UNIVERSITÀ DEGLI STUDI DI MILANO – BICOCCA

Dipartimento di Informatica, Sistemistica e Comunicazione

Corso di Laurea Magistrale in Data Science

GUIDA ALLA SCELTA CONSAPEVOLE DI UN AIRBNB IN NEW YORK CITY: A CIASCUNO, IL SUO!

Project Team:

Boschi Giulia - 804623

Inzadi Greta - 813649

Testa Luca - 816000

INDICE

Introduzione

Descrizione del project team

I componenti del gruppo sono stati strategicamente scelti per favorire una collaborazione efficace e compensativa. Di fatti, combinando un elemento avente una preparazione squisitamente informatica (Luca Testa), con due elementi prettamente statistici (Giulia Boschi e Greta Inzadi), è stato possibile colmare le rispettive lacune e andare a coprire le conoscenze in più campi possibili. Oltre alle variegate conoscenze pregresse, anche le competenze e le naturali attitudini dei singoli si sono ben amalgamate, concedendo a ciascuno di esprimersi ed occuparsi maggiormente della parte in cui eccelle, pur partecipando attivamente negli altri ambiti del progetto, consentendo così la creazione di una certa sinergia nel team che ha permesso di lavorare in assoluta armonia, sintonia e anche in un'ottica di apprendimento reciproco. Il principio cardine, su cui ci si è basati per favorire la buona riuscita del progetto, è stato la consolidata combinazione tra "mente e braccio" (ruoli assunti alternatamente a seconda del task da svolgere), la quale si è sempre rivelata funzionale, pratica e portatrice di ottimi risultati. Per quanto riguarda la stesura del codice e la parte più onerosa dal punto di vista della programmazione e computazione, il ruolo di "braccio" è stato assunto da Luca Testa, indiscutibilmente più preparato. Tuttavia, anche Giulia Boschi ("connubio tra braccio e mente") si è distinta per aver portato a termine la scrittura di algoritmi utili, soprattutto alla data cleaning e alla Sentiment Analysis, oltre ad aver ideato alcune query d'interesse. Greta Inzadi (in questo caso specifico, con la funzione preponderante di "mente", o meglio, di "supervisore"), ha collaborato con entrambi, ricercando metodi che ottimizzassero il codice scritto o che risolvessero problematiche che impedivano di procedere con le varie fasi di trattamento dei dati (ad esempio trovando soluzioni ad errori o a funzioni poco prestanti ed efficienti, che agevolassero lo svolgimento dell'altrui lavoro, d'altra parte ci si è approcciati ad un linguaggio con cui nessuno aveva molta dimestichezza), suggerendo anche analisi specifiche per lo scopo del progetto. La realizzazione delle infografiche ha impegnato, nella sua praticità, specialmente le due statistiche, le quali si suppone sappiano destreggiarsi meglio con le rappresentazioni grafiche delle indagini condotte e, in generale, con la data visualization (disciplina in cui tutto il gruppo è, bene o male, alle prime armi). Dal canto suo, l'elemento informatico è stato di grande sostegno, sollevando problemi, inesattezze ed elementi di disturbo o ambiguità che non soddisfacevano i canoni richiesti dalla consegna del progetto (è visibile, quindi, un ribaltamento dei ruoli). La scrittura del report finale è attribuibile per la maggior parte a Greta Inzadi, la quale ha consultato gli altri componenti e attinto dalle loro conoscenze e spiegazioni per

renderlo il più chiaro ed esauriente possibile. Infine, i tre elementi, a parità di ruolo, hanno convogliato le loro attività nella fase di valutazione delle infografiche da parte di terzi, così come nella rappresentazione dei risultati ottenuti e nella creazione della presentazione PowerPoint.

Descrizione preliminare del lavoro

I dataset scelti riguardano le prenotazioni Airbnb di New York City, la presenza di eventuali altri hotel in prossimità, le valutazioni dei ristoranti newyorkesi, dove sono allocati i principali food store, dove sono situate le fermate della metro nella città, quali sono le principali organizzazioni culturali della grande mela, dove si sono svolti i principali eventi ed infine quali sono le zone in cui sono stati commessi crimi negli ultimi anni.
L'analisi ha lo scopo di costruire una guida su come scegliere l'appartamento in cui soggiornare, basata sia sulle raccomandazioni degli ospiti precedenti, sia su informazioni esterne che potrebbero influenzare la decisione degli utenti. Inoltre, si andranno a determinare degli indicatori per stabilire quali Airbnb sono di maggiore successo.
Le V da noi scelte sono state, quindi, il volume e la varietà. Per la prima si è raccolto un quantitativo di dati considerevole (per un ammontare superiore a 2GB) considerando ciascun aspetto elencato in precedenza, i quali verranno gestiti tramite il framework Hadoop. Per la seconda, invece, si provvederà ad integrare i dataset trovati tramite timestamp e quartieri (o distretti), infatti, le informazioni relative a ciascun dataset potranno essere aggregate tramite la colonna "nta" (o "borough" laddove il quartiere non fosse stato reso noto).

Materiali e metodi

Materiali

Ci si trova in presenza di diversi dataset, appunto per andare a considerare i vari aspetti che possono influenzare la scelta di un alloggio.

NTA - Neighborhood Tabulation Areas
Shapefile ottenuto dal sito NYC Open Data contenente l'ampiezza di ogni quartiere di ciascun distretto, le loro coordinate geografiche e la loro ampiezza in metri quadrati (trasformata in km quadrati). Sommando le shape area dei quartieri di ogni distretto è stato possibile ottenere le shape area di questi ultimi e, sempre grazie a questa informazione, si sono calcolate le densità riportate nei file successivi.

BOUNDARIES
Si tratta di uno shapefile contenente le coordinate geografiche dei distretti rappresentati come poligoni.

AIRBNB

Il file "listing" conteneva dei valori di quartieri che non corrispondevano a quelli ufficiali usati per tutti gli altri file (necessari all'integrazione). Si è generata una corrispondenza tra i nomi ufficiosi nel file ottenuto da Airbnb e quelli ufficiali contenuti nel file "NTA". Per fare questo, si è verificato se il nome ufficioso fosse contenuto in quello ufficiale. A questo punto si è aggiunta una colonna indicante per ogni annuncio la lista di possibili nomi ufficiali che avrebbero potuto corrispondere a quello ufficioso assegnatogli da Airbnb. Sono state estratte dal dataset quelle osservazioni che avessero più di una corrispondenza tra i nomi ufficiali o che non ne avessero nessuna, e solo per queste osservazioni si è verificato in quale quartiere ricadessero, generando il punto geometrico usando la libreria "Geopandas" di Python e un ciclo "for" per mostrare all'interno di quale quartiere ricadessero (multipoligoni contenuti in "NTA").
Ottenuti i quartieri ufficiali per ogni osservazione si è eliminata la colonna contenente quelli ufficiosi e si è aggiunta la colonna contenente i codici per ogni quartiere. Procedendo con un'ulteriore verifica, si è osservato che un'osservazione corrispondente ad "Hell's Kitchen" non ha trovato alcuna corrispondenza. Si è presupposto che le sue coordinate fossero sbagliate e per questo è stata rimossa dall'analisi.

Sono stati scaricati dal sito di Airbnb il file "calendar", con dati dal 2016 al marzo 2020, al cui interno si trovano le varie prenotazioni effettuate, ovvero gli id delle case, la data e una colonna binaria (T/F) per la disponibilità. Nel file "listing" si trova invece la descrizione degli appartamenti, mentre nel file "review" le recensioni degli ospiti precedenti.
Degli alloggi si conosce il nome e l'id della struttura, il nome e l'id del proprietario, il quartiere, il vicinato, il tipo di stanza, il prezzo, la disponibilità, le coordinate geografiche, il numero minimo e massimo di notti, il numero e il testo delle recensioni e la data risalente all'ultima review effettuata. Per disporre dei dati aggiornati, si sono scaricati e integrati i molteplici update mensili.
In ogni file si ripetevano gli id di ogni annuncio nel file "listing" per ogni giorno di rilevazione fino ad un anno successivo ed indicavano se l'appartamento nel dato giorno fosse prenotato oppure no. Si è proceduto tagliando da ogni file solo le informazioni relative al mese che intercorre tra il giorno della rilevazione e il giorno della successiva così da ottenere informazioni il più possibile corrette (cioè riducendo al minimo la possibilità di avere informazioni riguardo all'occupazione di un appartamento che è stato poi cancellato oppure di perdere prenotazioni dell'ultimo momento). Sono stati concatenati tutti i file per ottenere i dati uniti per tutti i listing in tutti gli anni.
Si è calcolata una media ponderata dei giorni di occupazione della casa per ogni id in listing. La media così ottenuta da "calendar" è aggiunta al file "listing", per conoscere quanto una casa è mediamente occupata in un anno. Sono stati rimossi tutti quegli annunci che non hanno trovato una corrispondenza in "calendar" in quanto si tratta di annunci relativi ad anni precedenti al 2016 che sono stati comunque registrati in listing, perché si tratta di annunci che sono stati di Arbnb, ma non sono più disponibili per prenotazioni sul sito.
Si è svolta la sentiment analysis sulle review lasciate da ogni utente calcolando il valore di polarity e subjectivity.
Sono state mantenute solo le righe in cui la somma dei due valori stimati in valore assoluto fosse diversa da zero. Infatti, dato che polarity varia tra -1 e 1 e subjectivity tra 0 e 1, si osserva che quando entrambe sono pari a 0 è perché l'algoritmo della sentiment analysis non ha prodotto alcun risultato.
Si è calcolato il valore di polarity per ogni casa come media delle recensioni ricevute per ogni id e si è portato il valore in una scala da 0 a 100 per renderlo facilmente confrontabile con il voto riportato nella colonna "review_score_rating" (voto medio della casa da quando è su Airbnb lasciato anche da chi non ha scritto una recensione) di "Airbnb". Si è creato il dataset "scores" per determinare la qualità delle case in ogni quartiere.
Sono stati estratte dal file "review" solo le recensioni lasciate dal 2019. Successivamente è stata calcolata la media per id della polarity, così da poterla confrontare con "review_score_rating" in "Airbnb". Infine, è stato calcolato

l'indicatore qualità prezzo di ogni casa con la formula seguente: (punteggio / 100) / (prezzo / prezzo massimo).
Per ogni nta è stato ricavato quanti annunci fossero presenti in quel quartiere rispettivamente al loro tipo: quanti si riferissero a case intere, quanti a camere private, quanti a camere condivise e quanti a stanze di hotel. Il conteggio degli annunci, rapportato all'estensione del quartiere, ha determinato la densità di case Airbnb per chilometro quadrato.
Dal dataset "scores" sono state estratte solo le 7000 e più osservazioni che avessero ottenuto punteggio pari a 100 e ne se è ricercate le caratteristiche:

- che tipo di policy di cancellazione hanno: strict_14_with_grace_period (cancellazione gratuita per 48 ore, a condizione che questa avvenga almeno 14 giorni prima del check-in, dopo tale periodo, gli ospiti potranno cancellare la prenotazione fino a 7 giorni prima del check-in e ottenere un rimborso del 50% del prezzo giornaliero e delle spese di pulizia, ma non dei costi del servizio), flexible (cancellazione gratuita fino a 14 giorni prima del check-in, dopo tale periodo, gli ospiti possono effettuare la cancellazione fino a 24 ore prima del check-in e ottenere un rimborso del prezzo giornaliero e delle spese di pulizia, ma non dei costi del servizio), moderate (cancellazione gratuita fino a 14 giorni prima del check-in, dopo tale periodo, gli ospiti possono effettuare la cancellazione fino a 5 giorni prima del check-in e ottenere un rimborso del prezzo giornaliero e delle spese di pulizia, ma non dei costi del servizi), super_strict_60 (gli ospiti possono cancellare la prenotazione almeno 60 giorni prima del check-in e ottenere un rimborso del 50% del prezzo giornaliero e delle spese di pulizia, ma non dei costi del servizio, i costi del servizio Airbnb non sono rimborsabili e questi termini sono applicati solo su invito per alcuni host in circostanze speciali), strict (si applicano automaticamente alle prenotazioni di almeno 28 notti, le prenotazioni sono completamente rimborsabili nelle 48 ore successive alla conferma, a condizione che la cancellazione avvenga almeno 28 giorni prima del check-in entro le 15:00 nel fuso orario locale della destinazione, se non specificato, se trascorrono più di 48 ore dalla prenotazione, gli ospiti possono cancellarla prima del check-in per ottenere un rimborso totale, esclusi i primi 30 giorni e i costi del servizio), super_strict_30 (gli ospiti possono cancellare la prenotazione almeno 30 giorni prima del check-in e ottenere un rimborso del 50% del prezzo giornaliero e delle spese di pulizia, ma non dei costi del servizio, i costi del servizio Airbnb non sono rimborsabili e questi termini sono applicati solo su invito per alcuni host in circostanze speciali);
- qual è la tipologia che va per la maggiore tra: entire home, private room, shared room, hotel room;
- quante persone possono essere ospitate;
- in quale quartiere si trovano;

- quali servizi sono più apprezzati;

al fine di individuare i fattori che accomunano gli alloggi di successo su Airbnb.

HOTEL PROPERTIES CITYWIDE
In questo dataset si dispone di alcune informazioni relative alla posizione di alcuni hotel di New York, con nome del proprietario, quartiere, coordinate geografiche e vari codici che ne indicano l'indirizzo civico.
Si è rimosso dai nomi degli hotel caratteri speciali, poiché si è osservato che alcune catene, aventi quindi lo stesso nome, sono indicate in modi differenti. Sono state rimosse le osservazioni in cui non è presente il quartiere (latitudine e longitudine risultano mancanti e non reperibili). Si è inserito il codice univoco per quartiere, integrando il dataset con il file "NTA".
Sono stati contati quanti hotel ci fossero per tipologia nel quartiere. Se un turista vuole cercare un'alternativa ad un alloggio Airbnb nel quartiere che preferisce, può vedere che hotel sono presenti lì, facendo riferimento alla lista seguente:

- RH: hotel/boatel;
- H3: limited service; many affiliated with national chain;
- H2: full-service hotel;
- H4: motel;
- H8: dormitory;
- H1: luxury hotel;
- H9: miscellaneous hotel;
- HB: boutique: 10-100 rooms, w/luxury facilities, themed, stylish, w/full svc accommodations;
- HR: sro- 1 or 2 people housed in individual rooms in multiple dwelling affordable housing;
- HS: extended stay/suite: amenities similar to apt; typically charge weekly rates & less expensive than full-service hotel;
- H6: apartment hotel;
- H5: hotel; private club, luxury type;
- HH: hostels- bed rentals in dormitory-like settings w/shared rooms & bathrooms;
- H7: apartment hotel - cooperatively owned.

Per rendere la classificazione di più facile lettura, sono state raggruppate le varie tipologie nelle seguenti macrocategorie:

- RH, H2, H9 nella categoria "Hotel";
- H3 rimane "Limited service";
- H4, H8, HH nella categoria "Motel/Hostels/Dormitory";
- H1, HB, H5 sono rientrati nella categoria "Luxury/boutique hotel";

- HR, HS, H6, H7 hanno preso il nome di "Apartment hotel".

RESTAURANT INSPECTION NYC

Si tratta di una raccolta dati conseguente alle ispezioni sanitarie annuali (condotte nel 2017, 2018 e 2019, con quella del 2020 ancora in atto) a cui sono stati sottoposti numerosi ristoranti newyorkesi. Di essi si conosce il nome, il quartiere, le coordinate geografiche, l'indirizzo, il numero di telefono, il tipo di cucina, il giorno e il tipo dell'ispezione, il tipo e il codice di violazione, il punteggio ottenuto e la valutazione finale. Poiché non tutti disponevano di questo voto, essendo un metodo standard quello di assegnazione di un grade, si è provveduto all'inserimento di una colonna, "estimated_grade", sulla quale poter dare indicazioni, con all'interno i voti che, stando al punteggio ottenuto, i ristoranti si sarebbero meritati.
Sono state rimosse le osservazioni per quei ristoranti di cui non si conosceva né il distretto né il quartiere, si è aggiunto ai rimanenti il codice del quartiere corrispondente.
Si sono assegnate le valutazioni basandosi sullo score: 1 se ha voto A, 0.5 se ha voto B e 0 se ha voto C. Si è creata la variabile "index" in base al grado stimato (A = 1, B = 0.5, C = 0) e si è individuato il punteggio ottenuto più di recente per ciascun ristorante.
Si sono identificati i tipi di cucina che si possono trovare in un determinato quartiere, si è sommata la variabile "index" per quartiere e per cucina per sapere in quale nta si mangia meglio uno specifico tipo di cucina e da questo calcolo è possibile raggruppare ulteriormente per sapere quale quartiere ha ristoranti migliori in generale. Sono stati calcolati quanti ristoranti ci sono in ogni quartiere per poi determinare il voto medio di ognuno (somma voti/ conteggio). Infine, è stata ponderata la numerosità di ristoranti nel quartiere per l'ampiezza del quartiere per ottenere la "densità" di ristoranti in ciascun nta.
Per far sì che la visualizzazione fosse più leggibile, alcuni tipi di cucina sono stati raggruppati, in base alle loro caratteristiche, nel modo seguente:

- "Beagles/Pretzels", "Bakery", "Delicatessen", "Donuts" e "Pancakes/Waffles" sono rientrati nella categoria "Bakery";
- "Barbecue", "Chicken", "Hamburgers", "Hotdogs", "Hotdogs/Pretzels" e "Steak" sono stati rinominati come "Hamburger/Hotdog/Chicken";
- "Basque" e "Tapas" sono riesntrati nella categoria più generale "Spanish";
- "Bottled beverages; including water; sodas; juices; etc", "Cafè/Coffee/Tea", "Ice Cream; Gelato; Yogurt; Ices, Nuts/Confectionary", "Sandwiches", "Sandwiches/Salads/Mixed Buffet", "Soups" e "Soups & Sandwiches" sono stati rinominati "Cafes: beverage, ice cream, sandwiches, buffet";

- "Cajun", "Chinese/Cuban", "Chinese/Japanese", "Creole/Cajun", "Soul Food" e "Vietnamese/Cambodian/Malaysia" sono diventati un'unica categoria nominata "Fusion";
- "Caribbean", "Latin (Cuban; Dominican; Puerto Rican; South & Central American)" sono stati uniti nella categoria più generale "Latin: Carribean, Cuban, Dominican, Puerto Rican, South & Central American";
- "Fruits/Vegetables", "Juice; Smoothies; Fruit Salads", "Salads" e "Vegetarian" hanno creato un'unica categoria "Vegetarian: fruits, vegetables, juice, smoothies, salads";
- "Italian", "Pizza" e "Pizza/Italian" sono stati raggruppati in "Italian/Pizza";
- "Tex-Mex" è rientrato nella categoria più generale "Mexican".

Le osservazioni della categoria non listed, per cui non era possibile definire una tipologia di cucina, sono state rimosse.

SUBWAY ENTRANCE

Qui si trovano le entrate delle stazioni newyorkesi, il tipo, il luogo specifico, la linea che le attraversa, la divisione a cui appartengono, le coordinate geografiche sia dell'entrata che della stazione, la possibilità di acquistare o meno i biglietti sul posto, se sono attrezzate per persone con disabilità (servizi ADA), se vi è presente lo staff e in che modalità e altre informazioni dettagliate sulla localizzazione e il tipo di entrata e uscita delle stazioni. Sono state estratte quante stazioni sono presenti in ogni quartiere, quante sono servite per persone con disabilità per cui è stato possibile calcolare la percentuale di inclusività (come quantità di stazioni servite ADA/ stazioni totali per nta), oltre che la densità per quartiere, per capire quali sono quelli più forniti.

TURNSTILE

In questo dataset sono presenti i conteggi progressivi di entrate e uscite per ciascuna stazione ogni quattro ore (segnalato grazie ad un time stamp), di cui si conoscono anche altre informazioni aggiuntive, quali la divisione, il nome della linea e ulteriori informazioni inerenti alla localizzazione.

I nomi di stazioni riportati nel file "Subway entrance" (considerata come riferimeto dei nomi ufficiali delle stazioni metro di NYC) sono quelli presi come ufficiali, ma non corrispondono a quelli che si trovano nel file considerato, inoltre, solo "Subway entrance" riporta le coordinate per ogni stazione. L'obiettivo dell'integrazione dei due dataset è quello di stimare l'attività delle metro nel quartiere. A questo scopo, durante l'analisi sono state perse alcune informazioni (dovute alla scarsa corrispondenza tra i nomi delle stazioni nei due file), ma d'altronde, si voleva ottenere una semplice stima.

Sono state create due liste che contengono i valori unici delle stazioni che si trovano in "Turnstile" e in "Subway entrance". Si è usato un ciclo per implementare la

distanza di Jaro-Winkler per individuare i nomi delle stazioni in "Subway entrance" che più somigliano a quelli in "Turnstile" e si è creato un dataset contenente i confronti tra stringhe e la rispettiva distanza di Jaro-Winkler. L'obiettivo di questo passaggio è stato ottenere una corrispondenza univoca tra i nomi delle stazioni in "Subway entrance" e in "Turnstile". Sono stati analizzati più nel dettaglio i casi in cui i nomi delle stazioni nel file di riferimento si ripetevano più volte. Il dataset contenente i casi ripetuti è stato scremato e si sono tenute solo le corrispondenze che riportassero il valore di Jaro-Winkler distance più elevato. Dalle corrispondenze multiple, al netto dei doppioni, sono state eliminate quelle il cui valore massimo di Jaro-Winkler era inferiore a 0,86.
Dal dataset generale si è proceduto infine a rimuovere tutte le corrispondenze con distanza di Jaro-Winkler inferiore a 0,9.
Si è deciso di considerare solo il mese di novembre 2019, ritenuto il periodo più recente e meno influenzato dalla presenza di festività rispetto, ad esempio, a dicembre 2019.
Il dataset contiene una rilevazione ogni quattro ore per ogni tornello, di ciascuna stazione. Le rilevazioni iniziano tra mezzanotte e le tre del mattino in base alla stazione (cambiando anche nel corso del mese sulla stessa stazione) e si tratta di dati cumulati sia per le entrate, che per le uscite. Si è deciso di tenere una sola rilevazione al giorno, la prima della giornata (che tiene conto delle 24 ore precedenti) per ogni tornello.
Si è creato quindi il dataset contenente una sola osservazione al giorno per ogni tornello di ciascuna stazione e il suo valore di entrate e uscite.
Per ogni tornello è stata calcolata la differenza tra il numero di entrate/uscite tra il giorno precedente e il successivo, così da ottenere dei valori assoluti e non più cumulati.
Sono presenti dei casi anomali come quelli di dati mancanti e quelli con valore di entrate ed uscite inferiori a 0. Trattandosi di dati cumulati se i valori sono negativi significa che il dato del tornello è stato azzerato per qualche motivo (magari di un guasto o di un tornello nuovo/aggiuntivo), per cui è mancante il dato assoluto di entrata/uscita e per l'analisi diventa inutilizzabile.
Ottenuti i dati assoluti si sono sommati i valori per ogni stazione, così da aver i dati di movimento in tutto il mese per ogni stazione e non più divisi per tornello.
Si è calcolata una differenza tra il numero di uscite ed entrate nel mese. L'idea era quella di verificare se fosse una stazione con un maggior numero di entrate o di uscite.
Sono state sommate per ogni quartiere le entrate e le uscite totali e se ne è calcolata la percentuale, così da determinare quali quartieri fossero più "trafficati".
Il dataset con le entrate ed uscite per stazione è stato integrato con quello contenente le corrispondenze con i nomi nel file "Subway entrance" così da ottenere le coordinate geografiche di ogni stazione anche in "Turnstile".

Sono state perse però, le informazioni riguardo le stazioni che non hanno trovato corrispondenza nel file di riferimento o la cui corrispondenza era scarsa. Sul dataset così ristretto si è proceduto a creare i punti geometrici (grazie alla libreria "Geopandas") per ogni stazione ed individuare all'interno di quale quartiere si trovava.

NYPD CRIMINAL COURT SUMMONS HISTORIC

Qui sono raccolti alcuni dei reati commessi nella città di New York, divisi per distretto, ma non per quartiere.

Sono state rimosse le osservazioni in cui sono nulli sia i valori di latitudine e longitudine, che quelli del distretto. Sono state eliminate tutte le osservazioni che avessero come nome di distretto "New York". Per le molteplici osservazioni che possiedono coordinate, ma non quartiere, sono stati generati dei punti geometrici tramite la libreria "Geopandas" e si è verificato in quale poligono del file "NTA" ricadessero (sono stati creati cinque elementi contenenti ognuno un poligono di un distretto, poi sono stati creati cinque dataset indipendenti, contenenti ciascuno solo le osservazioni che ricadono nel distretto che dà il nome al dataset e alla fine sono stati concatenati tutti a quello originale privato delle righe con distretto nullo).

Sono state contate quante violazioni sono state compiute in ogni distretto dal 2017 (stesso periodo di osservazioni del dataset "Park events") che è stato poi ponderato con la numerosità della popolazione per distretto (fonte Wikipedia, stima dell'anno 2017) per stimare in modo più robusto i distretti con il maggior numero di "crimini" commessi.

PARKS DIMENSIONS

In questo dataset sono contenute le dimensioni dei parchi newyorkesi in acri, più qualche informazione aggiuntiva irrilevante per la nostra analisi. Non sono noti i quartieri, ma solo i distretti. ciò può essere giustificato dal fatto che uno stesso parco possa appartenere a nta differenti, a seconda della dimensione. Si sono sostituite le lettere identificative dei distretti con il nome esteso del quartiere in modo che i due dataset possedessero entrambi una colonna su cui poter effettuare un'integrazione. Si sono raggruppati i parchi per distretto e si è sommata la loro ampiezza totale, la si è trasformata in chilometri quadrati (per renderla confrontabile con le ampiezze dei distretti calcolate dal file "NTA"). Dal rapporto tra le due grandezze, moltiplicato per cento, si è ottenuta la percentuale di aree verdi per ogni distretto (quindi: area verde nel distretto / dimensione del distretto * 100).

PARKS SPECIAL EVENTS

In questo dataset si hanno informazioni relative ad eventi trascorsi e avvenuti nei principali parchi newyorkesi. Di questi si conoscono il nome dell'associazione che ha realizzato l'evento, il nome e il tipo di quest'ultimo, la data, la location e il distretto

in cui ha tenuto luogo, la categoria in cui rientra e la relativa classificazione, il tipo di pubblico a cui si è rivolto e il numero di spettatori presenti all'evento.
Per ogni distretto sono stati ricavati il numero di parchi e di eventi svoltisi tra il 2017 e il 2019 per sapere quale distretto sia più attivo. Si sono contati quanti eventi sono stati svolti in ogni distretto, sia per tipologia che per categoria, ciò per agevolare l'ipotetico turista nella scelta, in base a dove si svolgono solitamente certi eventi.

CULTURAL ORGANIZATIONS
In questo dataset sono riportate alcune delle principali organizzazioni culturali della città di New York. Di esse si conoscono: il nome, il tipo di disciplina, l'indirizzo, il quartiere e il distretto di appartenenza, il codice postale, il contatto telefonico e le coordinate geografiche.
Si è proceduto con la rimozione delle osservazioni che presentavano la colonna "nta" (quartiere) nulla e a coloro che la possedevano, è stato assegnato il codice corrispondente. Si è contato quante organizzazioni, appartenenti ad una determinata categoria, ci fossero in ogni nta e se n'è calcolata la densità per quartiere, questo per indirizzare l'ipotetico turista nella scelta del quartiere dove alloggiare, usando come criterio la vicinanza a centri che rispecchino i suoi interessi e passioni. Per realizzare una visualizzazione più leggibile, sono state create delle macrocategorie, raggruppando le categorie già esistenti nel modo seguente:

- "Architecture/design", "Museum" e "Crafts" hanno creato la categoria "Museum/Architecture/Design";
- "Botanical", "Science", "Zoo" sono rientrate in "Science/Botanical/Zoo";
- "Dance", "Folk arts", "Music" e "Theater" sono state unite nella categoria "Theater/Music/Dance";
- "Film/video/audio", "Humanities", "Literature", "Photography", "Visual arts" e "New media" hanno formato la categoria "Visual arts: photography, film, video/Literature/Humanities";
- "Multi-discipline, non perf", "Multidiscipline perf" e "Other" hanno preso il nome della più generale "Multi-Discipl, Perf & Non-Perf".

RETAIL FOOD STORE
In questo dataset sono riportati vari store presenti nella città di New York. Per ogni store si conoscono le dimensioni, il tipo di merce venduta (JAC -> Multiple operations- store- food manufacturer, A -> stores, JABC -> Multiple operations- store- backery - food manufacturer), le coordinate geografiche e il nome.
I valori di latitudine e longitudine per questo dataset erano contenuti in un dizionario nella colonna "location". Per ottenere i valori numerici di latitudine e longitudine divisi, è stata separata la colonna in due variabili. In questo modo è stato possibile generare un punto geometrico (utilizzando la libreria "Geopandas" di Python) per verificare all'interno di quale quartiere si trovasse, confrontando i punti con le

informazioni geografiche all'interno del dataset "NTA" (poligoni dei quartieri, stesso procedimento utilizzato precedentemente per i dati di "listing" di Airbnb).
Dato che non tutte le osservazioni si riferivano a New York, il procedimento di individuare all'interno di quale quartiere si trovassero le osservazioni, ha permesso di escludere quelle che non erano della città. Si è concluso il tutto con una join di tipo "inner" tra le osservazioni che sono risultate localizzate correttamente nella grande mela e il dataset intero.
Si è calcolato per ogni quartiere quanti food store ci fossero e si è ponderato per l'ampiezza degli nta per ottenere la densità e capire quali quartieri fossero più o meno serviti.

Metodi

Per quanto concerne la parte del volume, è stato utilizzato un file system distribuito (HDFS), appartenente ad Apache Hadoop, sfruttando la sua scalabilità e flessibilità per ottimizzare i tempi di gestione del quantitativo di dati a nostra disposizione. Effettuato il download dei file binari, per il settaggio di Hadoop e delle variabili d'ambiente, create le cartelle che andranno a contenere i datanode e il namenode, si è proceduto con la configurazione del framework. È stato necessario, quindi, modificare i quattro file seguenti: hadoop-env, core-site, hdfs-site e mapred-site, al fine di inserire delle proprietà del computer utilizzato, affinché il server funzionasse e fosse possibile utilizzare Hadoop localmente. A questo punto, non restava che inizializzare il namenode e testare il framework, lanciando il server.
Pyspark è stato scelto come linguaggio per estrapolare i dati e svolgere data processing, il quale si articola in tre diverse fasi: data preparation e quality, data analytics e data integration.
Per i dataset che possedevano le coordinate geografiche è stata utilizzata la libreria "Geopandas", messa a disposizione da Python, per individuare il distretto/il quartiere di appartenenza di ciascuna osservazione. Tramite lo shapefile "NTA" "Boundaries", si è potuto verificare in quale poligono (quartiere o distretto) rientrassero i punti geografici in questione e assegnare loro il nome del quartiere o del distretto. Tramite questa colonna, è possibile integrare i dataset di cui si desidera aggregare le informazioni.
La Sentiment Analysis, invece, è stata implementata con Pandas. Le review in lingua inglese degli annunci Airbnb sono state inserite, in apposite liste, le stringhe contenenti le review, pulite da punteggiatura, numeri, simboli, caratteri speciali ed emoticon. Sono state rimosse le cosiddette "stop words" e in seguito, si sono normalizzate le parole con i metodi di stemming (processo di riduzione della forma flessa di una parola alla sua forma radice) e lemmatizzazione (processo di riduzione di una forma flessa di una parola alla sua forma canonica).

Sono state calcolate per ogni review la polarity e la subjectivity. La prima varia tra -1 e 1 e indica la media dei punteggi ottenuti da ciuscuna parola all'interno della stringa, a seconda che esprima un giudizio negativo (-1) o positivo (1). La seconda invece varia tra 0 e 1, dove 0 significa "oggettivo", mentre 1 "soggettivo".
Le infografiche sono state realizzate con il software Tableau, mentre le rispettive valutazioni sono state eseguite tramite interviste e questionari online. Pandas ed Excel sono stati sfruttati per rappresentare graficamente i risultati delle valutazioni.

Criticità

Al fine di integrare i dataset sui quartieri (quindi ad un livello di granularità più specifico rispetto ai distretti come Manhattan, Bronx, ecc.), è sorta la necessità di stabilire se le coordinate geografiche di un'osservazione ricadessero in un multipoligono, piuttosto che in un altro. Trattandosi di punti geometrici, il type dell'attributo sarebbe stato "geometry", pertanto il csv è stato convertito in uno shapefile tramite Geopandas. Il problema si presentava nel momento in cui si è voluto gestire il volume, poiché l'unico risultato abbastanza soddisfacente ci è stato dato da Geopandas. Le soluzioni che sono state prese in considerazione sono le state le seguenti: Pypark, Geospark e Dask dataframe. Nel primo caso ci è stato impossibile leggere lo shapefile sull'HDFS (Pyspark, infatti, non riconosceva questo tipo di colonne), perciò si è tentato di trasformarlo in csv, perdendo inevitabilmente la caratteristica di "geometry" dell'attributo. Nel secondo caso non è andata a buon fine l'installazione di Geospark, poiché ci si è avvalsi di Pyspark e non Spark. Infine, nell'ultimo caso si è riusciti a leggere gli shapefile, ma l'analisi geospaziale è risultata ancora fallimentare. Si è valutato di utilizzare Hive con file shape e geojson, ma anche questo tentativo è risultato inconcludente. Una via ulteriore, che si è intrapresa, è stata di utilizzare MongoDB e Mongo Sharding. La criticità di questi è che necessitavano di un sistema Linux per lavorare e non disponendo di una tecnologia sufficientemente potente, oltre al conflitto che si creava tra MongoDB e Hadoop, si è dovuto rinunciare alle analisi geospaziali, poiché computazionalmente troppo onerose. Si è scelto, quindi, di limitarsi a considerare i vari quartieri di New York e a svolgere analisi in modo tale da stabilire le migliori/peggiori aree della città dove alloggiare e a determinare degli indicatori per i quali avere successo su Airbnb. Il volume è stato gestito tramite Hadoop, anche se per una più corretta trattazione dei quartieri mancanti si è dovuto ricorrere a Geopandas. L'integrazione è realizzabile tramite i quartieri.

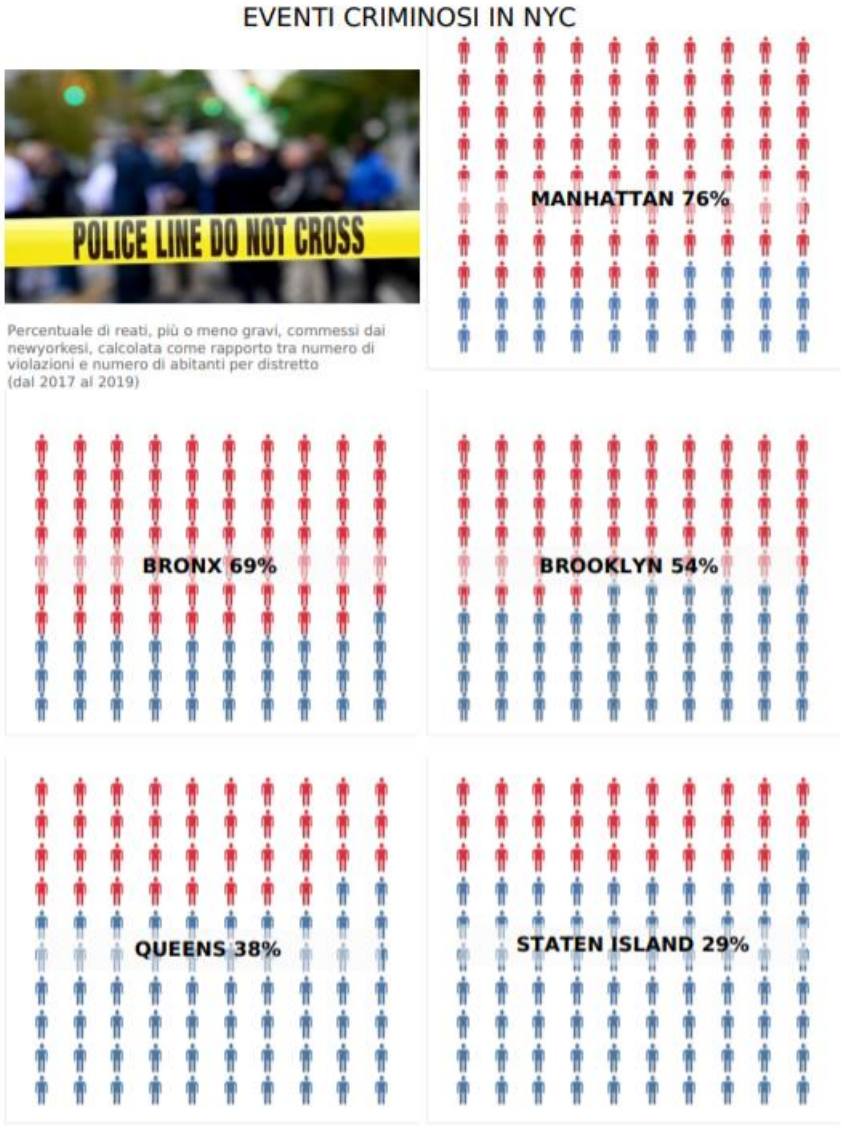
Risultati: Where to go or not to go in New York City?

Aree verdi

Si è calcolata la percentuale di aree verdi per ciascuno distretto come rapporto tra l'ampiezza dei parchi presenti in quel distretto e l'estensione del distretto stesso espressa in chilometri quadrati. Si è ottenuto che, proporzionalmente, è Manhattan a possedere la percentuale più elevata, pari all'1,26% con 103 parchi in 63,33km^2. Segue il Bronx, con lo 0,97% di aree verdi e 83 parchi in 118,66km^2, poi Staten Island, con lo 0,91% di aree verdi corrispondente a 41 parchi in 162,38km^2, dopodiché si posiziona il Queens, con lo 0,42% di aree verdi cioè 89 parchi in 304,59km^2 ed infine Brooklyn con lo 0,34%, 57 parchi in 193,76 km^2.

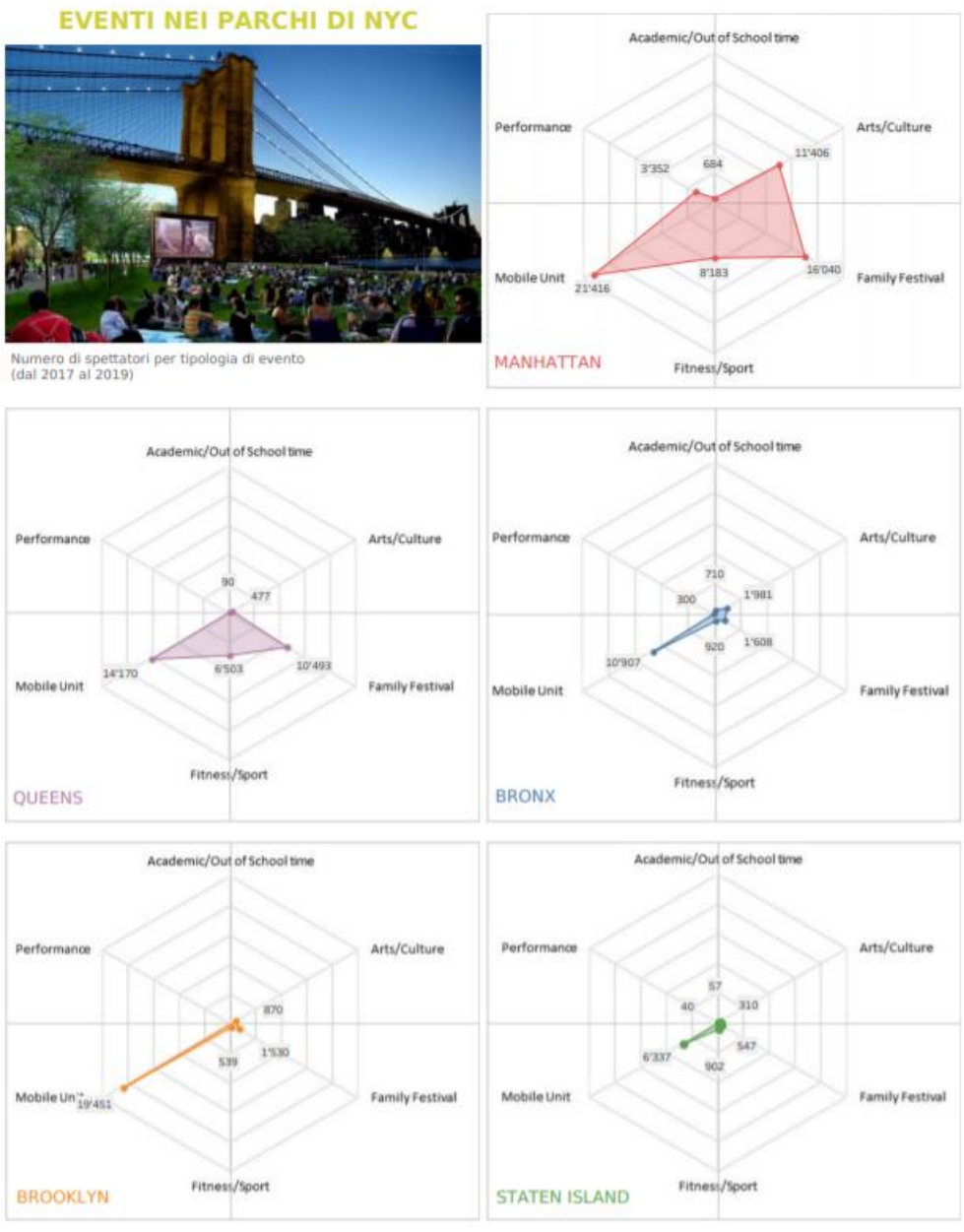
Eventi criminosi

Si è calcolata la percentuale di reati, più o meno gravi, commessi dai newyorkesi come rapporto tra numero di violazioni e numero di abitanti e si è ottenuto che Manhattan è il distretto in cui sono avvenute più trasgressioni (76%), seguita dal Bronx con il 69%, poi Brooklyn con il 54%, Queens con il 38% ed infine Staten Island con il 29%.



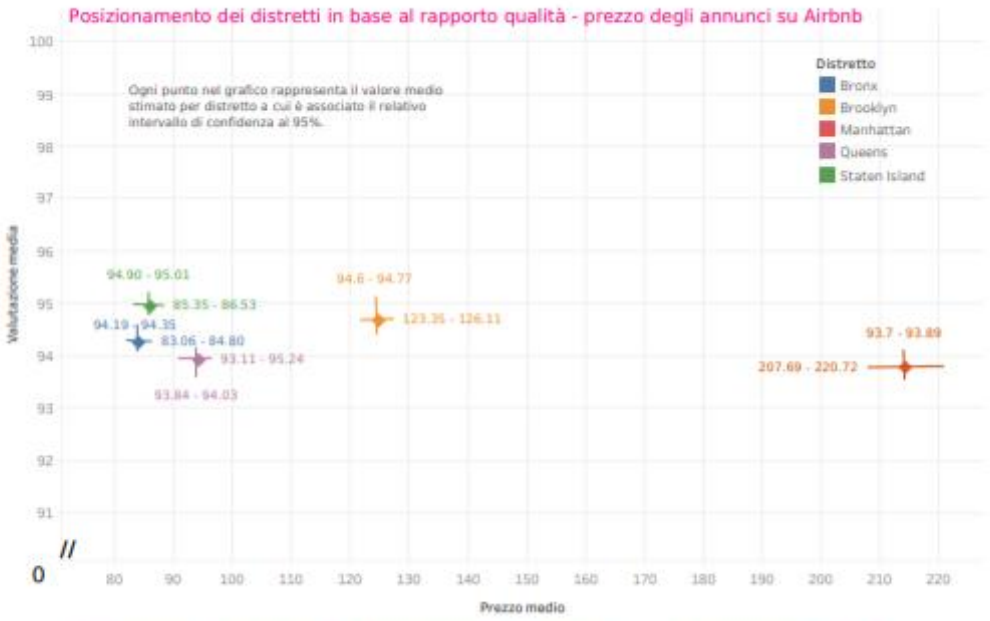
Eventi nei parchi

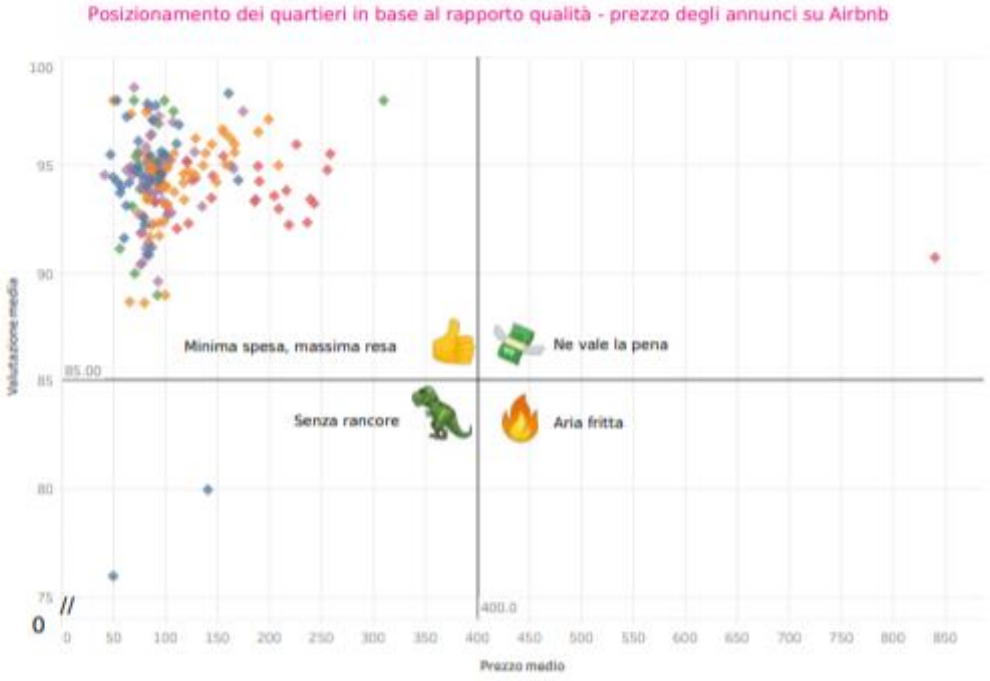
Per ogni distretto, sono stati catalogati i tipi di eventi avvenuti nei parchi newyorkesi e si è rappresentato a quali il pubblico tende a partecipare maggiormente. Per Manhattan è "Mobile Unit" il più frequentato, seguito da "Family Festival" e "Art/Culture". Per il Queens si ha una situazione analoga, solo che al terzo posto si classifica "Fitness/Sport". Nel Bronx, così come a Brooklyn e a Staten Island, solo "Mobile Unit" raggiunge un considerevole numero di spettatori.



Posizionamento dei distretti e dei quartieri in base al rapporto qualità-prezzo calcolato per gli annunci Airbnb

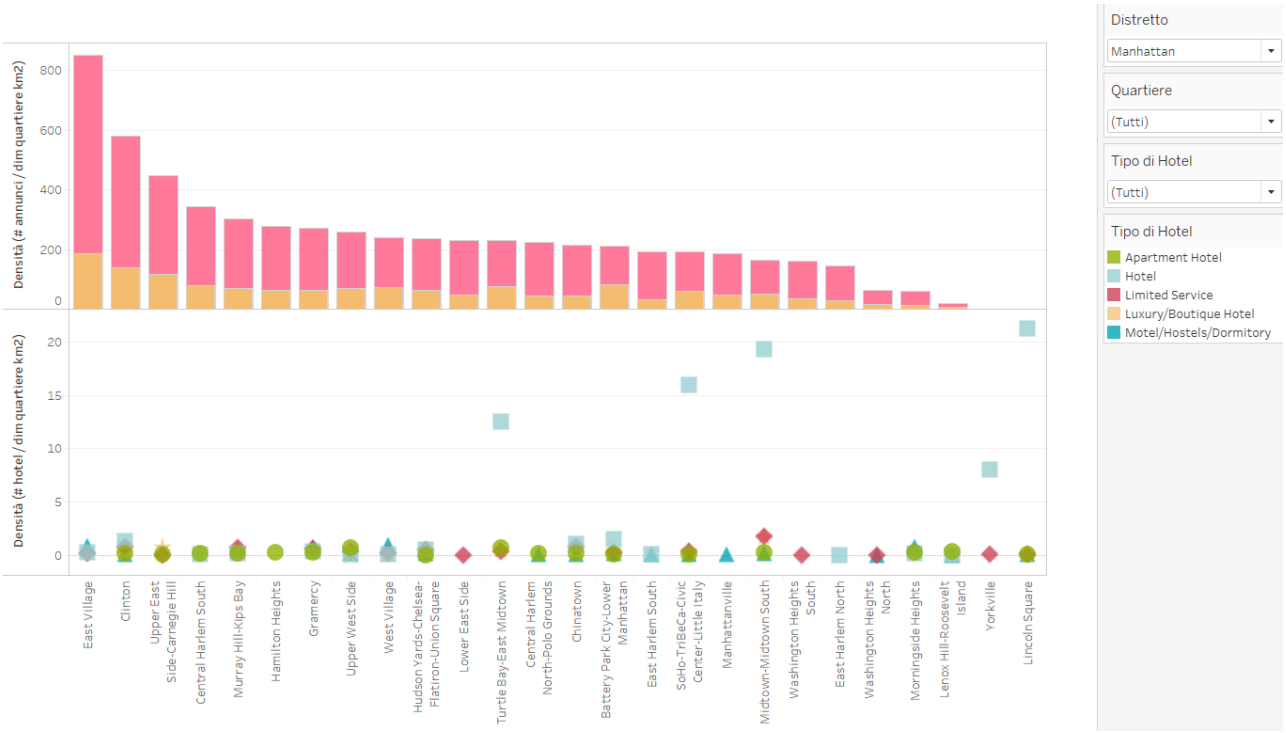
Si è realizzato uno scatterplot per rappresentare in primo luogo i distretti e in secondo luogo i quartieri, in base al valore medio di qualità prezzo degli annunci Airbnb per distretto e quartiere rispettivamente (nel primo caso con relativo intervallo di confidenza al 95%).





In quale quartiere si trovano gli annunci Airbnb? E i migliori? E se Airbnb non soddisfa… Hotel!

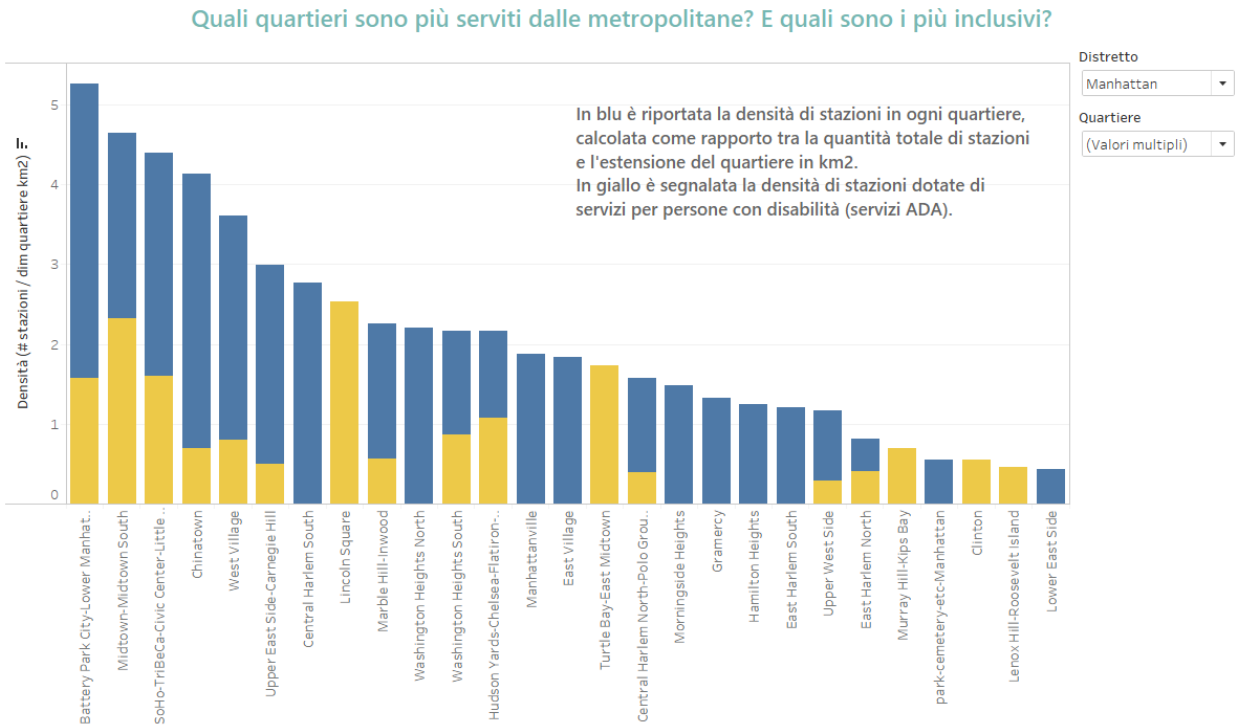
Si è calcolata la densità di annunci per ogni quartiere come rapporto tra numero di annunci su estensione dell'nta espressa in km^2. Tra questi si sono evidenziati quelli che si possono ritenere di maggiore successo, ovvero quelli con un punteggio pari a 100. Se l'offerta di Airbnb non dovesse soddisfare, è stata conteggiata la presenza in un quartiere per ciascuna tipologia di hotel e se n'è ricavata la densità allo stesso modo degli annunci Airbnb.



Es. Manhattan

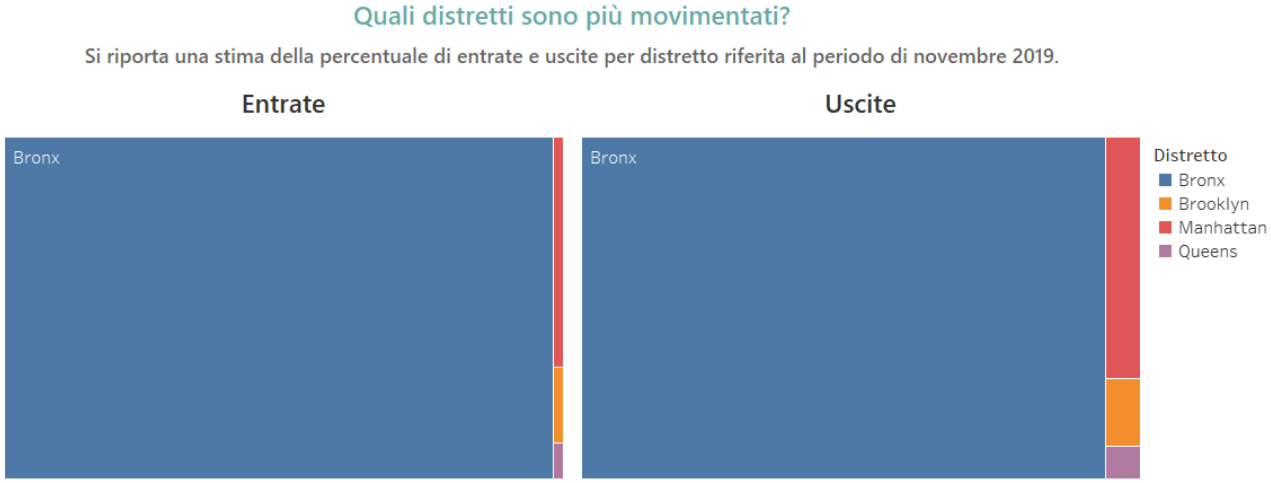
Stazioni metropolitane

Si è calcolata la densità di stazioni per ogni quartiere, dividendo il numero di stazioni per l'ampiezza del quartiere in chilometri quadrati. Si è messa in rilievo la densità delle stazioni inclusive (dotate di servizi ADA) per ogni quartiere.



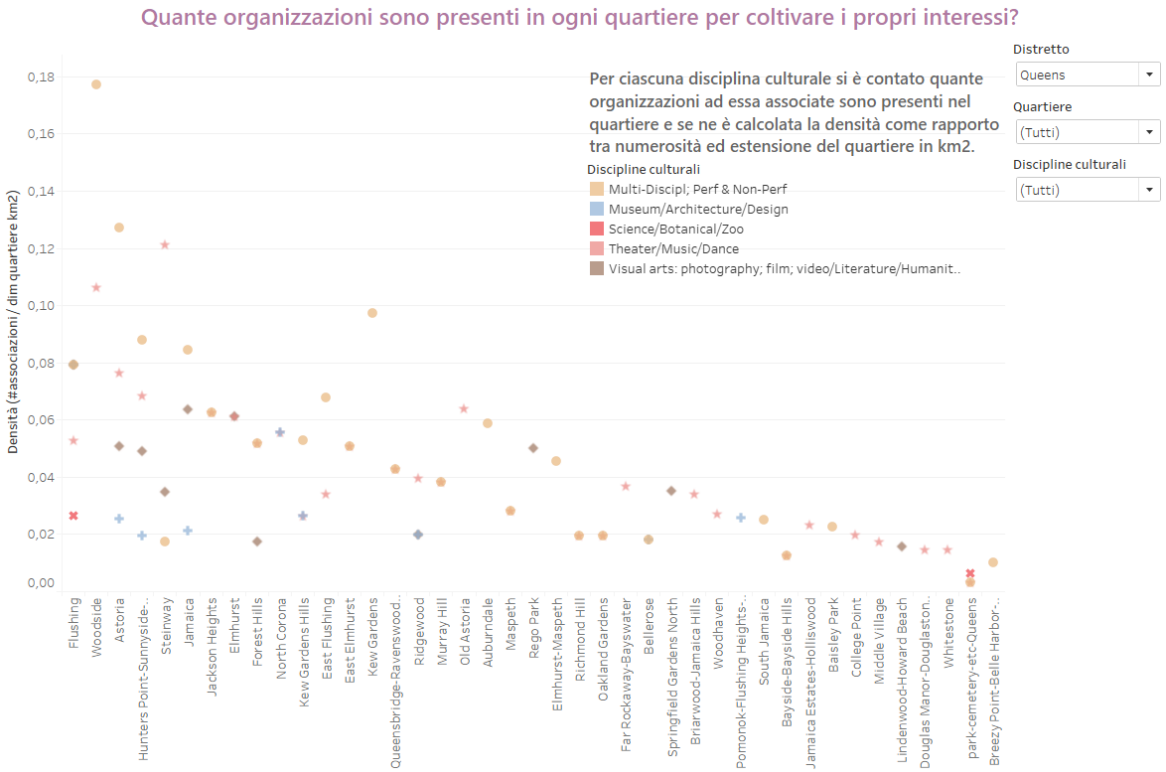
Es. Manhattan

Si è mostrata inoltre la percentuale di entrate e di uscite dai tornelli delle varie stazioni per ogni distretto e si può vedere che è il Bronx a risultare il più movimentato (ciò potrebbe essere giustificato dal fatto che è costituito per lo più da lavoratori pendolari).



Interessi culturali

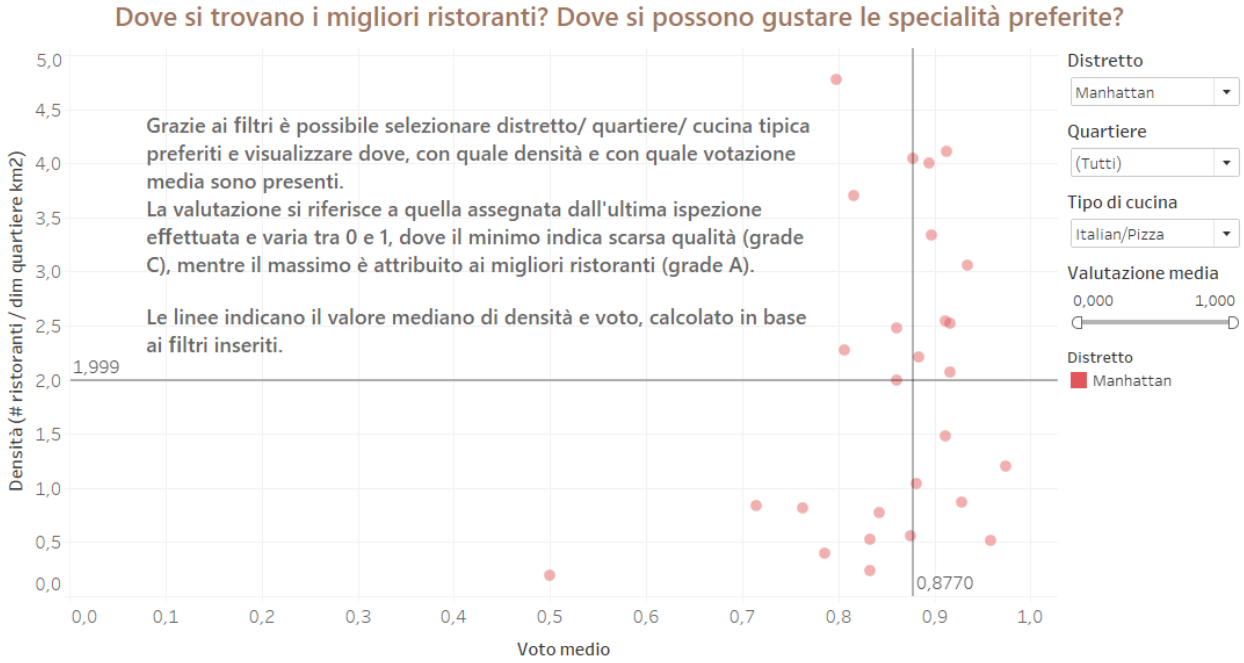
Sono state individuate le macrocategorie in cui si suddividono le principali organizzazioni culturali e si è calcolato per quartiere e distretto quante ce ne fossero per tipologia, così da ricavare la densità (rapporto tra la numerosità e l'ampiezza del quartiere in chilometri quadrati) di queste nei vari nta.



Es. Queens

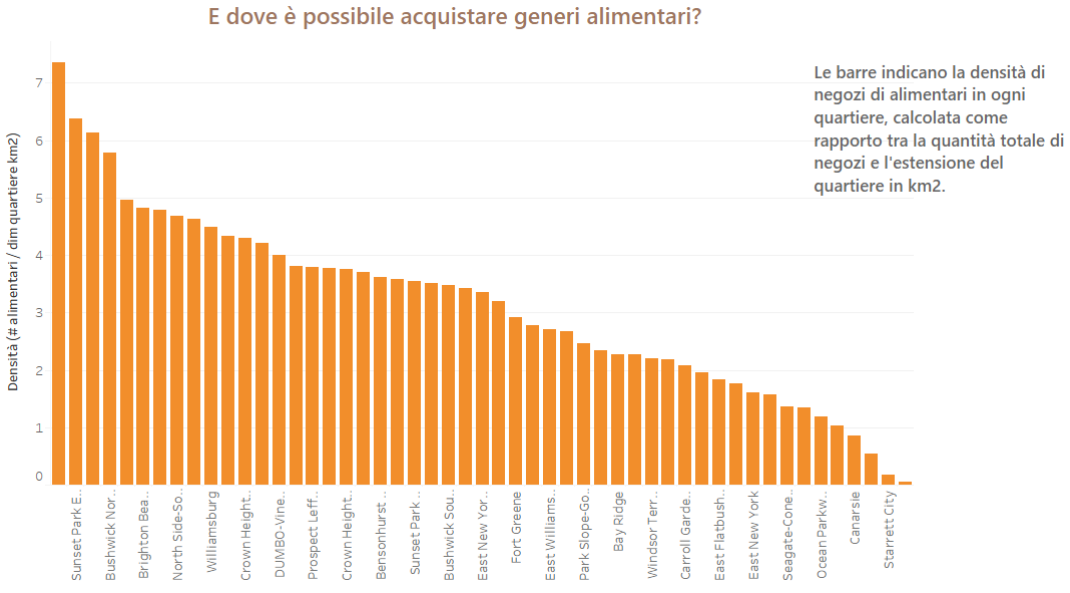
Dove e come mangiare

Nella rappresentazione è possibile visualizzare i ristoranti ispezionati per distretto, quartiere, tipo di cucina e valutazione media. In questo modo è facile individuare i quartieri dove è possibile assaporare specialità di una certa cucina meglio valutata.



Es. Cibo italiano a Manhattan

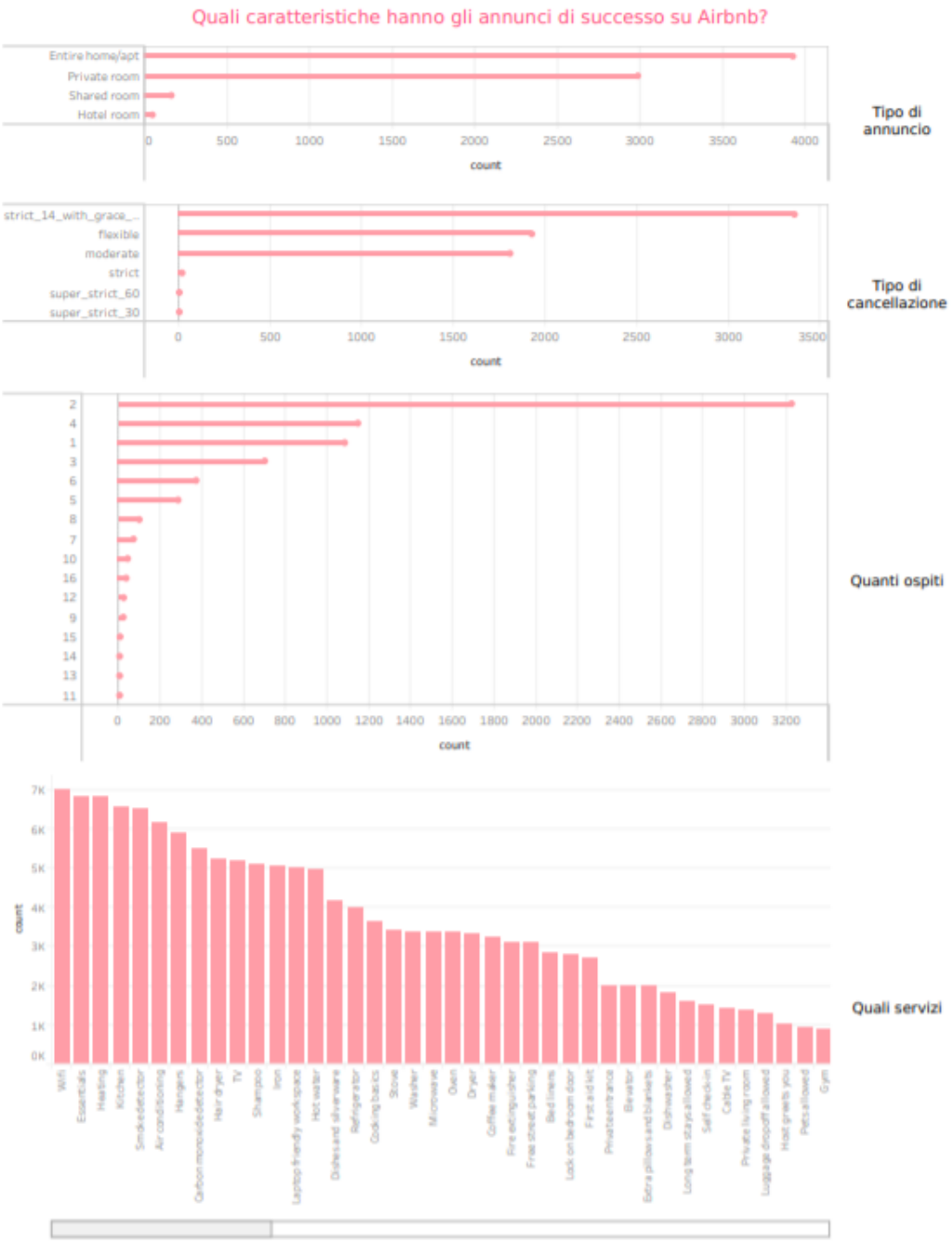
Se invece si preferisce acquistare il cibo in qualche food store, è a disposizione il grafico che mostra per ogni quartiere la densità di negozi alimentari (numero di food store rapportato all'estensione dell'nta in chilometri quadrati).



Es. Negozi alimentari a Brooklyn

Caratteristiche vincenti su Airbnb

Dopo aver estratto gli annunci qualificati come "eccellenti" (coloro che hanno ottenuto un punteggio pari a 100), se n'è individuate le caratteristiche per determinare quali accomunano gli annunci di successo su Airbnb. Come parametri si è tenuto conto del tipo di abitazione (predilette sono le case intere o gli appartamenti e le stanze private), il tipo di cancellazione (preferite sono "strict_14_with_grace_period", "flexible" e "moderate"), il numero di persone ospitabili (1,2,3,4 per lo più) e i tipi di servizi offerti dalla casa (wifi, essentials, riscaldamento, cucina, ecc).



Valutazione infografiche

Le infografiche sono state somministrate ad un campione di persone, al fine di ottenere un riscontro per apportare modifiche e migliorie al lavoro svolto. Per la valutazione euristica e gli user test, sono state intervistare rispettivamente 4 e 12 persone. Nel primo caso, agli utenti è stato chiesto di esprimere un giudizio, sia estetico che pratico, relativamente ai grafici mostrati. Nel secondo caso, invece, gli utenti sono stati sottoposti ad una vera e propria interazione con le visualizzazioni: sono state poste loro delle domande a cui dovevano rispondere servendosi dei grafici. Il questionario psicometrico è stato inviato in modo "virale" (circa una quarantina di persone raggiunte) e, alle persone che hanno partecipato, sono state chieste alcune loro generalità, oltre che di assegnare un voto alle infografiche tramite una scala di valori uguale per ciascun parametro considerato. Nel seguito si troveranno, per ciascuna tipologia di valutazione, sia le note di merito e gli aspetti positivi, che le problematiche e difficoltà riscontrate dagli utenti, oltre che a qualche rappresentazione grafica degli esiti, ove è stato possibile realizzarla.

Euristica

Si riportano le interviste effettuate per valutare l'impatto visivo e comunicativo della prima versione delle infografiche su terzi:

Fabiano, 57 anni, tecnico informatico specializzato

"Le visualizzazioni sono buone, molto chiare, molto belle, anche se io sono più abituato a grafici di sintesi. I commenti vanno benissimo, sono chiari. Gli scatterplot sono di difficile interpretazione, ma forse perché non ci lavoro. Mi sembra davvero un bel lavoro, anche molto utile e facile da utilizzare una volta acquisita dimestichezza."

Anna, 54 anni, segretaria

"Nella prime rappresentazioni lo schema è abbastanza chiaro, esteticamente è d'impatto visivo abbastanza immediato, si capisce subito. Nelle altre ho avuto più difficoltà, perché non sono abituata a questo veicolo informativo, però si capisce abbastanza. Lo "scatterplot" dei quartieri è un po' confuso. Quello dei distretti è poco visibile ma più chiaro rispetto al precedente, in generale si fa fatica a capire gli assi."

Roberto, 29 anni, libero professionista

"Nella prime rappresentazioni non capisco immediatamente il grafico, devo ragionare troppo sul significato delle percentuali e non mi piace il fatto di dovermici soffermare troppo, vorrei qualcosa di più intuitivo preferirei un grafico a torta, immediato. Inoltre, i quadratini grigi sono troppi rispetto a quelli verdi e ciò rende il grafico dispersivo. Anche il grafico delle entrate e uscite non è che sia così chiaro. Complessivamente, però, è una bella infografica e riesco a muovermi agilmente, sembra di sfogliare un catalogo."

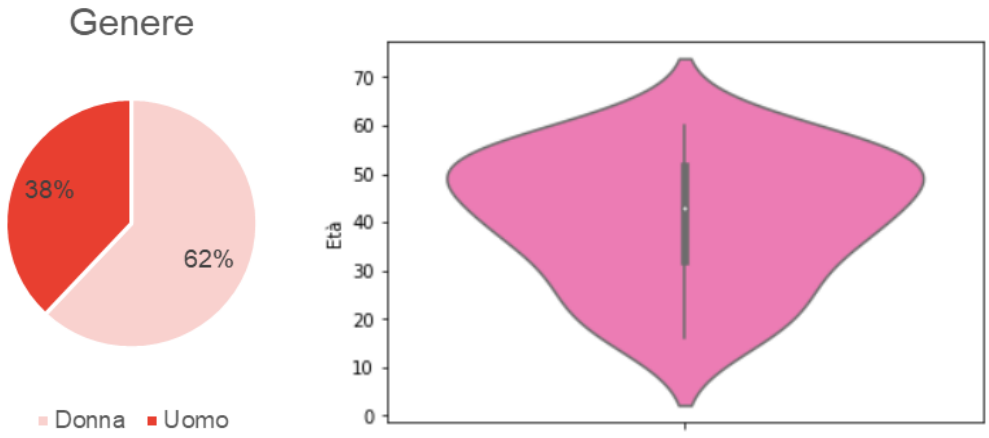
Giorgio, 24 anni, studente

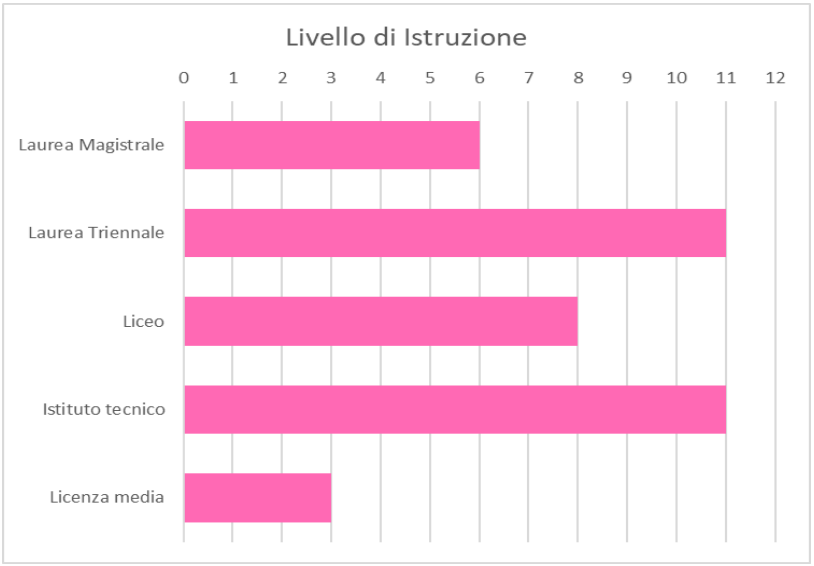
"Le visualizzazioni sono tutte molto buone, intuitive, semplici ma efficaci, soprattutto da utilizzare. Se posso muovere una critica, avrei utilizzato la mediana per suddividere in quadranti lo scatterplot del posizionamento degli annunci Airbnb per quartiere, perché indicatore per me più corretto. Inoltre, avrei rappresentato diversamente i servizi degli Airbnb di successo, perché non è chiaro che siano quelli degli annunci con punteggio pari a 100. Infine, modificherei il commento di entrate e uscite, perché impreciso e non si capisce cosa si volesse rappresentare con quei grafici."

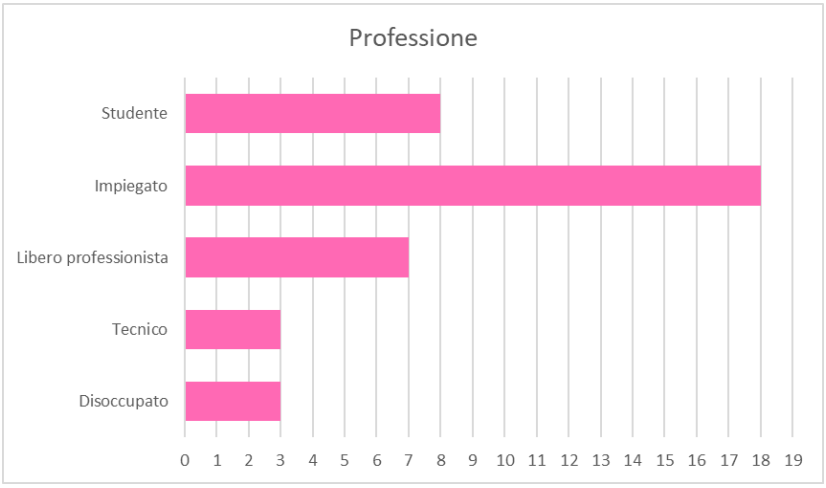
Questionari psicometrici

Al fine di valutare l'efficacia comunicativa della visualizzazione da noi realizzata, gli utenti sono stati invitati a rispondere a diversi quesiti. In una prima sezione si chiede di fornire alcune generalità, quali: genere, età, livello di istruzione e professione. Nelle due sezioni successive, sono mostrate e contestualizzate le due dashboard appartenenti alla prima versione delle infografiche. Perciò, si entra nel merito delle vere e proprie valutazioni e gli utenti sono stati tenuti ad esprimere il proprio parere riguardo a determinati aspetti delle visualizzazioni, quali: piacevolezza, comprensibilità, valore, appassionamento, prevedibilità, efficienza, chiarezza, ordine e probabilità.
Di seguito sono riportate le risposte al questionario online:

<u>Analisi Esplorativa</u>

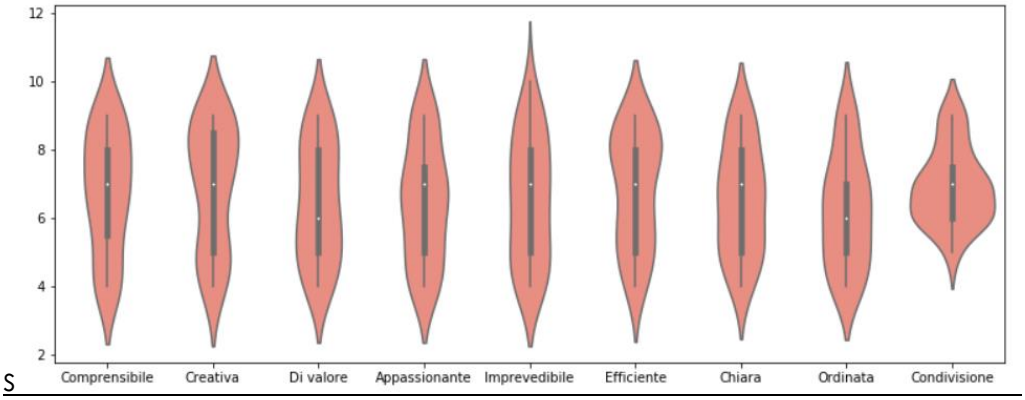






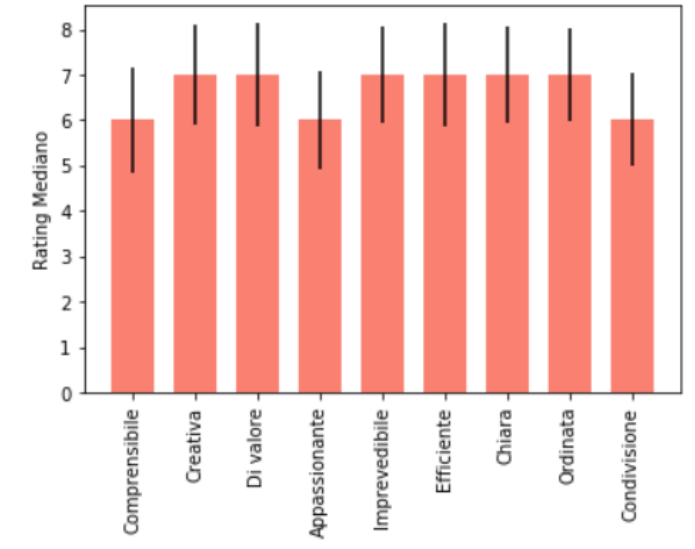
Valutazione della storia

Distribuzione della valutazione data per ogni caratteristica richiesta

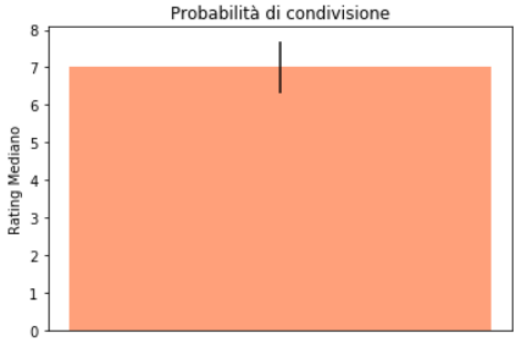


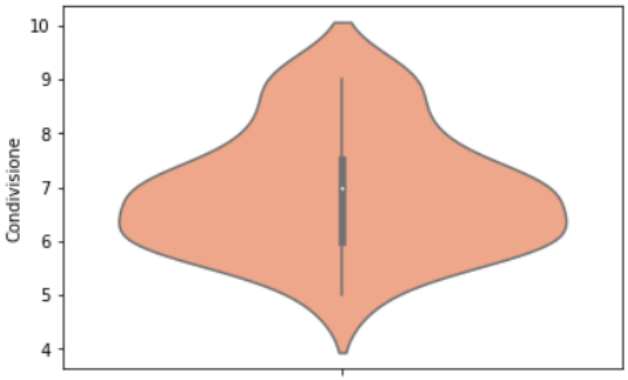
S

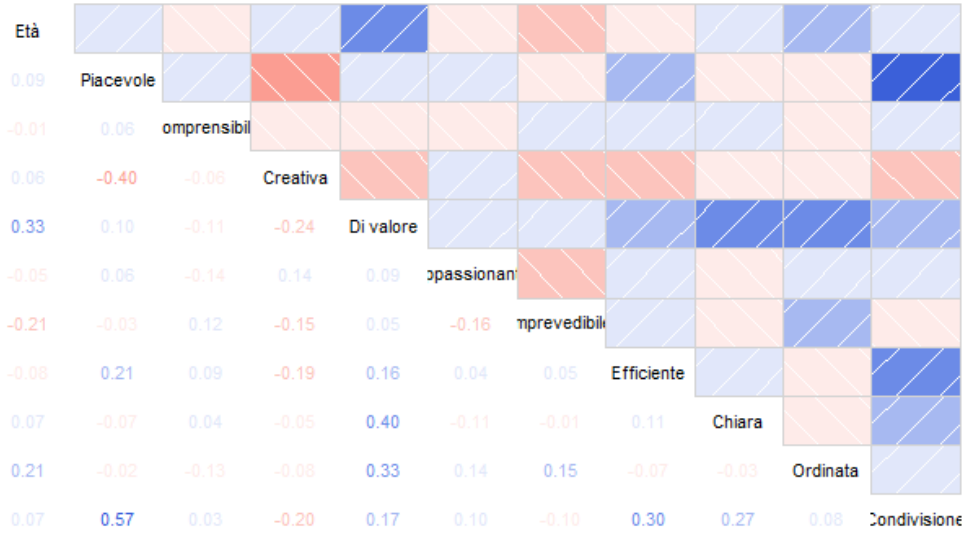
Valore mediano della valutazione espressa per ogni categoria con il relativo intervallo di confidenza al 95%



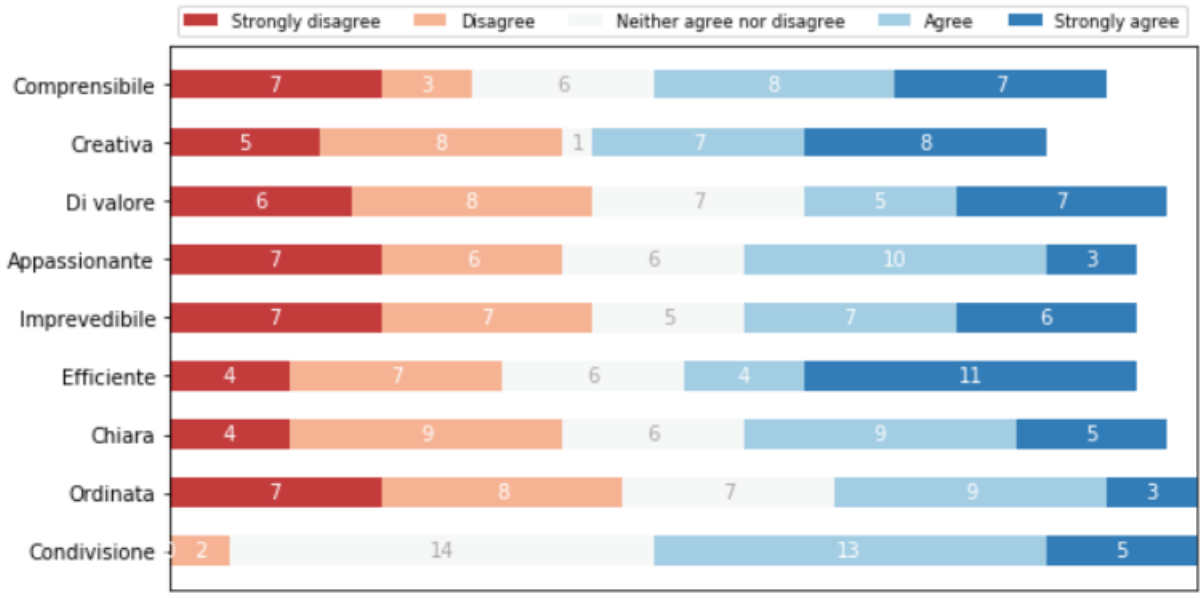
Valutazione complessiva espressa dagli utenti in termini di condivisibilità delle infografiche. A destra il valore mediano e il relativo intervallo di confidenza al 95%.







Valore assoluto delle risposte assegnate ad ogni categoria di valutazione



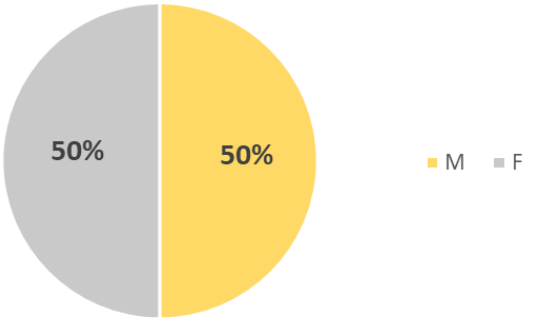
User test

Le domande, riferite alla prima versione delle infografiche, a cui è stato chiesto di trovare risposta sono le seguenti:

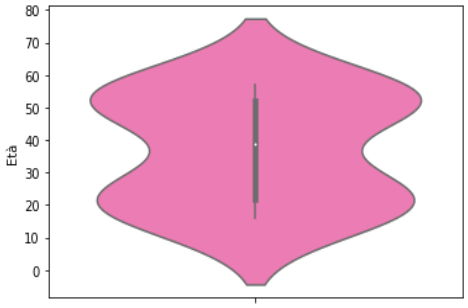
1. Cosa rappresentano i quadratini grigi?
2. Le percentuali all'interno dei grafici cosa rappresentano?
3. Quale evento è il più frequentato a Manhattan?
4. A quale quartiere di quale distretto appartengono gli annunci più economici, ma peggio valutati?
5. Quale distretto ha il numero maggiore di annunci Airbnb? E di hotel?
6. Trova il distretto con più servizi ADA.
7. Trova il distretto con meno organizzazioni per "Science/Botanical/Zoo".
8. In quale distretto si mangia meglio italiano?
9. Quale tipo di cancellazione è prediletto?

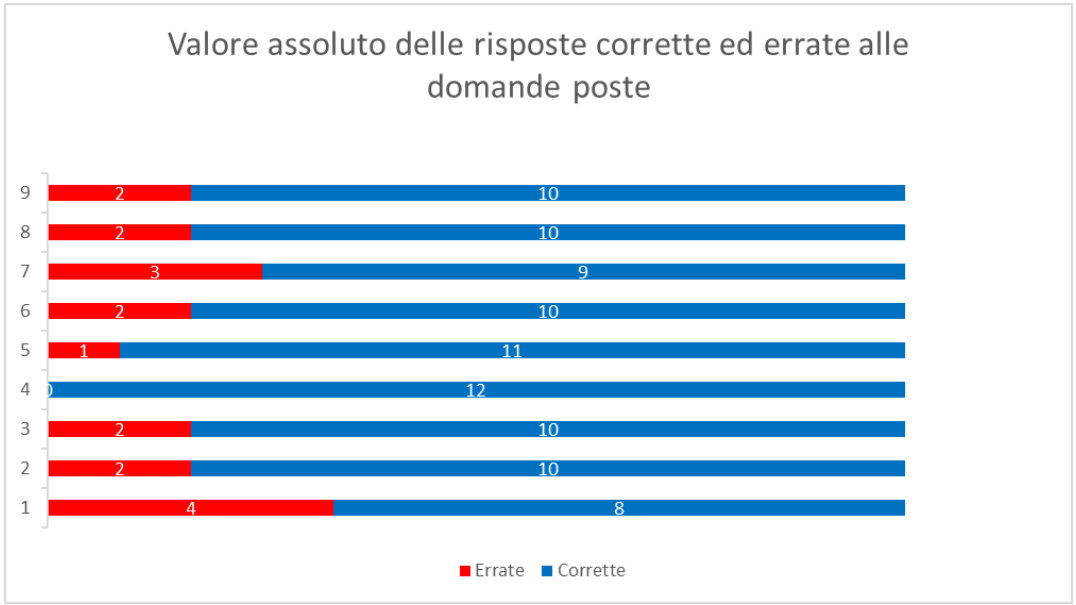
Malgrado le difficoltà riscontrate nell'interpretazione dei dati, una volta ottenute le delucidazioni necessarie, il feedback dello user test è abbastanza positivo, quasi tutti hanno risposto correttamente a buona parte delle domande. Si riportano in sintesi i risultati dell'indagine esplorativa e delle domande del test:

Genere



Età





Conclusioni e sviluppi futuri

Si è realizzata con successo una semplice guida per aiutare i turisti, che decidono di scegliere di alloggiare in un Airbnb, a selezionare il proprio appartamento (o stanza) in base alle proprie esigenze e ai propri gusti e interessi. Per gli host invece è stata stilata una lista di caratteristiche determinanti per aver successo su Airbnb da cui attingere per migliorarsi.

Tra gli sviluppi futuri rientra naturalmente la risoluzione delle criticità riscontrate, al fine di realizzare l'analisi geospaziale a cui si aspirava.