**PROGETTO SANREMO**

Link progetto:

<https://lookerstudio.google.com/reporting/e13417a7-7e64-46ee-b39c-7b54c912dc4d>

Schema di lavoro

Per questo progetto ho preso in considerazione tutto il materiale che ci è stato fornito.

Ho pensato di suddividere le analisi in: analisi dei vincitori, analisi dei partecipanti, analisi dei conduttori, analisi del podio, analisi dei dati Spotify e analisi Auditel.

Per la pulizia dei dati ho utilizzato Power Query ed Excel, e inoltre mi sono aiutata con l’AI per alcuni problemi di conversione dei dati, soprattutto nel file dei dati Auditel e di Spotify.

La mia analisi parte da Sanremo 1951 fino a Sanremo 2023, mentre per quanto riguarda gli ascolti i dati si hanno solo dal 1987, anno in cui l’auditel ha iniziato a registrarli.

Spiegazione step by step

Dal file Excel dati-festival-sanremo-1951-2023, ho estrapolato il file dati\_festival contenente anno, periodo, data finale, numero serate, sede, presentatore, direttore artistico, numero partecipanti, partecipanti nuove proposte, numero canzoni, vincitore, tipo e vincitore nuove proposte.

Questa sorgente dati l’ho utilizzata in Looker per visualizzare il numero delle edizioni del festival, il numero totale canzoni presentate, il numero totale dei partecipanti, il numero delle serate del festival, la tipologia di vittoria (singolo, duetto, gruppo o trio), il conteggio di tipo di vittoria per decennio, il numero minimo, medio e massimo di partecipanti per edizione, il numero minimo, medio e massimo di canzoni per edizione, il numero di partecipanti per anno e il numero di canzoni per anno.

Successivamente, dal file dati\_festival ho preso la colonna vincitori e ho splittato i nomi, qualora fossero presenti collaborazioni (duetti o trii) e ho spostato tutto nel file dati\_vincitori.

Il file Excel dati\_vincitori contiene anno, data finale, vincitore, genere, provincia di nascita, zona e età (alla vittoria). I dati di genere, età, provincia di nascita e zona l’ho recuperati chiedendo supporto all’AI e controllando successivamente il risultato.

Questa sorgente dati l’ho utilizzata in Looker per visualizzare il genere più rappresentato dai vincitori, l’età media, l’età media dei vincitori per anno, il vincitore più giovane e più anziano, l’età minima, media e massima dei vincitori per decennio, l’albo d’oro dei vincitori e la provenienza dei vincitori.

Infine, dal file dati\_festival ho preso la colonna conduttori e ho splittato i nomi, qualora fossero presenti conduzioni doppie o triple e ho spostato tutto nel file dati\_presentatori.

Il file Excel dati\_presentatori contiene anno, data finale, presentatore, sesso, provincia di nascita, zona e età (alla conduzione). I dati di sesso, età, provincia di nascita e zona l’ho recuperati chiedendo supporto all’AI e controllando successivamente il risultato.

Questa sorgente dati l’ho utilizzata in Looker per visualizzare il numero totale di conduttori, il genere più rappresentato dai conduttori, l’albo d’oro dei conduttori, l’età media, l’età media dei conduttori per anno, il conduttore più giovane e più anziano e la provenienza dei conduttori.

Il file Excel numeropartecipazioni contiene artista, edizioni e numero delle partecipazioni al festival. L’ho ottenuto facendo scraping dal sito web <https://it.wikipedia.org/wiki/Partecipanti_al_Festival_di_Sanremo> tramite Power Query, prendendo solo gli artisti con più di 2 partecipazioni.

Questa sorgente dati l’ho utilizzata in Looker per visualizzare il numero di artisti con più di 2 partecipazioni, per visualizzare i nomi degli artisti con più di 2 partecipazioni (in ordine decrescente di numero) con relativo filtro del nome dell’artista.

Il file Excel dati\_ascolti dal 1987 contiene l’edizione del Festival, l’anno, la data finale, il numero di telespettatori nella prima, seconda, terza, quarta e quinta serata del Festival, la media dei telespettatori nelle diverse serate per anno, la % di share nella prima, seconda, terza, quarta e quinta serata del Festival e la media della % di share nelle diverse serate per anno e il presentatore dell’anno considerato. Questo set di dati, l’ho ricavato facendo scraping dal sito web <https://it.wikipedia.org/wiki/Festival_di_Sanremo#Ascolti> tramite Power Query, aiutandomi con la conversione dei tipi di dato con l’AI.

Questa sorgente dati l’ho utilizzata in Looker per visualizzare il valore massimo, medio e minimo dei telespettatori (dal campo media\_telespettatori), l’andamento della media\_telespettatori per anno, il valore massimo, medio e minimo della % di share (dal campo calcolato media\_share % ottenuto facendo media\_share / 100 e visualizzando il risultato in percentuale), l’andamento della media % di share per anno, l‘andamento della % di share per anno e per serata con aggiunto un filtro per visualizzare le informazioni su uno o più anni specifici.

Dal file Excel dati-classifica-sanremo-1951-2023, contenente le classifiche delle canzoni degli anni del Festival ho estrapolato il file partecipanti\_sanremo contenente interprete, genere e tipo.

L’interprete l’ho ottenuto splittando i nomi dal campo interprete del file di partenza e rimuovendo i vari duplicati. Il campo genere e tipo (singolo, gruppo) l’ho recuperati chiedendo supporto all’AI e controllando successivamente il risultato.

Questa sorgente dati l’ho utilizzata in Looker per visualizzare il genere più rappresentato nei partecipanti, il tipo dei partecipanti (singolo o gruppo), il numero di partecipanti donne (solista femminile, gruppo femminile, trio femminile o duo femminile) e il numero di partecipanti uomini (solista maschile, gruppo maschile, trio maschile o duo maschile).

Per quanto riguarda l’analisi dei dati Spotify ho scelto di prendere in considerazione solo le canzoni classificatesi rispettivamente al primo, secondo e terzo posto negli anni del Festival.

Ho creato il file dati\_spotify ottenuto dal file dati-classifica-sanremo-1951-2023 dove ho estrapolato la posizione in classifica (solo 1-2-3), l’interprete, la canzone e l’anno unendolo al file dati-canzoni-spotify-sanremo-1951-2023 dove estrapolato la durata e la popolarità della canzone (grazie al cerca.vert). In alcuni casi, il cerca.vert mi ha restituito celle vuote (perché i valori non c’erano) e ho chiesto aiuto all’AI per riempire con dei valori verosimili quei dati.

Il file Excel dati\_spotify contiene il campo posto in classifica, anno, data finale, canzone, durata (min), popolarità e cantante.

Questa sorgente dati l’ho utilizzata in Looker per visualizzare le canzoni più brevi e più lunghe, la durata media delle canzoni sul podio, la durata media delle canzoni sul podio per decennio, la popolarità media delle canzoni sul podio, l’index score medio delle canzoni sul podio per decennio con relativo filtro al posto in classifica, la popolarità degli ultimi 10 anni e posto in classifica, la top 10 di canzoni più popolari, la distribuzione della popolarità per durata e la distribuzione della durata (divisa in range) delle canzoni sul podio.

Dal file dati-classifica-sanremo-1951-2023 ho creato il file Excel classifica\_primi\_secondi\_terzi dove ho estrapolato la posizione in classifica (solo 1-2-3), l’interprete, la canzone e l’anno, splittando l’interprete qualora fossero presenti collaborazioni, come duetti o trii (ad esempio Mahmood è arrivato sul podio sia da solista nel 2019 che con Blanco nel 2022 quindi mi deve risultare che sul podio è salito due volte).

Il file Excel classifica\_primi\_secondi\_terzi contiene il posto in classifica, l’anno, la data finale, il cantante e la canzone relativa.

Questa sorgente dati l’ho utilizzata in Looker per visualizzare il numero di podi per cantante.

In Looker ho creato l’unione join podio\_numeropartecipazioni, che vede protagoniste le tabelle classifica\_primi\_secondi\_terzi e numeropartecipazioni, unite dalla chiave Cantante (tab 1) – Artista (tab 2).

Ho fatto una left join della tabella classifica\_primi\_secondi\_terzi prendendo i campi anno, cantante, canzone, classifica con la tabella numeropartecipazioni prendendo i campi artista e numero. Il risultato ottenuto è una tabella che contiene solo gli artisti sul podio con il rispettivo numero di partecipazioni totali al Festival. Ovviamente gli artisti che sono sul podio avendo partecipato uno o due volte presentano un valore nullo sul campo numero, poiché nella tabella numeropartecipazioni ci sono solo coloro che hanno partecipato più di 2 volte al Festival. Questi dati mancanti l’ho esclusi con un filtro sul grafico interessato (numero > 2).

Questa join l’ho utilizzata per visualizzare una tabella che mostra il numero di partecipazioni, il numero di podi, il numero di primi posti (campo calcolato : sum(if(Classifica=1,1,0)), di secondi posti( campo calcolato : sum(if(Classifica=2,1,0)), di terzi posti (campo calcolato : sum(if(Classifica=3,1,0)) e la % di podio sulle partecipazioni totali (campo calcolato % Podi = count(Cantante)/Numero) per ogni artista, con un filtro per poter visualizzare le informazioni su uno o più artisti specifici.

Il file Excel classifica e genere\_merge l’ho ottenuto facendo un merge su Power Query dei file classifica\_primi\_secondi\_terzi e partecipanti\_sanremo sulla chiave Cantante (tab 1) e Interprete (tab 2).

Ho fatto una left join della tabella classifica\_primi\_secondi\_terzi prendendo i campi anno, cantante, canzone, classifica con la tabella partecipanti\_sanremo prendendo i campi interprete, genere e tipo.

Questa join l’ho utilizzata per avere una tabella contenente tutti gli artisti sul podio (1-2-3) con relativo anno, genere e tipo (singolo o gruppo).

Questa sorgente dati l’ho utilizzata su Looker per visualizzare il conteggio podi per genere, la % di podi femminili sul totale, la % di podi maschili sul totale, il numero di donne sul podio (soliste femminili, duo femminile, gruppo femminile, trio femminile) e il numero di uomini sul podio (solisti maschili, duo maschile, gruppo maschile, trio maschile).

Altre info:

* Il campo Anno per decennio è un campo calcolato in Looker che è stato utilizzato per suddividere gli anni del festival in decenni.
* Il campo Anno – cc è un campo calcolato in Looker, creato a partire dal campo data finale (che rappresenta la data dell’ultima serata del festival dell’anno considerato) tramite la funzione year(). Ho creato questo campo calcolato per raggirare il problema avuto con il campo Anno già presente nei file Excel in quanto non era possibile fare una conversione del tipo di dato.
* Per il conteggio di uomini e donne sul podio e del numero totale di partecipanti uomini e donne al Festival ho utilizzato i campi calcolati Conteggio uomini = SUM(IF(REGEXP\_MATCH(Genere, ".\*Maschile.\*"), 1, 0)) e Conteggio donne = SUM(IF(REGEXP\_MATCH(Genere, ".\*Femminile.\*"), 1, 0)). Ho utilizzato questo modo per prendere nel conteggio non solo il genere Maschile/Femminile ma anche Gruppo Maschile/Gruppo Femminile, Duo Maschile/Duo Femminile e Trio Maschile/Trio Femminile presenti campo genere della tabella partecipanti\_sanremo.