UNIVERSITÀ DI PISA

Le principali cause di decesso tra i giovani

E i fattori che impattano sulla loro salute

Un progetto di Data Journalism

Arianna Di Serio

Sofia Capone

Corso di laurea magistrale
in Informatica Umanistica

Sommario

1. OBIETTIVI E CONTENUTO DEL PROGETTO

Il nostro progetto ha come obiettivo lo sviluppo di un'analisi approfondita sulle cause principali di decesso tra i giovani in Italia, con particolare attenzione ai fattori che incidono sulla loro salute.
Attraverso l'utilizzo di dati statistici provenienti dall'ISTAT (Istituto Nazionale di Statistica), sono stati esaminati diversi indicatori, tra cui le cause di morte più comuni, le abitudini relative al fumo e all'alcol, nonché lo stile alimentare, per fornire una panoramica completa basata sui dati che possa contribuire a una maggiore consapevolezza riguardo alla salute dei giovani italiani.
Infine, è stato realizzato un confronto con un altro Paese dell'Unione Europea per studiare differenze e somiglianze e individuare le criticità che richiedono un intervento immediato da parte delle autorità competenti per aiutare i giovani a perseguire uno stile di vita sano ed equilibrato.

2. I DATASET UTILIZZATI

Per realizzare la nostra analisi abbiamo utilizzato i seguenti dataset:

1. **Dataset sulle principali cause di decesso in Italia**: l'ISTAT raccoglie e pubblica dati sulle cause di morte in Italia attraverso il Sistema Informativo di Mortalità (SIM). Questi dati includono informazioni dettagliate sulle cause specifiche di morte, suddivise in diverse categorie (ad esempio malattie cardiache, incidenti stradali, suicidi, ecc...).
I dati sono stati analizzati relativamente ai giovani, definendo la fascia di età specifica in base alle definizioni fornite dall'ISTAT. L'arco temporale coperto va dal 2015 al 2020.
I dataset sono 2: *cause decessi per età* (dimensione: 27000 records ca.) e *cause decessi per regione* (dimensione: 156000 records ca.).

2. **Dataset sull'abitudine al fumo in Italia (ISTAT)**: per comprendere l'impatto del fumo sulla salute dei giovani sono stati utilizzati dataset relativi all'abitudine al fumo in Italia. Questi dati includono informazioni sul numero di fumatori, suddivisi per fascia d'età, genere e regione. Sono stati esaminati i dati per la fascia di età dei giovani, al fine di identificare eventuali correlazioni tra il fumo e le principali cause di morte. I dati sono relativi agli anni 2021 e 2022.
La dimensione del dataset è di 24 records.

3. **Dataset sull'abitudine all'alcol in Italia (ISTAT)**: per valutare l'influenza dell'abuso di alcol sulla salute dei giovani, sono stati considerati dati sull'abitudine all'alcol in Italia.

Anche queste informazioni sono suddivise per fascia d'età, genere e regione e anche in questo caso sono stati analizzati i dati relativi ai giovani per individuare eventuali legami tra il consumo di alcol e le cause di morte più comuni. Il dataset riguarda il 2021 e 2022.
I dataset sono 2: *abitudine all'alcol per età* (dimensione: 25 records) e *abitudine all'alcol per regione* (dimensione: 39 records).

4. **Dataset sullo stile alimentare in Italia (ISTAT)**: per esaminare l'impatto dello stile alimentare sulla salute dei giovani, sono stati considerati dati relativi alle abitudini alimentari in Italia. Anche in questo caso i dati risalgono al 2021 e 2022.
La dimensione complessiva dei dataset è di 120 records.

5. **Dataset sulle principali cause di decesso in Spagna** (fino al 2022): attraverso i dati forniti da l'Instituto Nacional de Estadística, noto con l'acronimo INE (organismo che coordina i servizi statistici della Spagna è stato sviluppato un confronto, per individuare differenze e somiglianze nelle principali cause di decesso tra i giovani italiani e spagnoli.
I dati partono dal 2015 e, solo per la Spagna, arrivano fino al 2022.
La dimensione del dataset è di 340560 records.

3. PRIMA FASE DI ANALISI: PULIZIA DEI DATI

La prima fase del nostro progetto ha previsto, oltre alla selezione dei dataset più appropriati allo scopo, un'attività di pulizia e normalizzazione dei dati.

In particolar modo, per ogni dataset, sono stati:

- Individuati, se presenti, i **valori nulli**.

  Quest'ultimi erano presenti unicamente nel dataset sulle cause di decesso, poiché per ogni morte è riempita la cella relativa all'anno di decesso e le restanti presentano valori NaN.

  Questi valori sono stati sostituiti con "0".

```
df.isna().sum()
```

```
Sesso                        0
Descrizione sesso            0
Classe di età ( anni)        0
Mese                         0
Descrizione mese             0
Codice causa di morte        0
DescrizioneCausa             0
Decessi 2015              3270
Decessi 2016              3406
Decessi 2017              3396
Decessi 2018              3319
Decessi 2019              3462
Decessi 2020              3382
Causa_livello1               0
Causa_livello2             796
Causa_livello3            7789
```

- Verificata la correttezza del **tipo dei dati** per ogni colonna. Qualora il tipo non fosse corretto si è provveduto a modificarlo.

```
df['Decessi 2015'] = df['Decessi 2015'].astype('int64')
df['Decessi 2016'] = df['Decessi 2016'].astype('int64')
df['Decessi 2017'] = df['Decessi 2017'].astype('int64')
df['Decessi 2018'] = df['Decessi 2018'].astype('int64')
df['Decessi 2019'] = df['Decessi 2019'].astype('int64')
df['Decessi 2020'] = df['Decessi 2020'].astype('int64')
import numpy as np
obj_columns = df.select_dtypes(include=object).columns.tolist()
df[obj_columns] = df[obj_columns].astype('string')
df.dtypes
```

- Verifica, attraverso la funzione di *unique*, dei valori per ogni cella, per verificare che non ci fossero stessi valori ma in formati diversi.

  Per **normalizzare** questi dati ci si è serviti della funzione di *replace* o di espressioni regolari.

```
valori3 = ["AIDS_malattia da HIV","AIDS (malattia da HIV)"]
nuova_categoria3 = "AIDS"

valori4 = ["asma","Asma"]
nuova_categoria4 = "Asma"

valori5 = ["Altre malattie ischemiche del cuore","altre malattie ischemiche del cuore"]
nuova_categoria5 = "Altre malattie ischemiche del cuore"

valori6 = ["Altri accidenti", "altri accidenti"]
nuova_categoria6 = "Altri accidenti"

valori7 = ["Malattie del rene e uretere", "Malattie del rene e dell'uretere"]
nuova_categoria7 = "Malattie del rene e uretere"

valori8 = ["altre malattie croniche delle basse vie respiratorie", "Altre malattie croniche delle basse vie respiratorie"]
nuova_categoria8 = "Altre malattie croniche delle basse vie respiratorie"

valori9 = ["Abuso di alcool (compresa psicosi alcolica)", "Abuso di alcool -compresa psicosi alcolica"]
nuova_categoria9 = "Abuso di alcool (compresa psicosi alcolica)"

df5["DescrizioneCausa"] = df5["DescrizioneCausa"].replace(valori2, nuova_categoria2)
df5["DescrizioneCausa"] = df5["DescrizioneCausa"].replace(valori3, nuova_categoria3)
df5["DescrizioneCausa"] = df5["DescrizioneCausa"].replace(valori4, nuova_categoria4)
df5["DescrizioneCausa"] = df5["DescrizioneCausa"].replace(valori5, nuova_categoria5)
```

- **Filtraggio** dei dati per selezionare solamente quelli relativi alle classi di età 15-34 anni.

```
giovani = ['15-19', '20-24', '25-29', '30-34']
```

```
df3 = df[df['Classe di età ( anni)'].isin(giovani)].reset_index()
df3
```

In alcuni casi è stata realizzata anche un'operazione di **aggregazione dati**, soprattutto ai fine della visualizzazione grafica delle analisi. Ad esempio, le varie tipologie di tumori sono state aggregate in un'unica categoria generica.

```
valori = ["altri Tumori maligni del tessuto linfatico/ematopoietico",
          "Tumori non maligni (benigni e di comportamento incerto)",
          "Tumori maligni del fegato e dei dotti biliari intraepatici",
          "altri Tumori maligni",
          "Tumori maligni del rene",
          "Tumori maligni del cervello e del sistema nervoso centrale",
          "Tumori maligni della vescica",
          "Tumori maligni del colon, del retto e dell'ano",
           "Tumori maligni delle labbra, cavità orale e faringe",
           "Tumori maligni della tiroide",
           "Tumori maligni del pancreas",
           "Tumori maligni dello stomaco",
           "Tumori maligni della trachea, dei bronchi e dei polmoni",
           "Tumori maligni ovaio",
           "Tumori maligni esofago",
          "Tumori maligni dell'esofago",
            "Tumori maligni della cervice uterina",
            "Tumori maligni di altre parti dell'utero",
            "Tumori maligni della laringe",
            "Tumori maligni della prostata",
          "Tumori maligni del seno",
         "Tumori maligni del colon, retto e ano",
          "Tumori non maligni-benigni e di comportamento incerto",
          "Altri Tumori maligni del tessuto linfatico/ematopoietico",
          "Tumori maligni di altre parti utero",
          "Tumori maligni dell'ovaio",
          "Altri tumori maligni"
         ]

nuova_categoria = "Tumori (tutti i tipi)"
df5["DescrizioneCausa"] = df5["DescrizioneCausa"].replace(valori, nuova_categoria)
```

4. SECONDA FASE DI ANALISI: LISTA DI DOMANDE

In una seconda fase abbiamo sviluppato le domande su cui ci interessava soffermarci e che volevamo esplorare con la nostra analisi:

1. Quali sono gli aspetti più importanti da considerare per valutare lo stile di vita (Dieta, Sedentarietà, Alcol, Fumo…)?
2. Come e quanto incide ciascuno di questi fattori sull'individuo? Come e quanto incidono sui giovani rispetto agli adulti?
3. Qual è lo stile di vita dei giovani in Italia? Quanto è sano?
4. Quali sono le abitudini dannose dei giovani?
5. Quali sono le cause di mortalità più diffuse tra i giovani in Italia?
6. Quali sono le cause di mortalità principali in altri paesi Europei?
7. Quali sono le cause che portano i giovani più facilmente a questi decessi?
8. Quale fascia di età viene considerata quando si parla di giovani (15-34 anni)?
   La classificazione adottata è quella dell'ISTAT.
9. Quali periodi di tempo sono stati presi in considerazione?
10. Ci sono state variazioni nel tempo? Soprattutto in seguito alla pandemia? Se sì quali?

5. TERZA FASE DI ANALISI: VISUALIZZAZIONE, GRAFICI, SITO WEB

Infine, l'attività ha previsto una fase di visualizzazione, sviluppo dei grafici e realizzazione del sito web per presentare i risultati dell'analisi.

Ci siamo concentrate su tre aspetti:

1. La situazione attuale in Italia
2. Il confronto con l'Europa
3. La variazione nel tempo in Italia

In questa fase abbiamo sviluppato le nostre statistiche sui dati, cercando correlazioni tra le informazioni, aspetti che necessitavano di un ulteriore approfondimento, fonti per interpretare i risultati e capire il contesto.

Nello sviluppo dei grafici abbiamo cercato di capire innanzitutto quali fossero i dati rilevanti, valutando se aggregarli o meno, e di scegliere il tipo di grafico più adatto.

Abbiamo rimosso tutto il "rumore" non necessario, per rendere la visualizzazione il più comprensibile e pulita possibile, aggiungendo eventualmente delle annotazioni o immagini per dare risalto a dei punti specifici del grafico o per far riferimento al contesto.

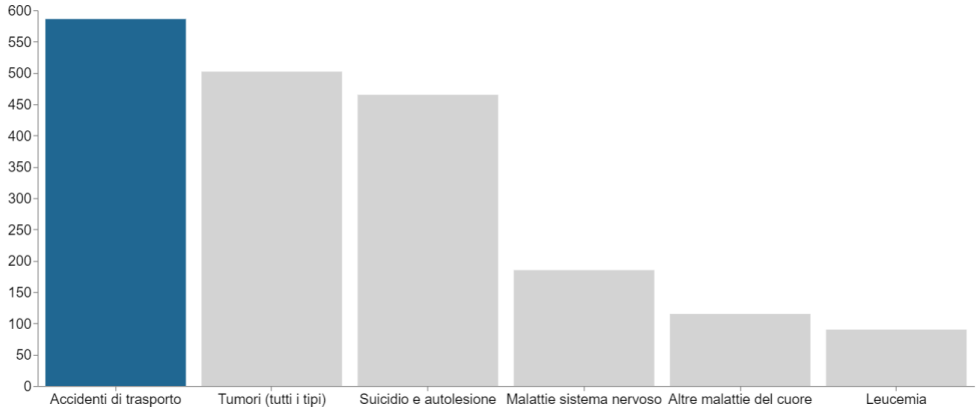
5.1 LA SITUAZIONE ATTUALE IN ITALIA

In questa sezione siamo andate ad analizzare la situazione a livello di cause principali di morti premature e fattori che le influenzano in Italia nel 2020, con alcuni dati relativi anche al 2021 e 2022.

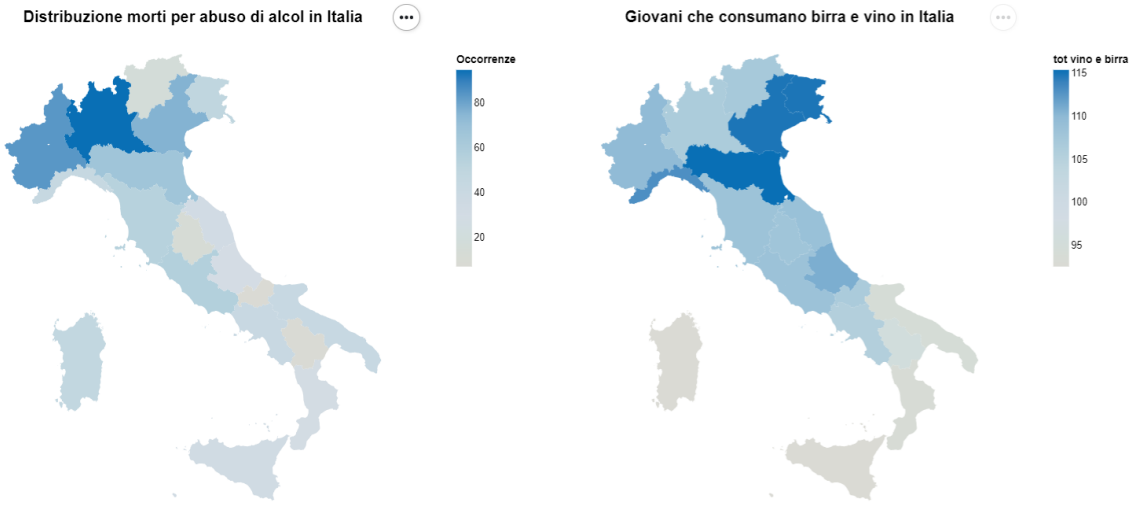
L'obiettivo è quello di individuare quali siano le cause di decessi principali tra i giovani e quali le abitudini rischiose che possano indurle, per indurre a maggiore consapevolezza nel pubblico e per sensibilizzare a tematiche come dipendenze e salute mentale.

Abbiamo sviluppato in particolar modo:

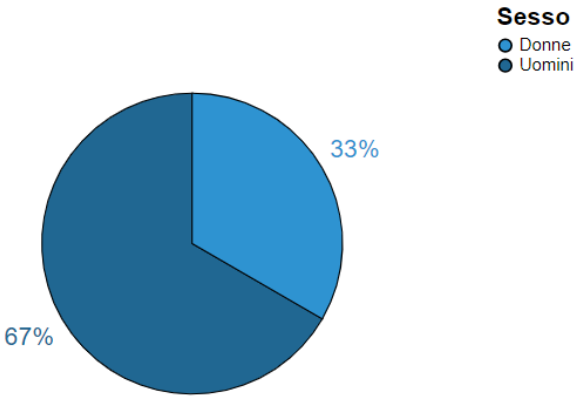
1. Grafici a barre, per avere un'idea del numero di morti per ogni causa



2. Mappe, per capire la distribuzione ad esempio di incidenti stradali e morti per abuso di alcol nelle regioni italiane



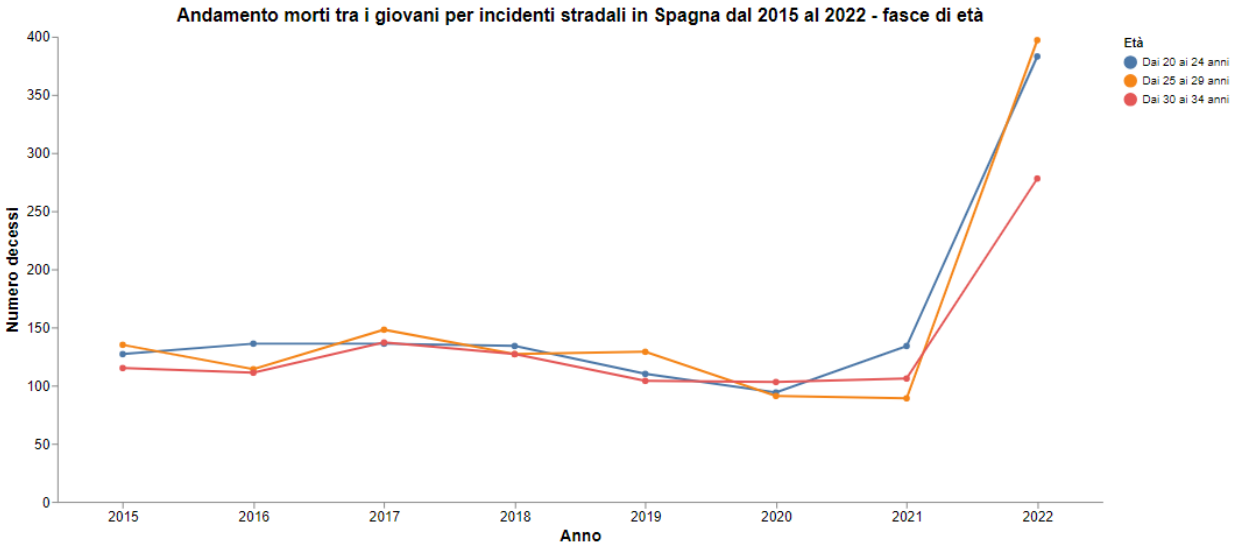
3. Grafici a torta, per avere informazioni sulle percentuali di uomini e donne relativamente alle varie cause di decesso



5.2 CONFRONTO ITALIA E SPAGNA

In questa seconda sezione invece si è voluto sviluppare un confronto tra l'Italia ed un altro paese europeo, per capire se ci fossero significative differenze nei decessi prematuri.

Anche in questo caso abbiamo adottato grafici a barre, grafici a torta, e anche linegraph, per mostrare l'andamento nel tempo del numero di morti.



5.3 L'ITALIA NEL TEMPO

Infine, l'ultimo aspetto preso in esame è la variazione nel tempo del numero di morti, tra il 2015 e il 2020, anni della pandemia di Covid-19.

L'obiettivo era quello di capire se effettivamente ci fosse stato, nel corso degli anni, un miglioramento a livello di salute fisica e mentale nei giovani, ed eventualmente quali fossero stati i motivi dietro questi cambiamenti.

FONTI

Gianinazzi, Andrea. Quaglia, Jacqueline. Inderwildi Bonivento, Laura. *Lo stress: un fattore di rischio tra i giovani*

Humanitas | Sintomi dello stress. *Consultato il 10/07/2023*

Humanitas | Sonno: perché è importante dormire bene. *Consultato il 10/07/2023*

ISTAT. *Consultato il 10/07/2023*

Istituto Superiore di Sanità | Alcol, epidemiologia e monitoraggio 2020. *Consultato il 10/07/2023*

Istituto Superiore di Sanità | Alcol, guida, sicurezza e salute. *Consultato il 10/07/2023*

Istituto Superiore di Sanità | Consumo alcolici tra la popolazione femminile. *Consultato il 22/07/2023*

Istituto Superiore di Sanità | Fumo e alcol tra gli adolescenti. *Consultato il 10/07/2023*

Istituto Superiore di Sanità | Incidenti stradali. *Consultato il 22/07/2023*

Istituto Superiore di Sanità | Sorveglianza e prevenzione dell'obesità. *Consultato il 10/07/2023*

Ministero della Salute | Alchol Prevention Day. *Consultato il 10/07/2023*.

Ministero della Salute | Salute della donna. *Consultato il 22/07/2023*

Politiche antidroga. *Consultato il 10/07/2023*.