



Università Politecnica delle Marche

Dipartimento di Ingegneria dell'Informazione

Corso di Laurea Magistrale in Ingegneria Informatica e dell'Automazione

Analisi dei video in tendenza su YouTube con Qlik, Tableau e Power BI

Corso di Data Science

Professore

Prof. Domenico Ursino

Gruppo 1

Michele Pasqualini
Denil Nicolosi
Eris Prifti

Anno accademico 2021-2022

Indice

| | | |
|----------|---|-----------|
| 1 | Introduzione | 3 |
| 1.1 | Dataset | 3 |
| 1.2 | Obiettivi dell'analisi | 4 |
| 1.3 | ETL | 5 |
| 2 | Qlik Sense | 5 |
| 2.1 | Caricamento dati e ETL | 5 |
| 2.2 | Analisi e descrizione delle visualizzazioni | 6 |
| 2.2.1 | Video per mese ed anno di pubblicazione e tendenza | 6 |
| 2.2.2 | Mappatura video mensili pubblicati e in tendenza dal 2017 al 2018 | 7 |
| 2.2.3 | Top 10 video con più commenti, likes e visualizzazioni | 8 |
| 2.2.4 | Distribuzione del tempo impiegato per andare in tendenza | 11 |
| 2.2.5 | Video che rimangono più in tendenza | 12 |
| 2.2.6 | Canali che sono stati più in tendenza | 13 |
| 2.2.7 | Analisi delle categorie in tendenza | 14 |
| 2.2.8 | Analisi geografica delle categorie e video in tendenza | 15 |
| 2.2.9 | Analisi temporale della pubblicazione dei video in tendenza | 15 |
| 3 | Tableau | 17 |
| 3.1 | Caricamento dati e ETL | 17 |
| 3.2 | Analisi e descrizione delle visualizzazioni | 18 |
| 3.2.1 | Video per anno di pubblicazione | 19 |
| 3.2.2 | Video con commenti e rating disabilitati | 20 |
| 3.2.3 | Top 10 dei video con maggior numero di likes, views e commenti | 20 |
| 3.2.4 | Top 5 categorie con maggior numero di video per paese | 21 |
| 3.2.5 | Top 5 categorie con maggior numero visualizzazioni per paese | 23 |
| 3.2.6 | Confronto categorie in base a likes, dislikes, views e commenti | 23 |
| 3.2.7 | Numero di video in tendenza e numero di canali per paese | 24 |
| 3.2.8 | Distribuzione dei likes, dislikes, commenti e views per paese | 25 |
| 3.2.9 | Media dei giorni in tendenza dei video per paese, categoria e canale | 30 |
| 3.2.10 | Distribuzione delle categorie in tendenza | 30 |
| 3.2.11 | Andamento delle categorie in funzione del mese di pubblicazione e di tendenza | 31 |
| 3.2.12 | Correlazioni | 32 |
| 3.2.13 | Previsioni delle misure (commenti, views, like e numero di video) | 33 |
| 3.2.14 | Numero di video in tendenza per giorni con previsione | 34 |
| 3.2.15 | Previsione delle misure per un video in tendenza | 35 |
| 4 | PowerBI | 40 |
| 4.1 | Caricamento dati e ETL | 40 |
| 4.2 | Analisi e descrizione delle visualizzazioni | 40 |
| 4.2.1 | Vista generale | 41 |
| 4.2.2 | Analisi su categorie | 41 |
| 4.2.3 | Analisi su canali | 42 |
| 4.2.4 | Numero di video in tendenza per paese | 43 |
| 4.2.5 | Visualizzazioni per paese e categoria | 45 |
| 4.2.6 | Analisi delle parole più utilizzate | 45 |
| 4.2.7 | Analisi ad albero delle views | 46 |
| 4.2.8 | Fattori di influenza sulle views | 47 |
| 4.2.9 | Media dei like per country nel tempo | 51 |
| 4.2.10 | Analisi delle misure per paese | 52 |

| | |
|---|-----------|
| 4.2.11 Correlazioni tra le misure | 53 |
| 4.2.12 Previsione del numero di video in tendenza | 54 |
| 5 Conclusioni | 57 |

1 Introduzione

1.1 Dataset

Per lo svolgimento del progetto di analisi è stato scelto il dataset open source reperibile al seguente link: www.kaggle.com/datasneak/youtube-new. Il dataset contiene un elenco dei video andati in tendenza sulla piattaforma YouTube nel periodo compreso tra Novembre 2017 e Giugno 2018, ed è stato raccolto utilizzando l'API di YouTube con il seguente codice sorgente github.com/mitchelljy/Trending-YouTube-Scraper. I vari file contengono informazioni sui video delle seguenti nazioni:

- Stati uniti
- Germania
- Messico
- Canada
- Regno Unito
- India
- Francia
- Giappone
- Korea
- Russia

Il dataset è costituito da file csv (uno per ogni paese) dove all'interno si trovano le informazioni relative ai vari video e un file in formato JSON in cui sono riportate le associazioni per le categorie dei video. I file csv sono formattati con i campi descritti nella seguente tabella [1].

| Nome campo | Descrizione |
|------------------------|---|
| video_id | ID alfanumerico del video |
| channel_title | Canale che ha pubblicato il video |
| tags | Tag associati ai video |
| likes | Numero di "mi piace" |
| dislikes | Numero di "non mi piace" |
| trending_date | Data in cui il video è stato in tendenza |
| title | Titolo del video |
| category_id | ID della categoria del video |
| publish_time | Data di pubblicazione del video |
| views | Numero di visualizzazioni |
| comments_disabled | Disabilitazione dei commenti del video |
| ratings_disabled | Disabilitazione dei like del video |
| video_error_or_removed | Se il video presenta errori o è stato rimosso |
| description | Descrizione del video |
| comment_count | Numero di commenti al video |
| thumbnail_link | Link alla copertina del video |

Tabella 1: Descrizione campi file video

Invece, per quanto riguarda i file JSON sulle categorie dei video, essi sono formattati come descritto nella seguente tabella [2].

| Nome campo | Descrizione |
|---------------|---|
| id | ID numerico della categoria |
| snippet/title | Titolo della categoria |
| assignable | Indica se la categoria è assegnata da YouTube |
| channelID | Identificatore per il canale |
| kind | Identifica il tipo di risorsa API |
| etag | Versione API utilizzata |

Tabella 2: Descrizione campi file category

1.2 Obiettivi dell’analisi

L’obiettivo della analisi è studiare quali sono i fattori determinanti da considerare per far sì che un video vada in tendenza su YouTube. Come riportato dal sito guida di Google [1]:

”La sezione Tendenze aiuta gli spettatori a seguire quello che accade su YouTube e in tutto il mondo. Le tendenze hanno come obiettivo quello di mettere in evidenza i video che potrebbero rivelarsi interessanti per un gran numero di spettatori. Alcune tendenze sono prevedibili, come la nuova canzone di un artista famoso o il trailer di un nuovo film. Altre invece possono sorprendere, come i video virali. La sezione Tendenze non è personalizzata e in ciascun paese mostra lo stesso elenco di video di tendenza a tutti gli utenti. Per questo motivo, nella sezione Tendenze potresti visualizzare anche video in una lingua diversa rispetto al tuo browser. In India, invece, le tendenze includono un elenco di video di tendenza per ciascuna delle nove lingue più parlate nel paese. L’elenco dei video di tendenza viene aggiornato ogni 15 minuti circa. A ogni aggiornamento, i video possono salire o scendere di posizione oppure mantenere la posizione originale.”

Sempre secondo la guida di Google [1], l’inserimento di un video tra le tendenze di Youtube prende in considerazione numerosi segnali, tra cui:

- numero di visualizzazioni;
- quanto velocemente un video ottiene visualizzazioni;
- provenienza delle visualizzazioni;
- data del video;
- rendimento del video rispetto ad altri caricamenti recenti dello stesso canale

Così facendo, i trend tendono a far emergere video che:

- sono accattivanti per una vasta gamma di spettatori;
- non sono fuorvianti; sensazionalistici o clickbait
- riflettono ciò che accade su YouTube e nel mondo;
- mostrano tanti creator diversi;
- idealmente, sorprendono o sono insoliti.

Dato il dataset a disposizione, non è possibile eseguire una analisi completa perché non sono disponibili i dati riguardo la provenienza delle visualizzazioni e come si comporta il video rispetto ad altri caricamenti recenti dello stesso canale. Allo stesso modo, non è possibile eseguire un confronto tra le misure dei video in tendenza e i video non in tendenza. E’ stato comunque possibile svolgere tipi di analisi volte ad individuare le differenze tra video che vanno una singola volta in tendenza e video che vanno più volte in tendenza.

1.3 ETL

La parte di ETL è stata svolta utilizzando il linguaggio Python con l'aiuto della libreria Pandas data la grande mole di dati a disposizione. Nei file che erano a disposizione sono state fatte le seguenti modifiche:

1. aggiunta della colonna *"Country"* in ogni file per poter interconnettere i vari dataset senza perdere l'informazione sul paese in cui è andato in tendenza il video;
2. cambiare il formato della codifica dei file csv in UTF-8;
3. nella tabella category sono stati lasciati solo gli attributi utili agli scopi della analisi, cioè: *"Id"*, *"snippet/title"* e *"assignable"*.
4. il formato della data del campo *"trending date"* è stato convertito da MM.DD.YYYY ad DD/MM/YYYY perché il formato iniziale non veniva riconosciuto da Qlik.

2 Qlik Sense

Qlik Sense è uno strumento di Business Intelligence tramite cui effettuare analisi, controllo e valutazione dei dati. Consente di creare visualizzazioni flessibili e interattive per prendere decisioni significative. Qlik Sense utilizza il cloud computing, permettendo di lavorare su qualsiasi dispositivo ed elaborando velocemente i dati. Una caratteristica molto interessante è la possibilità di cambiare dinamicamente le visualizzazioni delle dashboard sulla base di filtri e delle selezioni effettuati dall'utente.



Sense[®]

Figura 1: Logo Qlik Sense

2.1 Caricamento dati e ETL

Tra i membri del gruppo è stata creata un applicazione condivisa in modo da facilitare e velocizzare il lavoro svolto. Dopo la fase di ETL spiegata in precedenza, sono stati caricati tutti i file CSV sull'applicazione condivisa di Qlik. In seguito, sono stati concatenati tutti i file CSV relativi ai video in un'unica tabella e tutti i file CSV delle categorie in un'altra tabella, dando luogo di fatto ad una relazione come mostrato in figura.

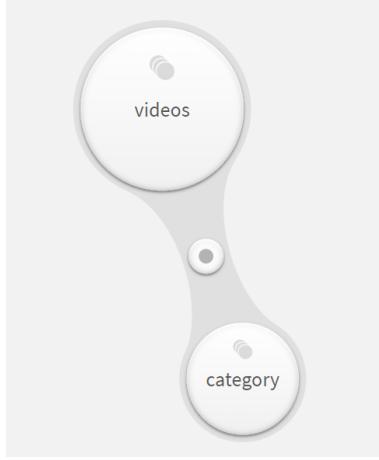


Figura 2: Associazione tra le due tabelle

"Videos" contiene tutte le tabelle relative ai video concatenate per ogni paese, mentre *"category"* concatena tutte la tabelle relative alle categorie di ogni paese. La connessione tra le due tabelle è stata effettuata usando l'attributo *"category_id"* di *videos* e l'attributo *"id"* di *category*. Questa relazione è stata suggerita da Qlik, in quanto è stata riconosciuta in automatico.

2.2 Analisi e descrizione delle visualizzazioni

Per la realizzazione delle viste, si è deciso di effettuare una suddivisione in diverse dashboard di seguito descritta:

- una vista che descrive il numero di video in base all'anno di pubblicazione e all'anno di tendenza;
- una vista che mostra una mappatura dei video mensili pubblicati e in tendenza dal 2017 al 2018;
- una vista che contiene alcuni KPI come visualizzazioni, likes e commenti, mostrando le top 10 per i rispettivi indicatori;
- una vista sulla distribuzione temporale dei video impiegata per andare in tendenza;
- una vista che mostra quali video rimangono più tempo in tendenza;
- una vista sui canali di maggiore tendenza;
- una vista sulle categorie di maggiore tendenza;
- una vista che contiene un analisi geografica delle categorie e dei video in tendenza;
- una vista sull'analisi temporale della pubblicazione dei video;
- una vista dei video che presentano errori il quale sono stati rimossi.

2.2.1 Video per mese ed anno di pubblicazione e tendenza

In questa dashboard sono riportate le informazioni temporali relative all'anno di pubblicazione dei video e l'anno in cui sono stati in tendenza. L'intero dataset contiene un totale di 186 289 video diversi. Essi sono stati selezionati in base al loro titolo attraverso il comando *count(distinct title)*, perché nell'intero dataset molti video vengono ripetuti, differenziandosi solo per alcuni attributi caratteristici.



Figura 3: Vista generale sui video pubblicati e in tendenza

In dettaglio, possiamo vedere i mesi di pubblicazione e tendenza, notando che come dichiarato sul sito del dataset[2], sono presenti i video andati in tendenza tra novembre 2017 e giugno 2018. Per le date di pubblicazione, si nota che la maggior parte dei video sono stati pubblicati nel medesimo periodo, a eccezione di pochi video la quale sono stati pubblicati nei periodi precedenti. Come verrà approfondito nelle analisi successive, si può subito intuire che solitamente i video vanno in tendenza in prossimità della pubblicazione, e raramente capita che un video molto più vecchio venga riscoperto e vada in tendenza.

2.2.2 Mappatura video mensili pubblicati e in tendenza dal 2017 al 2018

In questa dashboard, attraverso l'utilizzo di una griglia, è stata riportata una mappatura del numero di video pubblicati lungo le dimensioni *"Anno di pubblicazione"* e *"Mese di pubblicazione"*. Come facilmente intuibile, nel corso degli anni 2017 e 2018 troviamo una maggiore concentrazione dei video pubblicati. In particolare, per l'anno 2017, si può osservare un'elevata densità di video pubblicati nei mesi di novembre e dicembre, arrivando a picchi rispettivamente di 16 960 e 26 630. Invece, la mappatura lungo l'anno 2018 è più uniforme, in quanto dal mese di gennaio al mese di maggio, il numero di video pubblicati oscilla tra 26 000 e 29 000, toccando quota 10 000 nel mese di giugno. Nel grafico sottostante, in alto a destra, viene riportata la stessa mappatura ma in funzione del periodo di tendenza. Si può osservare che lungo i due anni di interesse, la distribuzione è simile alla precedente, questo conferma il fatto che molti video sono andati in tendenza subito dopo la loro pubblicazione. Infine, viene raffigurato un grafico a barre che cerca di riassumere graficamente quanto mostrato dalle viste precedenti, sottolineando il numero di video andati in tendenza nei mesi del 2017 (colore blu) e il numero di video andati in tendenza nei mesi del 2018 (colore viola).

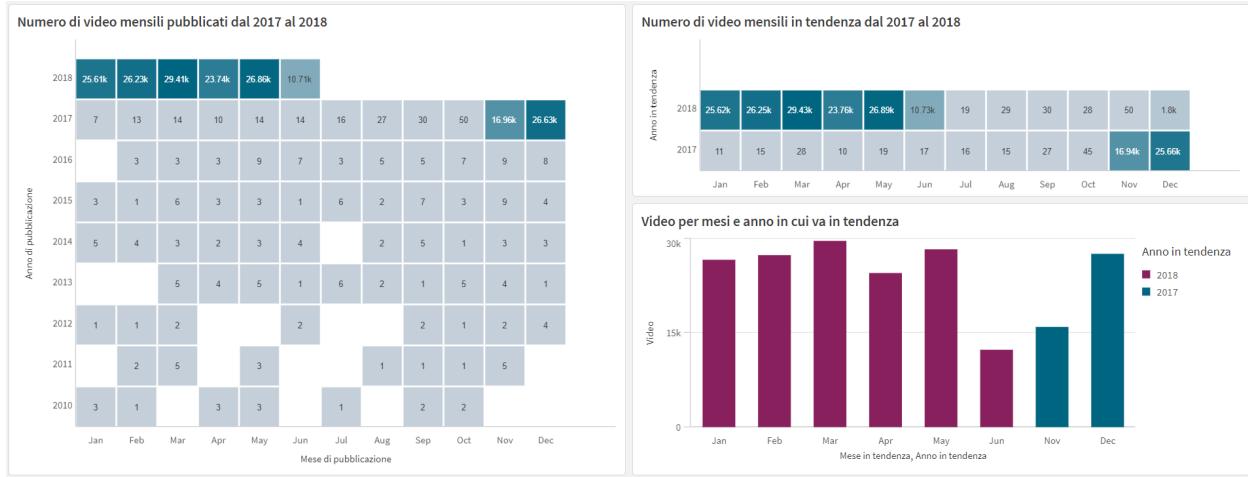


Figura 4: Vista sui video pubblicati e in tendenza mensilmente

2.2.3 Top 10 video con più commenti, likes e visualizzazioni

Dopo aver riportato il numero totale dei video, la vista mostra i tre KPI fondamentali per questo tipo di analisi: visualizzazioni medie, media dei likes e media dei commenti di ogni video. In base a questi indicatori, si può stabilire una top 10. Dall'analisi risulta che il maggior numero di visualizzazioni ottenute da un video è pari a 424.24 M, mentre il maggior numero di likes ammonta a 5.61 M e il maggior numero di commenti a 1.63 M. Infine, in basso a sinistra vengono riportati due grafici a torta per mostrare i commenti e i likes disabilitati su ogni video. E' interessante osservare che, rispetto all'intero dataset, solo il 2,5 % dei video ha commenti e likes disabilitati.

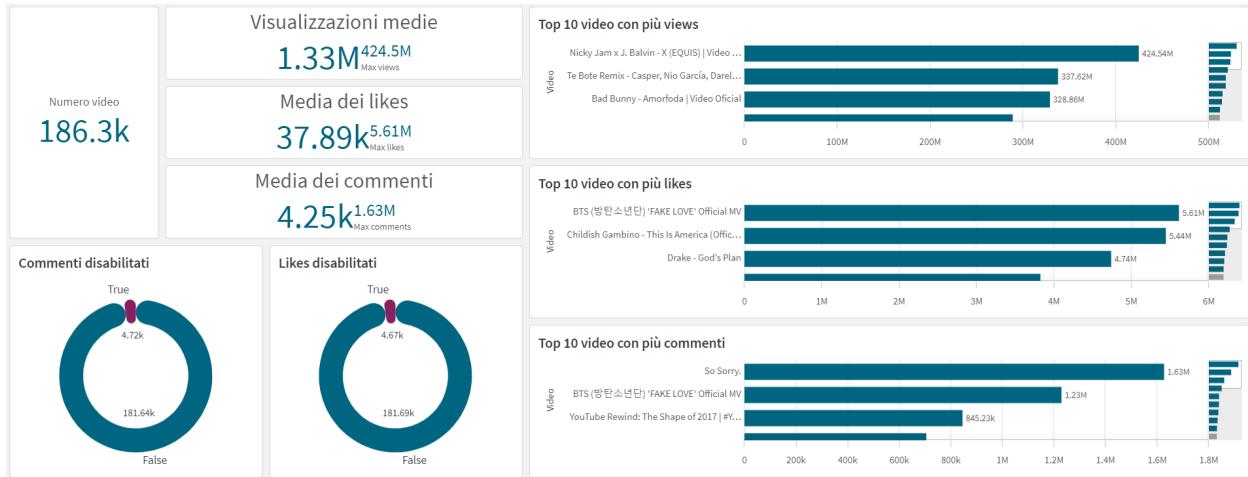


Figura 5: Vista sui video in Top 10

Di seguito, andiamo ad osservare nel dettaglio ogni singolo grafico. Il primo grafico a barre mostra i primi 10 video che hanno ricevuto più commenti. Si può osservare che il primo ha un numero di commenti di gran lunga maggiore degli altri. Analizzando e filtrando per le categorie a cui questi video appartengono, troviamo in testa quella di *"Entertainment"*, seguita da *"Music"* e da *"Nonprofits & Activism"*.

Top 10 video con più commenti

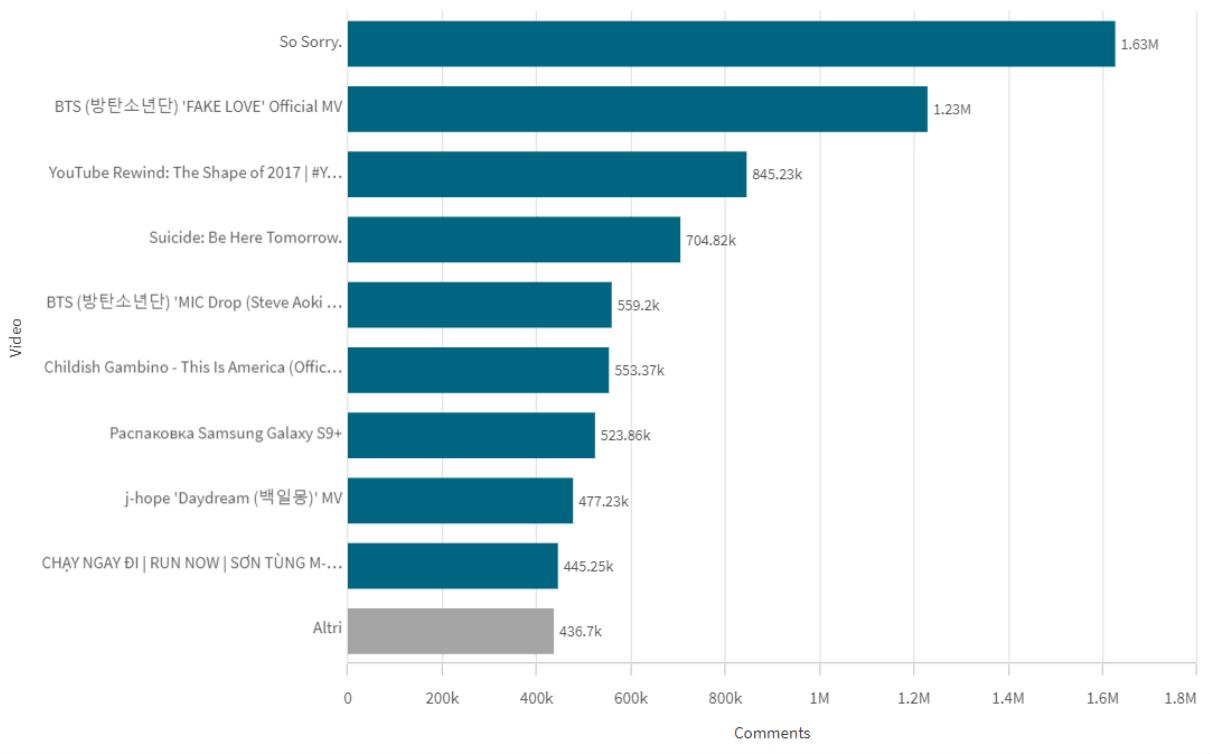


Figura 6: Top 10 video per commenti

Utilizzando lo stesso approccio, si può individuare che nella top 10 dei video con più likes, sono le categorie "Music" ed "Entertainment" a prevalere su tutte le altre.

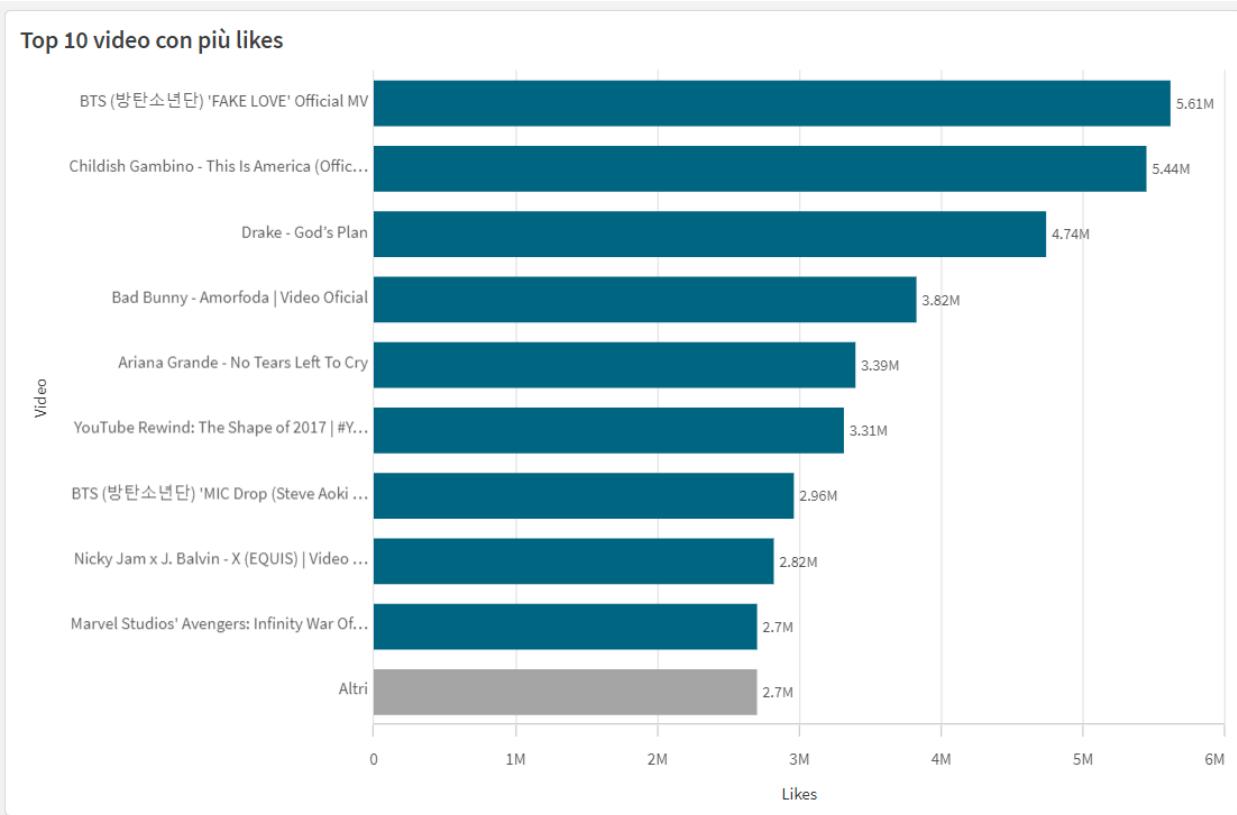


Figura 7: Top 10 video per likes

Analizzando il grafico a barre relativo alle visualizzazioni, la top 10 è predominata da video che appartengono alla categoria "Music", totalizzando un numero di visualizzazioni molto più grande rispetto alle altre categorie.

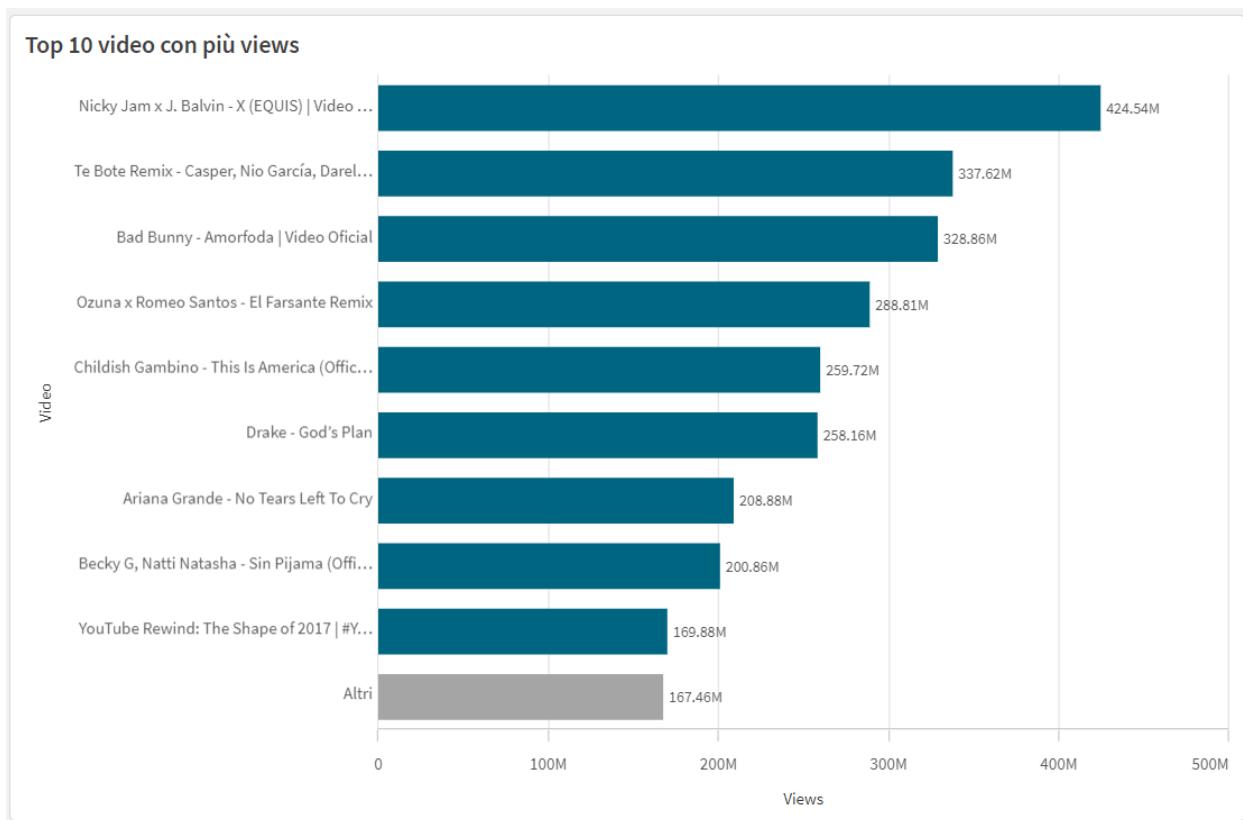


Figura 8: Top 10 video per views

2.2.4 Distribuzione del tempo impiegato per andare in tendenza

In questa dashboard si è analizzato il tempo impiegato dai video, a seguito della loro pubblicazione, per andare in tendenza. In questo grafico riportato in figura abbiamo nell'asse delle ordinate il numero di giorni impiegati per andare in tendenza e nell'asse delle ascisse la data di pubblicazione. Come riportato nella legenda, la densità del colore delle barre ci indica invece il numero di video che si collocano in quella posizione specifica. Dal grafico si può subito comprendere che la distribuzione di video aumenta nei giorni subito maggiori di zero e dopo pochi giorni diminuisce drasticamente la densità.

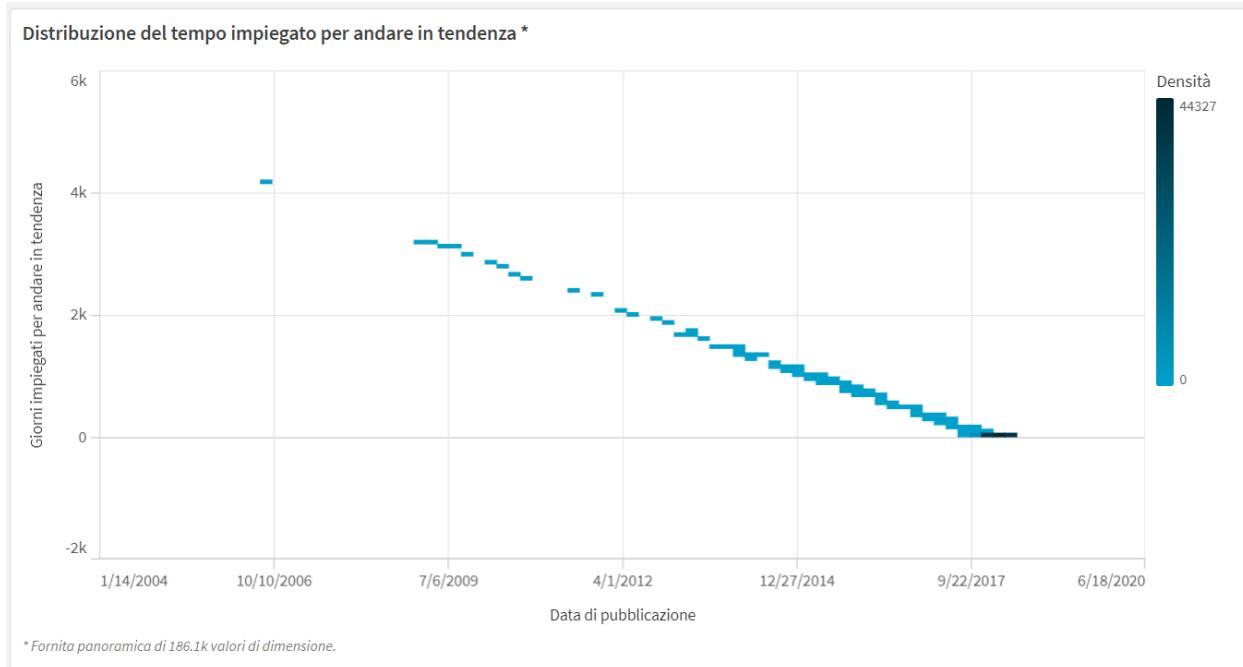


Figura 9: Distribuzione temporale impiegata dai video per andare in tendenza

2.2.5 Video che rimangono più in tendenza

Nella seguente dashboard analizziamo i video che rimangono più in tendenza, in quali mesi si concentrano e la relazione che c'è tra tempo impiegato per andare in tendenza e tempo di permanenza in tendenza.

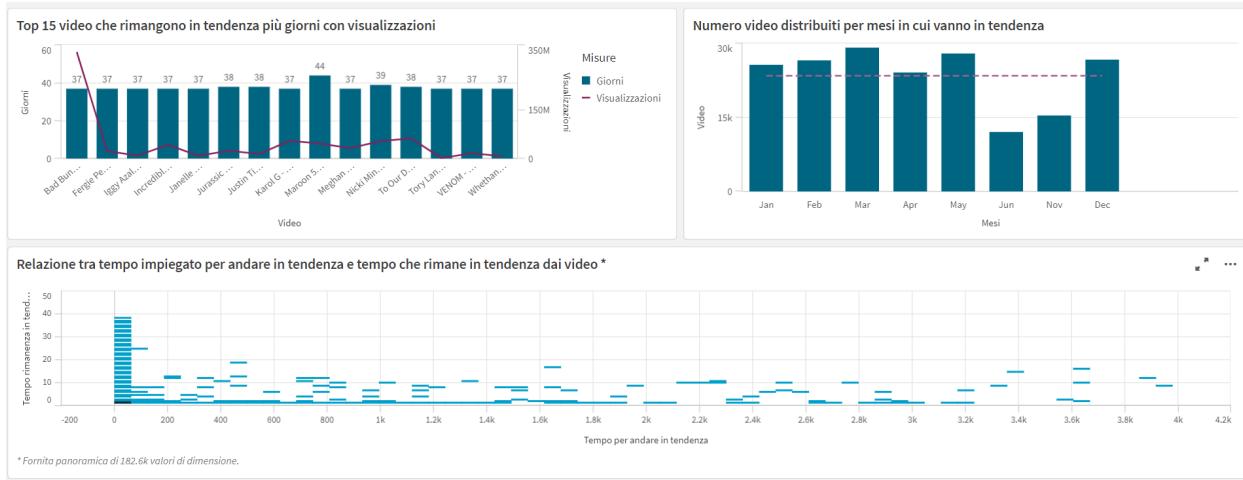


Figura 10: Vista sui video che rimangono di più in tendenza

In questo grafico in particolare abbiamo in ascissa il tempo impiegato per andare in tendenza e in ordinata il tempo di permanenza in tendenza. Da qui si nota subito che i video che rimangono per più tempo in tendenza

sono quelli che impiegano poco tempo per andarci. Di conseguenza, i video che impiegano molto tempo per andare in tendenza, molto probabilmente ci rimarranno per poco tempo.

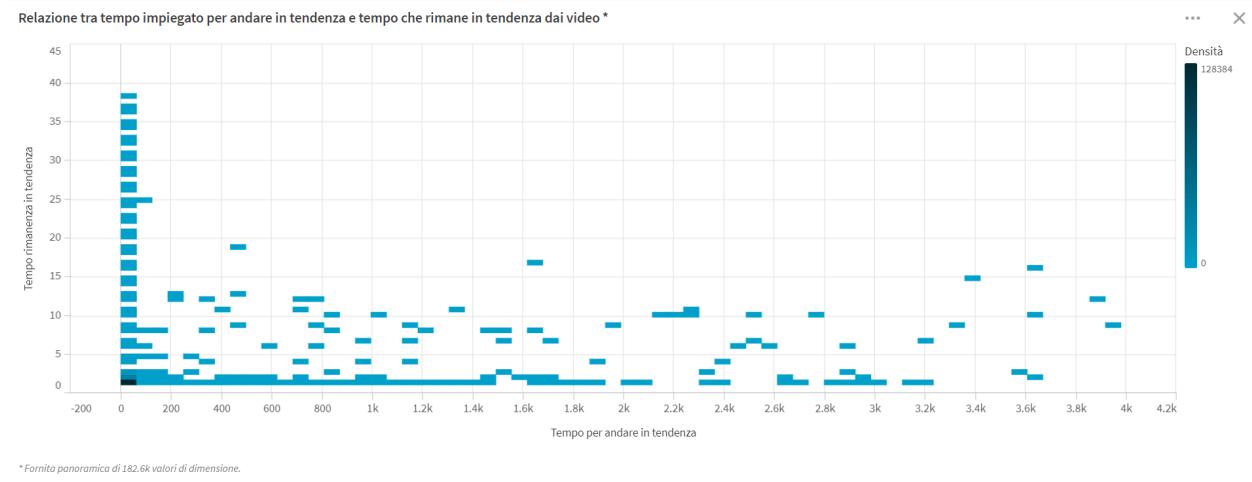


Figura 11: Confronto tra il tempo impiegato per andare in tendenza e il tempo che rimane in tendenza

Nel seguente grafico visualizziamo la top 15 video che sono rimasti in tendenza per più giorni, con le relative visualizzazioni. Il video che è rimasto più tempo in tendenza in assoluto è il video musicale "Maroon 5 - Wait" che è rimasto nei primati per ben 48 giorni. In ordine, viene seguito da "Nicki Minaj - Chun-Li" con 39 giorni di permanenza, da "Jurassic World: Fallen Kingdom - Official Trailer", "Justin Timberlake's FULL Pepsi Super Bowl LII Halftime Show!— NFL Highlights" e "To Our Daughter" con 38 giorni.

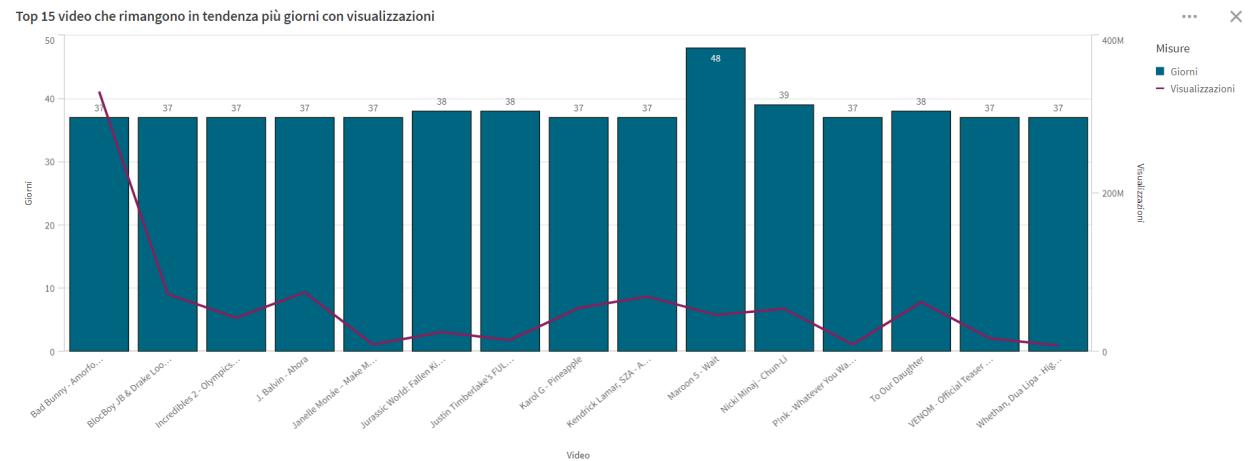


Figura 12: Top 15 video in tendenza per più giorni

2.2.6 Canali che sono stati più in tendenza

Uno degli obiettivi è stato andare ad analizzare i canali che sono andati più in tendenza, successivamente sono stati analizzati quelli che avevano più visualizzazioni e in seguito i canali con più video in tendenza.

Di seguito, nella dashboard vengono riportati i canali con le caratteristiche descritte sopra, nel periodo di tempo considerato. Dall'analisi fatta è stato osservato, per quanto riguarda il grafico dei canali che sono andati più in tendenza, che uno dei canali di maggiore tendenza è *"The Late Show with Stephen Colbert"* con 984 giorni in tendenza, a seguire si ha *"WWE"* con 804 giorni, infine anche *"Late Night with Seth Meyers"* con 773 giorni. Si è deciso, per quanto riguarda questo grafico, di riportare solo i primi tre con il numero di giorni maggiore. Dalla vista Top 10 dei canali con più visualizzazioni si è notato che la maggioranza dei canali presenti nel grafico appartengono alla categoria *"Music"*. Mentre per la Top 10 dei canali con il numero di video più in tendenza si è notato che la maggior parte di questi canali appartenevano alla categoria *"Entertainment"*.

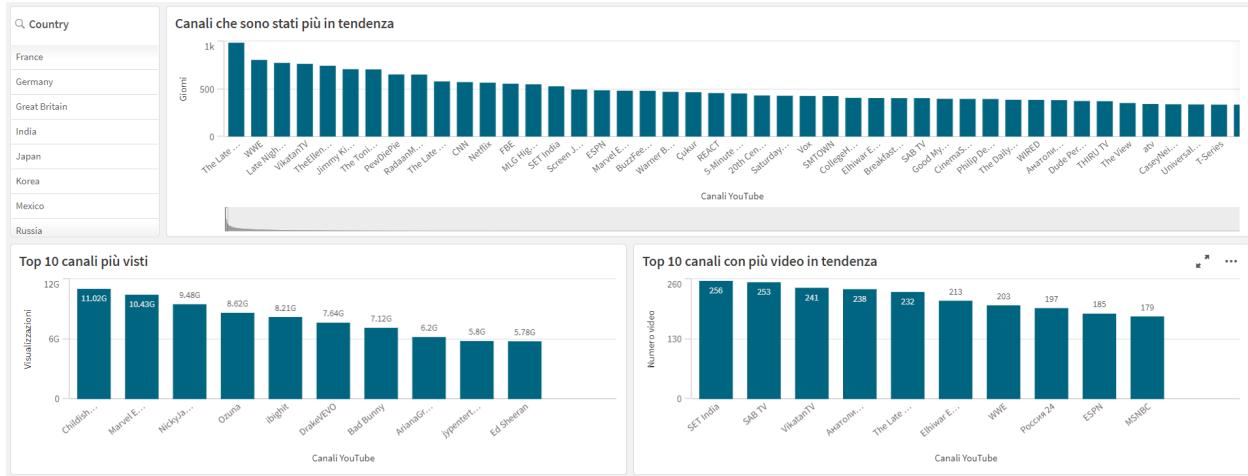


Figura 13: Vista generale sui canali di maggiore tendenza

2.2.7 Analisi delle categorie in tendenza

Nella seguente sezione si va ad analizzare quali categorie hanno più views e quali categorie con i rispettivi video vanno più facilmente in tendenza. Dal grafico in basso a sinistra si nota facilmente quali video appartenenti alle categorie hanno più views. In ordine, come prima categoria si ha *"Music"*, a seguire si ha *"Entertainment"* ed infine *"People & Blog"*. Come si può vedere nel grafico in basso a destra le tre principali categorie con i loro rispettivi video che vanno in tendenza sono *"Entertainment"* che ha il maggior numero di video in tendenza, a seguire si ha *"People & Blog"* e infine la categoria *"Music"*. Infine, nel grafico in alto vengono visualizzate le categorie con il maggior numero di video in tendenza con la relativa data di pubblicazione e si può notare come la principale categoria sia quella dedicata all'intrattenimento.

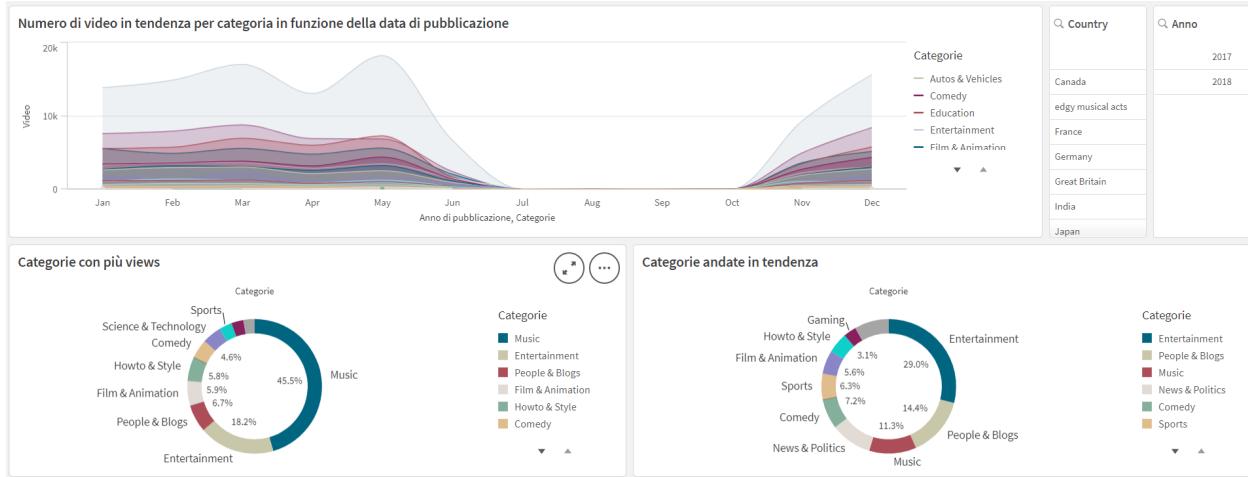


Figura 14: Vista generale sulle categorie di maggiore tendenza

2.2.8 Analisi geografica delle categorie e video in tendenza

Nella seguente dashboard viene effettuata una analisi geografica sul numero di video in tendenza e quali categorie vanno per la maggiore in ogni paese. Nel primi due grafici si può notare che la Russia è il paese con più video (34690), seguita con poco distacco dal Messico (33783), dalla Francia (30529) e dalla Germania (29610). Nel grafico a destra invece si vede come sono distribuite le categorie nei paesi, ad esempio nel Messico le categorie che vanno per la maggiore sono *"Entertainment"*, *"People & Blog"* e *"Sport"*, mentre in Russia la categoria principale è *"People & Blog"*, seguita a pari merito da *"News & Politics"* e *"Entertainment"*.

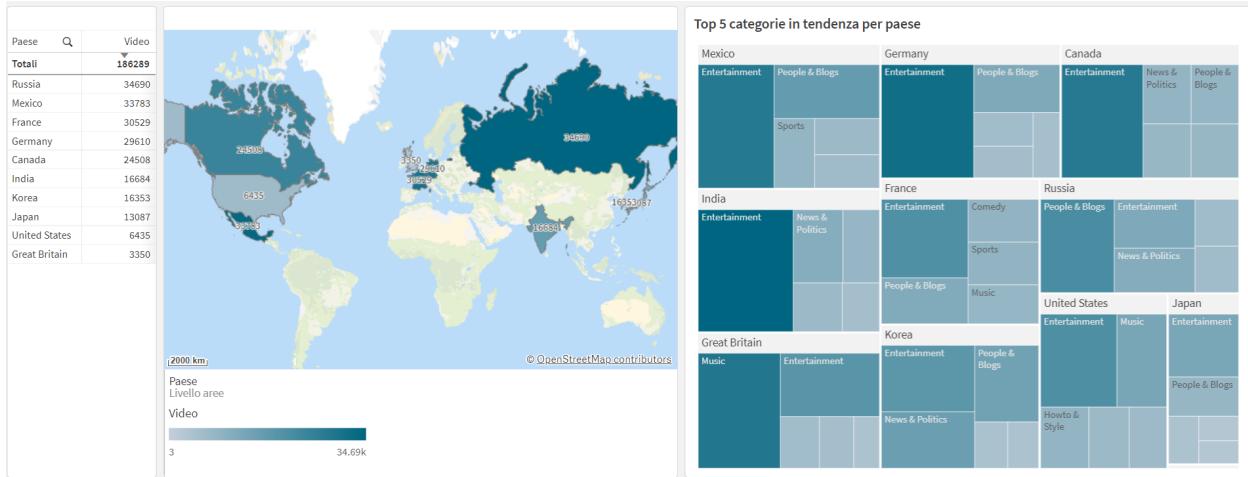


Figura 15: Distribuzione geografica delle categorie e dei video in tendenza

2.2.9 Analisi temporale della pubblicazione dei video in tendenza

Nella penultima dashboard, è stato ritenuto interessante descrivere l'andamento orario, giornaliero, mensile e annuale sulla pubblicazione dei video. Nel primo grafico in alto a sinistra vengono riportate tutte le ore che compongono una giornata. Troviamo un picco in prossimità delle ore 16, questo evidenzia il fatto che in

quell'orario viene pubblicato un elevato numero di video. In particolare, abbiamo un valore di 14.050 video. Spostandoci nel grafico in alto a destra, troviamo la stessa distribuzione, ma suddivisa in base a giorni settimanali. Si può osservare che il giorno in cui vengono pubblicati più video è il venerdì, con 29.410 video. Proseguendo, nel grafico in basso a sinistra, abbiamo riportato l'andamento in funzione dei giorni mensili. Qui si può osservare che la distribuzione è abbastanza uniforme, in quanto non ci sono particolari punti rilevanti. Possiamo considerare questo dato non influente ai fini dell'analisi. Invece, nell'ultimo grafico, riportiamo la distribuzione dei video in base all'anno di pubblicazione. Fino al 2016 il numero dei video pubblicati che sono andati in tendenza, è molto basso, tanto che viene approssimato quasi allo zero. Invece, dal 2016 in poi,abbiamo una crescita rilevante, questo perché il dataset si concentra proprio sui video andati in tendenza tra il 2017 e il 2018.

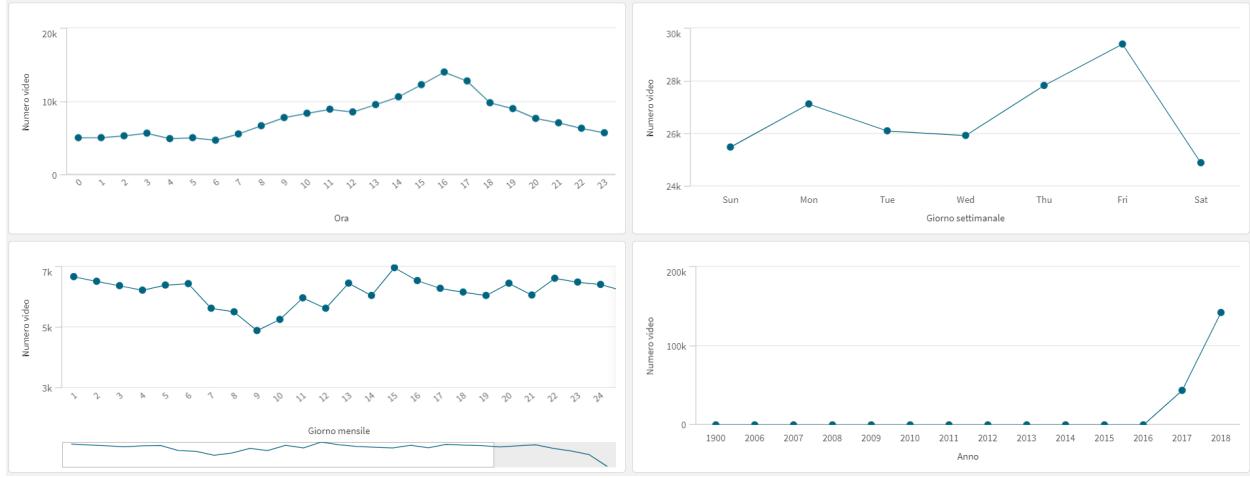


Figura 16: Vista sugli andamenti temporali delle pubblicazioni dei video

3 Tableau

Tableau è un software che può essere usato come programma desktop sia per Windows che per Mac che si adatta a tutti gli ambienti di dati. Tableau offre velocità, scalabilità, adattabilità ed affidabilità. Oltre a Tableau desktop, sono presenti:

- Tableau Online;
- Tableau Server;
- Tableau Prep;
- Tableau Manager.

Tableau permette la creazione di fogli personalizzati tramite i quali vengono costruiti dei grafici con cui è possibile creare delle dashboard, i quali possono contenere più fogli. Unendo le dashboard possiamo creare una storia proprio come su Qlik. Tra i tre programmi, Tableau permette una migliore analisi delle tendenze perché ha vari modelli configurabili (lineare, polinomiale, esponenziale, ecc.) che permettono di adattarsi al meglio alle serie temporali, ed automaticamente vengono calcolati R-squared ed p-value. La differenza principale che lo contraddistingue da Qlik risiede nel fatto che consente di effettuare anche delle analisi predittive tramite il forecasting.



Figura 17: Logo Tableau

3.1 Caricamento dati e ETL

Dopo che è stato creato un profilo su Tableau ed aver ottenuto la licenza da studente, è stata scaricata l'applicazione desktop. Successivamente sono stati importati i vari file CSV in Tableau come file di testo e sono stati concatenati per costruire il dataset. In particolare, Tableau offre la possibilità di creare una nuova unificazione, in cui al suo interno vengono inseriti tutti i file CSV riguardanti i video. Lo stesso approccio è stato utilizzato anche per le categorie, avendo quindi due tabelle denominate *"Video"* e *"Category"*. Queste sono state collegate tramite una relazione uno a molti tra gli attributi *"Video.category_id"* e *"Category.id"*.

☐ Video+ (Conessioni multiple)



Figura 18: Collegamento tabelle

Sono stati poi filtrati tutti i record che avevano campi null, come, ad esempio, *comment_count*, *likes*, *dislikeas* e *views*.

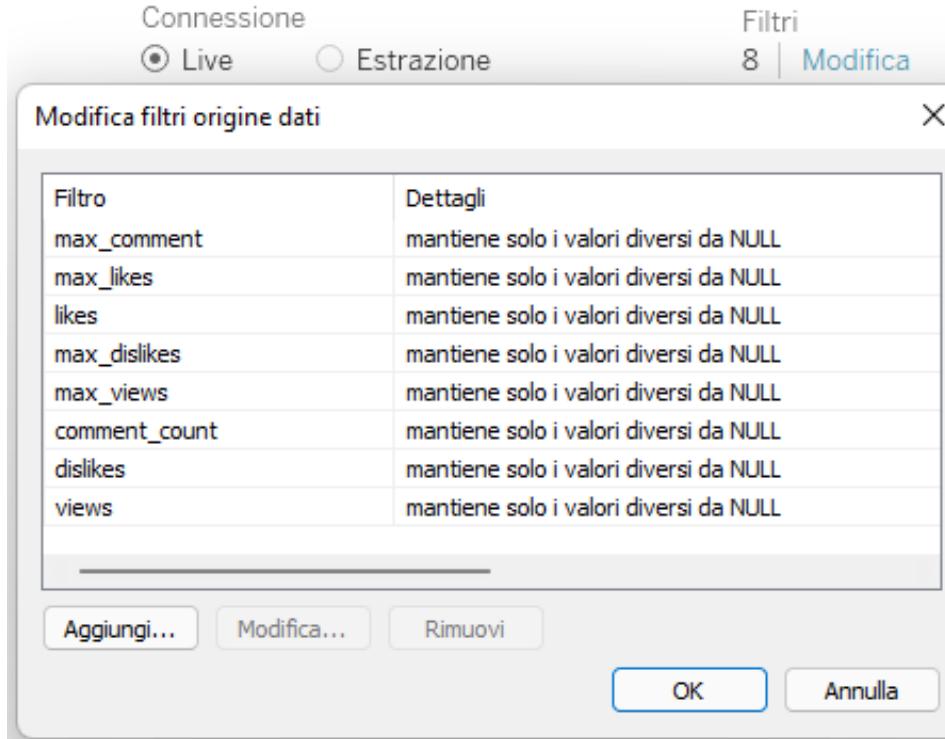


Figura 19: Filtri valori nulli

Dato che i video che vanno in tendenza per più giorni compaiono più volte nel dataset, sono stati creati dei campi calcolati per ottenere il valore più recente dei commenti, like, dislike e views. I campi calcolati sono descritti come segue:

- *Commenti: Fixed [title]: MAX([comment_count])*
- *Like: Fixed [title]: MAX([likes])*
- *Dislike: Fixed [title]: MAX([dislikes])*
- *Views: Fixed [title]: MAX([views])*

3.2 Analisi e descrizione delle visualizzazioni

Per quanto riguarda la realizzazione delle viste si è deciso di basarsi, generalmente, su una suddivisione delle viste simile a quella proposta nel progetto di Qlik, aggiungendo tuttavia delle nuove dashboard, che fanno uso delle feature che Tableau offre in più rispetto a quest'ultimo. Le viste realizzate in Tableau sono:

- una vista sui video per anno di pubblicazione;
- una vista sui video con commenti e rating disabilitati;
- una vista sulla top 10 dei video con maggiore numero di likes, views e commenti;
- una vista sulla top 5 delle categorie con maggior numero di video per paese;
- una vista sulla top 5 categorie con maggior numero di visualizzazioni per paese;
- una vista sul confronto delle categorie in base a likes, dislikes, views e commenti;

- una vista sul numero di video in tendenza e numero di canali per paese;
- una vista sulla distribuzione dei likes, dislikes, commenti e views per paese;
- una vista sulla media dei giorni in tendenza dei video per paese, categorie e canale;
- una vista sulla distribuzione delle categorie in tendenza;
- una vista sull'andamento delle categorie in funzione del mese di pubblicazione e di tendenza;
- una vista sulle correlazioni;
- una vista sulle previsioni delle misure;
- una vista sul numero di video in tendenza per giorni con previsione;
- una vista sulla revisione delle misure per un video in tendenza.

3.2.1 Video per anno di pubblicazione

La prima vista riguarda un'analisi sul numero di video per anno di pubblicazione. Per ogni anno sono state considerate le date di pubblicazione dei video ed è stato applicato un conteggio. Dal grafico si può vedere che, l'anno dove sono stati pubblicati il maggior numero di video è il 2018. Come descritto dal dataset, vengono trattati i video andati in tendenza tra novembre 2017 e giugno 2018. Infatti, la maggior parte di essi sono stati pubblicati proprio in questo intervallo temporale, ad eccezione di alcuni che, invece, sono stati pubblicati qualche anno prima, come nel caso di 62 video corrispondenti all'anno 2016.

Video per anno di pubblicazione

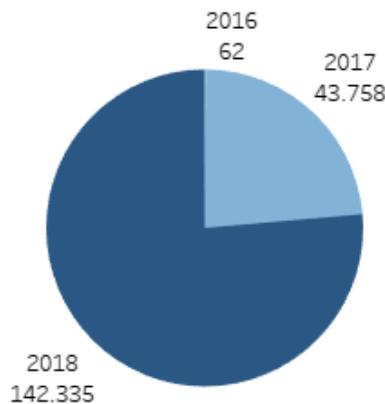


Figura 20: Video per anno di pubblicazione

3.2.2 Video con commenti e rating disabilitati

Per l'analisi dei commenti e rating disabilitati è stato utilizzato un grafico a torta che ci ha permesso di osservare quante volte, in percentuale, si verifica un commento disabilitato o un rating disabilitato all'intero del dataset. Abbiamo adottato due metriche per descrivere questo fenomeno, ovvero *true* e *false*. Dall'analisi fatta nell'immagine sottostante, si può notare che solamente il 2,54% dei video ha commenti disabilitati e il 2,51% ha rating disabilitati. Mentre la percentuale dei video che hanno entrambi i campi disabilitati equivale al 0,75%. Quindi, si può osservare che i video con rating e commenti disabilitati sono in quantità trascurabile rispetto all'intero dataset.

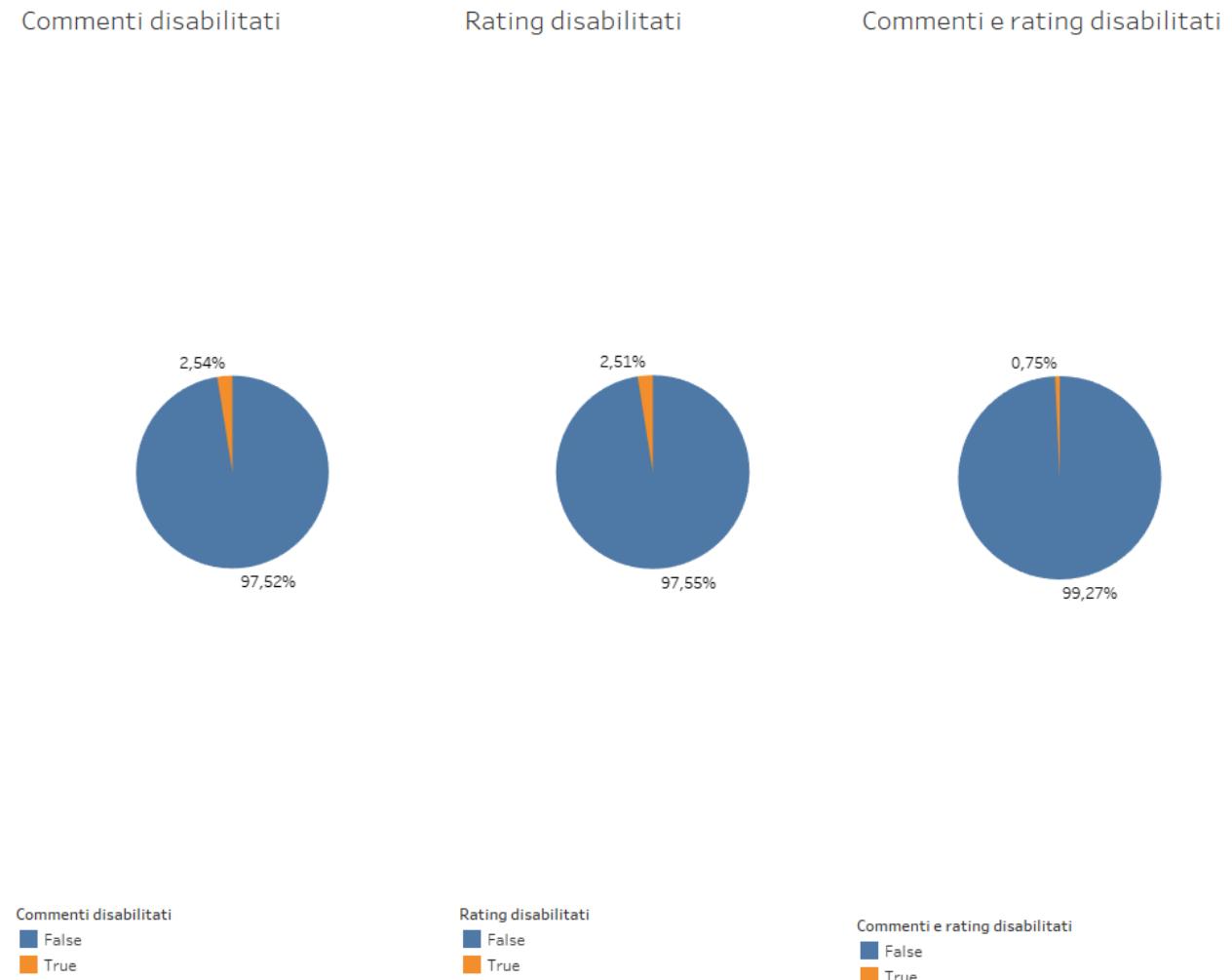


Figura 21: Video con commenti e rating disabilitati

3.2.3 Top 10 dei video con maggior numero di likes, views e commenti

La dashboard riporta quattro KPI fondamentali: il numero di video, la media dei likes, la media delle visualizzazioni e la media dei commenti. In base a questi ultimi indicatori si può stabilire una top 10. Il video con più like in assoluto è il video *"BTS 'FAKE LOVE' Official MV"* con 5.613.827 like, il video con più views è il video *"Nicky Jam x J. Balvin - X (EQUIS) — Video Oficial — Prod. Afro Bros Jeon"*

con 424.538.912 views, infine, il video con più commenti è sempre *"BTS 'FAKE LOVE' Official MV"* con 1.228.655 commenti.

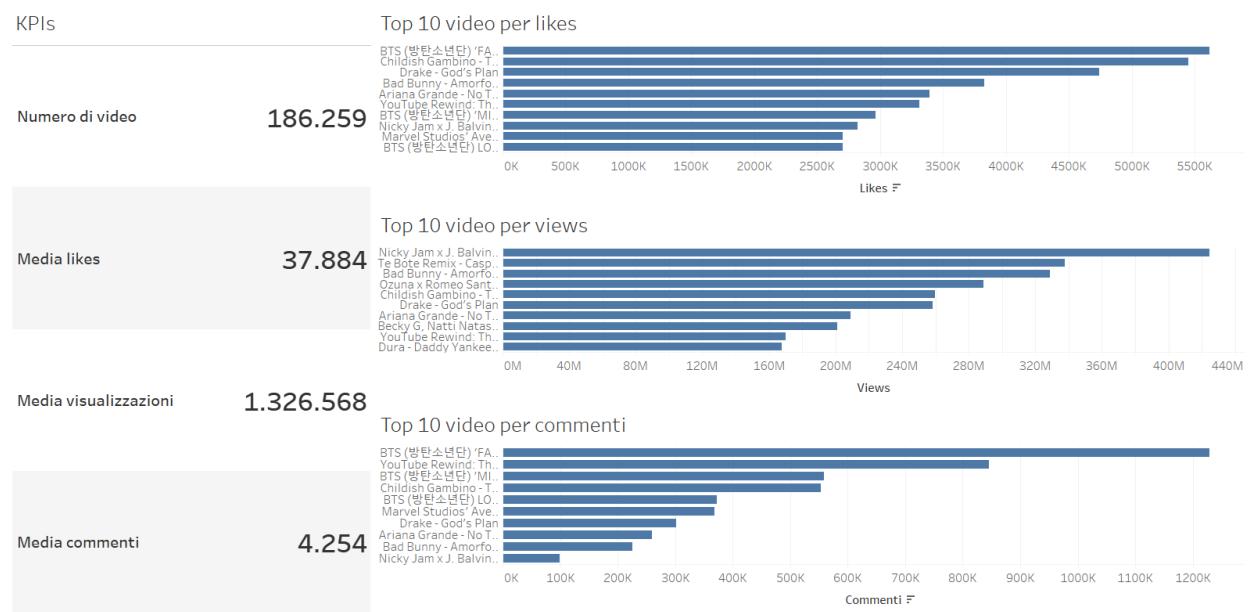


Figura 22: Top 10 dei video per likes, views e commenti

3.2.4 Top 5 categorie con maggior numero di video per paese

Nella seguente dashboard sono state selezionate le prime cinque categorie con più video in tendenza per ogni paese. Dal grafico a mappa, si può osservare che la categoria con maggior numero di video è quella di *"Entertainment"*, seguita poi dalla *"People & Blogs"* e *"Music"*. I paesi con maggior numero di video sono Messico e Russia.

Distribuzione delle categorie per paese

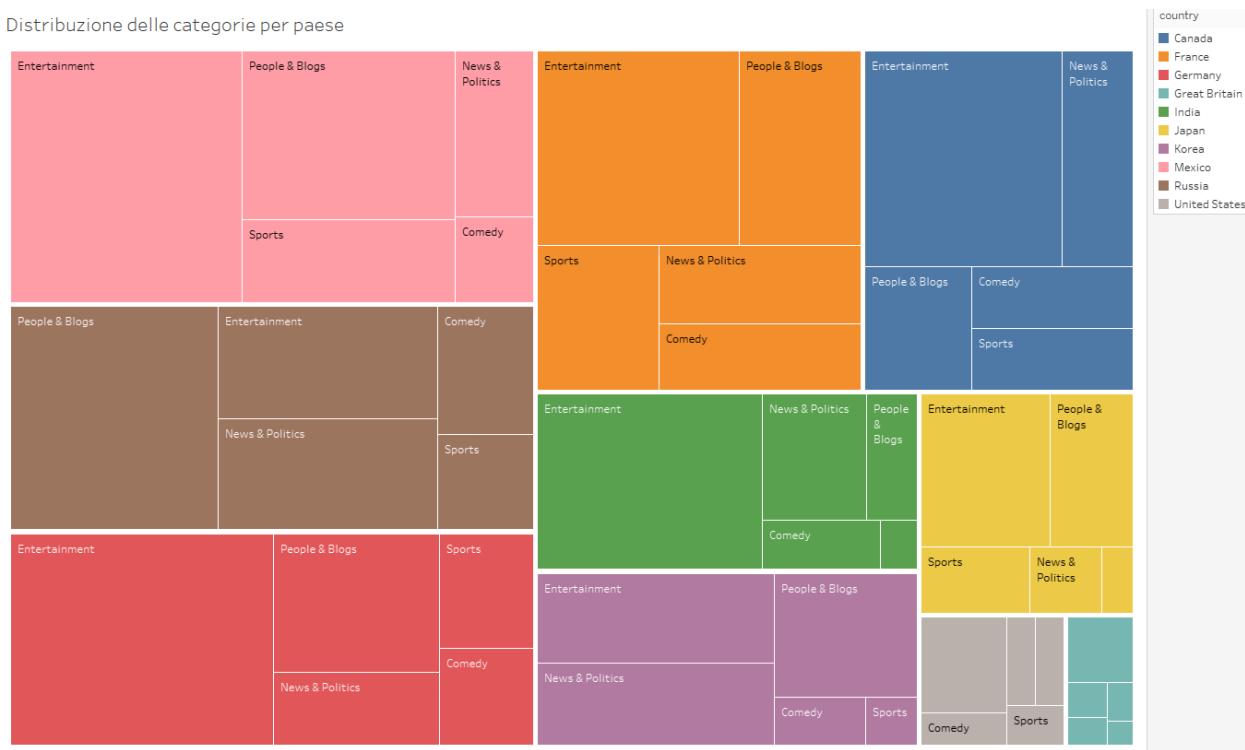


Figura 23: Top 5 categorie per numero di video per paese

3.2.5 Top 5 categorie con maggior numero visualizzazioni per paese

Nella seguente dashboard sono state selezionate le prime cinque categorie con maggior numero di visualizzazioni per paese attraverso un apposito filtro. Infatti, i diversi colori della dashboard rappresentano i diversi paesi, mentre la dimensione dei riquadri riguarda la dimensione delle visualizzazioni che riceve ogni categoria. Come si può notare, per ogni paese le due categorie dominanti sono *"Entertainment"* e *"Music"* che si alternano il primo e secondo posto, seguite poi dalle altre categorie. Possiamo affermare che questo grafico è strettamente collegato al grafico precedente, in quanto è possibile osservare un'analogia tra le categorie con il maggior numero di video e le categorie con il maggior numero di visualizzazioni, a meno della categoria *"Music"* che a differenza del grafico precedente qui occupa spesso le prime posizioni. Questo aspetto indica un fattore molto importante: benché la categoria della musica abbia meno video rispetto alle altre categorie, ha comunque un elevato interesse. Dunque si intuisce che i video musicali riscuotono sicuramente un maggiore successo rispetto a tutti gli altri.

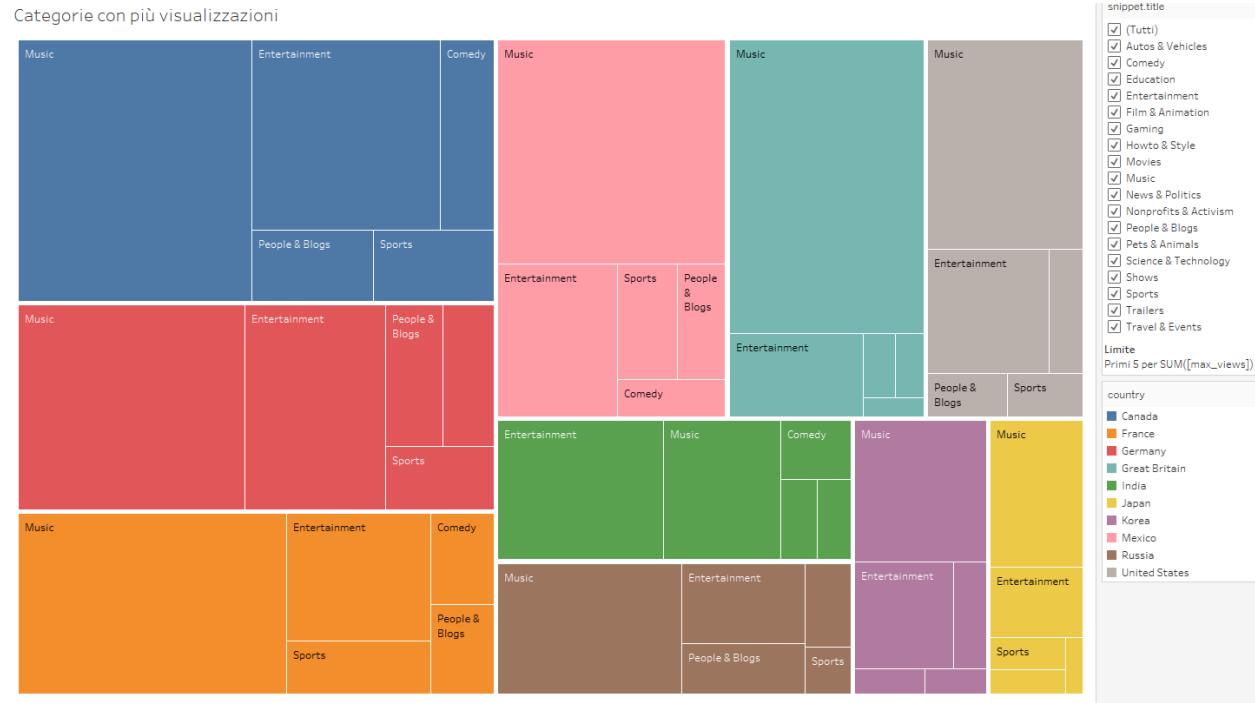


Figura 24: Top 5 categorie per visualizzazioni paese

3.2.6 Confronto categorie in base a likes, dislikes, views e commenti

In questa dashboard viene mostrato un confronto tra le categorie dei video in base a quattro misure calcolate con la funzione di aggregazione SOMMA. Queste sono: la somma del numero di likes, la somma del numero di dislikes, la somma del numero di views e la somma del numero di commenti. Il grafico a barre riportato in figura 25 mostra la distribuzione di questi parametri rispetto alla categorie che compongono l'intero dataset. Possiamo affermare che, ad esempio, la categoria *"Music"* è quella che riceve più visualizzazioni e likes, ma ha un ridotto numero di commenti. Mentre la categoria *"Entertainment"* è sicuramente più informe rispetto alla quantità di likes, dislikes, views e commenti in quanto su tutte queste misure presenta elevati valori. Questo potrebbe indicare che è la categoria che riceve maggiori interazione dagli utenti.



Figura 25: Likes, dislikes, commenti e visualizzazioni per categoria

3.2.7 Numero di video in tendenza e numero di canali per paese

In questo paragrafo si va ad analizzare il numero di video che vanno tendenza sui differenti paesi. Si può notare che la Russia è il paese con più video (34690), seguita con poco distacco dal Messico (33783), dalla Francia (30529) e dalla Germania (29610). Per evidenziare questa differenza, abbiamo applicato una colorazione in base alla densità dei video su ogni paese. Anche nella figura 27 viene riportato un grafico a mappa per mostrare il numero dei canali presenti nei differenti paesi. Si può notare che le mappe sono abbastanza simili, questo conferma il fatto che i canali e il numero di video in tendenza sono proporzionali, cioè i paesi dove i video vanno in tendenza sono gli stessi in cui abbiamo un elevato numero di canali YouTube.

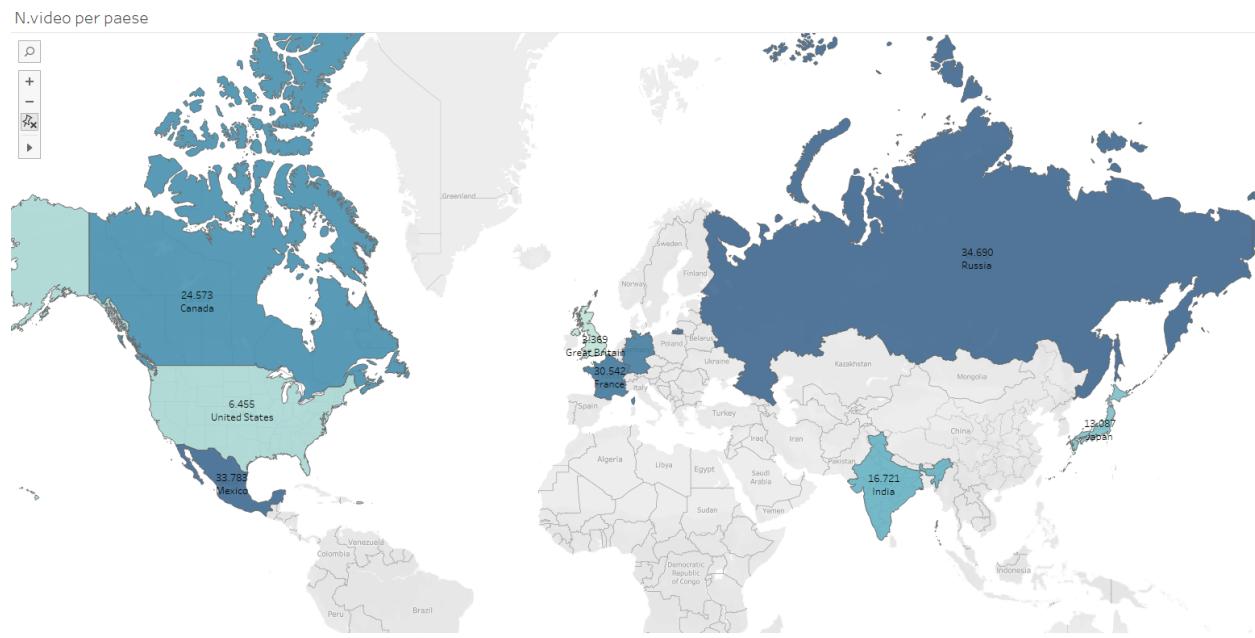


Figura 26: Numero di video in tendenza per paese

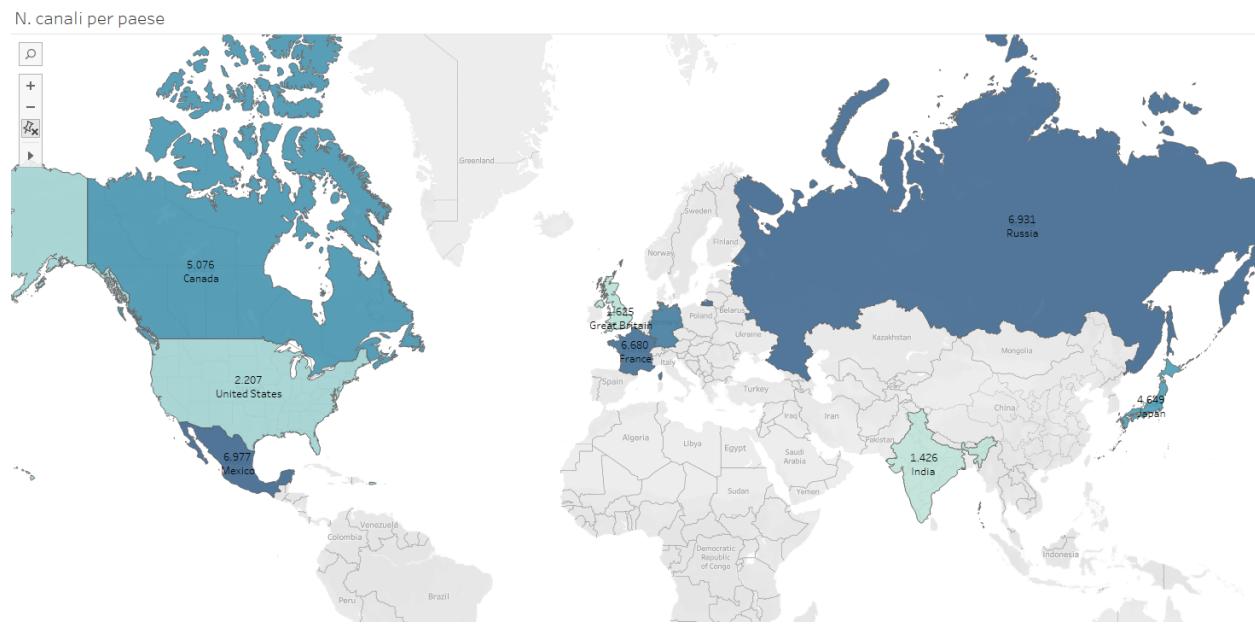


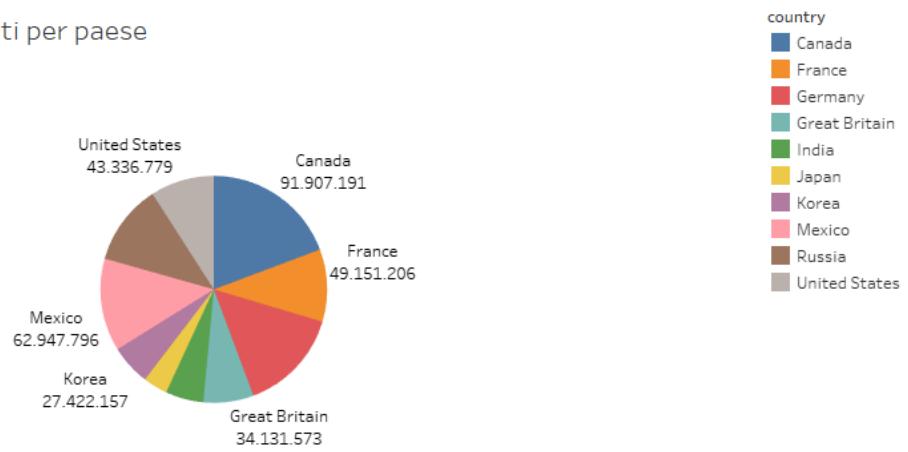
Figura 27: Numero di canali per paese

3.2.8 Distribuzione dei likes, dislikes, commenti e views per paese

In questa dashboard vengono riportate le diverse distribuzioni riguardanti i likes, i dislikes, i commenti e le views. Nel grafico sottostante si può osservare che il paese con il maggior numero di commenti è il Canada, a

seguire si ha il Messico, la Germania e poi gli altri paesi. Il secondo grafico a torta rappresenta il numero dei commenti su ogni video in base alla paese di appartenenza. Questo indicatore è stato calcolato attraverso l'operazione `COUNT(comment_count)/COUNT(title)`, in modo tale da avere una misura normalizzata su cui fare analisi. A differenza del precedente grafico a torta, qui prevalgono paesi come Great Britain, United States e via via gli altri.

Distribuzione dei commenti per paese



Rapporto commenti/video per paese

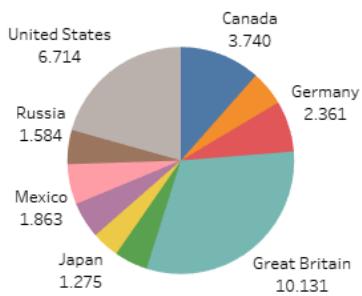
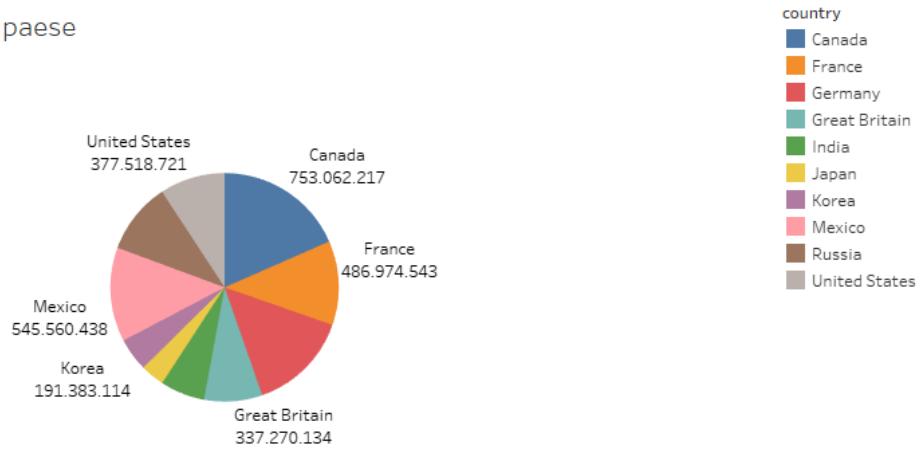


Figura 28: Video per anno di pubblicazione

Nell'immagine sottostante (Figura 29) si vanno ad analizzare la percentuale di likes e il rapporto likes su video in base al paese. Per quanto riguarda la distribuzione dei likes per paese, si può notare che ad avere il maggior numero di likes è il Canada, a seguire si ha la Germania e la Francia. Invece, nel grafico che mostra il rapporto likes su video, notiamo una gerarchia che vede prevalere sempre la Gran Bretagna, successivamente gli Stati Uniti e a seguire gli altri paesi.

Distribuzione dei like per paese



Rapporto like/video per paese

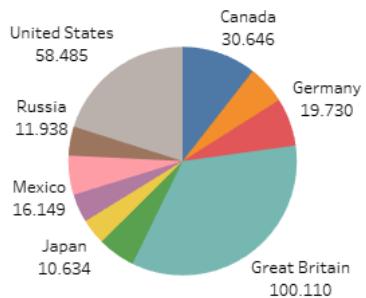
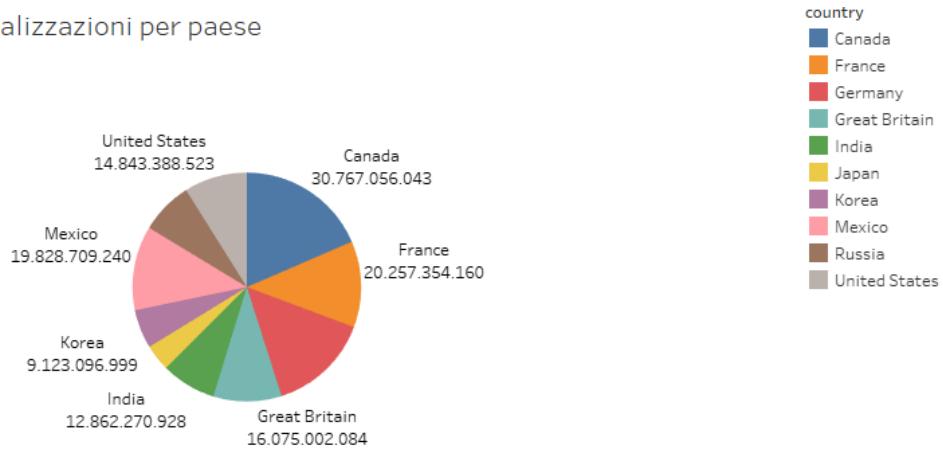


Figura 29: Video per anno di pubblicazione

La Figura 30 mostra la distribuzione delle views per paese e anche il rapporto views su video per paese. Si può notare che ad avere il maggior numero di views è il Canada, a seguire si ha la Germania e la Francia. Invece, il rapporto views per video, a differenza di prima, mostra che è la Gran Bretagna ad avere il maggiore numero di views su ogni video, a seguire si hanno gli Stati Uniti e il Canada, poi tutti gli altri paesi.

Distribuzione delle visualizzazioni per paese



Rapporto views/video per paese

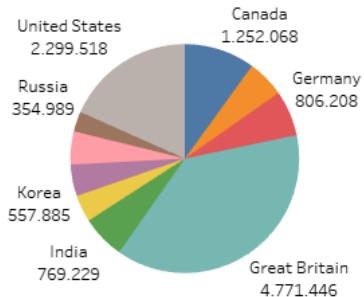
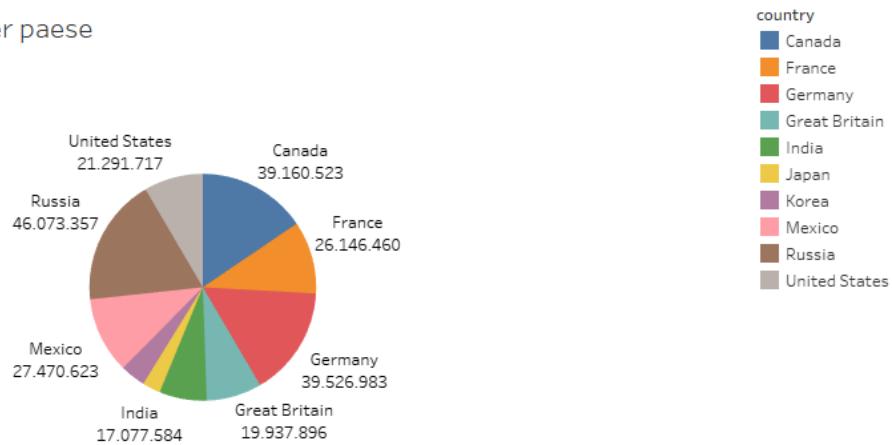


Figura 30: Video per anno di pubblicazione

Nella Figura 31 si vanno ad analizzare la distribuzione dei dislikes per paese e anche il rapporto dislikes su video per paese. Dal grafico della distribuzione dei dislikes per paese si può notare che ad avere il maggior numero di dislikes sui video è il Canada, a seguire si ha la Germania e la Francia. Invece, dall'grafico che mostra il numero di dislikes su ogni video, si evince che è ancora la Gran Bretagna ad avere il maggiore numero di dislikes sui video, a seguire si hanno gli Stati Uniti e il Canada.

Distribuzione dei dislike per paese



Rapporto dislike/video per paese

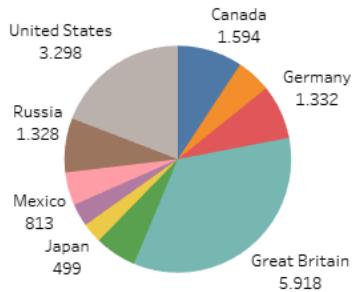


Figura 31: Video per anno di pubblicazione

Grazie all'analisi su questi quattro indicatori, si sono potuti realizzare quattro grafici a torta che descrive le percentuali di likes, views, dislikes e commenti e il loro rapporto tra questi indicatori e ogni video in base al paese di pubblicazione del video stesso. Da essi si nota che per la prima tipologia di grafico, le prime tre posizioni sono state sempre occupate dagli stessi paesi, ovvero Canada, Germania e Francia, mentre per quanto riguarda la seconda tipologia di grafico, i paesi che prevalgono sempre sono Gran Bretagna, Stati Uniti, Canada. Da qui si intuisce che sono quest'ultimi i paesi dove i video subiscono più interazioni da parte degli utenti. Questo è un aspetto fondamentale quando si vuole eseguire un'analisi statistica sui video pubblicati nella piattaforma YouTube, perché questo parametro costituisce una sorta di indice di monetizzazione da parte del content creator che ha pubblicato il video stesso.

3.2.9 Media dei giorni in tendenza dei video per paese, categoria e canale

In questa dashboard viene analizzato il tempo medio di permanenza in tendenza dei video. Nel primo grafico è stata effettuata un'analisi per paese, in cui si evince che in Gran Bretagna i video rimangono in tendenza per 12 giorni, seguita dagli Stati Uniti con una media di 8 giorni, mentre i restanti paesi hanno una media di 2 o meno giorni. Nell'analisi successiva viene calcolato il tempo medio in tendenza per categoria: quella rimasta per più tempo in tendenza è la categoria *"Music"*, con 3 giorni di media, seguita da *"Comedy"* con 2 giorni e tutte le altre categorie con dati compresi tra 1 e 2 giorni. Infine, viene effettuata un'analisi per tempo di tendenza medio per canale. Qui si osserva che i numeri sono molto più elevati e ciò può indicare che ci sono molti canali che vanno in tendenza molto più spesso a differenza di altri che ci vanno molto raramente. Il canale *"Hip-Hop Power"* ha una media di ben 39 giorni in tendenza, seguita con valori molto simili da tantissimi altri canali musicali. La conclusione di questa analisi è che la Gran Bretagna è il paese in cui i video rimangono in tendenza per più tempo e che la categoria più soggetta a rimanere nelle classifiche è la musica, confermato dal fatto che i canali con un maggior tempo medio in tendenza riguardano tutti quest'ultima categoria.

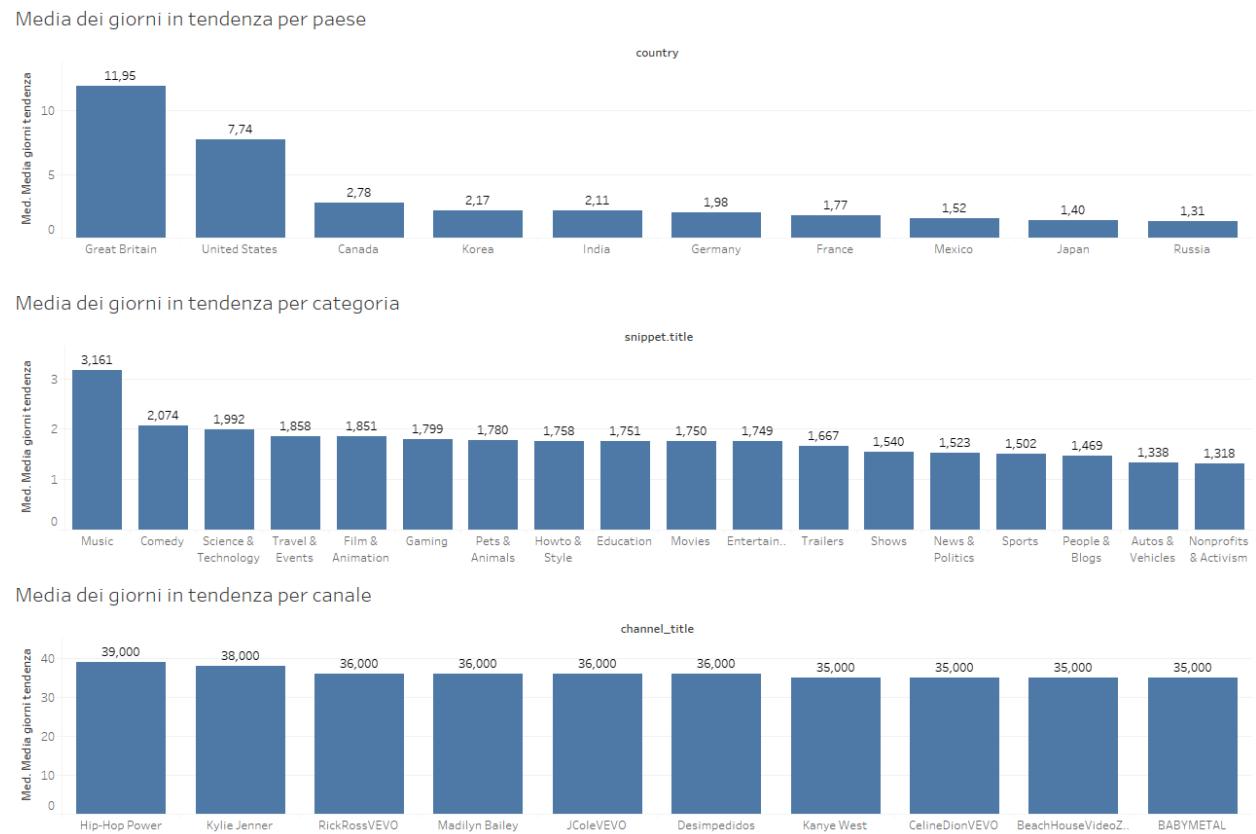


Figura 32: Media dei giorni in tendenza dei video per paese, categoria e canale

3.2.10 Distribuzione delle categorie in tendenza

In questo capitolo si analizzano le categorie con il numero di video che vanno principalmente in tendenza. Come si può vedere dall'immagine sottostante la categoria con il maggior numero di video in tendenza è *"Entertainment"*. Questo è stato un risultato inatteso, in quanto dalle analisi precedenti si intuiva che la categoria *"Music"* avrebbe avuto il numero maggiore di video in tendenza. Invece, questa categoria non

si avvicina nemmeno nei primi tre posti. Di fatto, ai primi tre posti della distribuzione delle categorie in tendenza si hanno *"Entertainment"*, *"People & Blogs"* e infine *"News & Politics"*.

Distribuzione delle categorie

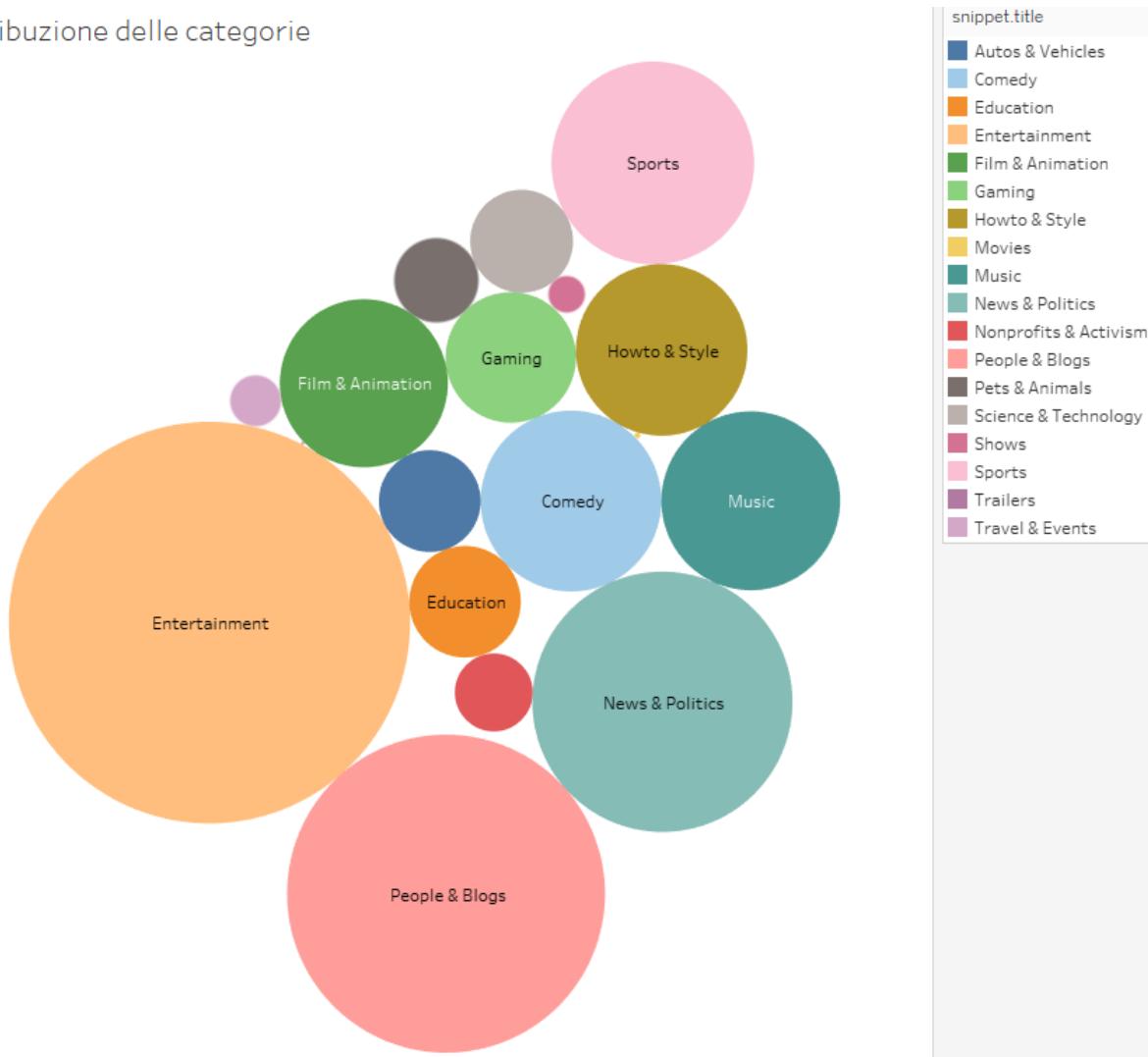


Figura 33: Distribuzione delle categorie in tendenza

3.2.11 Andamento delle categorie in funzione del mese di pubblicazione e di tendenza

In questa dashboard è stato analizzato il numero di video in tendenza per mese e come questi vengono distribuiti per categoria. Nel grafico a sinistra si può vedere come il mese con più video in tendenza è marzo con ben 29.607 video (in tutti i paesi soggetti all'analisi), che comunque non si discosta tantissimo da tutti gli altri mesi. I mesi giugno e novembre hanno invece dei valori molto più inferiori, dovuto al fatto che il dataset è stato raccolto dalla metà del novembre 2017 fino alla metà del giugno 2018, e quindi entrambi i mesi non sono pieni. Nei due grafici a destra invece si può notare che il maggior numero di video in tendenza sono stati pubblicati nel periodo di febbraio e maggio e che la distribuzione delle categorie è sempre rimasta più o meno costante. Anche visualizzando la distribuzione dei video per mese di tendenza, si nota una concentrazione più alta sempre nel medesimo periodo.

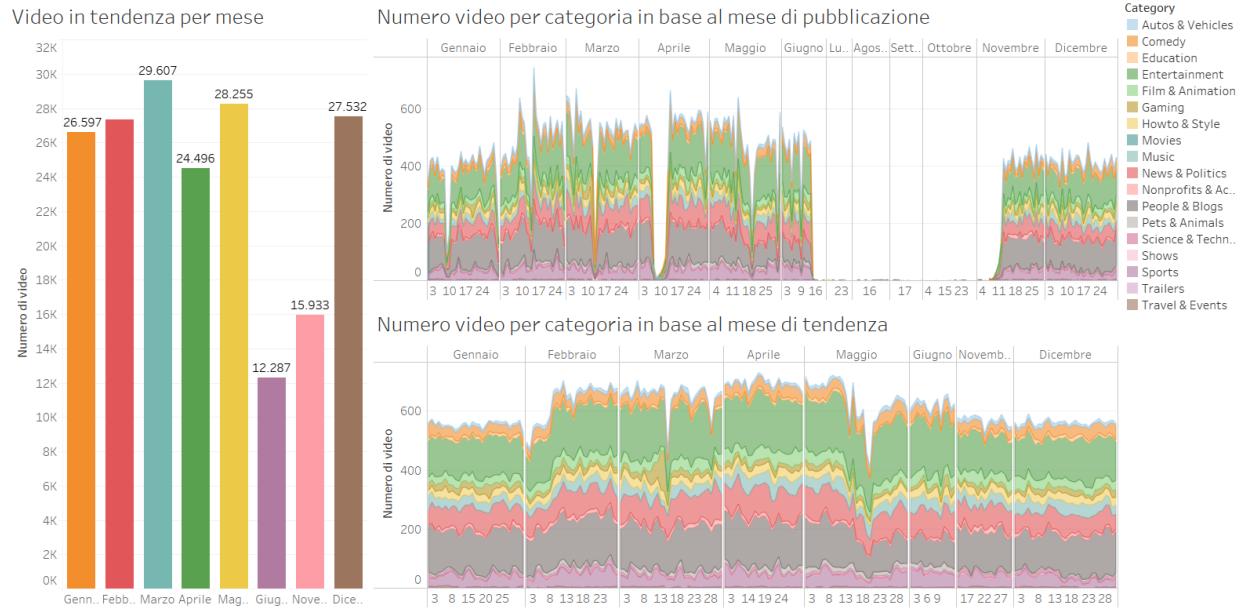


Figura 34: Video per categoria per mese di pubblicazione e tendenza

3.2.12 Correlazioni

A differenza di Qlik, Tableau permette la rappresentazione di tutti i punti nello scatter plot e soprattutto permette di individuare tendenze attraverso l'utilizzo di vari tipi di modelli calcolando anche dei valori come il p-value, R-squared, la media, la mediana, i percentili, i quartili, etc. Nei seguenti grafici si fa un analisi sulle correlazioni tra: views e likes, views e commenti, views e dislikes. Sono state utilizzate varie funzioni di tendenza, ma quelle che vengono utilizzate nei grafici sottostanti sono quelle che forniscono il valore R-squared maggiore. Ad ogni grafico, inoltre, è stato aggiunto la media delle varie misure per vedere il loro collocamento. Nel grafico in basso a sinistra, per la relazione tra likes e views, è stata utilizzata una funzione di tendenza di tipo potenza per calcolare la tendenza. R-squared ha un valore di circa 0.60 mentre il p-value ha valore inferiore a 0.0001. Nel grafico al centro, per la relazione tra views e commenti, è stato usato un modello polinomiale con grado 3 e il risultato per R-squared è circa 0.23 mentre il risultato del p-value è meno di 0.0001. Per finire, nel grafico a destra, per la relazione tra commenti e likes, è stato utilizzato un modello polinomiale che ha restituito un valore R-squared di 0.71, mentre per il p-value di 0.0001.

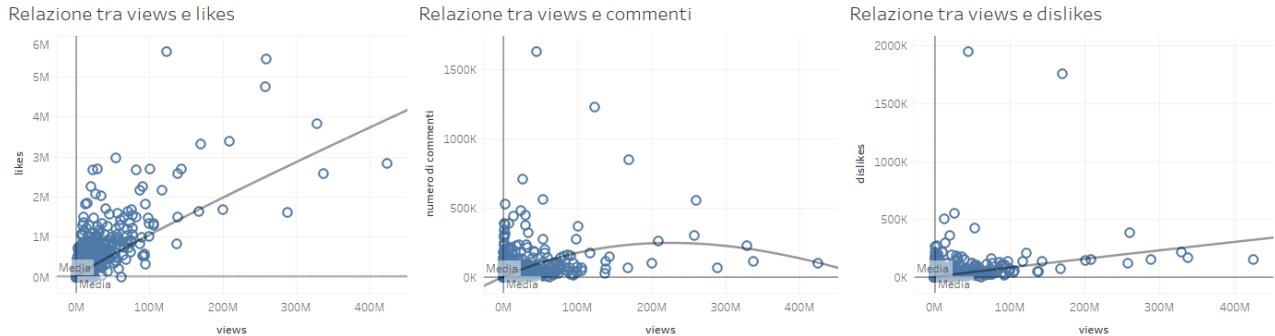


Figura 35: Relazione tra views e likes

Dall'analisi effettuata emerge che i valori del p-value sono molto buoni, essendo inferiori a 0.05. Invece, i valori di R-squared sono troppo bassi. Allora si può dedurre che i dati non sono certamente casuali, ma non hanno una tendenza rappresentabile accuratamente con un modello. Inoltre, osservando i vari grafici si può notare che la maggior parte dei dati è concentrata verso l'inizio degli assi e quindi i dati che si trovano lontani possono essere considerati come anomalie (outliers). Tutto questo è stato fatto per andare a vedere singolarmente le correlazioni presenti tra i vari campi. In particolare, come ipotesi nulla è stato preso il fatto che i campi, in tutti i casi, fossero tra loro correlati, ad esempio, ad alti valori di "likes" dovrebbero corrispondere alti valori di "view". Ma, come è visibile dai grafici sopra riportati, i valori sono scollegati. Questo viene testimoniato anche dal p-value il quale in tutti i casi ha un valore inferiore a 0.05. Ciò ci ha permesso di concludere che i campi presi in considerazione non sono correlati.

3.2.13 Previsioni delle misure (commenti, views, like e numero di video)

Nella figura 36 viene riportata la dashboard sulle previsioni delle misure views, likes, commenti e video. Per ognuna delle misure in questione vengono tracciate le linee di tendenza sia per l'andamento effettivo che per l'andamento stimato. Si può vedere che seppur avendo un andamento abbastanza aleatorio, le misure tendono a crescere. Questo indica che views, commenti, likes e numero di video aumentano di giorno in giorno. Invece, per quanto riguarda le previsioni, non sono molto precise perché rimangono costanti nel tempo, infatti l'intervallo di confidenza è ampio ed è centrato sul valore medio. Questo potrebbe essere dovuto alla scarsa estensione temporale dei dati. Infine, nei boxplot affianco, è possibile vedere come si distribuiscono i valori sulle misure in questione, avendo sempre una separazione tra i dati effettivi e le stime.

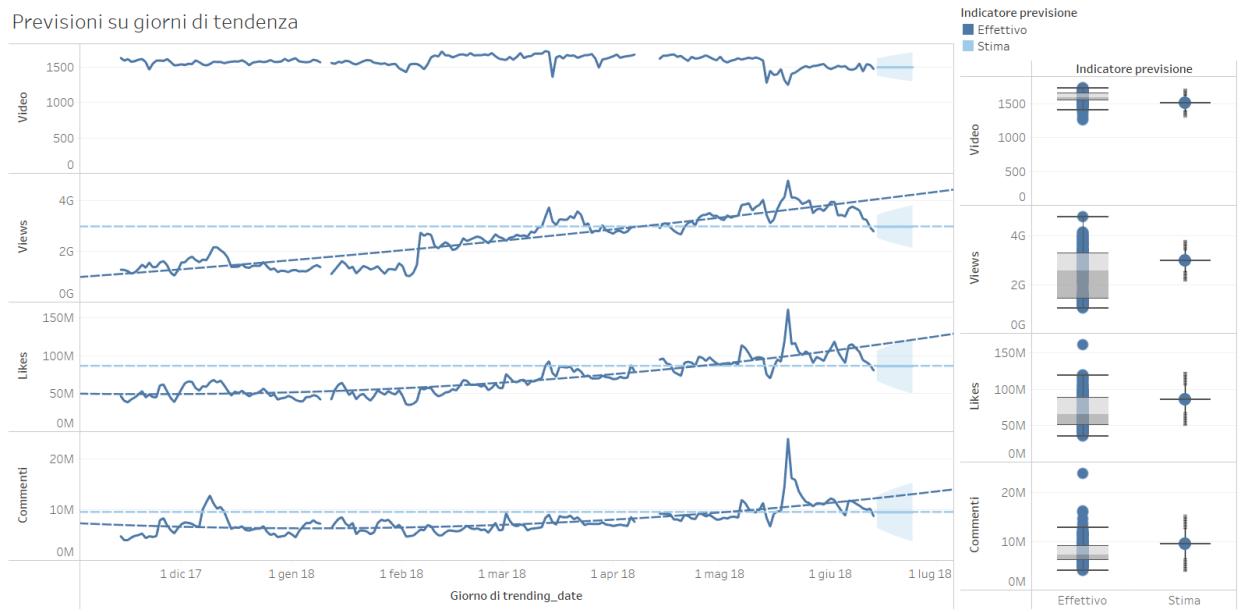


Figura 36: Previsioni delle misure

3.2.14 Numero di video in tendenza per giorni con previsione

Questa dashboard sfrutta la funzionalità "Previsione" di Tableau, che permette di stimare lungo una dimensione l'andamento di dati non disponibili, facendo uso dell'algoritmo dell'Exponential Time Smoothing. Nel caso in questione si è stimato il numero di video in tendenza in base ai giorni dell'anno. Per verificare la validità del sistema sulla serie temporale di interesse è stata fatta una previsione su un anno: dal 20 Novembre 2017 al 24 Settembre 2018, tenendo conto che tra il 2017 e il 2018 i dati sono contenuti nel dataset. Il primo grafico, presente in alto, mostra quindi la previsione, mentre il secondo, costruito a partire da tutti i dati effettivamente disponibili, permette di valutare l'accuratezza. Come si può notare la previsione del primo anno risulta essere compatibile con i dati del secondo grafico, dimostrando una buona capacità di stima da parte del sistema.

Numero video in tendenza per giorni con previsioni



Numero video in tendenza giornalieri



Figura 37: Numero di video in tendenza per giorni con previsione

3.2.15 Previsione delle misure per un video in tendenza

Anche nella seguente dashboard vengono eseguite delle previsioni, in particolare prendendo come riferimento il video *"BTS 'FAKE LOVE' Official MV"*, un video pubblicato in Korea. Nel dettaglio sono state eseguite delle previsioni sul numero di likes, views, commenti e dislikes. In tutti e quattro i grafici, si può notare che la distribuzione di questi indicatori crescerà in modo lineare, cioè con l'aumentare dei giorni aumenteranno il numero delle views, il numero di likes, e così via. L'unica eccezione possiamo trovarla nella distribuzione del numero di commenti, in quanto questa avrà un andamento costante rispetto alle altre misure.

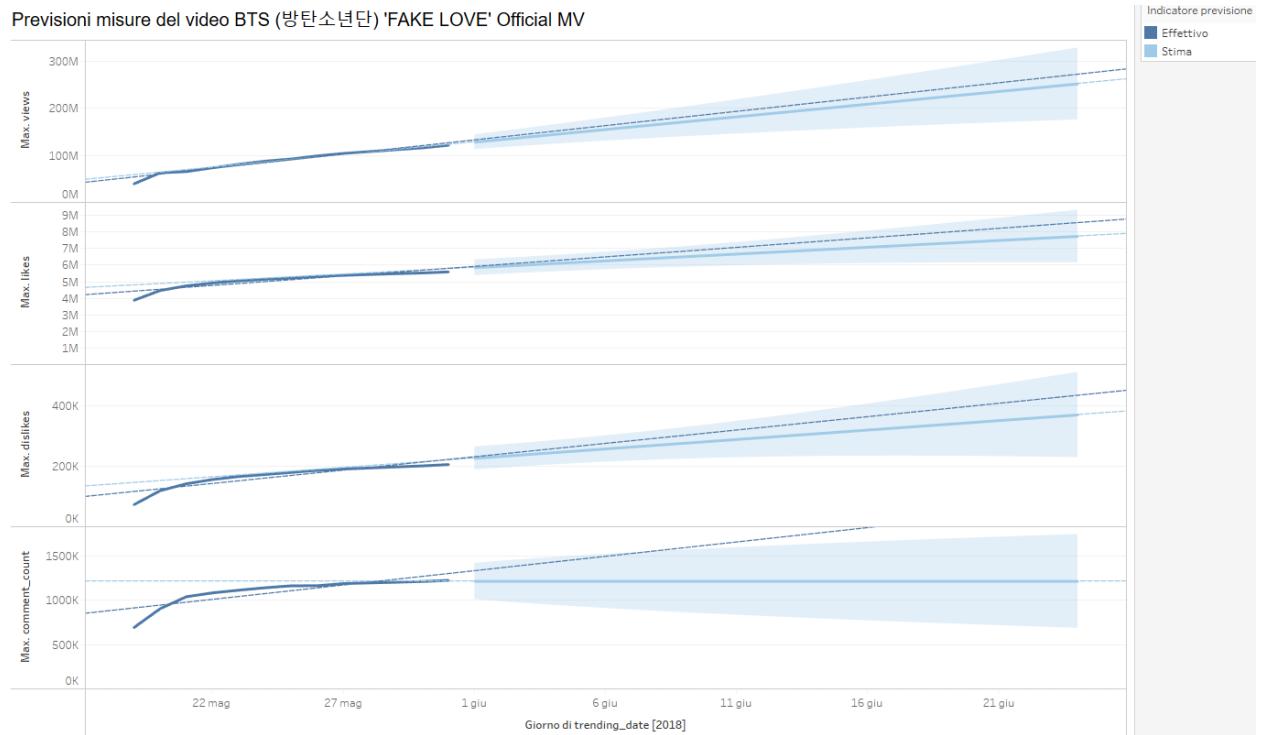


Figura 38: Previsione delle misure del video ”BTS ‘FAKE LOVE’ Official MV”

Per verificare la validità di questa tipologia di analisi, viene riportato uno screenshot del video in questione, in cui è possibile visualizzare alcuni dettagli circa le visualizzazioni, likes e commenti del video stesso alla data del 04/02/2022, per dimostrare che i valori attuali che ha il video sono comunque validi all'interno della previsione. Infatti nella figura 40 i valori della previsione sono stati estesi fino alla data del 01/04/2022 e si può vedere come i valori ammessi della stima si sono allargati considerevolmente, tanto da rendere molto difficile determinare una previsione.

BTS (방탄소년단) 'FAKE LOVE' Official MV

1.067.210.244 visualizzazioni • 18 mag 2018

18 MLN NON MI PIACE CONDIVIDI

HYBE LABELS 63,9 Mln di iscritti

BTS (방탄소년단) 'FAKE LOVE' Official MV

MOSTRA ALTRO

4.504.964 commenti ORDINA PER

Figura 39: Misure del video "BTS 'FAKE LOVE' Official MV" alla data del 04/02/2022"

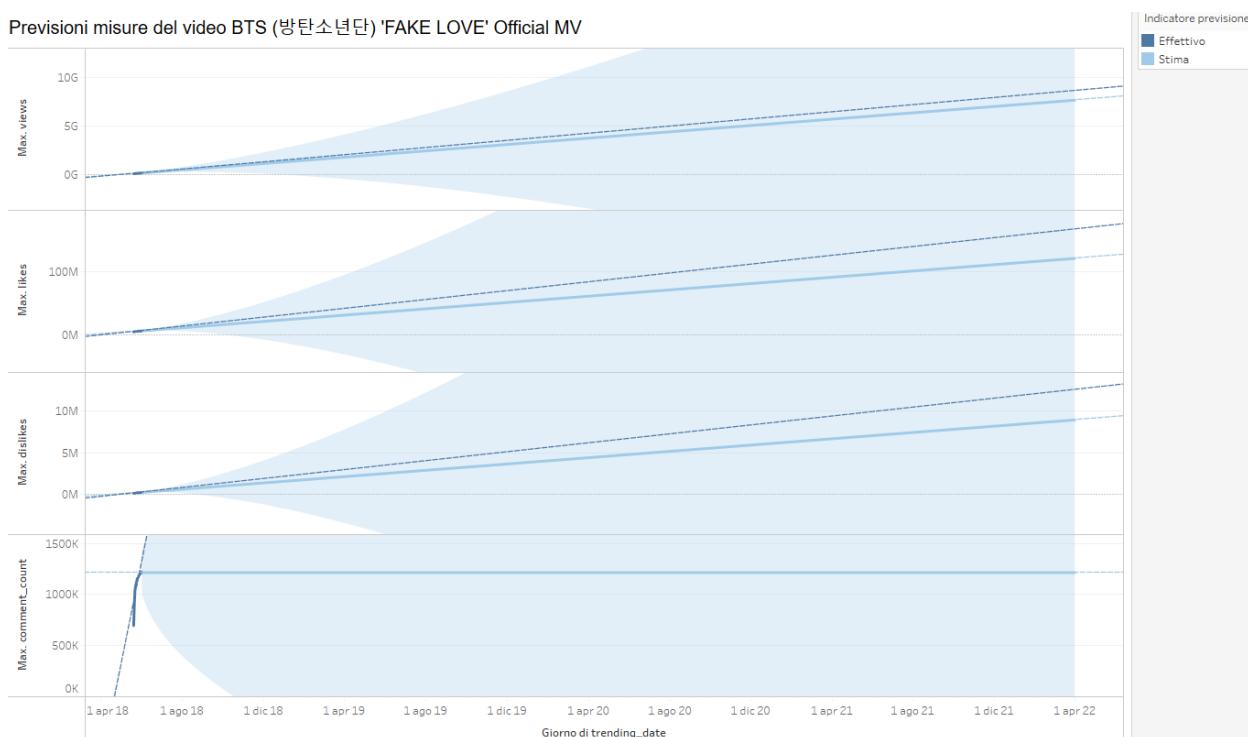


Figura 40: Previsioni estese del video "BTS 'FAKE LOVE' Official MV" fino alla data odierna

Anche per il video "YouTube Rewind: The Shape of 2017 - YouTubeRewind" è stata sviluppata una previsione delle misure. A differenza del video analizzato in precedenza che aveva un andamento quasi lineare, qui si osserva un andamento praticamente logaritmico (quindi con una crescita molto lenta dopo diverso tempo). Sono stati poi rilevate le misure alla data del 04/02/2022 in figura 42 per confrontarle con

quelle delle previsioni. Le misure dei like e dei commenti cadono in una finestra accettabile della previsione (per i dislikes non è stato possibile effettuare il confronto poiché recentemente YouTube ha nascosto questo valore agli utenti). Per quanto riguarda le visualizzazioni, si nota invece che la previsione rilevata da Tableau è errata. Infatti il video presenta quasi 240 milioni di visualizzazioni alla data del 04/02/2022, valore che è già fuori dalla stima alla data del 14/01/2018.

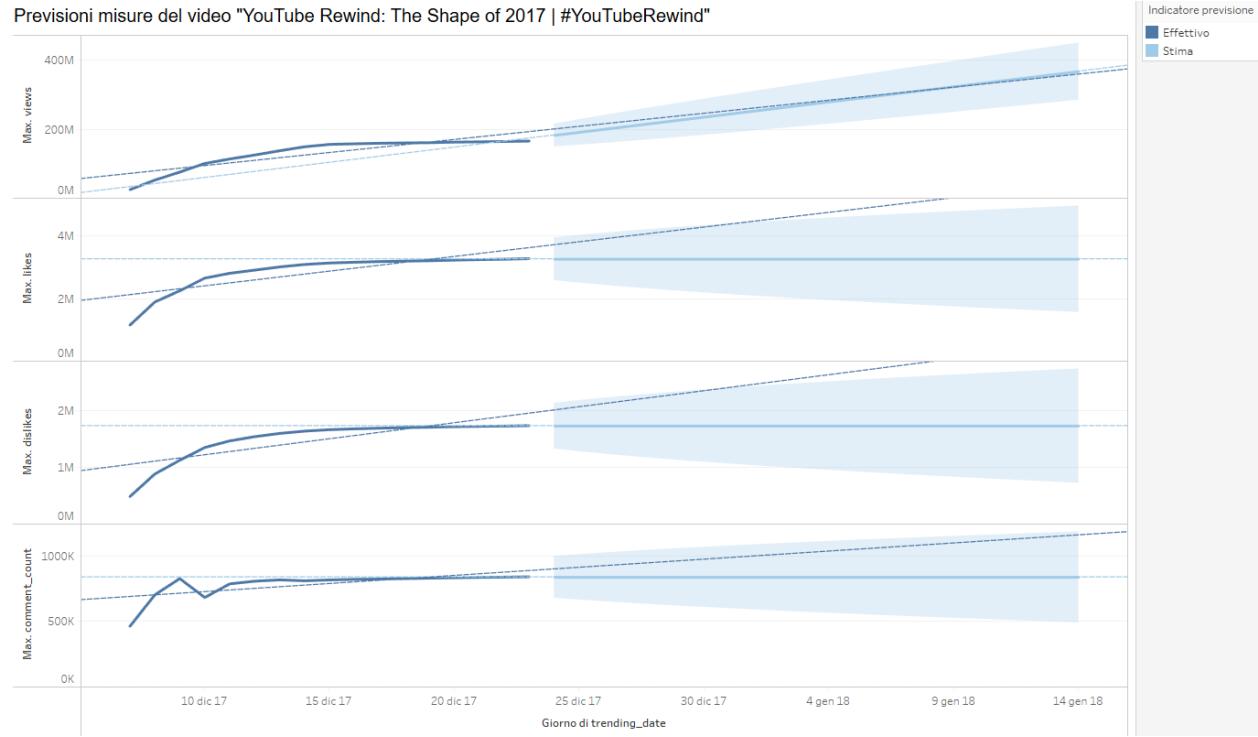


Figura 41: Previsione delle misure del video ”YouTube Rewind: The Shape of 2017 - YouTubeRewind”

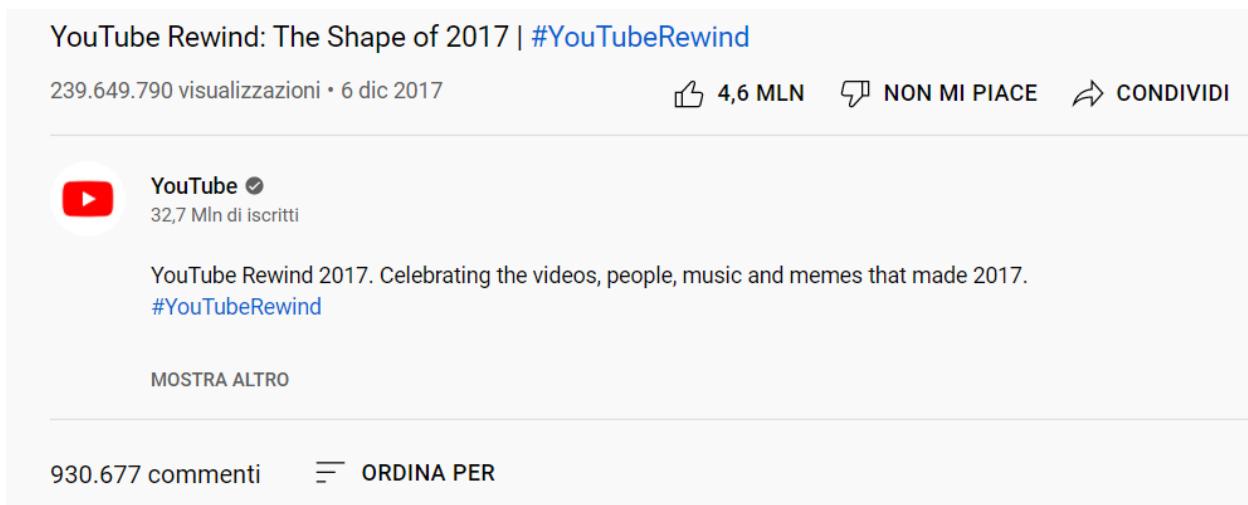


Figura 42: Misure del video "YouTube Rewind: The Shape of 2017 - YouTubeRewind" alla data del 04/02/2022"

4 PowerBI

Power Bi è l'ultimo strumento che è stato utilizzato per analizzare i dati. Il software è una raccolta di servizi software, app e connettori che interagiscono per trasformare i dati non correlati in un insieme di informazioni coerenti, interattive e graficamente gradevoli all'occhio. Attualmente, secondo il magic quadrant di Gartner, nel suo campo Power BI è il migliore tool in circolazione, sia per *"Completeness of vision"* che per *"Ability to execute"*. In questa sezione verranno replicate ed aggiunte alcune tipologie di analisi descrittive e predittive, fatte anche con Qlik e Tableau. Una caratteristica che distingue questo strumento dagli altri è la possibilità di importare grafici aggiuntivi attraverso una specie di marketplace accessibile direttamente dall'interfaccia dell'applicazione. Alcuni di questi strumenti richiedono un'installazione manuale delle librerie R utilizzando Rstudio. Infine, Power BI, dato che supporta la versione cloud, permette di lavorare da qualsiasi dispositivo dato che i dataset sono stati caricati nel cloud proprio come è stato fatto con Qlik.

4.1 Caricamento dati e ETL

Inizialmente è stato scritto uno script in python che legge tutti i file csv e li concatena assieme producendo un unico file di output, in cui i video appartenenti ai diversi paesi differiscono per il campo country. Successivamente, i dati sono stati caricati in Power BI per essere trasformati. In particolare, nei due csv risultanti, attraverso il linguaggio Power Query siamo andati a filtrare i dati e rimuovere tutte le inconsistenze presenti, in modo tale da avere due file completamente privi di errori. I filtri applicati sono stati:

- impostazione della seconda riga come intestazione del dataset;
- modifica del tipo e rimozione di alcune colonne;
- filtraggio delle righe vuote;
- normalizzazione di alcune colonne per correggere valori errati.

Infine, finita la fase di trasformazione, i dati sono stati caricati all'interno di Power BI per essere analizzati dettagliatamente.

4.2 Analisi e descrizione delle visualizzazioni

Le analisi effettuate sono di tipo descrittivo, diagnostico ed anche predittivo. Le dashboard che sono state realizzate in PowerBI, ricalcano in parte le viste realizzate in precedenza. Tuttavia, sono state apportate delle modifiche importanti, introducendo anche delle nuove viste, usate per previsioni di serie temporali. Le dashboard, che verranno descritte più in dettaglio nelle sezioni successive, includono:

- una vista generale;
- una vista dedicata all'analisi delle categorie;
- una vista dedicata all'analisi dei canali;
- una vista sul numero di video in tendenza per paese;
- una grafico a mappa sulle visualizzazioni per paese e categoria;
- una vista sulle parole più utilizzate;
- un grafico ad albero sulle views;
- una vista sui fattori di influenza sulle views;
- una vista sulla media dei likes per country in funzione del tempo;

- una vista sulle metriche principali in base al paese;
- una vista sulle correlazioni tra le misure;
- una vista sui fattori di influenza sulle views;
- una vista che fornisce diversi tipi di previsioni sul numero di video che andranno in tendenza.

4.2.1 Vista generale

In questa dashboard sono riportati i 4 principali KPI (numeri di video, media di likes, views e commenti), una breve analisi temporale relativa all'anno e al mese di pubblicazione e i video che sono rimasti di più in tendenza. Per quanto riguarda i 4 KPI, si confermano i valori ottenuti anche con Qlik e Tableau. Per l'analisi temporale sulla pubblicazione dei video, si osserva che i 3/4 dei video sono stati pubblicati nel 2018, il restante 1/4 nel 2017 e solo in piccolissima parte nel 2016 e altri anni precedenti. Nell'analisi temporale per mese si nota che la maggior parte dei video viene pubblicata nel periodo primaverile, mentre nel periodo giugno-ottobre vi è un drastico calo dovuto al fatto che il dataset contiene i video in tendenza nel periodo tra novembre 2017 e giugno 2018 e dato che nella maggior parte dei casi i video in tendenza sono stati pubblicati pochi giorni prima, questa distribuzione temporale viene riflessa anche in questo grafico. Per quanto riguarda il tempo in tendenza dei video, a destra vi è una top 10 dei video che vi sono rimasti per più tempo, mentre al centro si osserva che la media dei giorni in tendenza per ogni video è di 15 giorni.

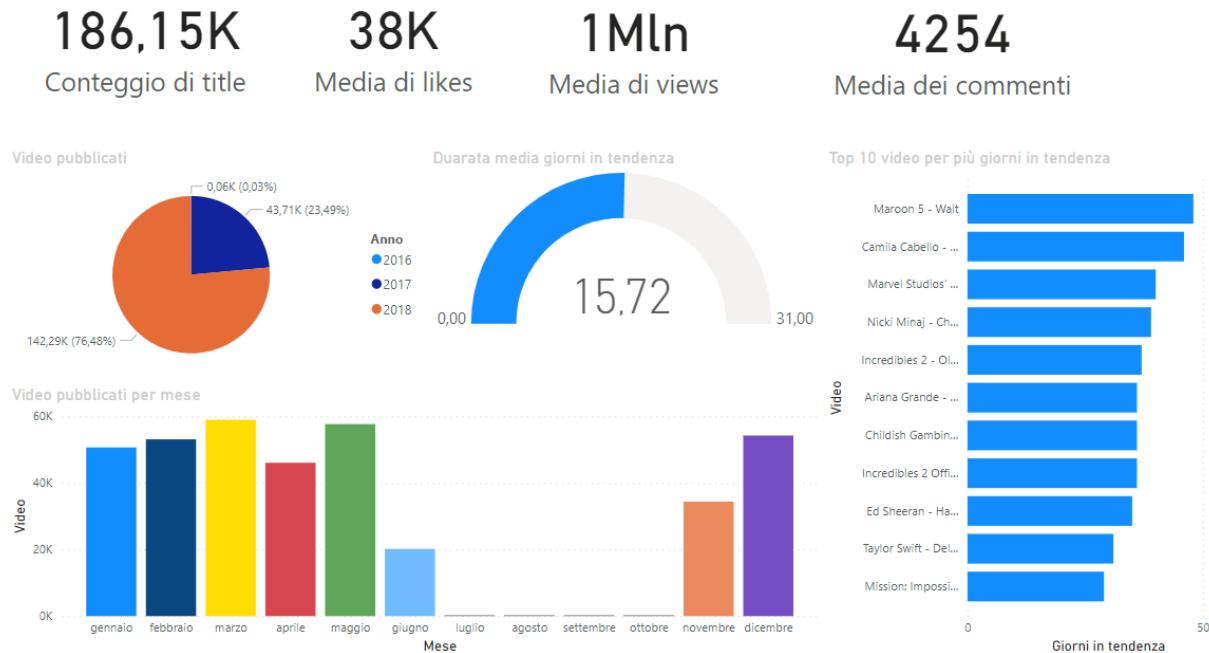


Figura 43: KPI e descrizione del dataset

4.2.2 Analisi su categorie

La dashboard sull'analisi delle categorie è stata ristrutturata utilizzando alcune nuove visualizzazioni offerte da Power BI. Il grafico in alto a sinistra mostra l'engagement delle categorie, ovvero quali sono le categorie che hanno ricevuto un maggior numero di likes, dislikes, commenti e views. Dal grafico emerge, quindi,

che la categoria con maggior interazioni è quella di *"Music"*. Il grafico a nastro in basso a sinistra mostra la media delle visualizzazioni che riceve ogni categoria rispetto al mese in cui va in tendenza. Qui si può notare come sia ancora la categoria *"Music"* ad avere una media delle views più elevata, il quale, però, diminuisce drasticamente nei mesi di novembre e dicembre. Invece, nella parte destra della dashboard, vengono riportati un grafico ad anello e un tabella che cercano di evidenziare rispettivamente alcuni numeri riguardanti le categorie, in modo tale da offrire una panoramica generale sui principali indicatori di analisi sulle categorie dei video.

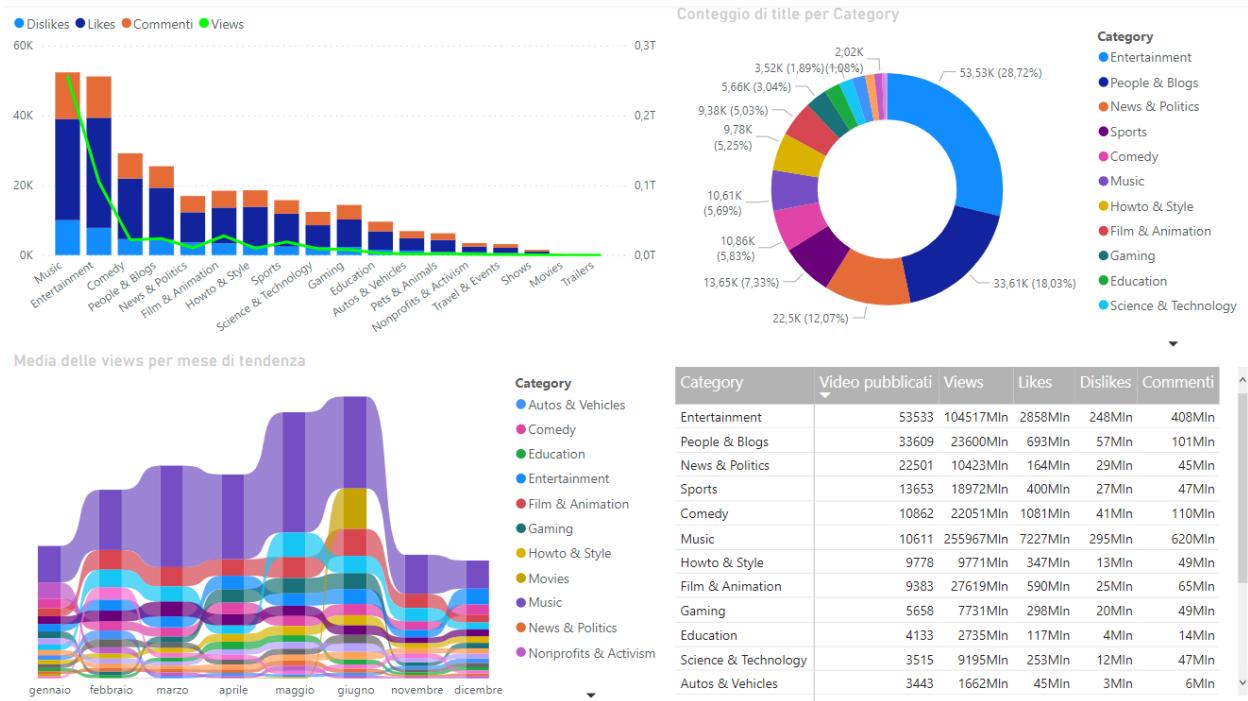


Figura 44: Analisi su categorie

4.2.3 Analisi su canali

La dashboard sui canali riporta alcuni grafici che evidenziano quali sono i canali in tendenza nel periodo preso in considerazione. In basso, i grafici a torta mostrano i video che hanno ricevuto più visualizzazioni e più likes. E' possibile notare che il video con il maggior numero di views non coincide con il video che ha il maggior numero di likes. Quindi, non possiamo affermare che esiste una relazione tra views e likes, ovvero non sempre il video che riceve più visualizzazioni è anche quello che riceve più likes, o viceversa. Il grafico a dispersione in alto a destra mostra come sono distribuite le visualizzazioni sui video pubblicati dai diversi canali YouTube. Vicino all'origine del grafico troviamo una forte densità dei punti, questo significa che i canali hanno pubblicato pochi video, ma soprattutto quei video hanno ricevuto poche visualizzazioni. Al contrario, sui punti isolati, troviamo un elevato numero di video che hanno ricevuto molte visualizzazioni. Per confermare quanto detto, affianco viene riportata una tabella che evidenzia tali dati, evidenziando, per ogni canale, i video pubblicati, le visualizzazioni e i likes.

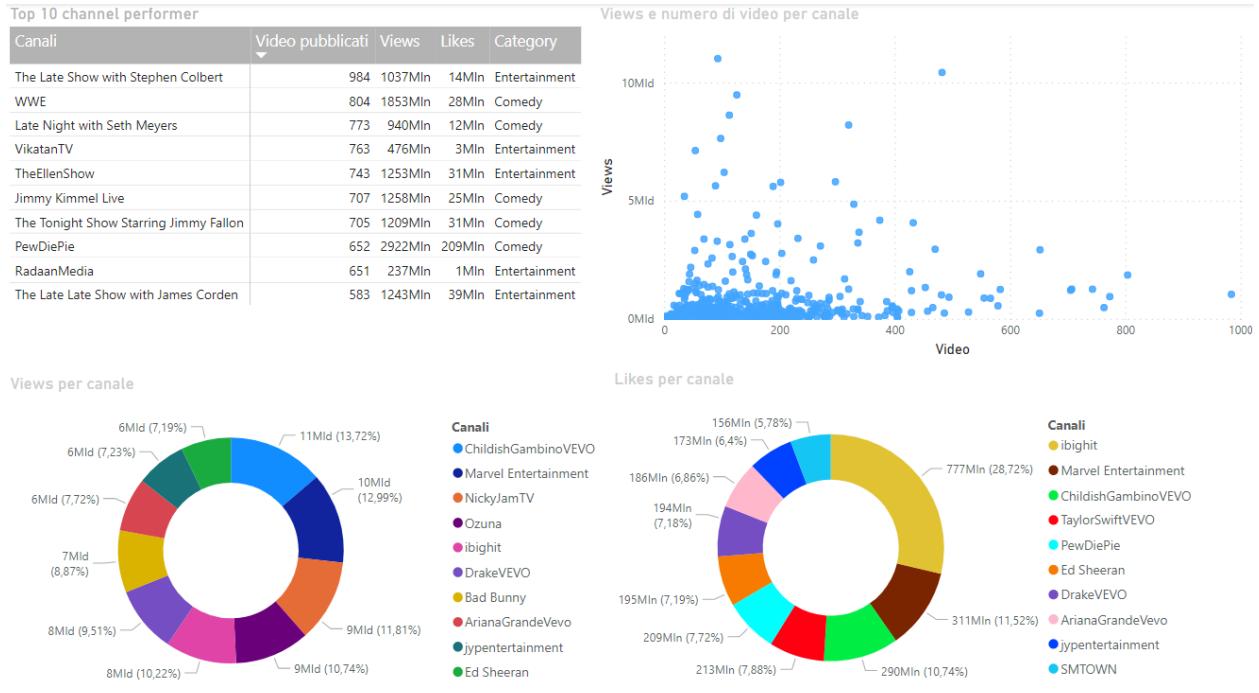


Figura 45: Analisi su canali

4.2.4 Numero di video in tendenza per paese

In questo paragrafo si va ad analizzare il numero di video che vanno in tendenza sui differenti paesi. Si può notare che come nelle altre analisi, è sempre la Russia a dominare la classifica, con 34295 video, seguita dal Messico con 33541 video e la Francia con 30005 video. Nonostante la Russia sia al primo posto, per numero di video in tendenza, non lo è in rapporto alla sua popolazione. Infatti in figura 47 è stato analizzato il rapporto video/popolazione per ogni paese, dal quale emerge che il Canada essendo anche uno dei paesi meno popolati, è il primo secondo questa classifica. Viene seguito poi subito dopo dalla Francia e dalla Germania, che ricoprivano già alte posizioni nella classifica precedente. Inoltre è interessante notare come l'India sia veramente molto popolata rispetto al numero di video in tendenza, ci sono solamente 12 video in tendenza per ogni milione di persone di popolazione.

Conteggio di title per country

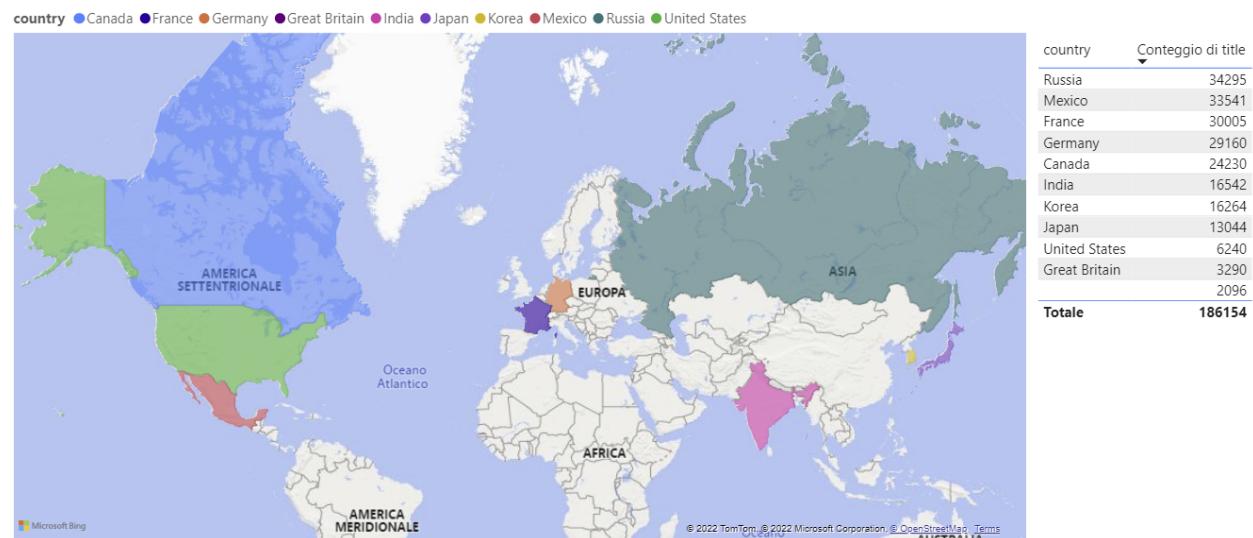


Figura 46: Numero di video in tendenza per paese

| Paese | Numero video | Popolazione (milioni) | Rapporto video/popolazione |
|---------------|--------------|-----------------------|----------------------------|
| Canada | 24230 | 38 | 637,63 |
| Francia | 30005 | 67 | 447,84 |
| Germania | 29160 | 83 | 351,33 |
| Korea | 16264 | 51 | 318,90 |
| Messico | 33541 | 128 | 262,04 |
| Russia | 34295 | 144 | 238,16 |
| Japan | 13044 | 125 | 104,35 |
| Gran Bretagna | 3290 | 67 | 49,10 |
| Stati Uniti | 6240 | 329 | 18,97 |
| India | 16542 | 1380 | 11,99 |

Rapporto video/popolazione per Paese

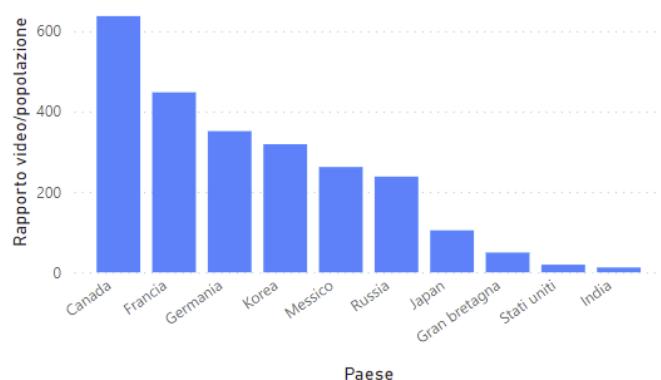


Figura 47: Rapporto video/popolazione per paese

4.2.5 Visualizzazioni per paese e categoria

Nel grafico a mappa sottostante viene riportata una visualizzazione delle categorie con più visualizzazioni con un apposito filtro, in base al paese di pubblicazione. Come in precedenza su Tableau, i diversi colori della dashboard rappresentano i diversi paesi, mentre la dimensione dei riquadri riguarda la dimensione delle visualizzazioni che riceve ogni categoria. Come si può notare, per ogni paese le due categorie dominanti sono *"Entertainment"* e *"Music"* che si alternano il primo e secondo posto, a seguire poi vengono le altre categorie.

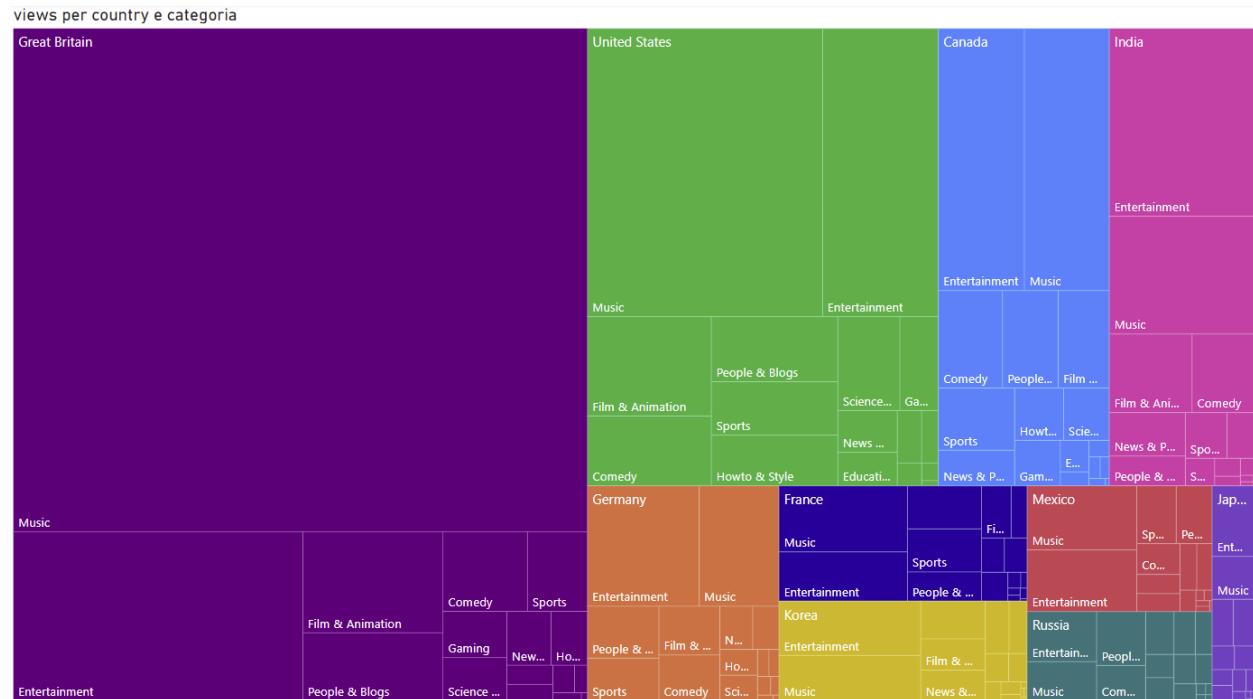


Figura 48: Visualizzazioni per paese e categoria

4.2.6 Analisi delle parole più utilizzate

In questa dashboard è stato scaricato l'oggetto visivo "Word cloud" dall'app source ufficiale di Power BI. Questo oggetto prende in input i campi di testo e individua quali sono le parole più frequenti. In particolare, sono stati analizzati i titoli dei video, le descrizioni, i tag e i nomi dei canali. Nell'analisi del titolo le parole che sembrano avere maggior rilievo sono *"2018"*, *"ENG"*, *"BTS"* e *"de"*. Nei tag compare sempre la parola *"2018"*, ma anche *"2017"* e *"Trump"* (dovuto forse al fatto che i video in tendenza riguardano proprio gli anni 2018 e 2017, date che coincidono anche con l'inizio della carica di Trump). Nelle descrizioni invece compaiono le parole *"https"*, *"www"* e *"com"* e ciò può significare che molto spesso in descrizione al video vengono riportati dei link. Nel titolo del canale le parole più utilizzate sono *"TV"*, *"news"*, *"Channel"* e *"Official"*, che sono effettivamente termini che hanno senso in questo contesto.

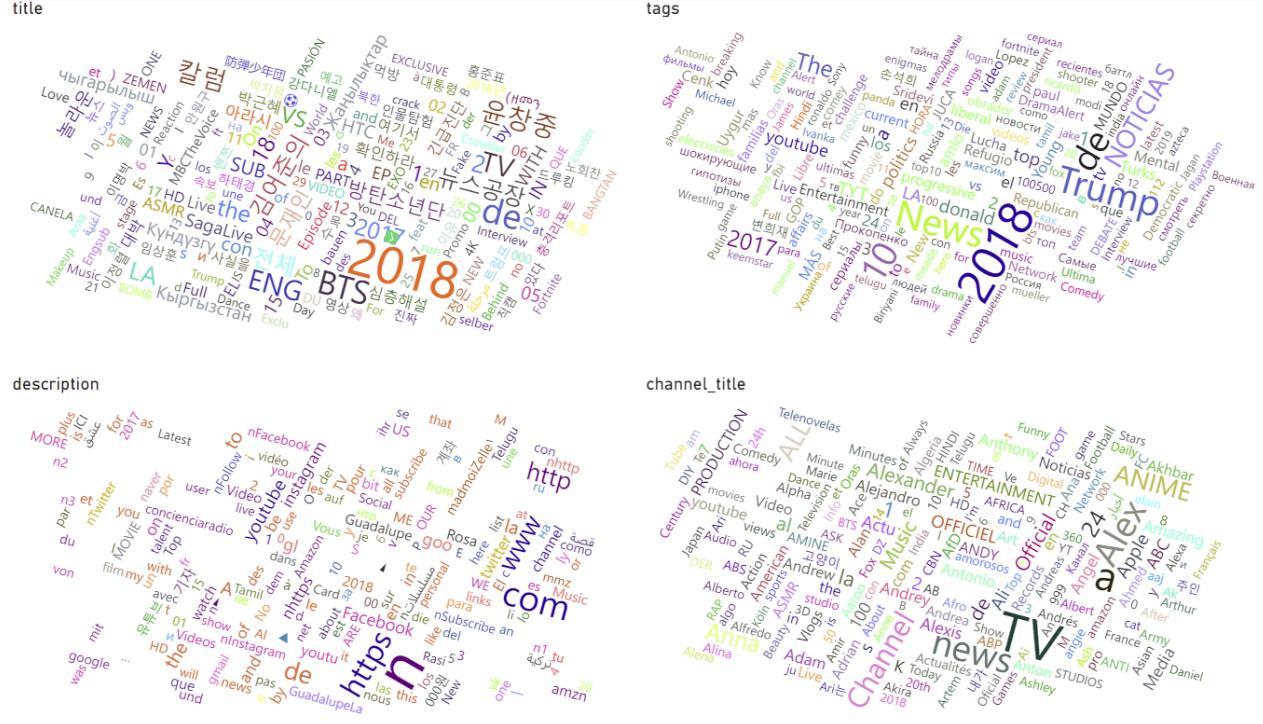


Figura 49: Analisi delle parole più utilizzate

4.2.7 Analisi ad albero delle views

In questa dashboard viene utilizzato un albero per analizzare le views. Nel primo livello dell'albero, le views vengono divise per paese. Una volta selezionato un paese, si può suddividere ancora per categoria e infine per nome del canale. La gerarchia dell'albero è facilmente intercambiabile e questo è molto utile quando si vuole analizzare una misura in diversi raggruppamenti, facendo così anche operazioni di drill down e roll-up.

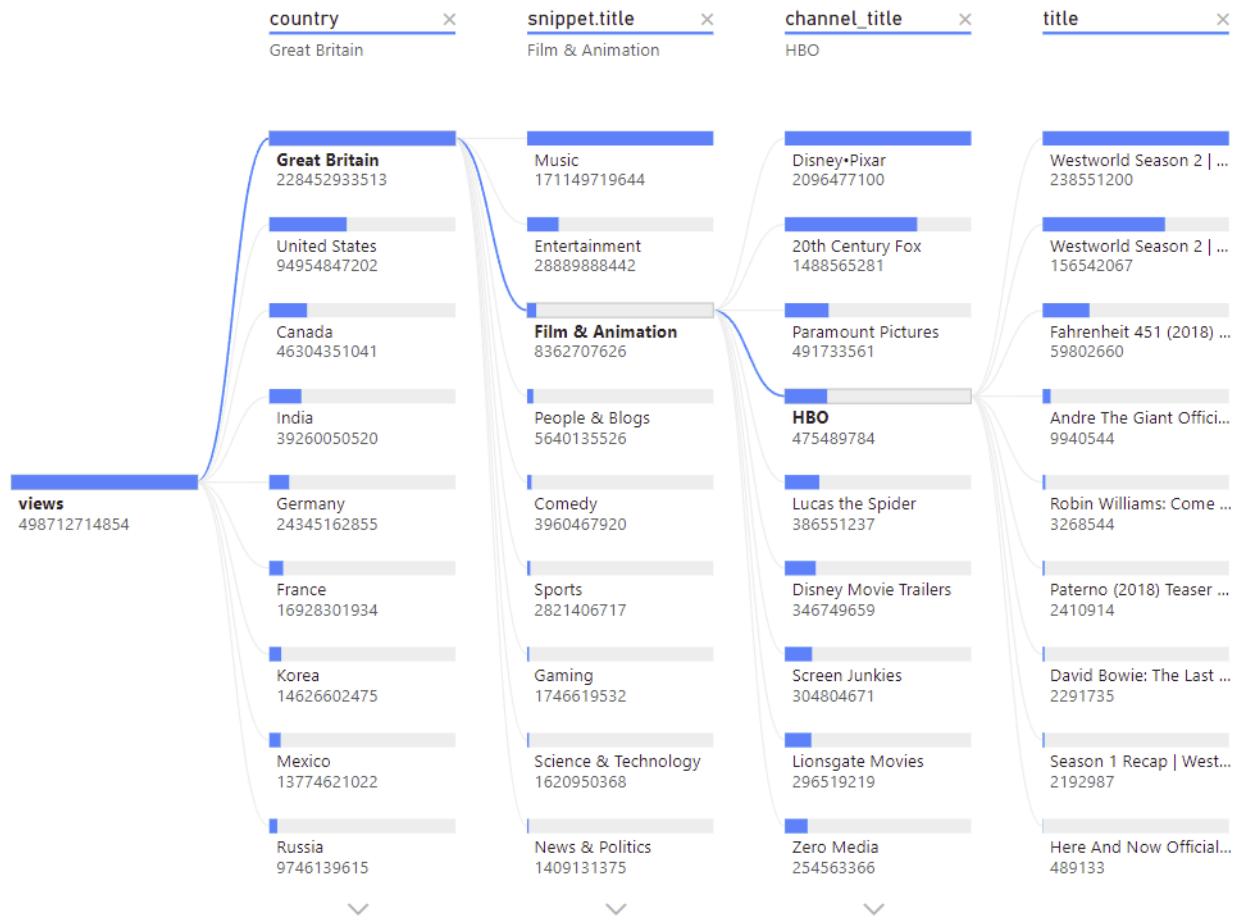


Figura 50: Analisi ad albero delle views

4.2.8 Fattori di influenza sulle views

Grazie a questo tool si possono individuare i fattori di influenza chiave per l'aumento delle views. In particolare, il tool afferma che quando i like aumentano di 170 mila, la media di views aumenta di circa 7 milioni. Inoltre mediamente in Gran Bretagna la media di views è più alta di 5,5 milioni rispetto agli altri paesi. Un aspetto particolare che emerge è che quando il numero di commenti è più basso di 20 mila rispetto alla media, allora si registrano anche 3 milioni di visualizzazioni in più.

Fattori di influenza chiave Segmenti principali



Fattore che influisce su views in modo che sia Aumenta ▾ ?

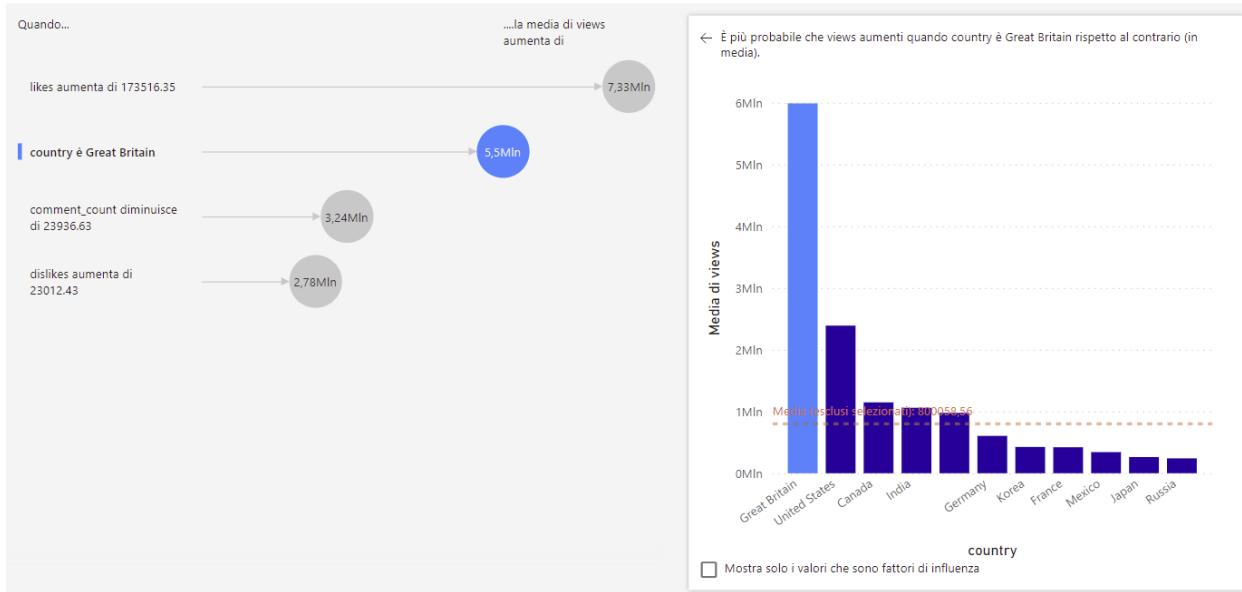


Figura 51: Fattori di influenza sulle views

Questo tool esegue anche operazioni di segmentazione. Dai dati selezionati, sono stati trovati 2 segmenti che sono stati classificati per media di views rispetto a tutto il dataset.

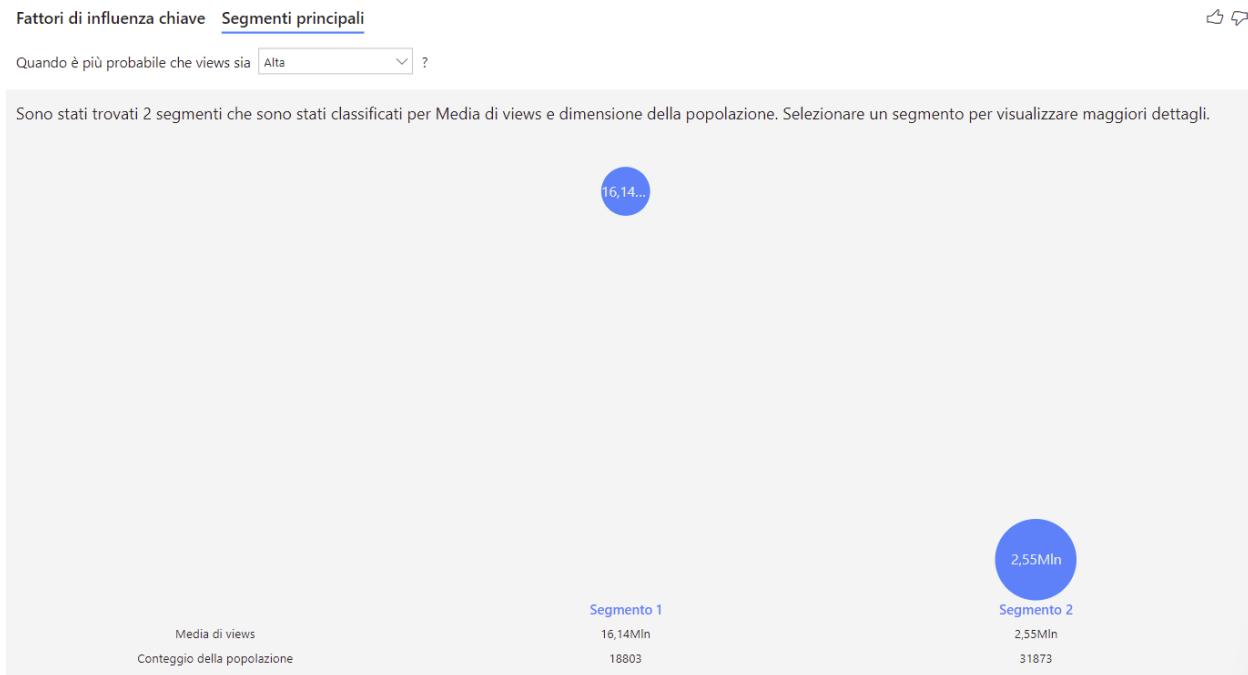


Figura 52: Segmenti principali

Il segmento 1 contiene 18803 video (per un 5% sul totale) che hanno un numero di like maggiore di 161929. In questo segmento il valore della media delle views è di 16 milioni, quasi 15 milioni in più rispetto alla media.

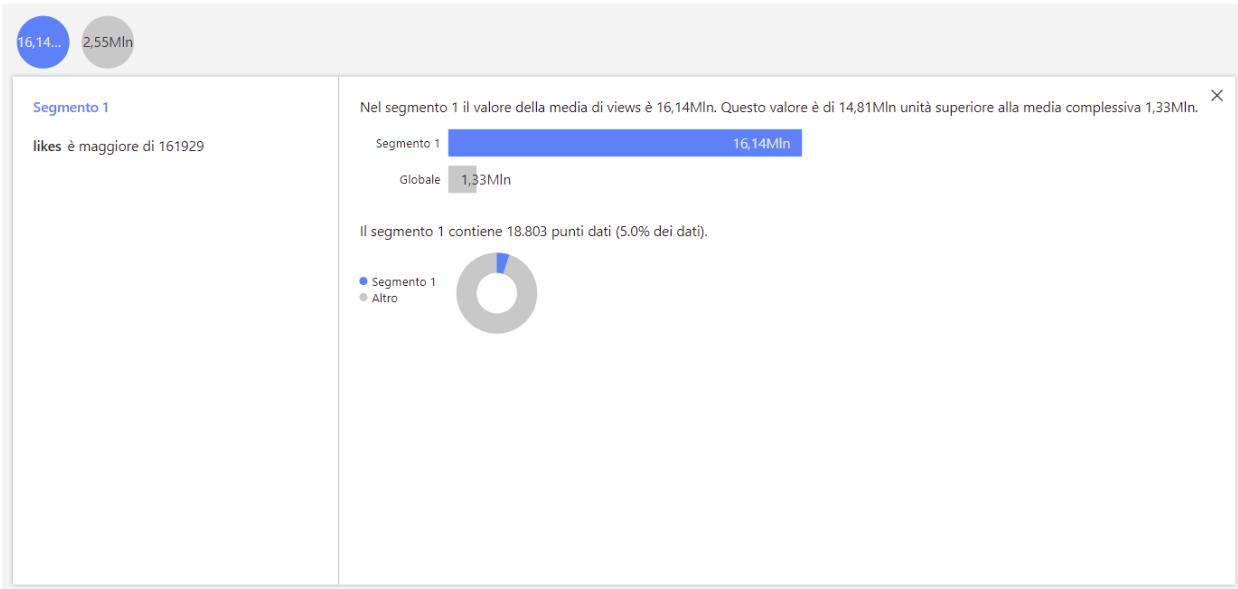
Quando è più probabile che views sia ?

Figura 53: Segmento 1

Il segmento 2 contiene 31873 video (per un 8.5% sul totale) che hanno un numero di like compreso tra 46707 e 161929. In questo segmento il valore della media delle views è di 2.55 milioni, 1.22 milioni in più rispetto alla media complessiva di 1.33 milioni.

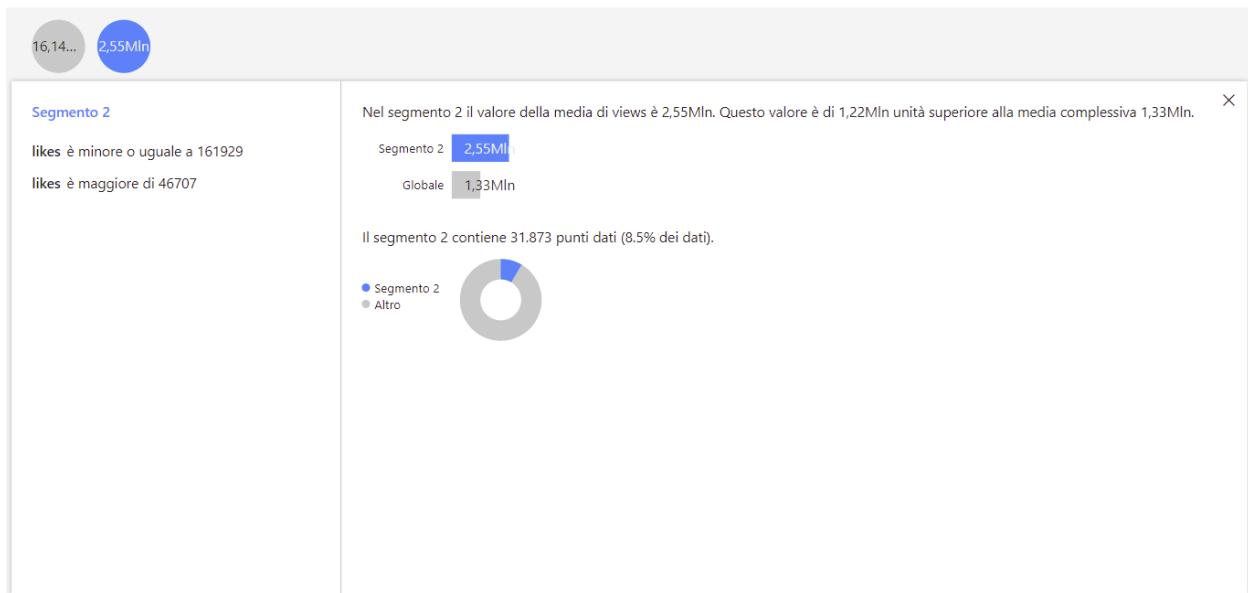
Quando è più probabile che views sia ?

Figura 54: Segmento 2

4.2.9 Media dei like per country nel tempo

In questo grafico viene riportato un istogramma che indica la media dei likes ricevuti dai video in funzione del mese in cui vanno in tendenza. La legenda mostra che ad ogni paese è associato un colore differente. L'andamento mostra un aumento generale dei likes con il passare del tempo. E' visibile sul grafico un picco elevato tra i mesi di maggio 2018 e giugno 2018, il quale riguarda il paese del Giappone. Dalla dashboard emerge che i paesi con la distribuzione dei likes più elevata sono la Gran Bretagna e gli Stati Uniti.

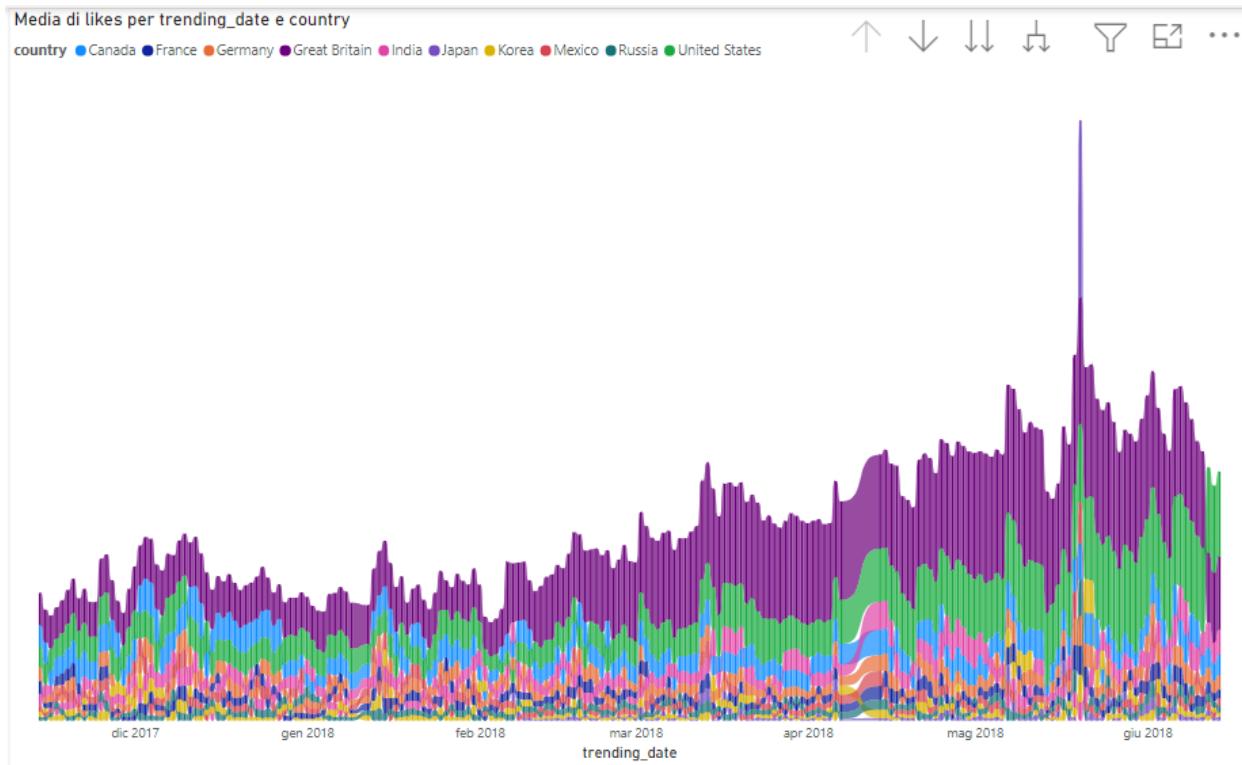


Figura 55: Media dei like per country nel tempo

4.2.10 Analisi delle misure per paese

In questa dashboard andiamo a mostrare una visualizzazione differente sulle principali misure per paese, in particolare vengono riportati quattro grafici ad imbuto che descrivono le quantità di likes, dislikes, commenti e visualizzazioni in ogni paese. Si può notare come la forma piramidale emerge in tutti i grafici tranne nel grafico in basso a sinistra, dato che si ha una distribuzione non uniforme. Mentre il grafico della quantità dei commenti per paese ha un andamento uniforme e questa è un'osservazione interessante.

Come si può notare dai grafici, il paese che ha il numero maggiore di likes, dislikes e views è la "Gran Bretagna", mentre per quanto riguarda il numero di commenti, al primo posto si ha il "Canada" e la "Russia", con le stesse quantità.

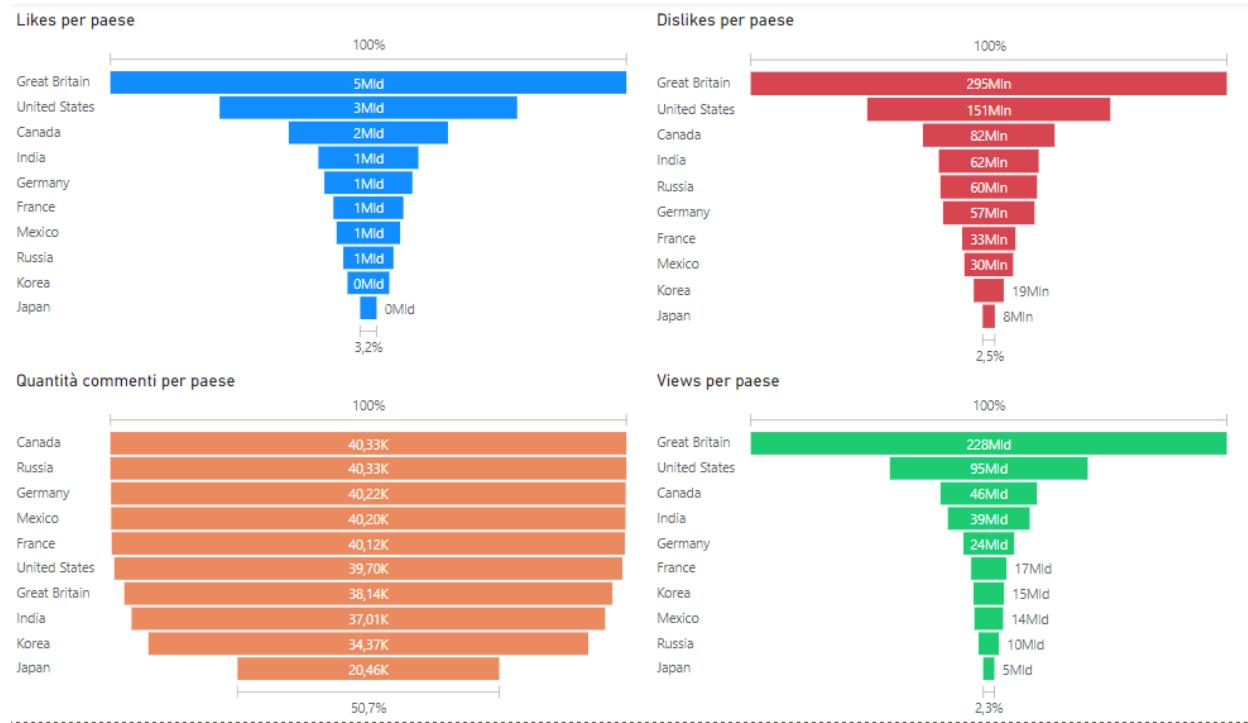


Figura 56: Analisi delle misure per paese

4.2.11 Correlazioni tra le misure

In questa dashboard vengono mostrati quattro grafici a dispersione in cui ogni punto colorato rappresenta la correlazione tra le principali misure di un video stesso, in particolare le più significative sono le correlazioni tra: likes e views, commenti e views, likes e dislikes, e dislikes e views.

A differenza di Tableau, Power BI non permette di specificare il modello con cui calcolare la tendenza avendo a disposizione solo il modello lineare, inoltre non permette di calcolare p-value e Rsquared direttamente.

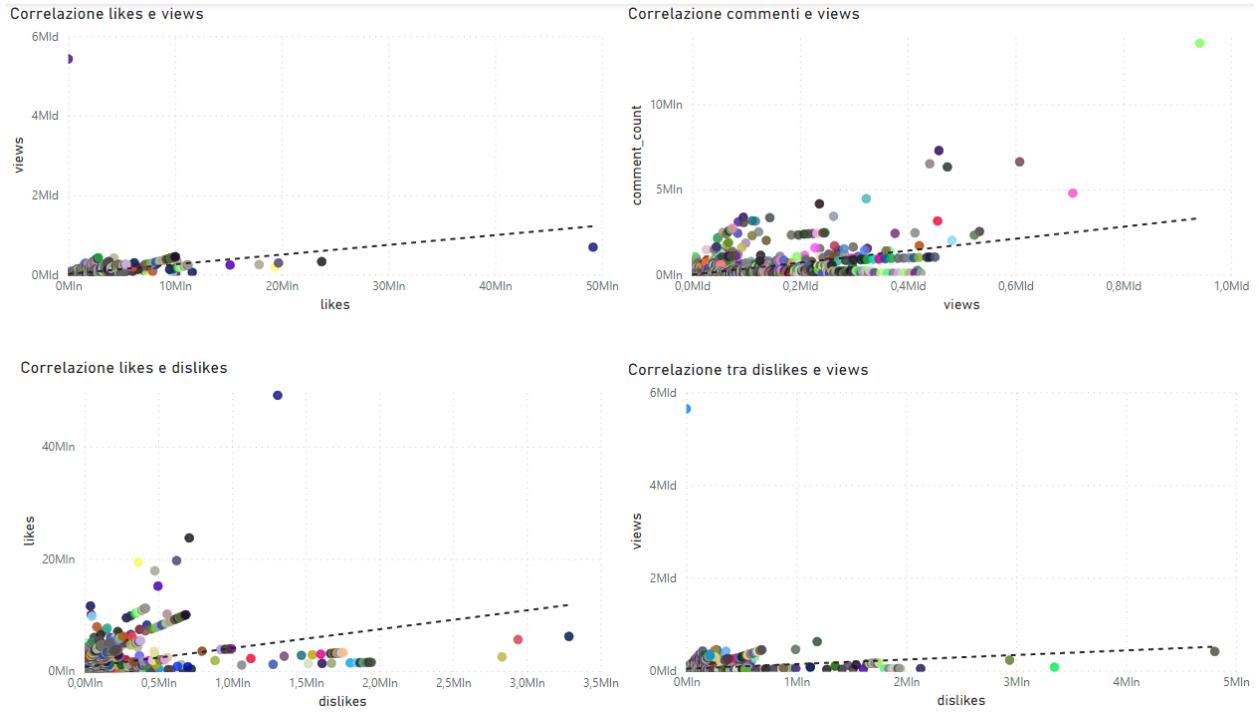


Figura 57: Correlazioni

4.2.12 Previsione del numero di video in tendenza

Per quanto riguarda le previsioni relative alle serie temporali, si sono dedicate tre viste, ognuna dei quali contiene un grafico che effettua previsione della serie temporale fino al mese successivo, ed un secondo, più descrittivo, che raffigura una sua decomposizione nelle componenti di trend e stagionalità. Le tre dashboard rappresentano rispettivamente: la previsione del numero di video in tendenza con il semplice forecasting di Power BI, con ARIMA e la previsione con l'algoritmo Exponential Time Smoothing (ETS), utilizzato anche da Tableau, che in questo caso fornisce delle buone previsioni, catturando in maniera appropriata sia la stagionalità della serie temporale che il suo trend.

Il primo grafico di forecasting (figura 58) è stato realizzato dal pannello analisi dell'oggetto visivo, opzionando la scheda forecasting. In essa sono state impostate le unità come mesi, mentre la lunghezza delle previsione è di 1 unità. Infine, l'intervallo di confidenza è stato impostato al 95%. Il risultato ottenuto mostra che la serie temporale è abbastanza irregolare, ciò è probabilmente dovuto all'alta variabilità dei video che vanno in tendenza, quindi una caratteristica abbastanza aleatoria. Invece, la decomposizione della serie temporale è stata eseguita grazie ad un tool di Power BI denominato "Time Series Decomposition". La decomposizione di serie temporali è il processo di acquisizione dei dati di serie temporali e di separazione in più componenti sottostanti. Nel nostro caso, suddivideremo le serie temporali in componenti trend e stagionale. La componente trend è utile per dirci se le nostre misurazioni aumentano o diminuiscono nel tempo. La componente stagionale è utile per dirci quanto pesantemente le nostre misurazioni siano influenzate da intervalli di tempo regolari. Possiamo vedere che il trend della serie mostra un andamento crescente, questo ci dice che il numero di video pubblicati che entrano in tendenza aumenta con il passare del tempo, in particolare esso ha una lieve crescita a partire dal mese di febbraio. Invece, dalla stagionalità della serie, è possibile osservare alcuni picchi nei mesi di gennaio e aprile mentre si ha una forte presenza di picchi negativi in prossimità dei mesi dicembre, marzo e luglio.

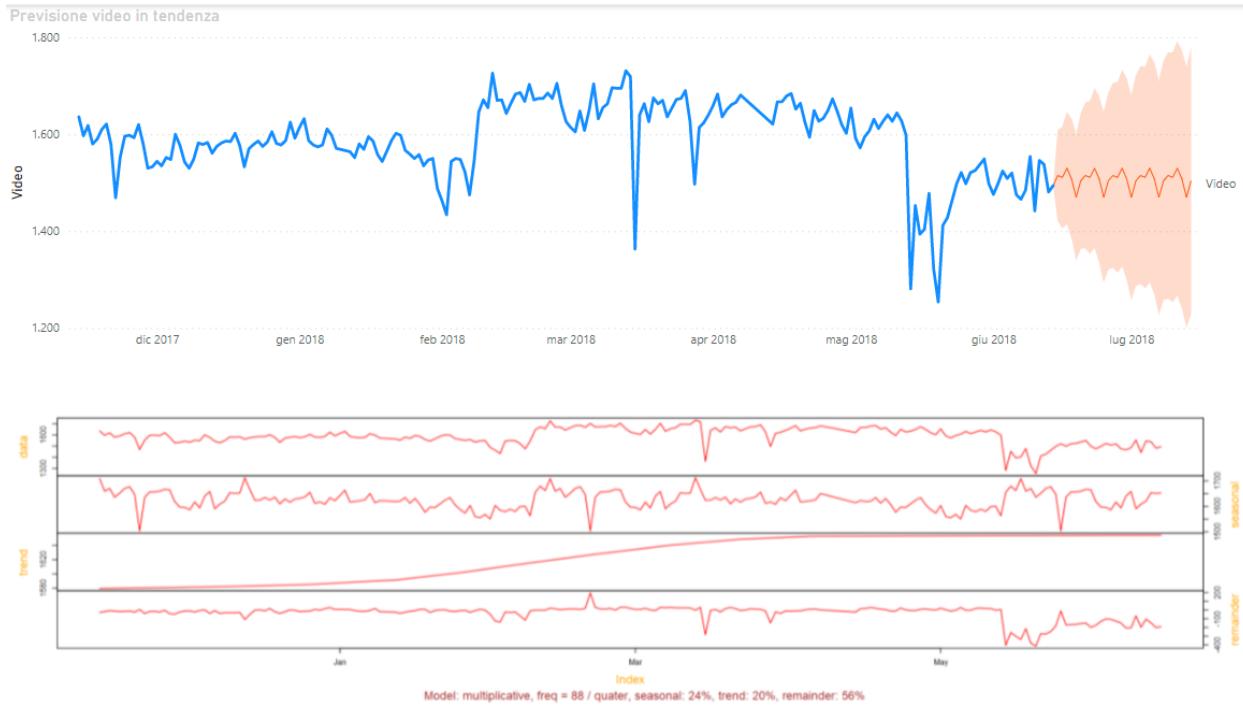


Figura 58: Previsione del numero di video in tendenza

Nella seconda dashboard (figura 59) relativa alle previsioni, si è utilizzato l'algoritmo ARIMA con parametri $p=2$, $d=1$, e $q=0$. Il primo grafico, infatti, mostra una previsione che cattura l'andamento della serie sempre nel periodo di tempo preso in considerazione. Anche qui è presente una variabilità molto alta. La previsione ottenuta con ARIMA non sembra essere molto accurata con i parametri sopra descritti. Dall'altra parte però riesce a catturare abbastanza bene il trend e la stagionalità che rimangono simili a quelli ottenuti nella previsione precedente.

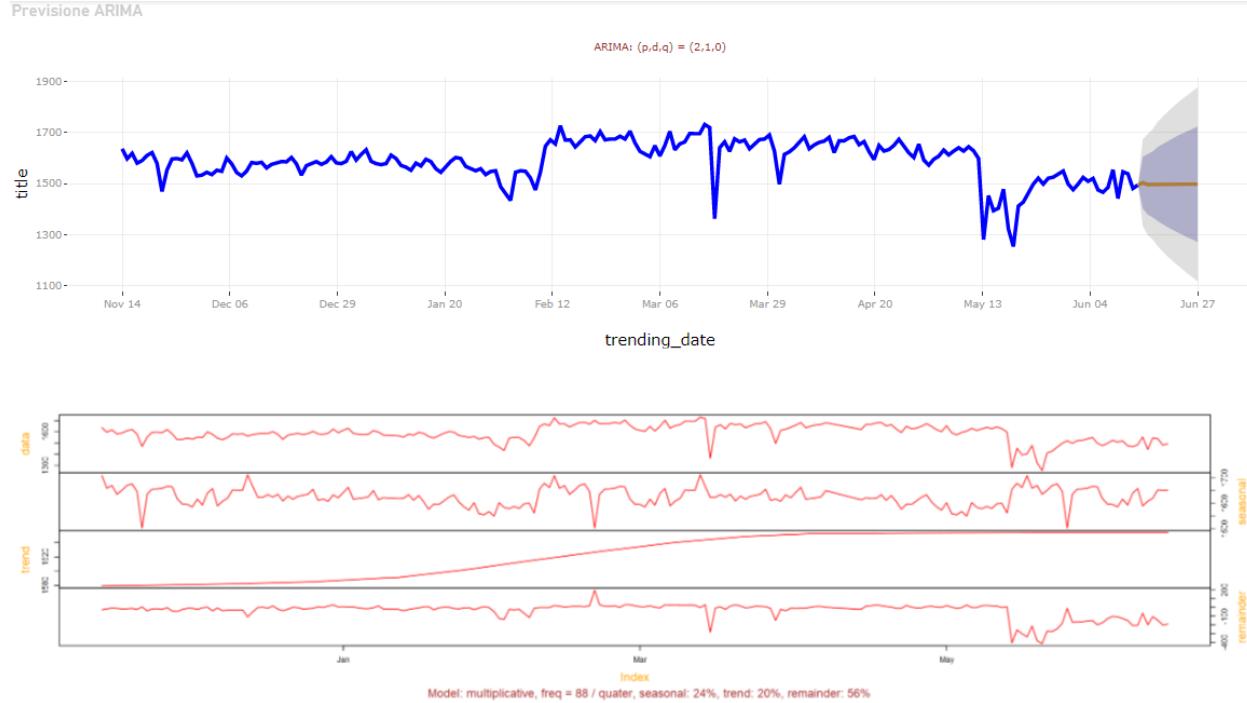


Figura 59: Previsione del numero di video in tendenza con ARIMA

Nella terza dashboard (figura 60), la previsione fa uso dell'algoritmo Exponential Time Smoothing (ETS). Esso è un algoritmo statico locale frequentemente utilizzato per previsioni con serie temporali. I risultati ottenuti dal forecasting con ETS sembrano essere migliori di quelli ottenuti con ARIMA, in quanto la previsione sembra più accurata. Lavorando, anche in questo caso, con una decomposizione delle stagionalità per giorni, è possibile notare che essa ha sempre gli stessi picchi positivi e negativi i corrispondenza degli stessi mesi. Anche il trend della serie temporale rimane invariato, infatti dal grafico possiamo vedere come anche in questo caso esso aumenti.



Figura 60: Previsione del numero di video in tendenza con ETS

5 Conclusioni

Nel corso di questo progetto sono state effettuate varie analisi descrittive e diagnostiche (come quelle che motivano particolari picchi nei grafici). E' stata effettuata anche una analisi predittiva di segmentazione dei dati. Nel dataset i dati a nostra disposizione riguardano solo i video in tendenza e ciò non consente di fare interessanti confronti fra video in classifica e non. Dalle diverse analisi è risultato che la piattaforma YouTube è utilizzata principalmente per video di musica e intrattenimento. Nella maggior parte dei casi, in video vanno in tendenza per pochi giorni, e questo si verifica subito dopo la pubblicazione sulla piattaforma. Sono stati effettuati vari tentativi per cercare delle correlazioni particolari tra le varie misure a disposizione (views, likes, dislikes e numero di commenti) ma sono stati ottenuti risultati poco significativi. Sono state effettuate anche analisi predittive, in particolare è stato utilizzato il forecasting classico sia su Tableau che Power BI, e in quest'ultimo è stato utilizzato anche l'algoritmo ARIMA che ha ottenuto risultati migliori. Sfortunatamente il dataset non si presta molto ad analisi di tendenza, perché le misure non hanno una tendenza e non dipendono solo dal tempo.

Dopo l'utilizzo in maniera molto approfondita dei tre tools, concludiamo con delle brevi valutazioni su di essi. Qlik risulta essere semplice ed intuitivo sia nella parte di caricamento dei dati che nell'utilizzo. Esso mette a disposizione una serie di fogli con cui si possono costruire delle dashboard chiare e di facile comprensione. Qlik riesce a trasmettere nella maniera più efficace la visione d'insieme, in quanto filtrando un qualche attributo su una dashboard, il filtro viene mantenuto per tutte le altre dashboards. Quindi, possiamo dire che ha diversi punti di forza e di debolezza, ma messo a confronto con altri software di data analytics risulta essere il meno potente. Infatti, in questo progetto, esso è utilizzato solo per effettuare analisi descrittive.

Tableau è un software molto potente, non a caso si trova sopra Qlik nel magic quadrant di Gartner, fornisce velocità di analisi, funziona su tutti i dispositivi in cui riescono ad arrivare i dati, permette di colla-

borare, ma ha un'interfaccia grafica che risulta essere poco intuitiva. Per realizzare un app in Tableau, ogni grafico viene sviluppato su un foglio diverso ed infine si costruiscono le dashboards raggruppando insieme specifici fogli.

Microsoft Power BI, infine, risulta essere il miglior compromesso tra i due. Fornisce un'interfaccia stile suite Office, quindi più familiare; le operazioni di ETL sono facilmente manipolabili grazie allo strumento Power Query, il quale permette di trasformare e pulire i dati in modo semplice ed efficace. Tra i tre software, Power BI è sicuramente il migliore, infatti occupa la posizione più in alto nel magic quadrant di Gartner. Esso non si limita a fornire strumenti per l'analisi descrittiva, predittiva e diagnostica, ma offre anche la possibilità di estendere di molto le proprie funzionalità grazie all'integrazione nativa con R, ed alla possibilità di scaricare tools perfettamente integrati con l'ambiente dal Marketplace proprietario. Il vero problema di Power Bi risiede nella sua capacità di effettuare forecasting, infatti appoggiandosi a strumenti esterni l'integrazione molto spesso non è molto esatta. Inoltre, rispetto a Tableau, non permette di poter capire se la predizione effettuata sia affidabile. Possiamo comunque concludere che, data la potenza messa a disposizione e le sue capacità pressoché infinite grazie all'integrazione di script esterni, è il tool più completo tra i tre.

Riferimenti bibliografici

- [1] Guida di youtube: https://support.google.com/youtube/answer/7239739?hl=it&ref_topic=9257501
- [2] Dataset: www.kaggle.com/datasneak/youtube-new