Buccoliero Antonio 776830
Carnati Emanuele 781231
De Cicco Alessandro 834225

Analisi esplorativa dei suicidi negli Stati Uniti d'America

Sommario

Buccoliero Antonio 776830
Carnati Emanuele 781231
De Cicco Alessandro 834225

I ~ INTRODUZIONE

Il fenomeno del suicidio è tra le realtà più tristi radicate nella società moderna. Nel 1998, l'Organizzazione Mondiale della Sanità ha classificato il suicidio come *dodicesima causa di morte nel mondo;* quasi un milione di persone si tolgono la vita in un anno solare, più di quelle assassinate e di quelle cadute in guerra. Inoltre, sempre l'OMS ha stimato che, se da un lato solo il 5% dei tentativi di suicidio viene portato a termine, dall'altro esiste un pericolosissimo potenziale, ben venti volte maggiore. Tra gli Stati che hanno registrato la maggiore crescita del tasso di suicidi vi sono gli Stati Uniti. Secondo un rapporto pubblicato dal CDC (Centro per la prevenzione ed il controllo delle malattie), il tasso dei suicidi è aumento da 10.5 soggetti su 100.000 abitanti del 1999 a circa 14, una crescita elevatissima, pari a circa il 24%. Inoltre, nel periodo 2006-2014 il tasso ha registrato una crescita ancora maggiore, raddoppiando di anno in anno, la cui principale ragione risiede, secondo alcuni esperti, nella Grande Recessione economica degli ultimi anni, che ha lasciato molto cittadini americani senza lavoro e, in alcuni casi, senza un'abitazione.
Il seguente elaborato propone un'analisi esplorativa delle variabili (fra quelle disponibili) più significative concernenti i suicidi negli Stati Uniti dal 2010 al 2015.

II ~ DBMS E FONTI UTILIZZATE

Il database utilizzato per lo storage dei dati è MongoDB. Si tratta di un DB NOSQL document based, nel quale ogni singolo elemento è salvato come documento BSON, quindi come lista di coppie chiave-valore eventualmente annidate. Si è preferito questo tipo di database in quanto gli attributi dei documenti, seppure per la maggior parte combacino, sono a volte differenti. Ricorre quindi l'esigenza di una struttura duttile, evitando la rigidità del modello relazionale, che sarebbe pieno di valori Null. Altro aspetto che abbiamo considerato nella scelta è il grande supporto e la relativa semplicità d'uso di cui gode MongoDB rispetto ad altri sistemi come Cassandra o Neo4j; ad esempio, MongoDB supporta nativamente lo *sharding*, rendendo facile ed efficiente eliminare i single *point of failure* che affliggono i database distribuiti. La scelta delle fonti è ricaduta sui datasets *Death in the United States* dal 2010 al 2015 forniti dal CDC (acronimo di *Centers for Disease Control and Prevention*). Di questi, 6 datasets sono in formato .csv e contengono dati del 2010, 2011, 2012, 2013, 2014 e 2015 relativi alle modalità, causa, luogo e data di morte, nonché alcune generalità degli individui, tra cui età, sesso, grado di educazione e stato civile. Gli altri 6 datasets si presentano in formato JSON e contengono il "vocabolario" dei file .csv; molti dei dati relativi a modalità di morte, malattie, cause incrociate ecc., sono rappresentati da un codice, il cui significato (in stringa) è deducibile nel rispettivo file JSON (Tab.1).

Buccoliero Antonio 776830
Carnati Emanuele 781231
De Cicco Alessandro 834225

```
{"resident_status":
        {"1": "RESIDENTS",
        "2": "INTRASTATE NONRESIDENTS",
        "3": "INTERSTATE NONRESIDENTS",
        "4": "FOREIGN RESIDENTS"},

"education_1989_revision":{"00": "No formal education"
                        "01-08": "Years of elementary school",
                        "09": "1 year of high school",
                        "10": "2 years of high school",
                        "11": "3 years of high school",
                        "12": "4 years of high school",
                        "13": "1 year of college",
                        "14": "2 years of college",
                        "15": "3 years of college",
                        "16": "4 years of college",
                        "17": "5 or more years of college", "99": "Not stated"},

"manner_of_death": {"1": "Accident",
                "2": "Suicide",
                "3": "Homicide",
                "4": "Pending investigation",
                "5": "Could not determine",
                "6": "Self-Inflicted",
                "7": "Natural",
                "Blank": "Not specified"},
```

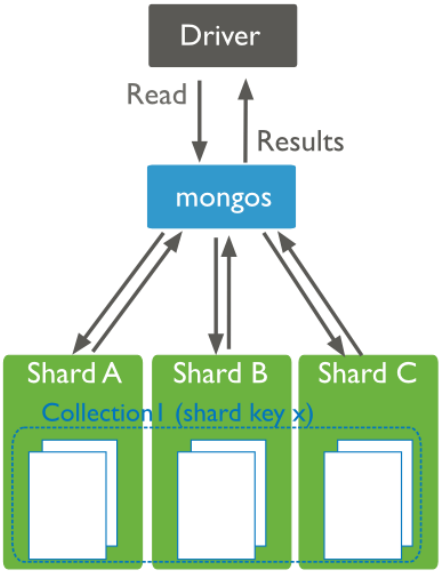
Tabella 1. Uno dei dataset in formato JSON contenenti i relativi attributi, con il significato descritto per ogni modalità che può assumere la variabile.

III ~ ARCHITETTURA DEL DATABASE

Il database utilizzato è MongoDB, un database NoSQL scalabile e riproducibile, implementato in un sistema Linux (OSx).
Per rendere il dataset scalabile e riproducibile, è stato implementato lo sharding, tecnica per la creazione di un database distribuito (anche detto sharded cluster) in cui ogni nodo contiene una porzione del database.
In MongoDB per realizzare un cluster in sharding sono necessari tre componenti, almeno uno per tipo:

- Config server: si tratta di un'istanza di *mongod* che contiene informazioni sull'architettura del cluster. Il config server è molto importante perché rappresenta l'entità che "sa" come i dati sono distribuiti nel cluster, e quindi come reperirli o a quale nodo inviare le operazioni da effettuare.

Buccoliero Antonio 776830
Carnati Emanuele 781231
De Cicco Alessandro 834225



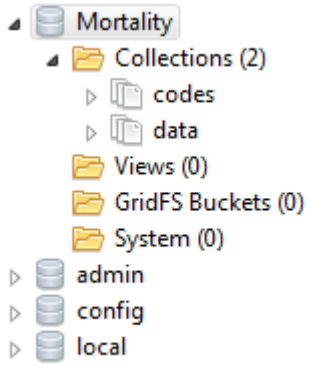
- Router mongos: un nodo speciale che non contiene dati, ma funge da interfaccia alle applicazioni client. Il router mongos comunica direttamente con le porzioni di shard e con il config server.
- Shard: può essere o una istanza di MongoDB (*mongod*) oppure un intero Replica Set.

Nel nostro caso, abbiamo creato tre shard sulle porte "37017, 47017, 57017" comunicanti con un config server alla porta 57040; abbiamo poi lanciato il router mongos tramite il config server.
A scopo di test, tutti gli shards e replica set sono stati avviati sulla stessa macchina: in fase di utilizzo reale, i vari shards e replica set sarebbero distribuiti su numerose macchine, comunicanti fra loro tramite rete.
Di seguito una parte dello script di creazione del dataset:

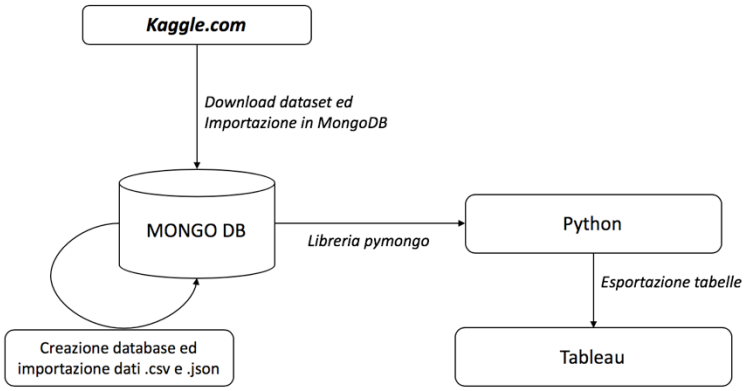
```
sudo mongod --replSet s0 --logpath s0-r0.log --dbpath /data/shard0/rs0 --port 37017 --fork --shardsvr --smallfiles

mongo --port 37017
config = { _id: "s0", members:[
        { _id : 0, host : "localhost:37017" }]};
rs.initiate(config)

sudo mongod --replSet s1 --logpath s1-r0.log --dbpath /data/shard1/rs0 --port 47017 --fork --shardsvr --smallfiles
sudo mongod --replSet s2 --logpath s2-r0.log --dbpath /data/shard2/rs0 --port 57017 --fork --shardsvr --smallfiles
sudo mongod --replSet cs --logpath cfg-a.log --dbpath /data/config/config-a --port 57040 --fork --configsvr --smallfiles
sudo mongos --logpath mongos-1.log --configdb "cs/localhost:57040" --fork

mongo

db.adminCommand( { addshard : "s0/"+"localhost:37017" } );
db.adminCommand( { addshard : "s1/"+"localhost:47017" } );
db.adminCommand( { addshard : "s2/"+"localhost:57017" } );
```

Buccoliero Antonio 776830
Carnati Emanuele 781231
De Cicco Alessandro 834225

IV ~ FLUSSO E SALVATAGGIO DEI DATI



I dati sono stati scaricati dal sito https://www.kaggle.com, all'indirizzo https://www.kaggle.com/cdc/mortality, contenente inizialmente le caratteristiche di ciascun decesso fra il 2005 ed il 2015 negli Stati Uniti. L'analisi effettuata impatta sugli anni compresi tra il 2010 ed il 2015, per un totale di 2,25 GB di dati processati.
I dati scaricati sono stati salvati all'interno di un database in MongoDB, chiamato "Mortality". Il database contiene due collezioni, "data" e "codes" che rappresentano rispettivamente i file csv dal 2010 al 2015 ed i file json dal 2010 al 2015. I dati sono stati caricati dal prompt, tramite il comando "mongoimport" (es. *sudo mongoimport -d=Mortality -c=data --type csv --file="/Users/emanuelecarnati/Desktop/PROGETTO DATA MANAGEMENT/mortality/csv/2005_data.csv" --headerline --host "localhost:27017") /headerline*).
Una volta caricati I file csv e json, è stato utilizzato python e la libreria pymongo per interrogare il dataset ed estrarre delle tabelle che poi diventeranno la nostra base dati per i grafici illustrativi, realizzati tramite il programma Tableau.
Di seguito uno schema riassuntivo:



Buccoliero Antonio 776830
Carnati Emanuele 781231
De Cicco Alessandro 834225

V ~ MANIPOLAZIONE DEI DATI

La manipolazione dei dati è stata gestita tramite script di Phyton, ed ha riguardato sostanzialmente gli aspetti di *integrazione* e di *filtraggio.*

- *Integrazione*: Le diverse modalità numeriche e/o in codice di alcuni dei 77 attributi dei file .csv sono stati sostituiti dalla corrispondente descrizione "stringata", contenuta nei file JSON.

- *Filtraggio*: Ai fini della nostra analisi, si è ritenuto opportuna, prima dell'estrazione, una sintetizzazione di alcuni attributi ridondanti in un unico attributo (es. *education_1989_revision* e *education_2003_revision* in un unico attributo *"education").* Successivamente, abbiamo filtrato l'attributo *"Manner Of Death"* ricavandoci solo le morte causate da suicidio.

VI ~ ESPLORAZIONE E RAPPRESENTAZIONE DEI RISULTATI

Estratto il nuovo dataset con i dati relativi a tutti i casi registrati di suicidio, si procede con l'analisi esplorativa. Tra tutti gli attributi esplorati, si è scelto di rappresentare quelli che offrono i risultati più significativi e possono offrire ulteriori spunti di ricerca.

1. ***Distribuzione per età e sesso dei suicidi in USA [2010-2015]***

   Dall' infografica sottostante (Fig.1) emerge l'abissale differenza fra i suicidi per genere. La popolazione maschile americana ha, mediamente, una probabilità di suicidio 3,5 volte maggiore di quella femminile. In particolare, la popolazione maschile registra una crescita vertiginosa del tasso di suicidio nell'età adolescenziale, per poi calare dai 22 anni fino ai 38-40 anni, mentre il sesso femminile cresce più linearmente dall'età adolescenziale fino ai 38-40 anni. L'andamento risulta essere simile, in proporzione, dai 40 anni in su, con picchi a 52 anni e successiva costante diminuzione.

Buccoliero Antonio 776830
Carnati Emanuele 781231
De Cicco Alessandro 834225

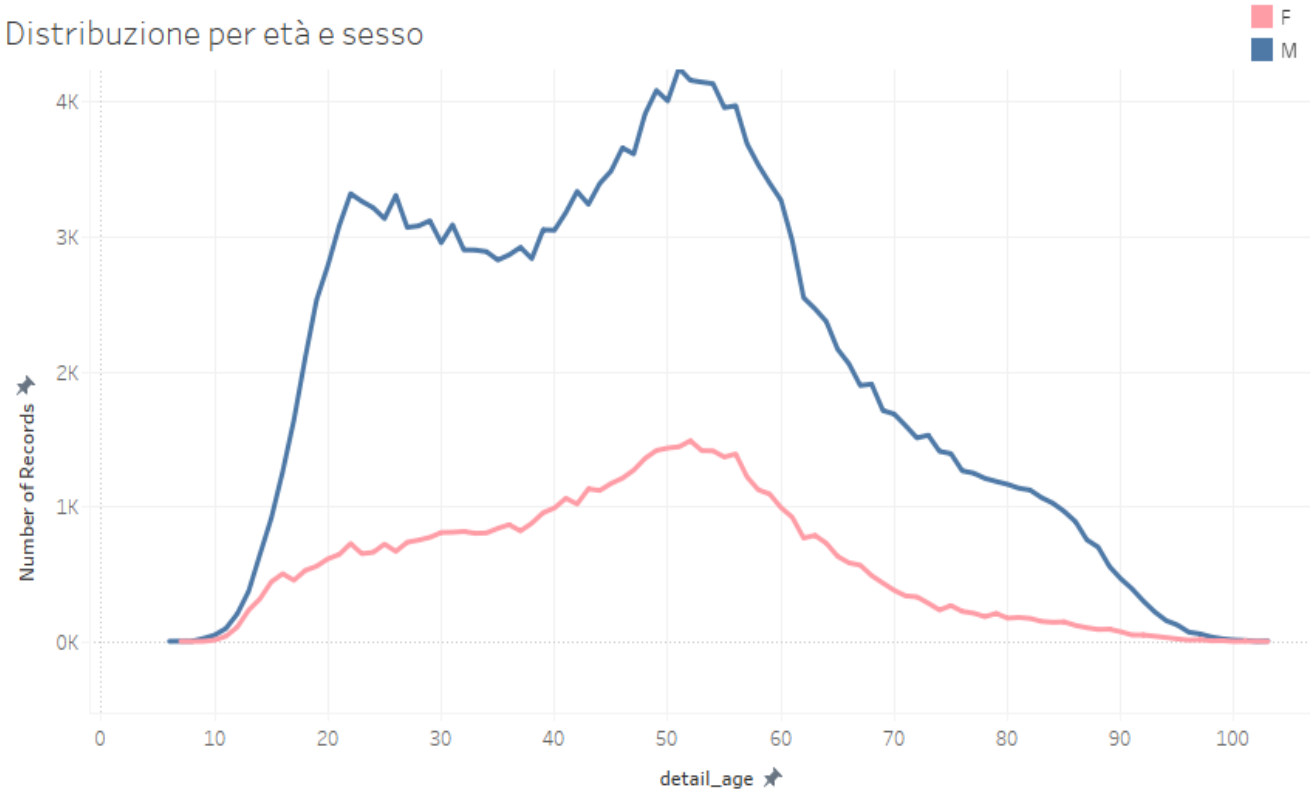
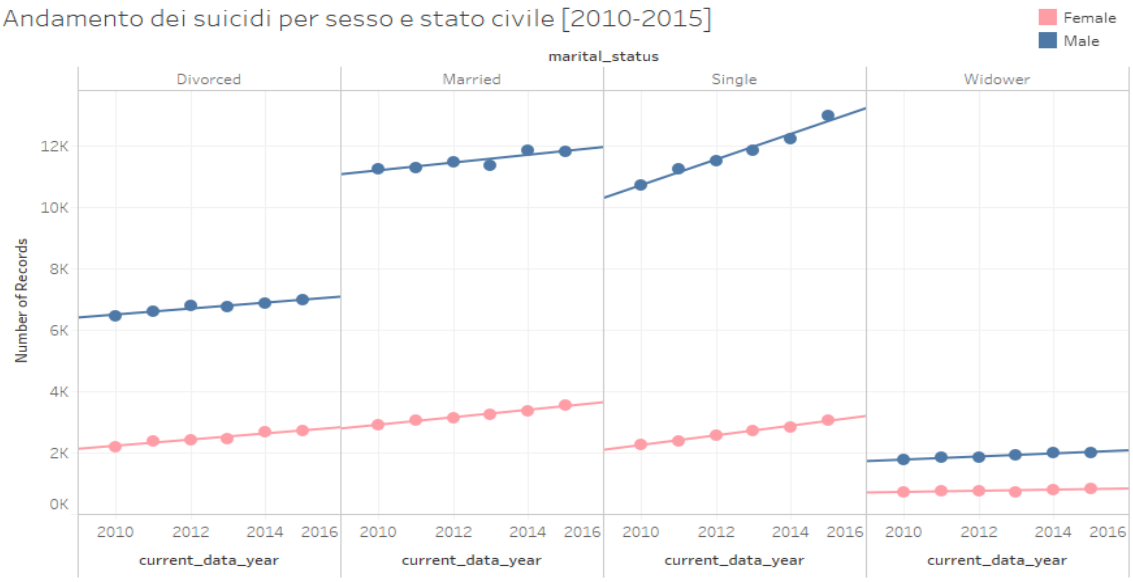


Figura 1. La distribuzione per età e sesso dei suicidi registrati in USA nel lustro 2010-2015

2. *Andamento dei suicidi per sesso e stato civile [2010-2015]*

I suicidi in USA sono, purtroppo, in forte crescita. Analizzando la popolazione per stato civile, emerge che tutte gli stati civili per ogni sesso sono in crescita (Fig.2) Supponendo logicamente che i divorziati si dividano nella stragrande maggioranza equamente tra persone di sesso maschile e femminile, è interessante notare che mentre nel 'gentil sesso' non si registrano significative differenze nei suicidi tra divorziate e sposate, negli uomini si registra un numero nettamente inferiori di suicidi dei divorziati rispetto agli sposati; non abbiamo, però, dati a sufficienza per poter stabilire se questa differenza relativa si giustificata da una maggior tendenza delle donne divorziate oppure da una minor tendenza degli uomini divorziati all'atto estremo.

Buccoliero Antonio 776830
Carnati Emanuele 781231
De Cicco Alessandro 834225



Individual trend lines:

Panes		Color	Line		Coefficients					
Row	Column	sex	p-value	DF	Term	Value	StdErr	t-value	p-value	R-Squared
Number of Records	Divorced	Male	0,0019714	4	current_data_year	97,2	13,4986	7,20076	0,0019714	0,928
					intercept	-188868	27165,9	-6,9524	0,0022489	
Number of Records	Divorced	Female	0,0015499	4	current_data_year	100,371	13,0784	7,67461	0,0015499	0,936
					intercept	-199523	26320,3	-7,58059	0,0016239	
Number of Records	Married	Male	0,0163811	4	current_data_year	126,829	31,8548	3,98146	0,0163811	0,799
					intercept	-243717	64107,7	-3,80168	0,019076	
Number of Records	Married	Female	< 0,0001	4	current_data_year	122,229	6,88174	17,7613	< 0,0001	0,988
					intercept	-242771	13849,5	-17,5292	< 0,0001	
Number of Records	Single	Male	0,0002301	4	current_data_year	418,714	33,2953	12,5758	0,0002301	0,975
					intercept	-830894	67006,9	-12,4001	0,0002431	
Number of Records	Single	Female	< 0,0001	4	current_data_year	157,686	7,04858	22,3713	< 0,0001	0,992
					intercept	-314701	14185,3	-22,1851	< 0,0001	
Number of Records	Widower	Male	0,001706	4	current_data_year	50,3714	6,73203	7,48235	0,001706	0,933
					intercept	-99474,3	13548,2	-7,34224	0,0018321	
Number of Records	Widower	Female	0,02715	4	current_data_year	18,6286	5,47081	3,40508	0,02715	0,744
					intercept	-36726	11010	-3,33569	0,0289517	

Figura 2. Andamento dei suicidi per sesso e stato civile con dati statistici descrittivi sottostanti. I suicidi nei single registrano la crescita maggiore, con un coefficiente di correlazione di 0.975 e 0.992 per sesso maschile e femminile.

Altro aspetto tragicamente rilevante è la crescita dei suicidi nel mondo dei single. L'ascesa non risparmia nessun sesso, anche se le differenze restano marcatissime. Dato che i single abbracciano qualsiasi fascia d'età, per capire se esiste una fascia d'età più o meno contributiva abbiamo esplorato aggiunto una terza variabile, ovvero l'età alle 2 sovrastanti, come si evince dall'infografica sottostante (Figura.3). Emerge un preoccupante fenomeno che riguarda i giovani adolescenti di sesso maschile, che sono gli sfortunati protagonisti di questo triste dramma che è la crescita del tasso di suicidio.

Buccoliero Antonio 776830
Carnati Emanuele 781231
De Cicco Alessandro 834225

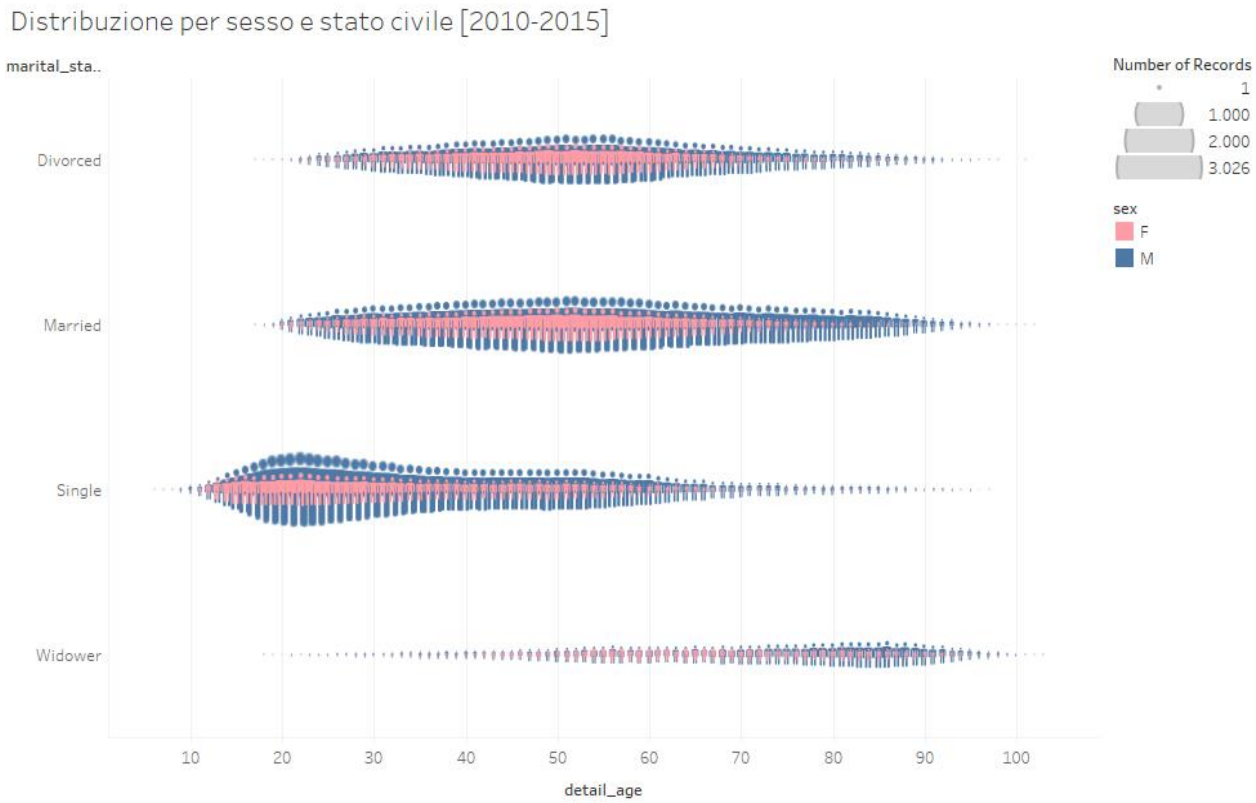


Figura 3. Distribuzione d'età per sesso e stato civile. Emerge, purtroppo, che la più grande componente che contribuisce alla crescita del tasso di suicidio è data dagli adolescenti di sesso maschile.

3. Modalità di suicidio per sesso e per età

E' risaputo che in USA esiste una certa facilità nel reperimento di armi da fuoco. Nella maggior parte dei 51 Stati è sufficiente aver compiuto tra i 18 ed i 21 anni, esibire un documento di riconoscimento che permetta di registrare l'arma al soggetto ed avere una fedina penale nella norma. Non si crede sia quindi un caso che le armi da fuoco siano il mezzo più utilizzato per compiere l'atto estremo. Come s'evince dalla Figura 4, il 55,69 % dei maschi ed il 30,75 % delle femmine utilizza un'arma da fuoco per suicidarsi, ed il 50.5% dei suicidi totali è stato commesso con un'arma da fuoco, ovvero circa 141.000 suicidi in 5 anni. Inoltre emerge un'altra coincidenza non di poco rilievo: nel sesso maschile, i suicidi per arma da fuoco diventano la causa principale dai 17 anni in su, ovvero nella vicinanza della legalità ad acquisire un'arma da fuoco, mentre per nelle donne dai 33 anni a salire. Altre cause più comuni sono l' impiccagione e soffocamento (25,63% maschi e 20,84% femmine) e l'avvelenamento (4,92 % maschi e ben 24,36% per le femmine).

Buccoliero Antonio 776830
Carnati Emanuele 781231
De Cicco Alessandro 834225

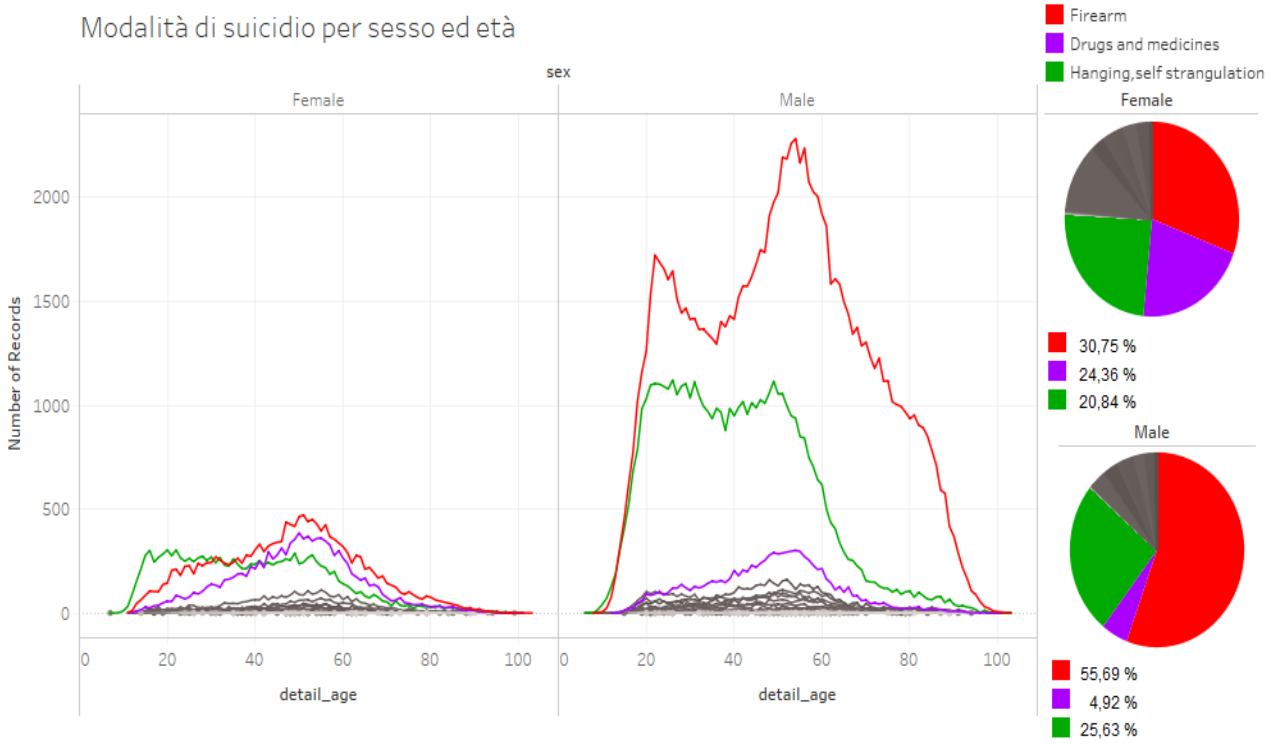


Figura 4. Modalità di suicidio per sesso ed età nel lustro 2010-2015.

4. Distribuzione settimanale e mensile dei suicidi

Tra gli aspetti meno scontati e sicuramente più interessanti, emerge la distribuzione settimanale dei suicidi. Differentemente da quella mensile, la quale è incorrelata con il numero di suicidi registrati, quella settimanale manifesta un andamento che apre a più considerazioni (Figura 5). *Il numero dei suicidi mostra un lineare andamento decrescente rispetto alla settimana lavorativa.* Il giorno in cui si verificano maggiori suicidi è il lunedì, poi linearmente tutti descrescenti fino al Sabato, per finire con la Domenica con un leggero rialzo, quasi a dimostrare l'ansia e la tensione pre-settimanale. Ciò offre spunti interessanti per analisi più approfondite che riguardino i rapporti tra dell'individuo nella sfera sociale americana, che sia lavorativa, formativa (scolastica), ricreativa e familiare.

Buccoliero Antonio 776830
Carnati Emanuele 781231
De Cicco Alessandro 834225

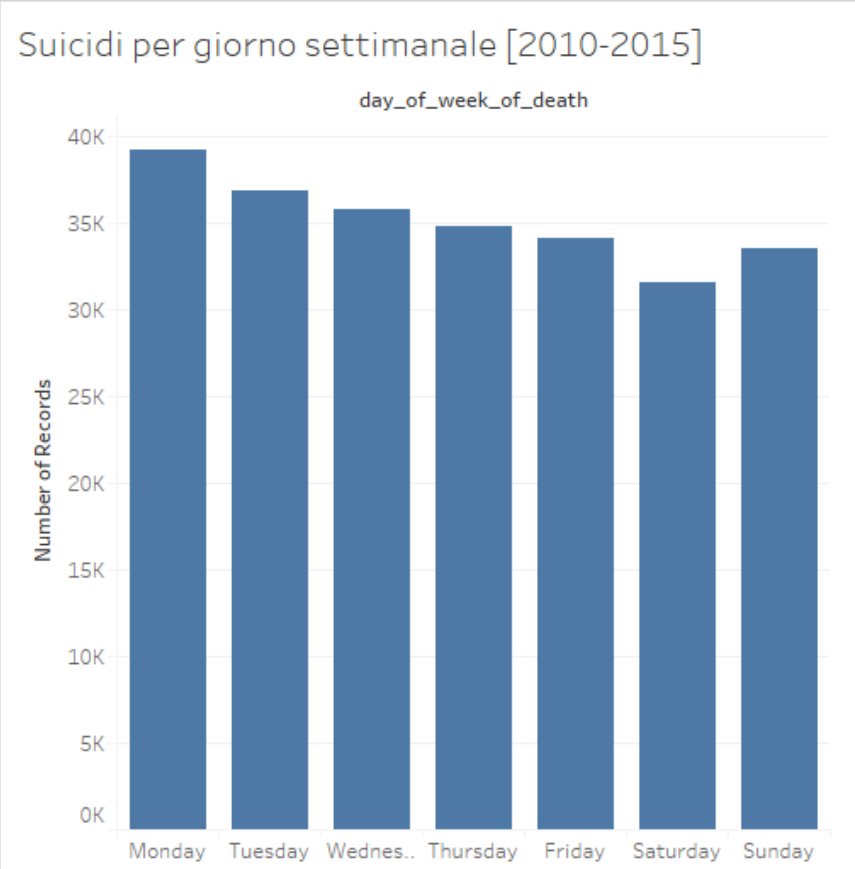


Figura 5. Suicidi per giorno settimanale [2010-2015]

Considerazioni

Anche se è logico ipotizzare che un atto tanto estremo derivi da situazioni psicologiche radicate nell'individuo, visto che la modalità di suicidio più frequente derivi da uno strumento che permetta una realizzazione praticamente istantanea del suicidio (arma da fuoco 50,5%) e visto che la manifestazione dell'evento è molto accentuata nei giorni iniziali della settimana lavorativa e formativa, è lecito supporre che la pressione psicologica nel breve termine possa giocare una differenza significativa nella realizzazione del suicidio, specie se si è in possesso di un' arma da fuoco e soprattutto se la proprietà e/o il possesso sia così facilmente permesso.