si hanno specifici Stati o regioni a livello mondiale dalle quali le donne nominate provengono? Qual è l'identikit della donna nominata?
3. Nel corso degli anni, la risonanza dell'evento è aumentata? Si è modificata la tipologia di donne inserite nella lista?
4. La lista della BBC ha permesso ad alcune donne, sperabilmente quelle meno conosciute, di poter aumentare la loro popolarità su Twitter e, conseguentemente, diffondere più efficacemente lo scopo del loro impegno nel mondo?

sono state create tali data viz al fine di fornirne una risposta:

- Visualizzazione 1: Heatmap, per mostrare l'andamento del numero di tweet nel corso degli anni dal 2015 al 2019;
- Visualizzazione 2: Lollipop, per analizzare da un punto di vista descrittivo la composizione delle categorie stabilite dalla BBC, l'età mediana e il numero di donne per ciascuna categoria;
- Visualizzazione 3: Alluvial diagram, per visualizzare la relazione tra categorie, provenienza geografica e popolarità delle donne;
- Visualizzazione 4: Scatterplot, per mostrare più nello specifico le caratteristiche degli Stati di provenienza delle donne nominate nella lista e i dati relativi ai Paesi. Vengono nominate più donne provenienti da stati ricchi/più sviluppati o viceversa?
- Visualizzazione 5: Barplot e Donut chart, per indicare l'impatto della nomination rispetto al numero di tweet per ciascuna donna pre-pubblicazione e post-pubblicazione.

Di seguito verranno quindi mostrate le singole visualizzazioni create, la domanda di ricerca alla quale ciascuna di esse ha voluto rispondere e la relativa valutazione euristica con le modifiche estrapolate dai commenti dagli utenti.

### 3.1.1 Visualizzazione 1 - Heatmap

#### Point Heatmap

Come prima visualizzazione è stata scelta una heatmap che rappresentasse, riferito allo specifico anno, un calendario contenente il numero di tweets per singolo giorno considerando particolari hashtags di riferimento (*#BBC100Women* e *@BBC100Women*). Tale infografica ha avuto lo scopo di mostrare come, prima della pubblicazione delle nomine, tali argomenti non avessero molto seguito (almeno su Twitter), cosa invece molto diversa il giorno stesso della pubblicazione e, indicativamente, per le successive tre settimane nelle quali, generalmente, la BBC ha promosso eventi collegati al tema per pubblicizzare e comunicare più efficacemente tutto il lavoro svolto per la creazione di tale lista e per mettere in mostra le personalità di spicco delle donne ivi presenti.

#### Valutazione euristica - Heatmap

I problemi riscontrati durante la valutazione euristica riguardante la prima visualizzazione sono riportati in tabella. Inoltre, sono stati anche definite le modifiche svolte come parte integrante dell'*assessment formativo* nella logica del *Ciclo di Deming* riferito però alla produzione della data visualization.

HEATMAP	ASPETTI PROBLEMATICI	ASPETTI MIGLIORATI
	- Il context inizialmente inserito (tramite annotazioni sulla heatmap stessa) ha confuso gli utenti che hanno dichiarato di ricevere troppe informazioni prima ancora di poter comprendere quale fosse il <i>point</i> della visualizzazione;	- Tali annotazioni sono state rimosse dalla visualizzazione su Tableau perché troppo confusionarie, e inserite in modo più ordinato nelle visualizzazioni statiche qui riportate;
	- Poiché a seguito dell'uscita della lista seguono tre settimane di eventi in favore della comunicazione dell'evento, tali giorni sarebbero dovuti essere maggiormente evidenziati in quanto gli utenti non erano in grado di capire esattamente il giorno dell'uscita e il periodo successivo.	- Per risolvere il problema nelle visualizzazioni statiche sono stati aggiunti dei glifi per indicare il periodo di inizio e quello di fine degli eventi post pubblicazione.





## Quanto se ne parla su Twitter?

Sono stati presi in considerazione i tweets con l'hashtag #BBC100women e la menzione @BBC100Women

**100 Women** Un evento organizzato annualmente dalla BBC che a partire dal 2013 ha come obiettivo quello di valorizzare il ruolo della donna nella società moderna. L'idea nasce in seguito allo stupro di gruppo nei confronti di una ragazza 23enne di Delhi avvenuto nel 2012. A seguito di numerose manifestazioni di protesta, alcuni personaggi di rilievo della BBC decisero di creare una serie incentrata sulle questioni e le conquiste delle donne nella società di oggi.

Nei mesi finali dell'anno viene solitamente definita la **lista delle 100 donne** che meglio si sono distinte e che sono riuscite a fare la differenza nel corso dell'anno in modo tale che possano essere di ispirazione per il mondo intero.

Dopo la pubblicazione della lista prende il via la **"stagione delle donne della BBC"**, della durata di 3 settimane: trasmissioni, documentari, dibattiti online incentrati sul tema delle donne.

Le donne di tutto il mondo sono incoraggiate a partecipare tramite **Twitter**, commentando la lista, nonché seguendo le interviste e i dibattiti che si svolgono dopo la pubblicazione di essa.

Di seguito viene mostrata la risonanza mediatica dell'evento e una panoramica descrittiva delle donne nominate in due anni distinti considerandone caratteristiche e provenienza. Infine, viene analizzata l'influenza dell'evento sulla popolarità delle donne su Twitter.

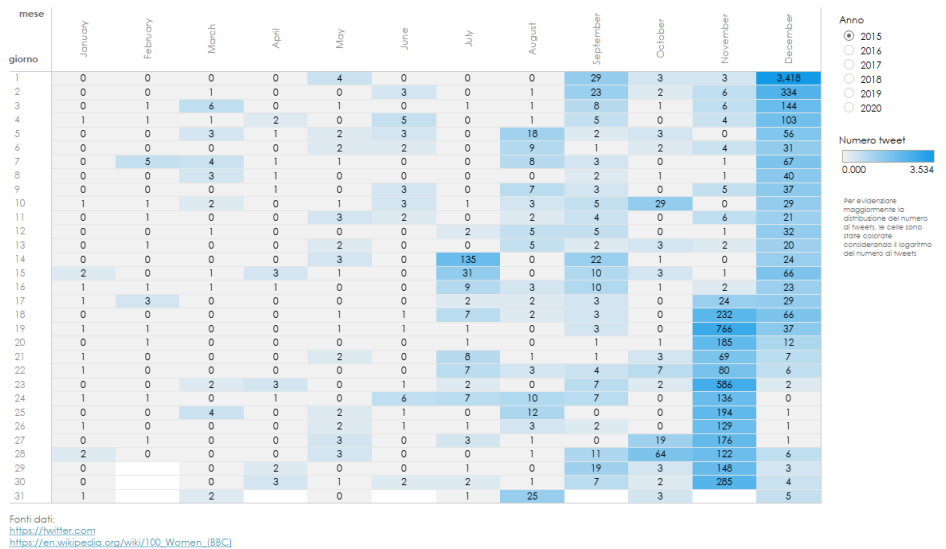


Figura 3.1: Quanto se ne parla su Twitter?

Infine, con l'intento di evidenziare il context e sottolineare le informazioni di interesse dal 2015 al 2020, vengono di seguito presentate le Heatmap per ogni anno di dati disponibili (Figure 3.2, 3.3, 3.4, 3.5, 3.6, 3.7).



## Nel 2015 su Twitter quanto si parla dell'evento 100 Women?

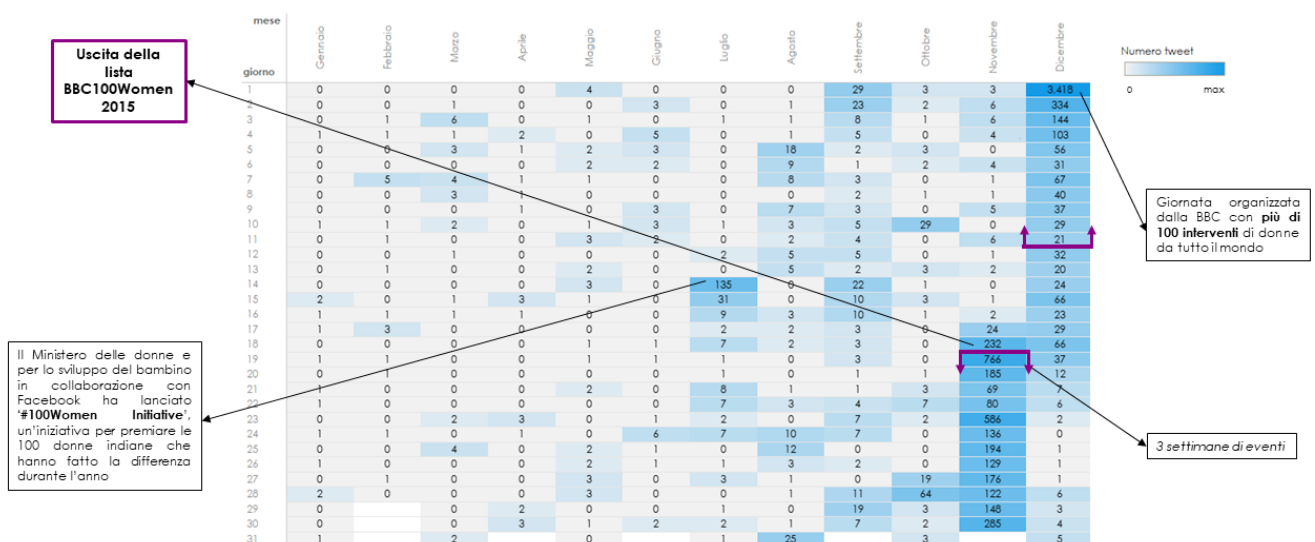


Figura 3.2: Nel 2015 su Twitter quanto si parla dell'evento 100Women?

Nel 2016 su **Twitter** quanto si parla dell'evento **100 Women** ?

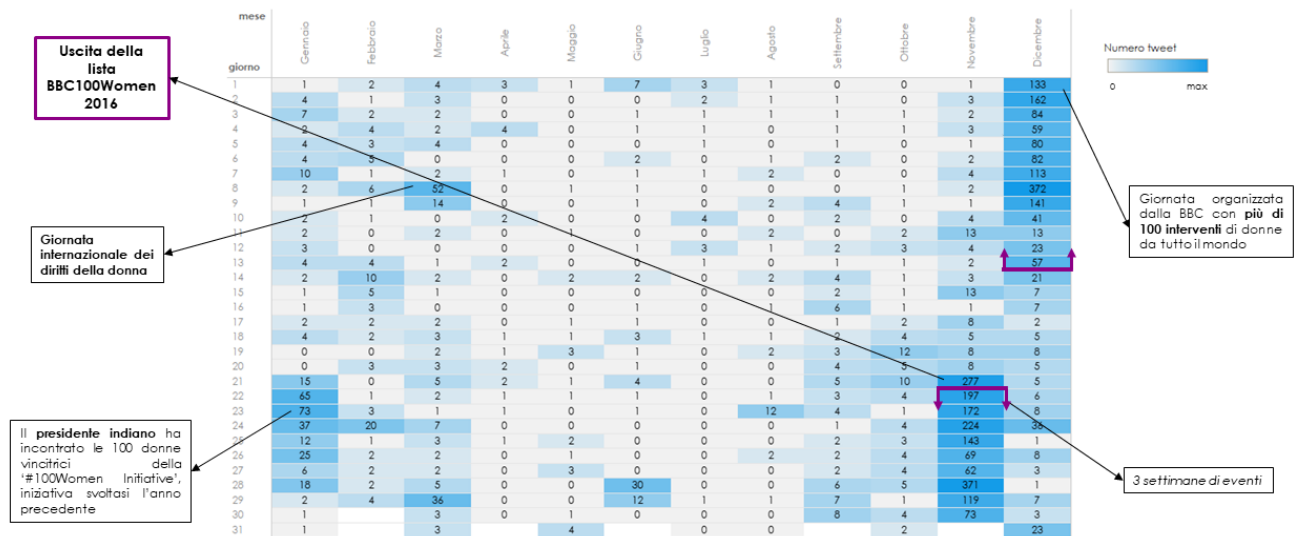


Figura 3.3: Nel 2016 su Twitter quanto si parla dell'evento 100Women?

Nel 2017 su **Twitter** quanto si parla dell'evento **100 Women** ?

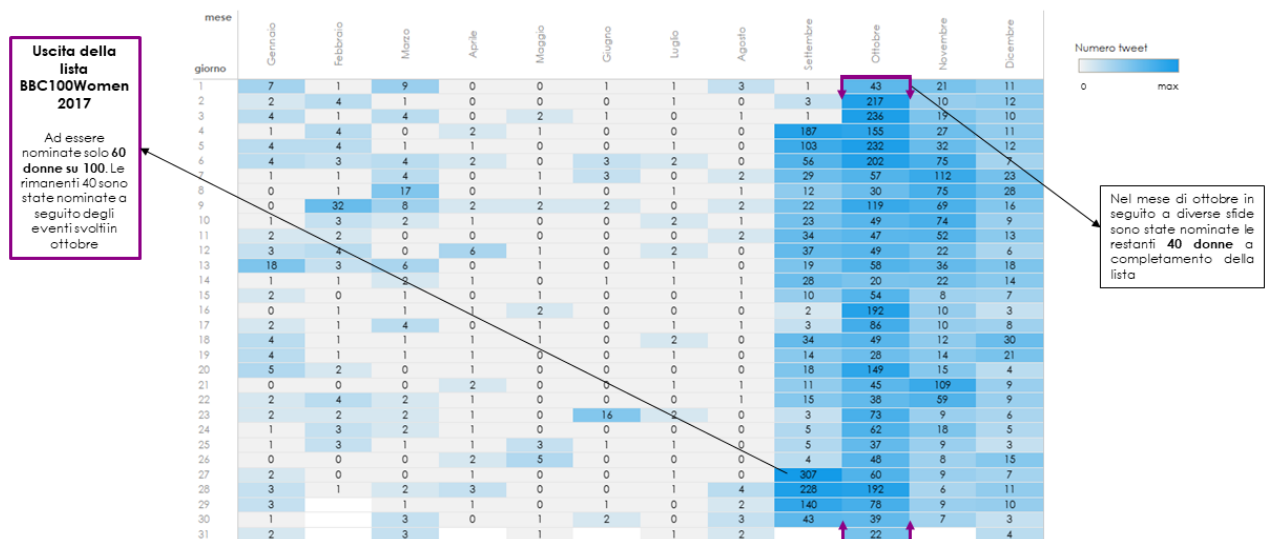


Figura 3.4: Nel 2017 su Twitter quanto si parla dell'evento 100Women?

### 3.1.2 Visualizzazione 2 - Lollipop

#### Point Lollipop

Il *point* del Lollipop chart è stato quello di *mostrare*, per gli anni 2015 e 2019, una prima descrizione della tipologia di donne scelte e nominate nella lista delle *100 Women* che hanno fornito il loro maggiore impatto sulla società in tali anni. Vengono quindi specificate le categorie create dalla BBC come elementi caratterizzanti la singola personalità e la misura della *mediana* dell'età. Questo ultimo aspetto ha fornito

Nel 2018 su **Twitter** quanto si parla dell'evento **100 Women** ?

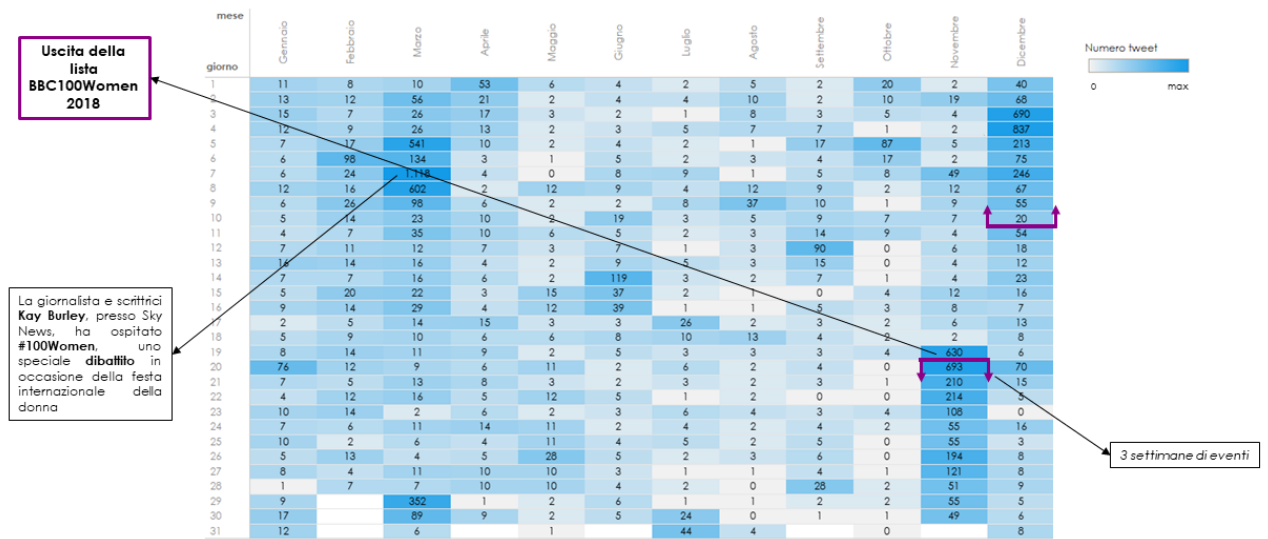


Figura 3.5: Nel 2018 su Twitter quanto si parla dell'evento 100Women?

Nel 2019 su **Twitter** quanto si parla dell'evento **100 Women** ?

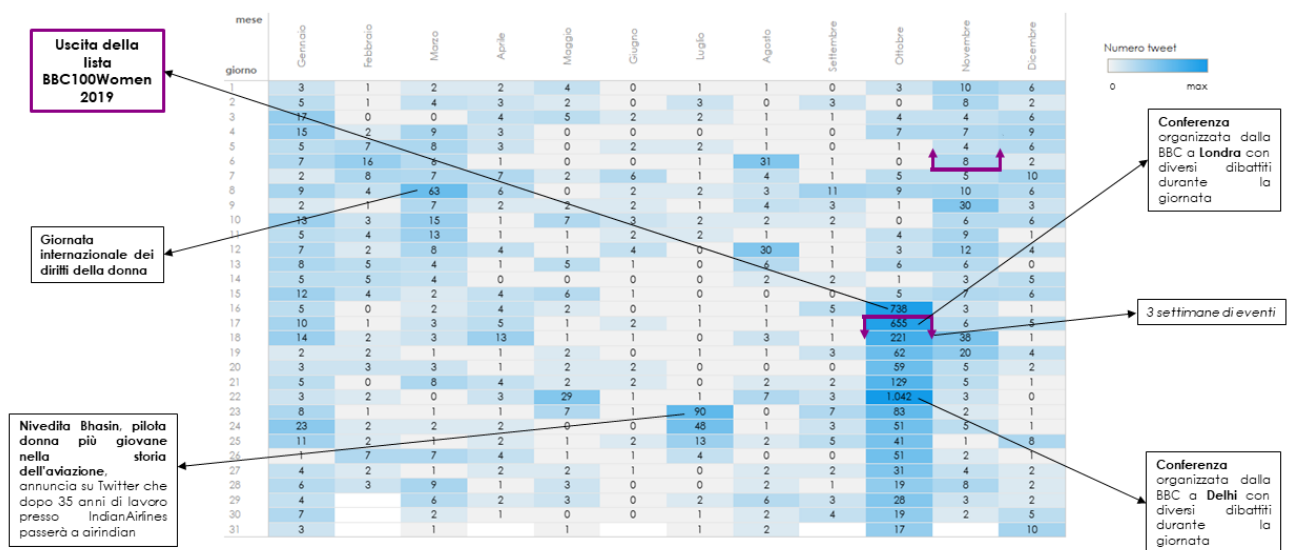


Figura 3.6: Nel 2019 su Twitter quanto si parla dell'evento 100Women?

un'informazione aggiuntiva (qual è la fascia d'età, oltre la categoria, che risulta essere più rappresentativa? Viene a modificarsi al variare dell'anno?), uscendo dalla logica classica dell'utilizzo della media, indicatore di tendenza centrale non robusto agli outlier.

Nel 2020 su **Twitter** quanto si parla dell'evento **100 Women** ?

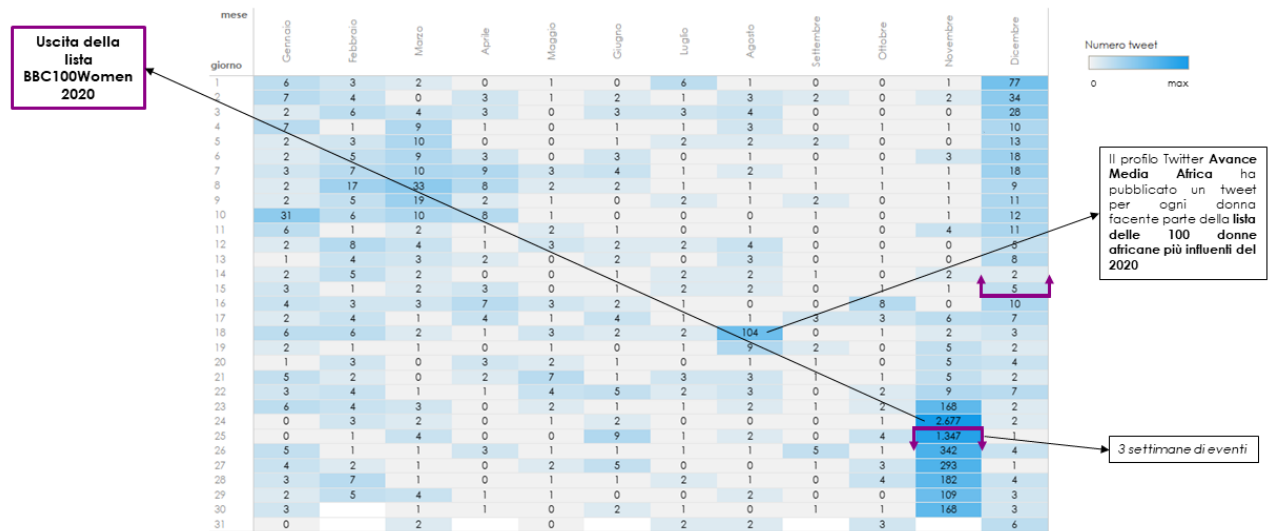


Figura 3.7: Nel 2020 su Twitter quanto si parla dell'evento 100Women?

#### Categorie rappresentative ed età

Qual è la classe più rappresentativa delle donne in lista?  
Quali sono le età mediane\* per ciascun gruppo?

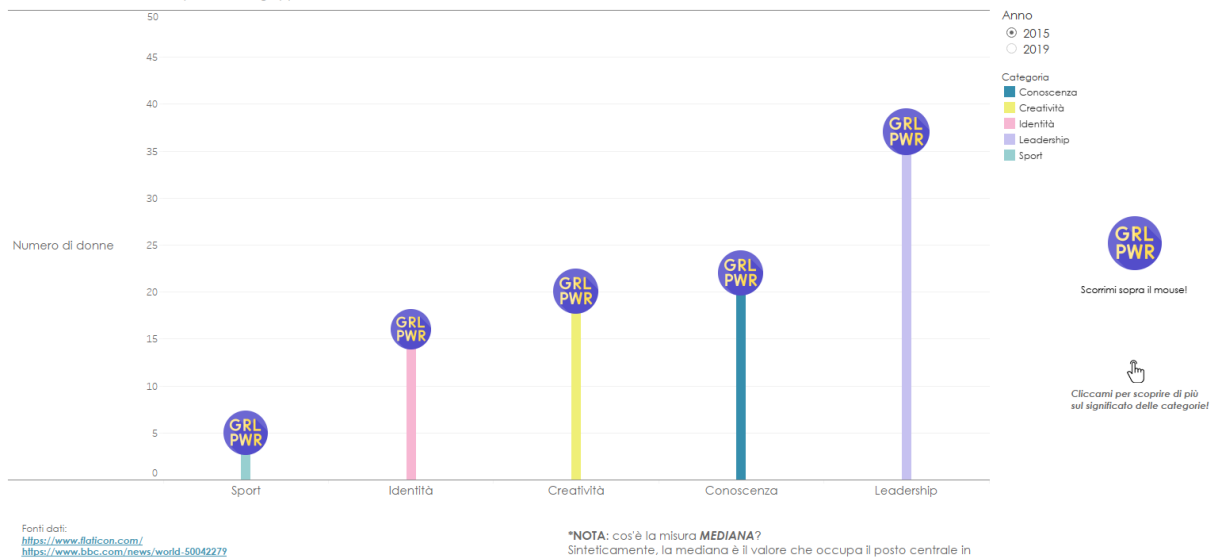


Figura 3.8: Quali sono le categorie di riferimento? Qual è l'età mediana?

### Valutazione euristica - Lollipop

Riguardo la seconda visualizzazione, sono riportati in tabella i problemi maggiormente riscontrati e risolti.

LOLLIPOP	ASPETTI PROBLEMATICI	ASPETTI MIGLIORATI
	- Un problema di particolare importanza riscontrato è stato quello di trasmettere all'utente tipo il concetto di <i>mediana</i> , di non così immediata comprensione come sperato. Infatti, non sono stati pochi coloro che hanno confuso tale concetto con quello di <i>media</i> , interpretando in maniera incorretta quanto vedevano;	- Per risolvere tale problematica, è stata aggiunta una nota in asterisco che potesse rimarcare il vero significato di tale indicatore, rendendo più esplicativo ed autoevidente quanto mostrato;
	- Inoltre, un elemento importante che non era stato compreso dai più ad una prima ispezione, sono stati i significati delle categorie presentate. Molti utenti hanno segnalato l'incomprensione, nello specifico, della categoria <i>identità</i> .	- Per rendere più chiaro ciò a cui fa riferimento la suddivisione scelta dalla BBC, si è voluto integrare il <i>Lollipop chart</i> con un'informazione inserita in secondo piano, tramite un glifo a forma di <i>mano</i> da poter cliccare. Si possono così ottenere tutte le necessarie spiegazioni in merito per poter afferrare i significati delle singole categorie.

### 3.1.3 Visualizzazione 3 - Alluvial

#### Point Alluvial

L'alluvial diagram ha la capacità di riassumere diverse dimensioni tramite l'altezza delle tre barre presenti ma anche grazie ai fasci colorati. In questo caso ci si riferisce alle categorie di appartenenza delle donne, alle regioni di provenienza e alla popolarità delle stesse su Twitter. L'obiettivo di tale visualizzazione è stato quello di far capire a colpo d'occhio, a seconda dell'anno scelto, quale potesse essere la categoria più influente (come già visto nel Lollipop), ma anche da quale regione a livello mondiale le donne nominate provenissero.

#### Valutazione euristica - Alluvial

L'Alluvial diagram ha presentato i problemi e le successive migliorie di seguito riportate:

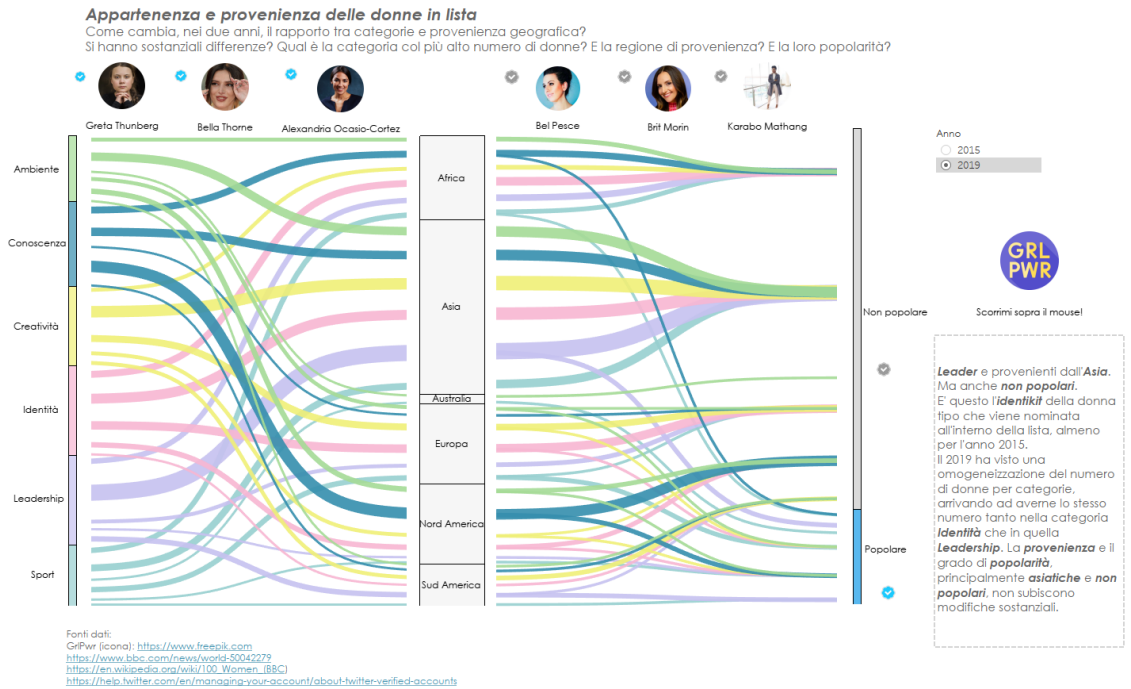


Figura 3.9: A quali categorie appartengono le donne in lista? Da quali regioni mondiali provengono?

ALLUVIAL	ASPETTI PROBLEMATICI	ASPETTI MIGLIORATI
	<ul style="list-style-type: none"> <li>- Si è notata una certa difficoltà nel comprendere tanto il point quanto la composizione della visualizzazione;</li> </ul>	<ul style="list-style-type: none"> <li>- Per risolvere tale problema, oltre alla già presente spiegazione dell'utilità e dell'obiettivo della visualizzazione presente all'interno del logo <i>GRLPWR</i> (<i>Girl Power</i>) al passaggio del mouse al di sopra, è stato inserito più chiaramente quale scopo si prefiggesse tale visualizzazione poco sotto al logo stesso. In tal modo vengono fornite informazioni immediate e chiarificatrici rispetto al <i>point</i> e alle differenze rispetto ai differenti anni;</li> </ul>
	<ul style="list-style-type: none"> <li>- Una seconda problematica è emersa in una prima progettazione nella quale erano stati scelti i termini di <i>Conosciuta/Sconosciuta</i> riferiti alla parte destra della dataviz. Per gli utenti meno avvezzi all'utilizzo dei social network in generale e, nello specifico a Twitter, tale distinzione non era chiara in un primo momento senza un'ulteriore spiegazione.</li> </ul>	<ul style="list-style-type: none"> <li>- Per specificarlo e chiarire tale punto, si è presa ispirazione dal concetto di <i>account verificato</i> presente in Twitter: si sono quindi inseriti e spiegati i glifi corrispondenti e sono stati modificati i termini in <i>Popolare/Non popolare</i>, a primo impatto più comprensibili.</li> </ul>

### 3.1.4 Visualizzazione 4 - Scatterplot

#### Point Scatterplot

Lo scatterplot ha voluto riassumere diversi indici con i quali si può avere un'idea della qualità della vita di una specifico Stato (pil pro capite, indice di sviluppo umano, *gender gap*). Tale visualizzazione ha avuto come obiettivo quello di mostrare come, a seconda dei valori degli indici assunti da ciascun elemento, gli Stati si potessero distribuire fra i quattro quadranti che si vengono a creare con l'inserimento delle fasce di normalità: quelli più ricchi e maggiormente egualitari dal punto di vista della parità di genere fra uomini e donne vengono a disporsi in alto e a destra dello scatterplot; viceversa per quegli Stati che presentano valori inferiori tanto dell'indice di sviluppo umano che del *gender gap*. La domanda a cui si è voluto rispondere è stata: da quali Stati provengono le donne nominate dalla BBC? Ne sono presenti alcuni in particolare che hanno un numero di donne nominate maggiore di altri? Si tratta di Stati che presentano particolari condizioni di povertà o, al contrario, sono quelli più ricchi che hanno una più alta prevalenza di donne nominate? E infine, confrontando i due anni, si possono notare forti differenze a livello anche di Stati che hanno migliorato la loro condizione in termini di indici?

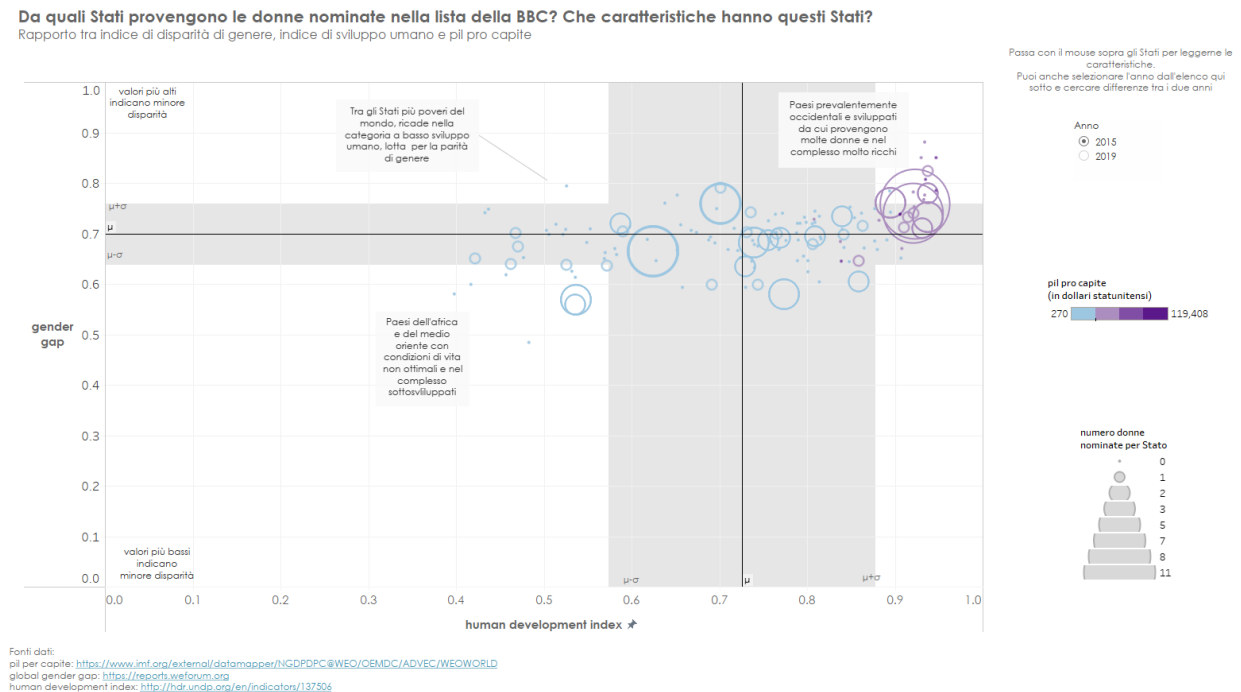


Figura 3.10: Da quali Stati provengono le donne nominate dalla BBC?

#### Valutazione euristica - Scatterplot

I problemi riscontrati nella visualizzazione dello Scatterplot vengono presentati nella seguente tabella:

SCATTERPLOT	ASPETTI PROBLEMATICI	ASPETTI MIGLIORATI
	Nel caso dell'indicatore <i>Gender Gap</i> non era risultato subito chiaro che valori più alti (e quindi vicini ad 1) indicassero una minore disparità di genere. Per molti utenti, tale indicazione era apparsa controuintiva.	Per risolvere tale problema si è indicato vicino ai valori 0 ed 1 il significato specifico.

### 3.1.5 Visualizzazione 5 - Boxplot/Barplot - Barplot e Donut chart

#### Point Boxplot/Barplot

L'ultima visualizzazione ha l'obiettivo di dimostrare quale sia stata la differenza in termini di numero di tweets post lista e numero di tweets pre lista: ovvero, dato il numero di tweets che la singola donna ha accumulato nel periodo precedente rispetto alla pubblicazione delle nomine, si sono avuti incrementi importanti nell'interesse della comunità femminile su Twitter, per quella specifica donna, a seguito della pubblicazione della BBC delle nomine? Se sì, questa differenza la si è avuta principalmente sulle donne che, su Twitter, erano già popolari, oppure maggiormente su quelle non popolari?

In sintesi: quali sono le donne che hanno ottenuto un boost di popolarità dall'uscita della lista?

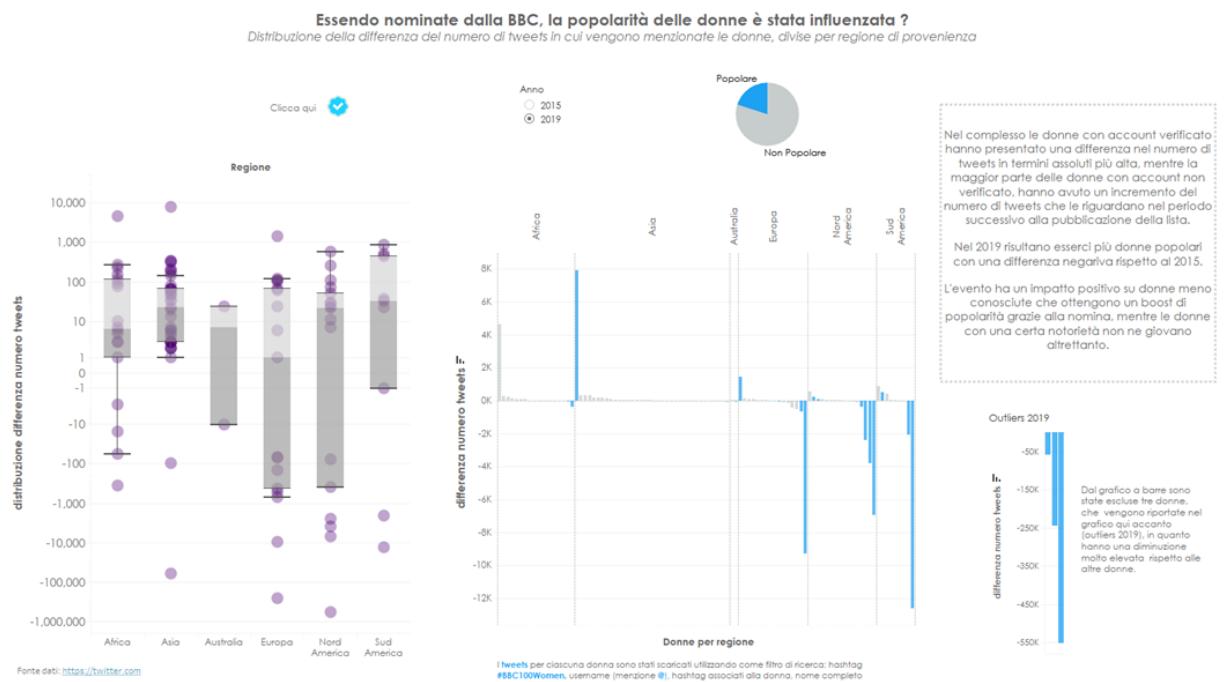


Figura 3.11: Come cambia la popolarità delle donne in lista?

#### Valutazione euristica - Boxplot/Barplot

Infine, riguardo il Boxplot/Barplot sono stati fatti i seguenti commenti:



BOXPLOT/BARPLOT	ASPETTI PROBLEMATICI	ASPETTI MIGLIORATI
	- Si è notata una difficoltà di esplorazione iniziale senza l'ausilio del componente del gruppo: il <i>Clicca qui</i> presente di fianco al badge non veniva considerato;	- E' stata semplificato l'accesso alle informazioni sulla modalità di esplorazione della visualizzazione: ovvero, il badge è stato convertito nel simbolo di Twitter di più grandi dimensioni ed è stata resa più visibile e comprensibile l'indicazione testuale;
	- Difficoltà da parte dell'utente nel comprendere la natura e struttura del Boxplot, vista la poca familiarità con la tipologia di visualizzazione;	- E' stato quindi sostituito il boxplot tramite uno 100% Stacked Bar Chart, di più facile comprensione, in cui viene mostrata la numerosità delle donne per Regione;
	- Infine, riguardo il pie chart, se gli utenti non hanno avuto problemi nel capirlo, in una ulteriore revisione, è stato deciso per un suo miglioramento per evitare un'inutile spreco di colore.	- Il pie chart è stato rimpiazzato da un donut chart con l'aggiunta del colore per evidenziare maggiormente la distinzione tra utenti verificati e non verificati.

Come si può notare dallo user test di Figura 3.19 e dall'ultimo violin plot di Figura 3.24, i cui dati sono da attribuire alla Figura 3.11, il tipo di visualizzazione proposta non è stata in grado, in termini di efficacia e di efficienza di trasmettere il proprio *point*.

Per questo motivo si è deciso di semplificare la data visualization per mostrare un'alternativa qualitativamente più immediata da comprendere e riordinata. Si è riscontrato, sottoponendo tramite valutazione euristica la data viz ad un ristretto numero di utenti, che tale nuova infografica (Figura 3.12) sia stata di più semplice comprensione.

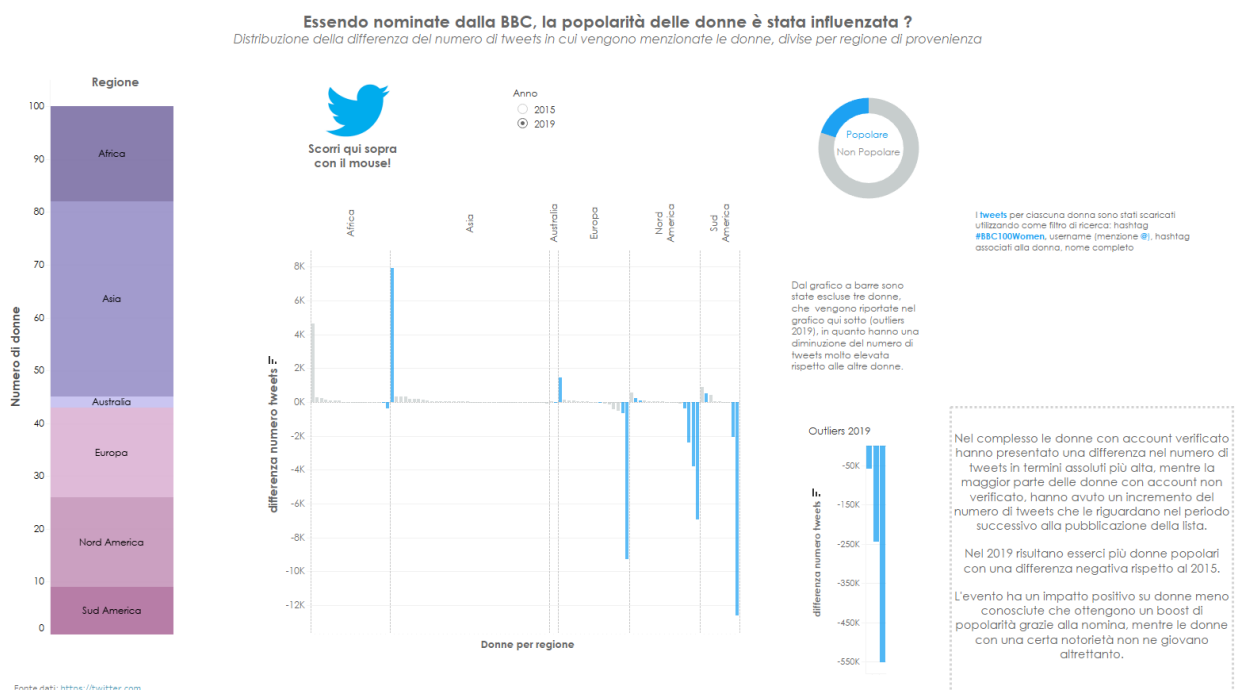


Figura 3.12: Come cambia la popolarità delle donne in lista?

L'utilizzo di un barplot è stato considerato come di più immediata comprensione, mentre il precedente pie chart è stato convertito in un donut chart per migliorarne la presentazione grafica. Tale visualizzazione, sebbene non sottoposta a valutazione, è stata comunque scelta come parte integrante di tale progetto.

## 3.2 Questionario Psicometrico

Il *questionario psicometrico* è stato sottoposto a 30 utenti che hanno accettato di valutare la qualità delle visualizzazioni tramite l'utilizzo della *Cabitzza-Locoro scale* a 6 valori. Si tratta di una tipologia di questionario che tenta di codificare le opinioni degli utenti (costrutti mentali e psicologici) in costrutti misurabili legati alle preferenze degli utenti stessi, tramite una *likert scale*.

Tale questionario è suddiviso in due parti di cui, la prima, la valutazione dell'infografica da parte dell'utente si è basata sull'attribuire un valore da 1 (pochissimo) a 6 (moltissimo) ai seguenti attributi:

Utile	Chiara	Informativa	Bella	Intuitiva
-------	--------	-------------	-------	-----------

La seconda sezione invece, contiene un'unica altra domanda relativa al *valore complessivo* che l'utente si sente di assegnare alla *data visualization* appena valutata, avendo a disposizione sempre i valori da 1 (pochissimo) a 6 (moltissimo).

Vengono ora presentati i risultati del questionario tramite due elementi grafici:

- Correlogramma (di Figura 3.13): utile per stabilire la presenza di correlazioni significative tra i diversi aggettivi scelti e, tra questi e il valore complessivo fornito dall'utente. Nel caso specifico, il correlogramma qui utilizzato, riassume i valori ottenuti in tutte le visualizzazioni considerate complessivamente;
- Reverse Horizontal 100% Stacked Bar Chart: specialmente utile quando è possibile estrapolare anche le risposte centrali, quelle 'neutre', per valutarne la tendenza verso la positività o negatività. Col tale visualizzazione è possibile mostrare l'andamento del gradimento degli utenti, il loro disappunto e la loro neutralità.

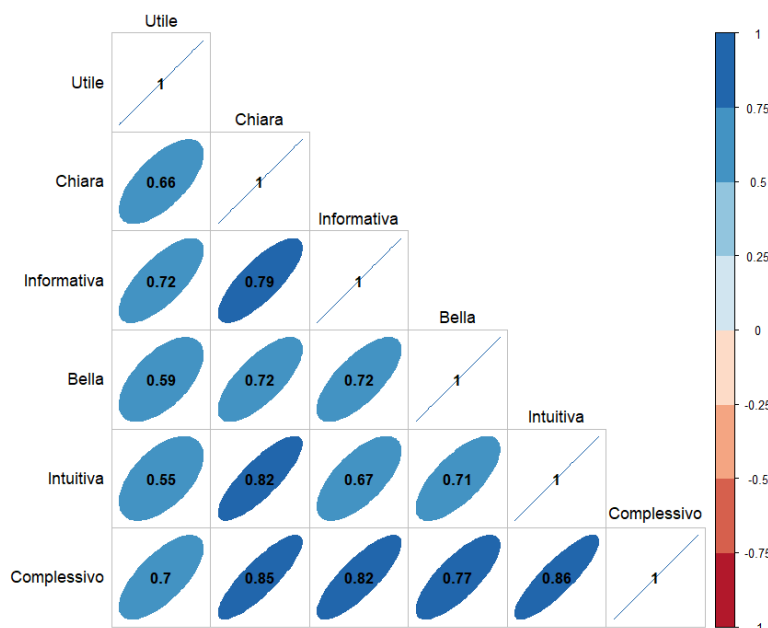


Figura 3.13: Correlogramma del questionario psicometrico

Dal correlogramma è possibile ottenere immediatamente alcune informazioni molto importanti e i seguenti sono i risultati più significativi:

- si nota una elevata correlazione positiva (maggiore di 0.8 o di poco al di sotto) fra il valore complessivo e le dimensioni di *chiarezza*, *informatività*, *bellezza* e *intuitività* a sottolineare che, complessivamente, le visualizzazioni sono apparse tanto chiare quanto informative ma anche belle e intuitive;
- L'*intuitività* correla molto positivamente con la *chiarezza* (0.82) e quasi allo stesso modo si comporta l'*informatività*.
- Le restanti correlazioni sono sempre positive anche se, generalmente, inferiori o di poco superiori allo 0.7, indicando comunque una propensione alla correlazioni fra gli aggettivi.

Infine, dal *Reverse Horizontal 100% Stacked Bar Chart*, creato per ogni singola visualizzazione è stato possibile ottenere i seguenti risultati:

### Heatmap

Come è possibile vedere dal *likert chart* di Figura 3.14, si nota come l'*utilità*, l'*informatività* e l'*intuitività* siano gli aggettivi che meglio hanno performato dal punto di vista dei rispondenti neutri (ovvero coloro che hanno dato come risposta 3 e 4). Quelli più positivi (che hanno quindi ricevuto come risposte 5 e 6) sono state l'*informatività* e il *valore complessivo*.

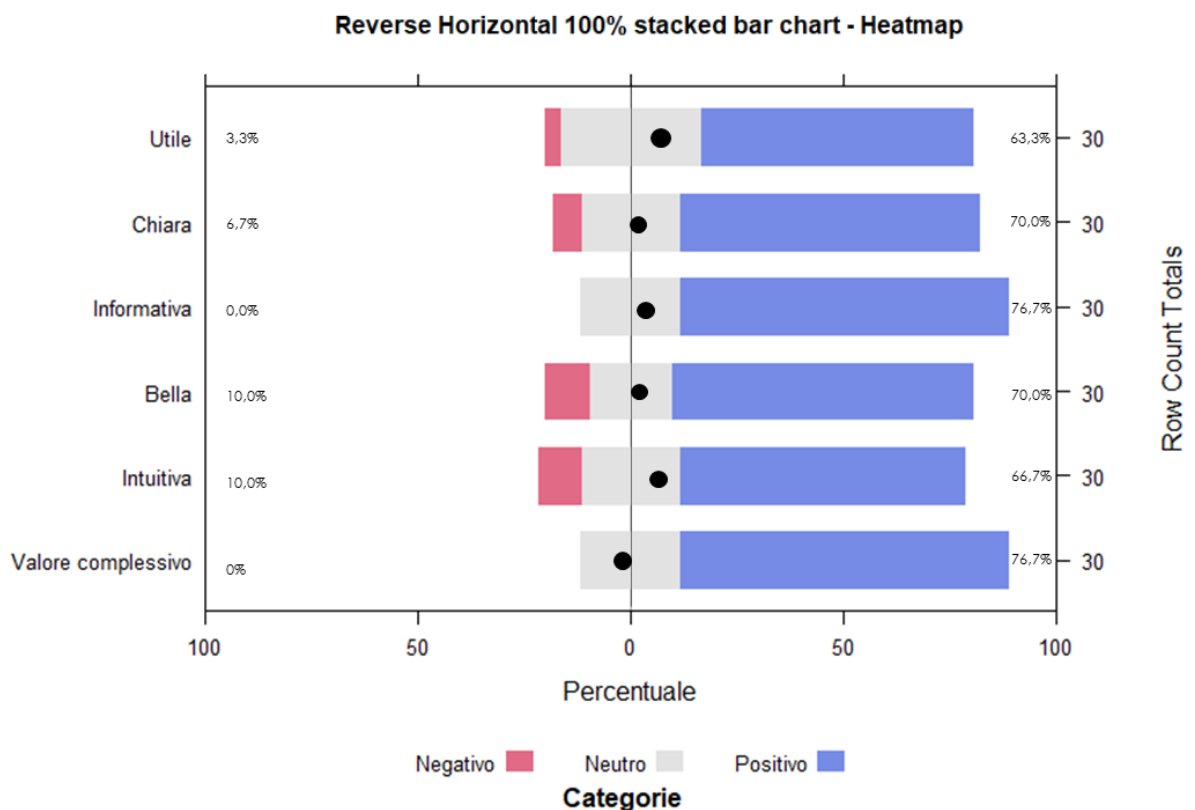


Figura 3.14: Likert chart - Heatmap

### Lollipop

Dal *likert chart* di Figura 3.15 si ha la capacità di comprendere come l'*utilità*, la *chiarezza*, l'*informatività* e la *bellezza* abbiano teso maggiormente verso la positività di coloro che è rimasto più neutro nella valutazione della dataviz. L'*intuitività* e la *bellezza* sono stati, invece, gli aggettivi che hanno ricevuto maggior consenso di positività.

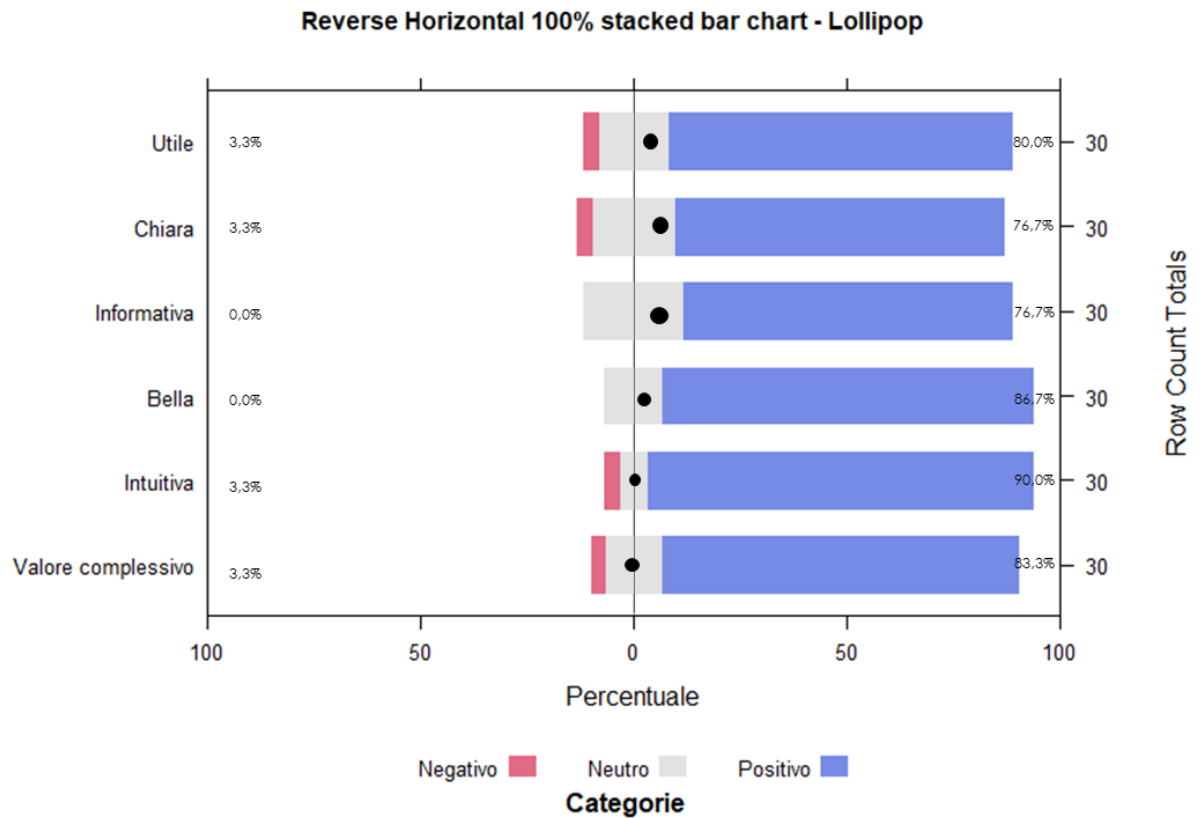


Figura 3.15: Likert chart - Lollipop

### Alluvial

Il *likert chart* di Figura 3.16 in riferimento all'*alluvial diagram* ha suscitato reazioni meno incoraggianti sotto alcuni punti vista dei rispondenti negativamente: la *chiarezza* e l'*intuitività* sono stati considerati negativamente da circa un quinto degli utenti, mentre i valori maggiormente positivi si sono ottenuti dal lato dell'*informatività* ma anche della *bellezza*.

### Scatterplot

Dal *likert chart* di Figura 3.17 i valori positivi e quelli neutri ma tendenti alla positività sono la totalità. Solamente nel caso dell'*utilità*, della *chiarezza* e dell'*intuitività* si sono registrate risposte negative, ma mai al di sopra del 10% degli utenti.

### Barplot e Donut chart

Infine, dal *likert chart* di Figura 3.18 in riferimento al Barplot/Boxplot, come con l'*Alluvial diagram*, gli utenti hanno espresso un giudizio più indeciso. Fra coloro che hanno fornito risposte neutre, i valori tendenti più verso il negativo sono stati quelli relativi alla *chiarezza* e *intuitività*. I risultati positivi (i valori fra il 5 e il 6) non hanno superato il 60% lasciando un'ampia fetta di rispondenti più verso una posizione neutra che una netta accettazione o diniego. A livello di *valore complessivo* però, non ci sono state risposte negative.

## 3.3 User test

Lo *user test* è un test di tipo quantitativo utile nella misurazione di due dimensioni fondamentali per la valutazione della data visualization:

- **Efficacia:** essa misura, tramite il *tasso d'errore*, la quantità di risposte incorrette alle domande che l'intervistatore ha posto all'utente finale. Il valore convenzionale scelto è quello del 95%: non dovrebbero essere presenti più del 5% di risposte errate;

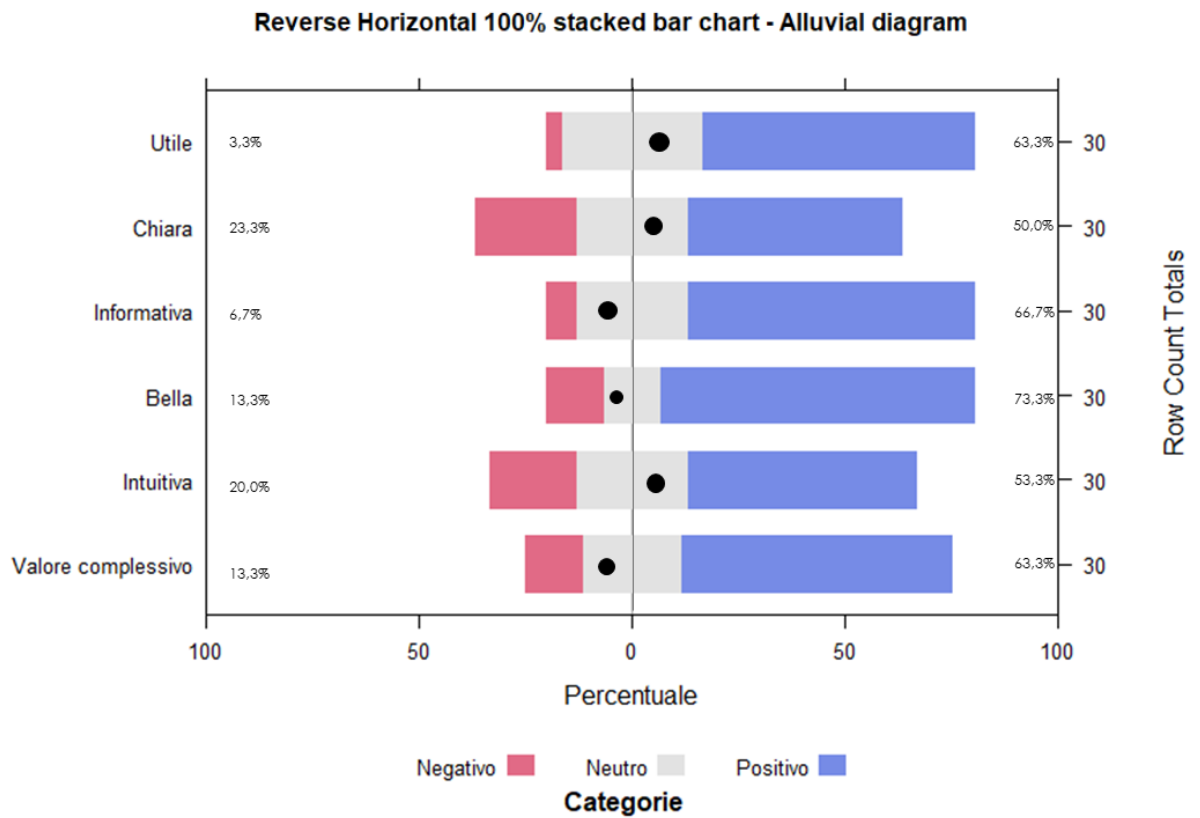


Figura 3.16: Likert chart - Alluvial diagram

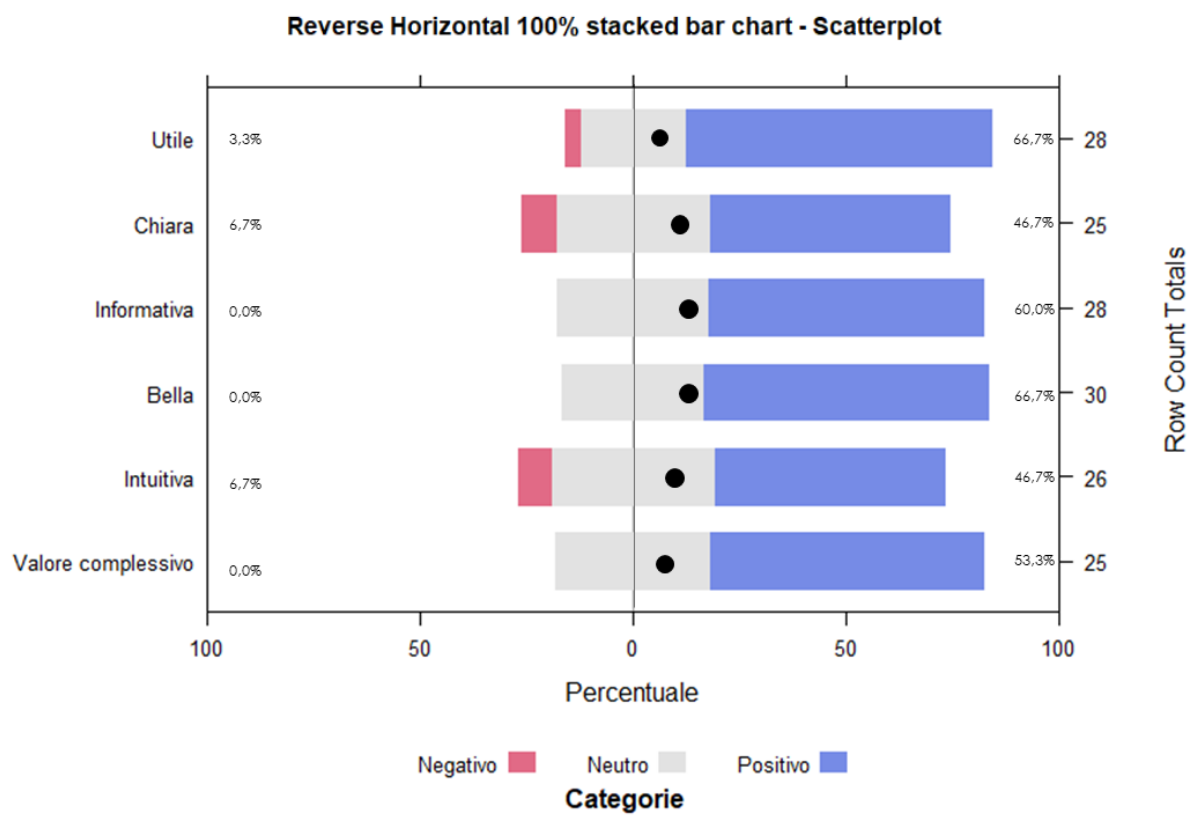


Figura 3.17: Likert chart - Scatterplot

- Efficienza: essa misura l'*execution time* di ogni utente rispetto alle singole risposte fornite. Confrontandolo con un valore di benchmark, si è valutato il tempo di risposta degli utenti.

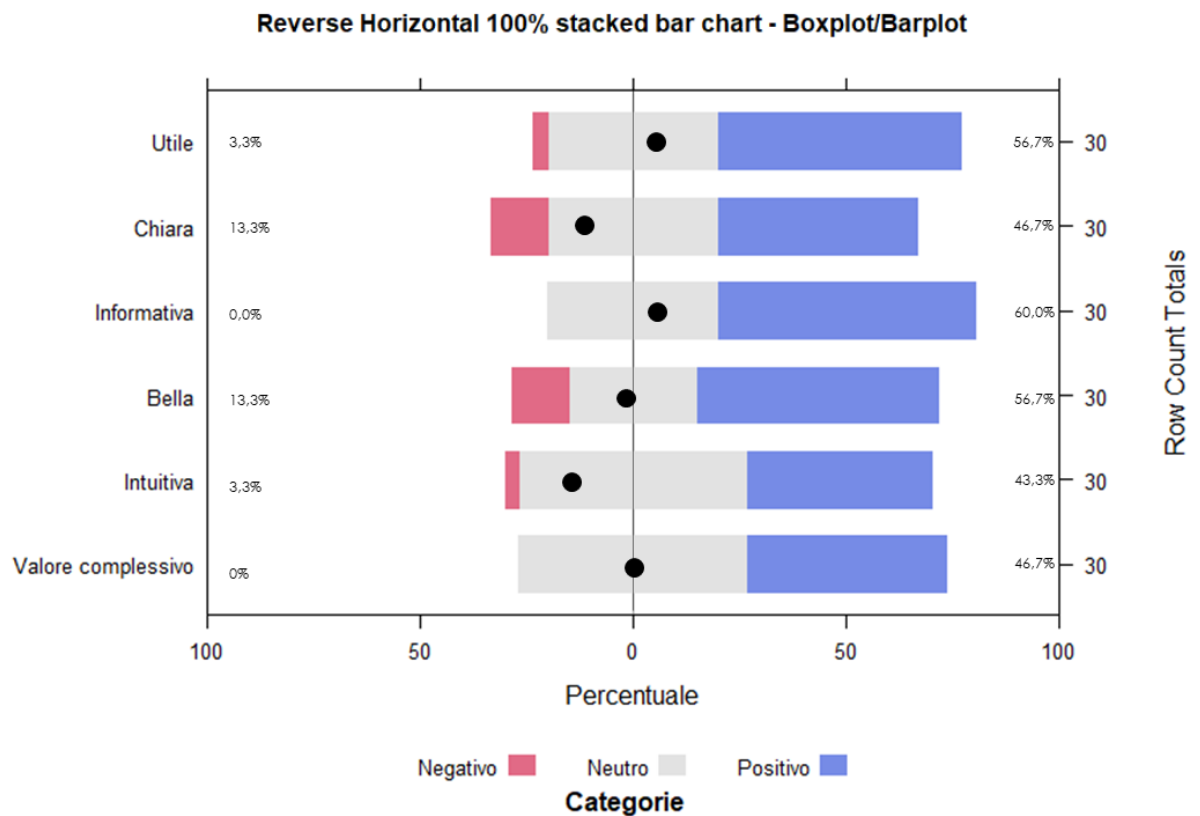


Figura 3.18: Likert chart - Boxplot-Barplot

Le possibili visualizzazioni per mostrare l'efficacia ed efficienza sono due:

- L'**efficacia** è rappresentabile attraverso l'Horizontal 100% stacked bar chart, con l'aggiunta dell'intervallo di confidenza al 5%, considerato appunto il livello di *error rate* accettabile;
- L'**efficienza**, invece, può essere rappresentata tramite il Violin plot con *stripplot e jittered data*, dove sull'asse delle ordinate vi si trova il tempo di esecuzione ed è presente la fascia di normalità insieme alla disposizione delle singole osservazioni.

Il numero di utenti raggiunti è stato di 12, ovvero 4 per ogni componente del gruppo. La scelta dell'esposizione dei risultati si espone tramite le due visualizzazioni di qualità appena citate, ovvero l'Horizontal 100% stacked bar chart e il *violin plot* con l'aggiunta ridondata delle risposte corrette e incorrette.

I task sottoposti agli utenti sono stati uno per ogni visualizzazione, ovvero:

1. Task 1 - Heatmap: *In che giorno e anno c'è stato il picco massimo del numero di tweets?* (Figura 3.20)
2. Task 2 - Lollipop: *In che anno la categoria leadership ha avuto il maggior numero di donne?* (Figura 3.21)
3. Task 3 - Alluvial: *Delle donne rientranti nella categoria leadership e asiatiche, quante sono popolari e quante non popolari nel 2019?* (Figura 3.22)
4. Task 4 - Scatterplot: *Nel 2019 qual è lo Stato con più donne nominate?* (Figura 3.23)
5. Task 5 - Boxplot/Barplot: *Nell'anno 2015 e in Europa qual è la percentuale o il numero di donne popolari?* (Figura 3.24)

Di tali domande, i primi quattro task hanno avuto un'efficacia del 100%, come visibile tanto dall'Horizontal 100% stacked bar chart (Figura 3.19) che dai seguenti violin plot, in cui sono presenti solo

risposte corrette. Il task finale relativo alla visualizzazione Boxplot/Barplot, invece, ha raccolto 7 risposte incorrette su un totale di 12, con un tasso d'errore pari all'58,3%.

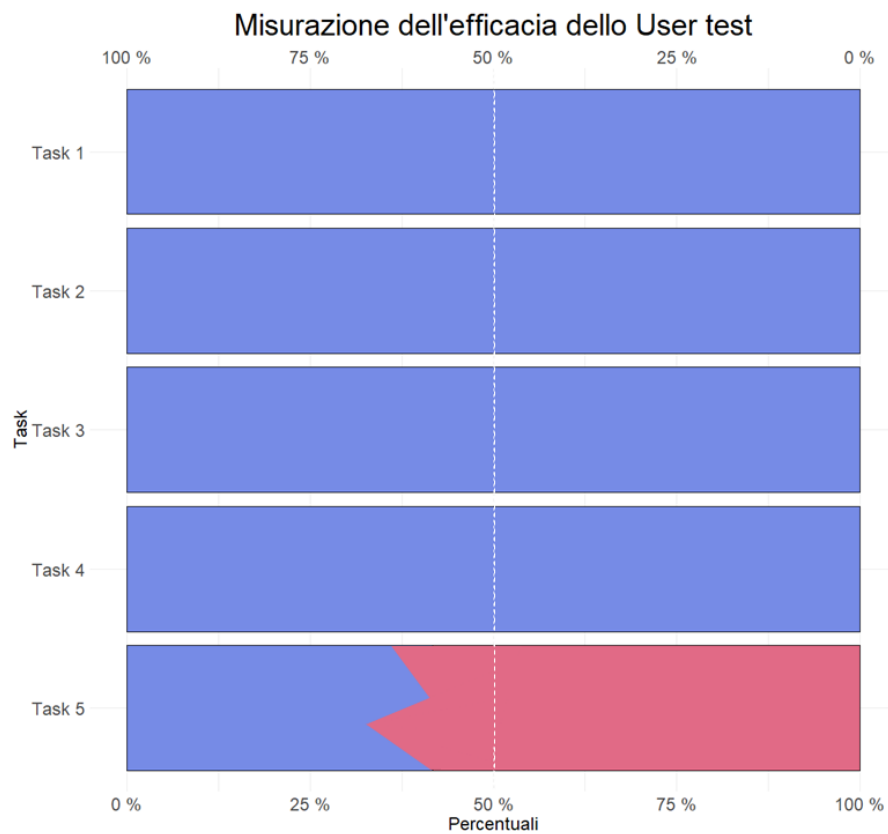


Figura 3.19: User test - Efficacia

Di seguito si mostrano i violin plot dei singoli task.

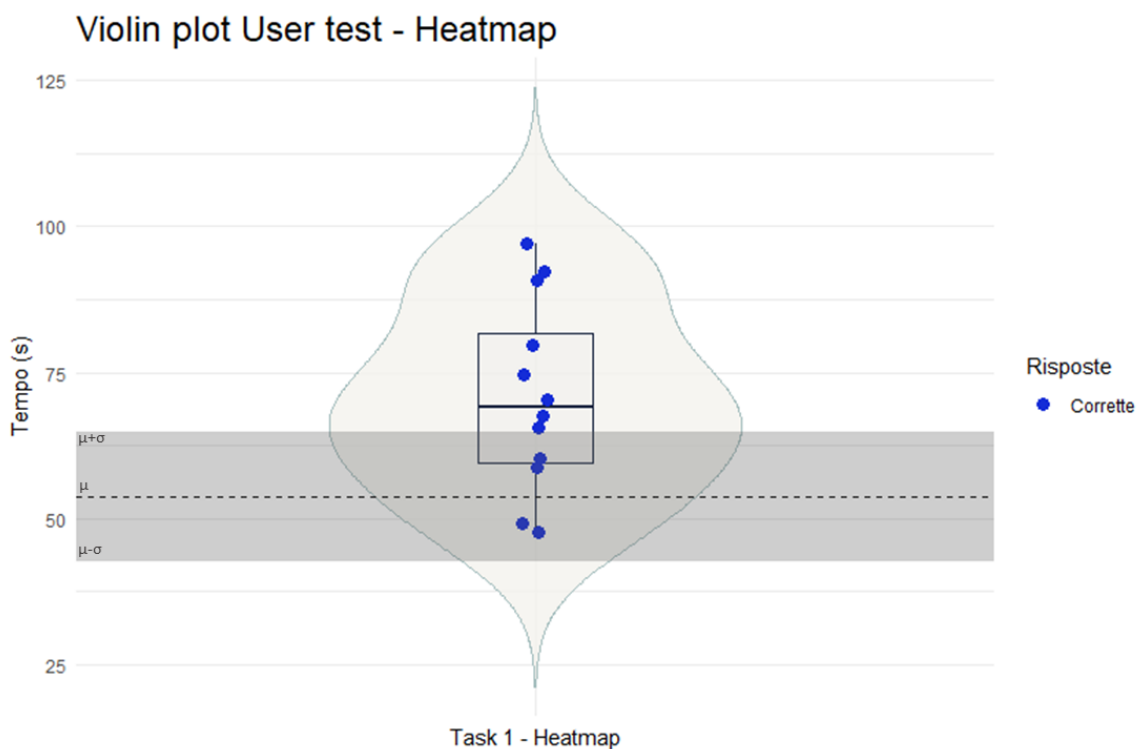


Figura 3.20: Violin plot User test - Heatmap

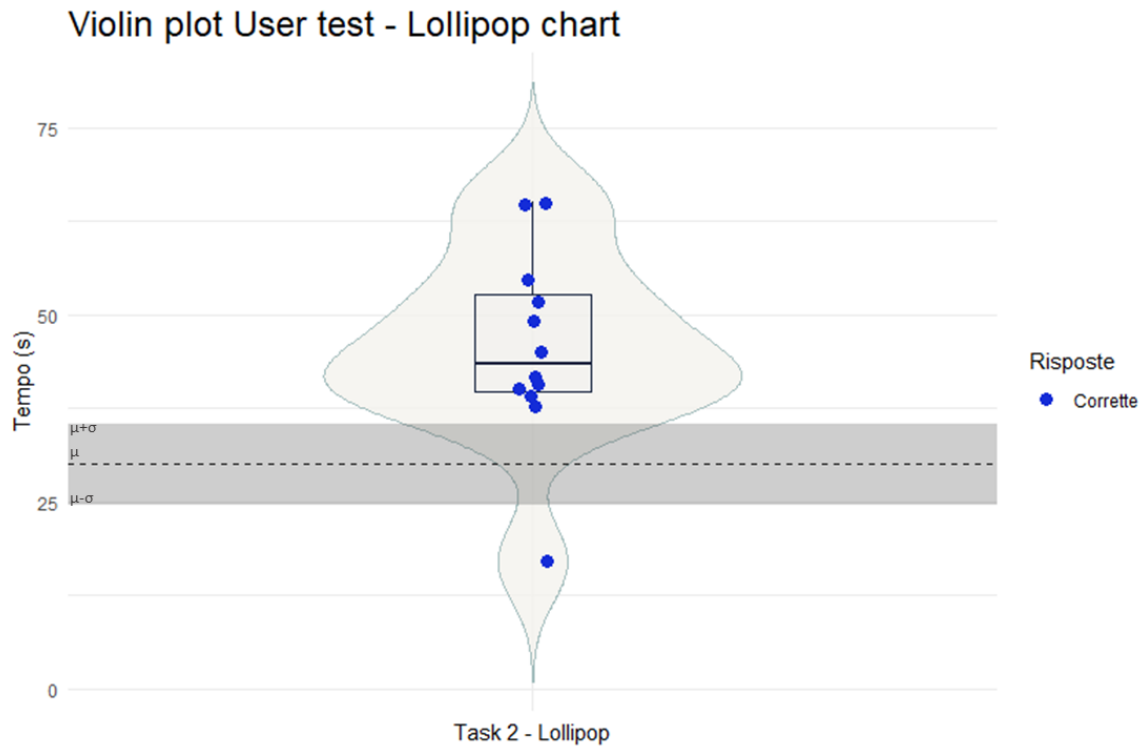


Figura 3.21: Violin plot User test - Lollipop

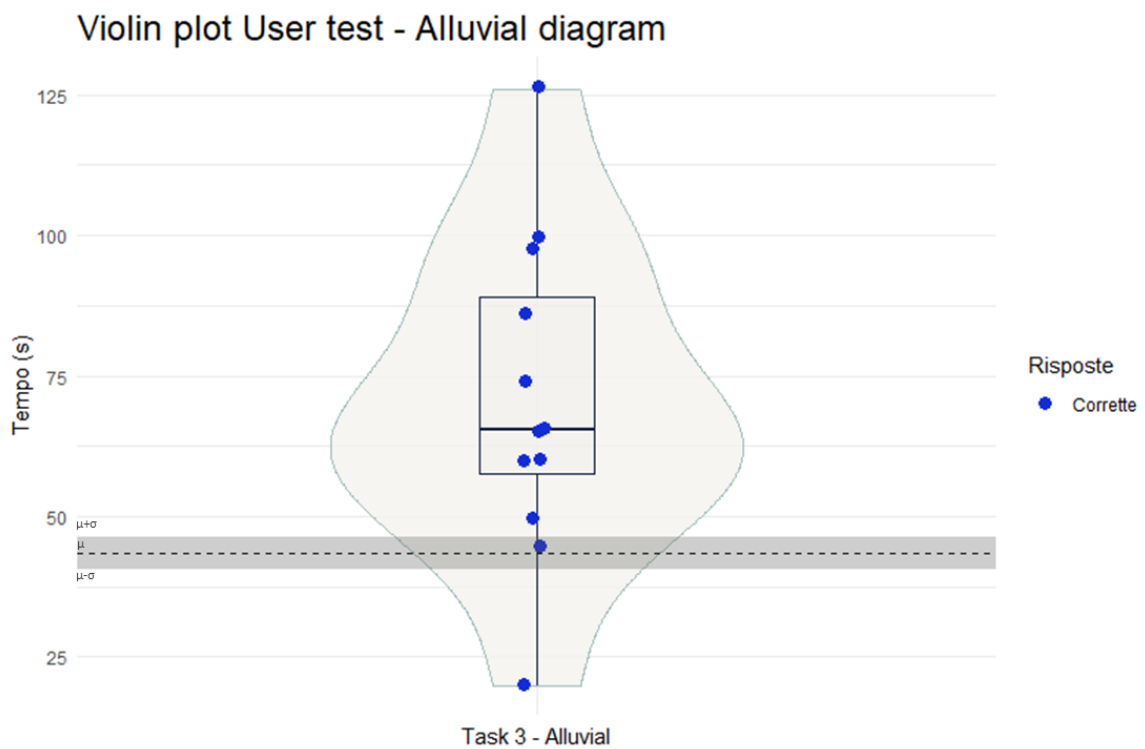


Figura 3.22: Violin plot User test - Alluvial diagram

In generale, si rileva come, eccetto nel caso della prima visualizzazione (Figura 3.20), i rispondenti abbiano impiegato più tempo del *normale* per fornire una risposta. Ciò potrebbe dovuto alla non familiarità con il tipo di visualizzazioni proposte, a differenza della prima nella quale è immediatamente intuibile una forma a calendario.

Infine, come si può notare proprio dal plot finale e unicamente in questo (Figura 3.24), sono presenti alcune risposte negative. E' possibile che l'errore sia dovuto, presumibilmente, ad una eccessiva fretta



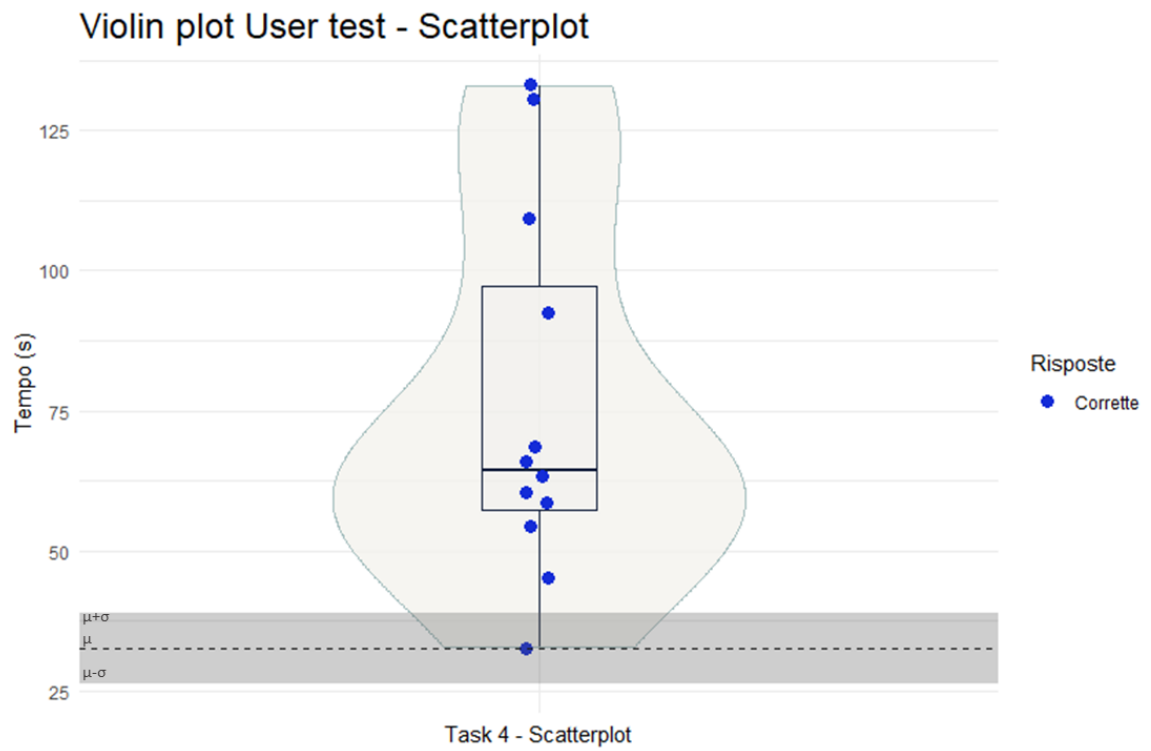


Figura 3.23: Violin plot User test - Scatterplot

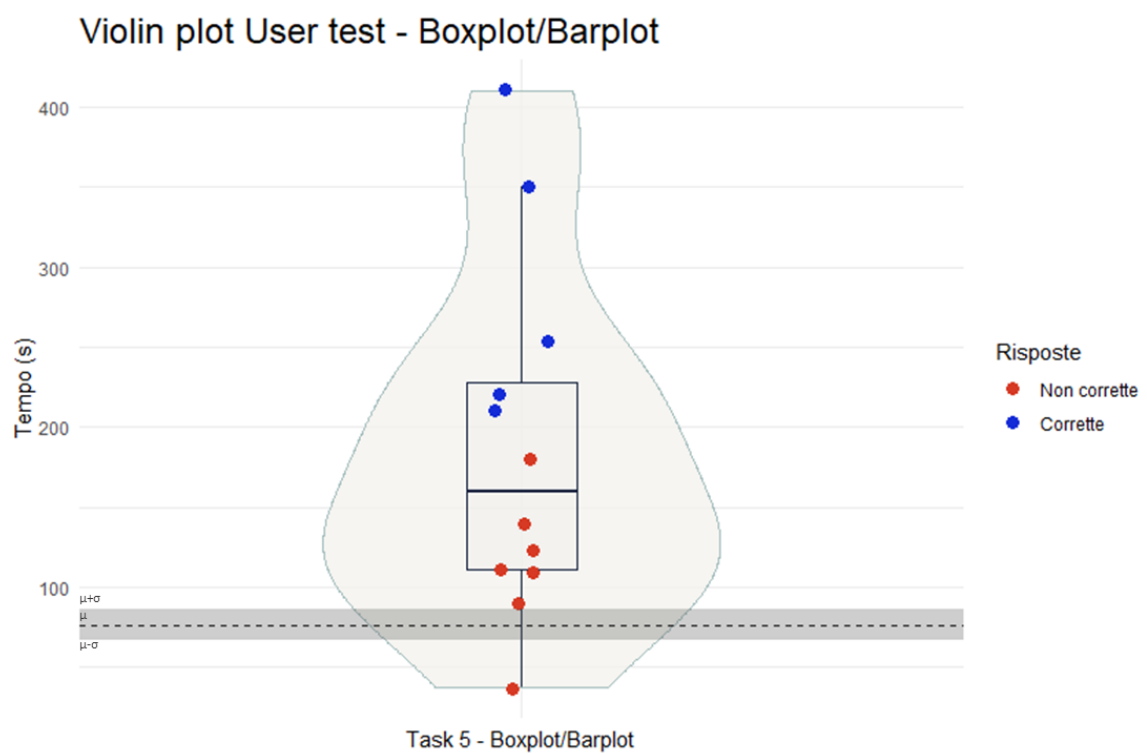


Figura 3.24: Violin plot User test - Boxplot-Barplot

nel rispondere, a differenza invece di chi ha fornito un responso corretto.

## CONCLUSIONI E SVILUPPI FUTURI

I risultati ottenuti hanno evidenziato come, dal momento dell'uscita della lista e per le successive tre settimane in cui sono stati organizzati documentari, trasmissioni e dibattiti online sul tema, il numero di persone che ne ha discusso tramite Twitter ha avuto un notevole balzo in avanti, rispetto ai precedenti periodi dell'anno. Per aumentare l'interesse nei primi mesi dell'anno potrebbe essere utile disporre eventi inerenti al tema. Inoltre, sembra che in alcuni anni ci sia maggior coinvolgimento su Twitter, probabilmente dovuto alla presenza in lista di donne molto conosciute a livello globale.

Rispetto al 2015, in cui vi era una forte preponderanza verso la categoria leadership, nell'ultimo anno analizzato le donne premiate si distribuiscono in modo equo tra le varie categorie presenti. Inoltre, è stata inserita la categoria ambiente, simbolo dell'estrema importanza che oggi hanno temi come la sostenibilità e la salvaguardia ambientale.

Considerando la mediana dell'età delle donne per ciascuna categoria risulta che in tutte le categorie l'età mediana sia di circa 30 anni. È interessante notare che la categoria identità ha età mediana più bassa rispetto alle altre categorie nel 2015, mentre la conoscenza ha un'età mediana notevolmente più alta in confronto alle restanti categorie nel 2019.

La maggior parte delle donne premiate ha origine asiatica, ma non mancano africane, europee, nord e sud americane. Nel complesso però, risulta che ad essere nominate sono poche donne per Stati mediamente ricchi e mediamente sviluppati, mentre ve ne sono pochissime nominate provenienti da Paesi poveri e sottosviluppati. Al contrario circa il 30% delle donne nominate che, sia nel 2015 sia nel 2019, provengono da Stati molto ricchi e molto sviluppati (soprattutto Europa e Stati Uniti).

Per entrambi gli anni presi in considerazione, la percentuale di donne non popolari è molto significativa: nel 2015 ben l'87% delle donne era sconosciuta ai più, mentre nel 2019 la percentuale è scesa all'80%. In generale, in termini di differenza del numero di tweets riferiti ad una specifica donna tra prima e dopo l'uscita della lista, le donne conosciute hanno presentato una differenza nel numero di tweets in termini assoluti più alta, mentre la maggior parte delle donne non popolari, hanno avuto un incremento del numero di tweets che le riguardano nel periodo successivo alla pubblicazione della lista. Nel 2019 risultano esserci più donne popolari con una differenza negativa rispetto al 2015. Si potrebbe concludere che l'evento abbia un impatto positivo su donne meno conosciute che ottengono un boost di popolarità grazie alla nomina, mentre quelle con una certa notorietà non ne giovano altrettanto.

Durante lo svolgimento del lavoro sono emersi limiti e potenzialità. I primi riguardando le ricerche manuali svolte su Twitter, molto dispendiose in termini di tempo e non scalabili, e lo scaricamento stesso dei tweets, che vista la mole dei tweets ha necessitato la suddivisione tra i componenti del gruppo delle parti da scaricare. Dei risultanti tweets si è conservato solo un conteggio, ma è sorto spontaneo pensare che in futuro l'analisi si potrebbe estendere alla valutazione del sentiment dei tweets stessi per comprenderne la natura positiva o meno nei confronti dell'evento e delle donne nominate. Per quanto riguarda le ricerche dei tweets si potrebbe considerare una selezione più accurata degli hashtags, eventualmente automatizzandone il processo. Altro aspetto che potrebbe essere analizzato più nel dettaglio riguarda la scelta e l'analisi degli indicatori degli Stati. La scelta è stata soggettiva ma si potrebbe valutare di inserire più indici e tra i più variegati in modo tale da avere una visione ancora più ampia del fenomeno. Un aspetto non trattato all'interno del progetto ma che potrebbe essere rilevante per sviluppi futuri, riguarda le categorie lavorative delle singole donne. La mancanza è spiegata dalla necessità di dover proporre una divisione delle attività lavorative, spesso, di difficile attuazione per rendere al meglio il dato reale in maniera oggettiva. Ultima ma non meno importante prospettiva di lavoro riguarda la riproposizione delle analisi svolte ogni anno, così da avere una storicità nello studio della lista della BBC.

# Bibliografia

- [1] Country Code <https://www.iban.com/country-codes>
- [2] BBC 100 Women <https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2020/2020-12-08/women.csv>
- [3] GDP Country <https://www.imf.org/external/datamapper/NGDPD@WEO/OEMDC/ADVEC/WEOORLD>
- [4] GDP per capita Country <https://www.imf.org/external/datamapper/NGDPDPC@WEO/OEMDC/ADVEC/WEOORLD>
- [5] GDP per capita Country <https://databank.worldbank.org/source/gender-statistics/preview/on>
- [6] Gender Global Gap 2020 <https://reports.weforum.org/global-gender-gap-report-2015/rankings/>
- [7] Gender Global Gap 2020 <https://reports.weforum.org/global-gender-gap-report-2020/the-global-gender-gap-index-2020-rankings/>
- [8] Human Development Index <http://hdr.undp.org/en/indicators/137506>
- [9] Cody Zacharias, *Twint Project*, 2.1.20, 2014. <https://github.com/twintproject/twint.git>
- [10] The apache software foundation, *Apache Kafka*, 2021. <https://kafka.apache.org/powered-by>
- [11] Nifi Team, *What is apache nifi?*, 2015. <https://nifi.apache.org/docs.html>
- [12] Tableau Software, *What is Tableau?*, 2003-2021. <https://www.tableau.com/why-tableau/what-is-tableau>
- [13] Python Software Foundation, Python Language Reference, version 3.9. *Python 3.9*. <http://www.python.org>
- [14] Microsoft Azure, *Microsoft Azure*. <https://azure.microsoft.com/it-it/services/virtual-machines/>
- [15] MongoDB, *MongoDB*. <https://www.mongodb.com/>
- [16] Mongo Express, *Mongo Express*. <https://github.com/mongo-express/mongo-express>