Progetto di Analisi Dati
Alice Perin n. matricola 626005
Silvia Villafranca n. matricola 628302
Anno Accademico 2020/2021

TITANIC: analisi della tragedia

Indice

1. Introduzione progetto e storia del Titanic

Il 10 aprile del 1912, dopo mesi e mesi di investimenti, pubblicità e preparazione, il Titanic salpò da Southampton per il viaggio inaugurale con destinazione New York. A partecipare al viaggio e accertarsi che le cose andassero per il meglio ci furono: Joseph Bruce Ismay, imprenditore britannico alla quale diede egli stesso il nome della nave e Thomas Andrews, l'ingegnere che progettò il colossale e leggendario Titanic.
Il Titanic attraversò il canale della Manica fino al suo primo scalo, Cherbourg, in Normandia. In seguito, si diresse verso il porto di Queenstown (oggi Cork, Irlanda) per far imbarcare gli ultimi passeggeri prima di addentrarsi nell'oceano. A bordo viaggiavano più di 2.400 persone.
Di seguito, una breve ricostruzione di ciò che successe il fatidico 14 aprile.

Il capitano Edward Smith, un esperto marinaio della White Star che in precedenza aveva guidato l'Olympic, nave gemella del Titanic e del Britannic, ordinò un cambio di rotta per evitare le zone in cui sapeva che gli iceberg si muovevano alla deriva. Verso le dieci di sera, il capitano si ritirò in cabina e il comando della nave rimase nelle mani del primo ufficiale William Murdoch. Mancavano 20

Fig. 1 Locandina originale pubblicitaria del Titanic
(Storica, National Geographic-storica)

minuti alla mezzanotte quando la vedetta Frederick Fleet avvertì la presenza di un iceberg nelle vicinanze. Fleet informò immediatamente Murdoch, che ordinò di virare a babordo e, subito dopo, di fermare i motori. In questo modo fu possibile evitare la collisione: il ghiaccio e l'acciaio si sfiorarono appena sul lato di tribordo. Eppure le conseguenze di quel leggero contatto furono fatali. Sebbene Smith fosse stato informato subito, si iniziarono a prendere misure di salvataggio solo circa trenta minuti dopo l'impatto, quando l'ingegnere Andrews confermò, in base ai suoi calcoli, che al Titanic rimanevano due ore scarse di vita in acqua. Alle due e cinque del mattino fu calata l'ultima scialuppa e il panico trasformò la tranquillità rarefatta che si era vissuta fino a quel momento in un dramma spaventoso. In meno di mezz'ora più di un migliaio di persone sarebbero morte, tutte perfettamente coscienti del fatto che non potevano fare nulla per evitarlo (Calle, 2020).

In questo progetto, abbiamo analizzato un dataset contenente una porzione di dati riguardanti i passeggeri della nave Titanic (fonte: Kaggle). Abbiamo reso la visualizzazione dei dati più semplice, per mezzo di grafici adatti in base al tipo delle singole variabili prese in esame. Abbiamo così messo in luce le relazioni tra gli eventi di quella tragica notte, ricavandone delle informazioni preziose.

2. Modalità e organizzazione del lavoro e del codice

Abbiamo diviso il codice in più sezioni per facilitarne la lettura e comprensione. In primo luogo, abbiamo installato le librerie necessarie all'esecuzione del codice per la visualizzazione dei dati: Pandas, Seaborn, Numpy, Matplotlib, Scipy, Plotly e Cufflinks. Dopo aver importato le librerie abbiamo importato il dataset in due variabili differenti nominate *df* e *df2*. Per mezzo dei comandi ".shape" e ".tail()" iniziamo a dare un'occhiata alla forma del dataset e ai valori che lo compongono. Successivamente passiamo al data cleaning, operazione importante che consente di pulire i dati dal

“rumore” e dai numerosi valori nulli. Inizialmente questa operazione l’avevamo gestita in modo classico, ovvero eliminando tutte le righe che presentavano dei NaN; nello specifico tra le variabili ‘Age’, ‘Cabin’ e ‘Embarked’ ma, in questo modo il dataset si era ridotto a sole 183 righe anziché le 891 di partenza, falsando così molte proporzioni tra i valori delle singole variabili. Perciò, come accennato precedentemente, abbiamo optato per la doppia importazione dello stesso dataset, in modo da eliminare in uno, le consistenti 687 righe riferendosi ai passeggeri per cui non è stata indicata una cabina; e nell’altro i 177 dell’età e i 2 porti d’imbarco. In questo modo, nel momento della stesura del codice per la creazione dei grafici, abbiamo usato a convenienza uno dei due dataset contenenti le variabili che ci interessavano.

Abbiamo usato la parte del progetto “Sezione modifica variabili e stile” per compiere delle piccole operazioni: inizializzazione di due variabili contenenti i codici colore da assegnare ai grafici (*colorsbarre*, *colorstorta*), modifica delle variabili qualitative “male” e “female” rispettivamente in 0 e 1 a scopo di facilitare il calcolo della correlazione e una sostituzione dei codici cabine con la loro relativa posizione sulla nave per facilitare la comprensione dell’informazione. Per quest’ultima modifica, che abbiamo applicato alla colonna “Cabin”, ci siamo servite della sostituzione delle stringhe mediante le espressioni regolari.

Nella sezione dedicata alla statistica, abbiamo creato per ognuna delle colonne (tranne che per “Ticket”, “Name” e “PassengerID”) dei grafici adatti alla visualizzazione del tipo di valori delle sue variabili. Per le variabili qualitative “Survived”, “Pclass”, “Sex” ed “Embarked” abbiamo utilizzato i grafici a torta per la visualizzazione delle percentuali di tali variabili e grafici a barre per mostrare la distribuzione da un altro punto di vista. Per le variabili con più di tre valori come “Parch” e “SibSp”, abbiamo utilizzato solamente la distribuzione in grafici a barre. Per quanto riguarda le variabili quantitative come “Age” e “Fare”, l’analisi statistica diventa più sfaccettata, in quanto è possibile calcolare la media aritmetica, quartili, mediana (secondo quartile), deviazione standard, valore massimo e minimo presenti per quella variabile nel dataset. Per queste due variabili è stato necessario gestire gli outliers, al fine di normalizzare i dati per rendere più veritieri i calcoli statistici della popolazione osservata. Per rappresentare la distribuzione dell’età dei passeggeri e le tariffe pagate, abbiamo usato istogrammi e boxplot (uno per popolazione completa e uno eliminando gli outliers). Per la variabile qualitativa “Cabin” abbiamo optato invece per un istogramma interattivo utilizzando iplot, per mostrare la distribuzione dei passeggeri all’interno della nave.

Successivamente, siamo passate al calcolo delle correlazioni tra tutte le variabili quantitative del dataset, compresa la variabile “Sex” binarizzata appositamente all’inizio del progetto. La visualizzazione di tali correlazioni è stata possibile attraverso un heatmap e successivamente tramite un pairplot. Una volta analizzati questi due grafici, abbiamo individuato le variabili con interpolazione lineare (positiva e negativa) tramite il coefficiente di Pearson visibile tramite l’heatmap stesso. Abbiamo quindi scelto le variabili più significative con correlazione maggiori di -0.37 e maggiori di 0.38. Le variabili correlate interessanti rappresentate con uno scatterplot sono:

“costo del biglietto e classe” (-0.55), “età e classe” (-0.37), “età e sopravvissuti” (-0.08), “sesso e sopravvissuti” (0.54), “genitori/figli e fratelli/sorelle/coniugi” (0.38).
Per completezza, abbiamo inoltre, incrociato alcuni grafici correlando tre variabili: “costo biglietto - età - sopravvissuti”, “classe - cabina - sopravvissuti”, “classe -numero figli/genitori-sopravvissuti”, “classe - sesso - sopravvissuti”.

3. Analisi dei dati

Seppur avendo in mano un dataset ristretto, che non prende in considerazione molte informazioni sui passeggeri del Titanic (vedi valori nulli/mancanti), abbiamo potuto, a nostro avviso, svolgere delle analisi di questi dati “campione” che potrebbero essere proporzionate ai risultati di quelle dell’intera popolazione. Dalla nostra analisi del dataset, si evince la risaputa ed enorme differenza di probabilità di sopravvivenza in base al genere e in misura minore alla classe.

I set di dati, in quanto tali, mostrano numeri e valori, non riescono da soli a modellare bene le dinamiche dell’evento. La spiegazione del contesto e delle dinamiche dell’evento, trasformano i dati in informazione. Di seguito le nostre intuizioni principali sui risultati del progetto in base ai coefficienti di correlazione tra le variabili sono:

- Nonostante ci siano più uomini a bordo, è più alta la probabilità di sopravvivere tra le donne;
- c’è più alta probabilità di sopravvivere tra i passeggeri di prima classe, nonostante l’inferiorità numerica rispetto alla classe più numerosa (terza classe) ma con tasso di sopravvivenza più basso;
- presenza di correlazioni tra genere e classe: abbiamo constatato una grande differenza di sopravvissuti all’interno della variabile “donne” in base alla classe di appartenenza. Si nota il decremento delle sopravvissute andando verso le classi più basse. Invece il numero di sopravvissuti uomini è sempre più basso rispetto a quello delle donne;
- correlazione costo biglietto e sopravvivenza, proprio perché i passeggeri di prima classe avevano più probabilità di sopravvivere;
- correlazione tra posizione del passeggero nella nave (ricavata attraverso il numero e posizione della cabina) e la sua sopravvivenza;
- correlazione tra numerosità famiglia e sopravvivenza;
- l’età non è statisticamente determinante per la sopravvivenza, nonostante la regola “prima donne e bambini”

4. Conclusioni

Le osservazioni elencate sopra, a partire dai grafici di riferimento, ci hanno fatto riflettere sulle motivazioni di tali correlazioni. Il focus centrale del progetto (e dell'evento in sé) è centrato sulla sopravvivenza dei viaggiatori e quindi è normale chiedersi qual è il tipo di passeggero che ha più probabilità di sopravvivere al naufragio del Titanic.
Il primo punto che salta all'occhio è la probabilità di sopravvivenza nettamente più alta tra le donne. Si suppone che questo sia portato dalla regola vigente "prima donne e bambini", che ha regolato le precedenze per accedere alle scialuppe, nonostante alcune fonti riportino che inizialmente le donne fossero restie a salire sulle navi di salvataggio, e quindi si pensa che nelle prime scialuppe ("prima" fase) ipoteticamente si siano salvati più uomini. Un altro fatto importante è che le donne di prima classe hanno un tasso di sopravvivenza ancora più alto di quelle delle altre classi. Si può notare dalle figure 2 e 3 sottostanti, che la prima classe potrebbe essere stata agevolata principalmente dalla sua posizione nella nave. Il colore verde chiaro indica la posizione delle cabine di prima classe, che si trovavano in prossimità del ponte superiore e delle scialuppe, a cui avevano sempre accesso. Alcune fonti indicano anche che la prima classe abbia avuto relativamente più sopravvissuti perché aveva ricevuto comunicazione di evacuare immediatamente grazie alla presenza di più steward rispetto alle altre classi che hanno ricevuto l'informazione in ritardo (Gupta et al., 2018).

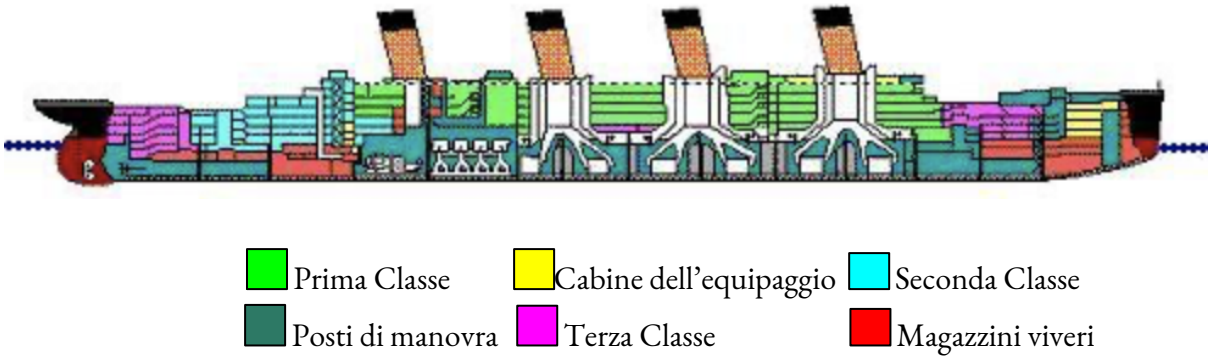
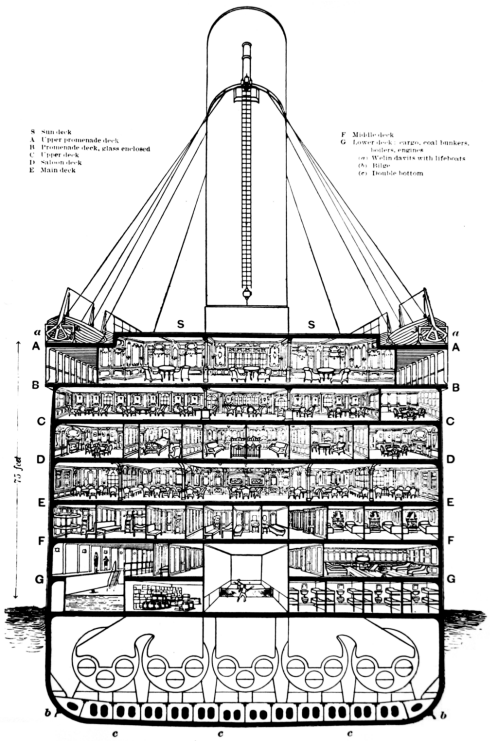


Fig.2 Legenda distribuzioni classi (Il Titanic di Claudio Bossi)

Se si guarda l'ultimo grafico sulla correlazione tra sopravvissuti per classe e sesso, si può notare anche che la categoria che ha risentito di più della regola "prima donne e bambini". Questo perché i passeggeri di seconda classe sono arrivati sul ponte superiore in una "seconda" fase, in cui questa regola è stata applicata con più rigore e fermezza, escludendo gli uomini dall'accesso alle scialuppe



(Stolz et al., 2019). La terza classe è quella che ha risentito di più della perdita di passeggeri sia per il fatto che le cabine erano prua (e poppa), in prossimità della falla, che secondo le ricostruzioni ufficiali è stata la prima parte della nave ad affondare. Inoltre, non avendo di norma accesso al ponte superiore, i passeggeri di terza classe non erano a conoscenza della posizione dello stesso, e di come arrivarci.

Come si può vedere dal grafico "Classe - numero figli/genitori - sopravvissuti", si può vedere una relazione tra le famiglie numerose e la sopravvivenza. In primo luogo, la causa potrebbe essere proprio la regola che ha portato donne e bambini in salvo per primi, e in secondo luogo perché le famiglie numerose godevano di una velocità di comunicazione più alta grazie al "passaparola" tra i membri della famiglia stessa (Gupta et al., 2018).

Fig. 3 Legenda distribuzioni dei ponti
(Wikipedia, RMS Titanic)

Ciò che ha influito maggiormente su questa analisi dei dati è sicuramente il dataset non completo. Infatti, il numero di vittime che viene riportato sulle fonti ufficiali sono più di 1500 morti in confronto ai nostri 424. In più ci sarebbero state molte variabili interessanti da prendere in considerazione per ricostruire ancora meglio la storia e arricchirla di particolari interessanti: l'equipaggio, dati più completi della posizione precisa del numero di cabine e passeggeri, il numero esatto di bambini, genitori, sposi, fratelli e sorelle, il numero delle scialuppe al momento dell'imbarco.

In conclusione, il tipo di passeggero che è stato più "privilegiato" alla sopravvivenza sono le donne di prima classe, che sono 80 in tutto rispetto alle 68 di seconda classe e alle 47 di terza classe.

5. Bibliografia/Sitografia

Bossi, C., www.titanicdiclaudiobossi.com .

Calle, E., “La tragedia del Titanic”, Storica National Geographic, 2020.

Gupta, Kshitiz, et al., “Surviving the Titanic Tragedy: A Sociological Study Using Machine Learning Models”, Suma de Negocios, vol. 9, n. 20, dicembre 2018, pag. 86–92.

Stolz, Jörg, et al. “Sociological Explanation and Mixed Methods: The Example of the Titanic”, *Quality & Quantity*, vol. 53, n. 3, maggio 2019, pagg. 1623–43.