Il bike-sharing nella Grande Mela ai tempi della pandemia

Analisi trasversale dell'evoluzione del noleggio di biciclette nella città di New York tra 2019 e 2020

Anna Mattiroli[1,2], Laura Rapino[1,3], Francesco Fustini[1,4], Guglielmo Muoio[1,5]

[1] *Università degli Studi di Milano-Bicocca, CdLM Data Science*

[2]matr. 826381, [3]matr. 831346, [4]matr. 830697, [5]matr. 826029

Giugno 2021

Abstract

Partendo dall'impiego di tecnologie e metodi analitici adatti alla gestione ed elaborazione dei dati sull'utilizzo del servizio di bike-sharing nella città di New York, si è condotta un'analisi trasversale che tiene conto della stagionalità del fenomeno e che si pone come obiettivo quello di mostrare come la pandemia di COVID-19 nell'anno 2020 abbia avuto un ruolo nel cambiamento della mobilità su due ruote della città. I risultati dell'analisi mostrano che l'utilizzo del servizio di bike-sharing, dopo il periodo di lockdown nella primavera 2020, ha recuperato e superato i livelli pre-pandemici. Gli utenti e le loro abitudini di utilizzo sono cambiati, il servizio è stato riscoperto da molti come un modo per uscire, fare esercizio e anche evitare l'affollamento sui mezzi di trasporto pubblici.

Keywords: Big Data, New York, Bike-sharing, COVID-19, MongoDB, Kafka, Dataviz, Tableau

Indice

1 Introduzione

Negli ultimi anni il fenomeno del bike-sharing ha assunto un ruolo importante nella rete dei trasporti di numerose città. CitiBike è un servizio privato per il noleggio di biciclette che serve la città di New York [1]. I dati sull'utilizzo del servizio sono messi a disposizione pubblicamente in forma anonima. Dal sito di CitiBike è infatti possibile reperire facilmente la cronologia dei dati di viaggio fino al 2017. Il set di dati include gli orari di inizio e di fine di ciascun viaggio, la stazione di partenza e di destinazione, l'età, il sesso dell'utente e se si tratta di un utente abbonato o occasionale.
Si è scelto di lavorare esclusivamente con i dati relativi agli anni 2019 e 2020, anno di diffusione dell'infezione da SARS-CoV-2. L'utilizzo delle biciclette a noleggio è per natura un fenomeno stagionale e fortemente influenzato dalle condizioni meteorologiche, per questo sono stati ottenuti e utilizzati nell'analisi anche i dati delle temperature e precipitazioni giornaliere dal sito Meteostat [2]. Inoltre, al fine di studiare l'impatto che la pandemia ha avuto sulla mobilità nella città di New York, sono stati recuperati anche i dati riguardanti il numero di contagi giornalieri.

2 Domanda di ricerca

Si è partiti dall'osservazione della stagionalità del fenomeno di bike-sharing, prendendo in considerazione il meteo nella zona di New York, per poi concentrare la domanda di ricerca su come la pandemia di COVID-19 nell'anno 2020 abbia influenzato l'utilizzo del servizio. In particolare, ci si è anche chiesti come sono cambiate le abitudini di utilizzo nel periodo di riapertura, con un focus sul confronto del traffico tra il mese di Settembre 2020 e 2019. Gli utenti più "affezionati" al servizio sono ancora gli stessi? Le stazioni più frequentate sono cambiate? Le persone utilizzano le bici per ragioni diverse da prima? E la mobilità cittadina segue gli stessi andamenti orari dei periodi pre-pandemici?

3 Data Management

Con il termine Big Data si fa riferimento a grosse quantità di dati informatici così grandi, veloci o complessi che per essere elaborati richiedono necessariamente l'impiego di tecnologie e metodi analitici specifici poiché non possono essere gestiti con metodi tradizionali. I fattori che li identificano sono dunque principalmente tre: Volume, Varietà, Velocità.
L'attività di bike sharing nella città di New York genera ogni anno decine di milioni di log in tempo reale derivanti dal noleggio delle biciclette. Questi dati sono caratterizzati da due delle tre V dei Big Data: Volume e Velocità.
Inoltre, nello sviluppo del progetto sono stati utilizzati anche i dati relativi alle condizioni meteo della città di New York ed i numeri giornalieri di casi confermati di COVID-19. È stata quindi effettuata un'integrazione spazio-temporale dei dati, andando così ad affrontare anche l'aspetto della Varietà dei Big Data.
I dati sono stati raccolti dal sito del servizio CitiBike, tramite la libreria Requests [3] di Python.
I dati ottenuti in formato JSON si presentavano completi e puliti, sono state necessarie solo minori operazioni di preprocessing al fine di ottimizzare la memorizzazione. Si è provveduto a selezionare solo le variabili strettamente necessarie e scartare quelle ridondanti, come ad esempio la latitudine e la longitudine delle stazioni di partenza e di arrivo le quali sono state memorizzate separatamente e reintegrate tramite ID nella fase di visualizzazione, solo quando necessarie. Un'altra operazione che ha permesso di ridurre la dimensione di ciascun documento è stata quella di rinominare le variabili attribuendogli nomi della lunghezza massima di due caratteri. Infatti, a differenza dei dati in formato tabellare, ciascun documento di un file JSON ha memorizzati al suo interno anche i nomi delle variabili. Lavorando con una mole di dati notevole, questa semplice operazione ha permesso di ridurre il volume totale dei dati memorizzati.
Alla fine della fase di preprocessing i singoli documenti erano strutturati nel seguente modo:

```
{'S': 4065,
'E': 3959,
'ST': datetime.datetime(2020, 12, 1, 0, 0, 3, 777000),
'ET': datetime.datetime(2020, 12, 1, 0, 11, 38, 316000),
'B': 49454,
'U': 'Customer',
'BY': 1988,
'G': 2,
```

```
'D': 694}
```

dove 'S' è il numero della stazione di partenza, 'E' il numero della stazione di arrivo, 'ST' comprende il giorno e l'ora di partenza del viaggio, 'ET' il giorno e l'ora di arrivo, 'B' è l'ID della bicicletta utilizzata, 'U' una variabile categoriale che assume valore 'Customer' quando la persona che noleggia la bicicletta è un utente occasionale, oppure 'Subscriber' nel caso di un utente abbonato. 'BY' è l'anno di nascita dell'utente e 'G' il suo genere (0=sconosciuto; 1=uomo; 2=donna). Infine 'D' rappresenta la durata totale del viaggio espressa in secondi.
I dati sono stati poi gradualmente caricati in una collezione di MongoDB [4] secondo lo schema di lavoro in Figura 1. Si evidenzia il fatto che le V di Volume e Velocità sono state affrontate solo in modo simulato. Nello specifico è stata implementata per la gestione del volume una partizione orizzontale, sviluppata in locale tramite dei container Docker [5] che simulano la presenza di più macchine. Apache Kafka [6] invece è stato utilizzata per simulare l'acquisizione di una parte dei dati in tempo reale. Entrambe le simulazioni sono approfondite nelle rispettive sezioni 3.1, 3.2.
Una volta memorizzati, i dati sono stati estratti all'occorrenza con delle queries scritte tramite la libreria PyMongo [7]. Infine, sono stati aggregati con le informazioni provenienti dalle altre fonti dati e salvati in formato CSV, più adatto alla fase di visualizzazione finale con Tableau [8].

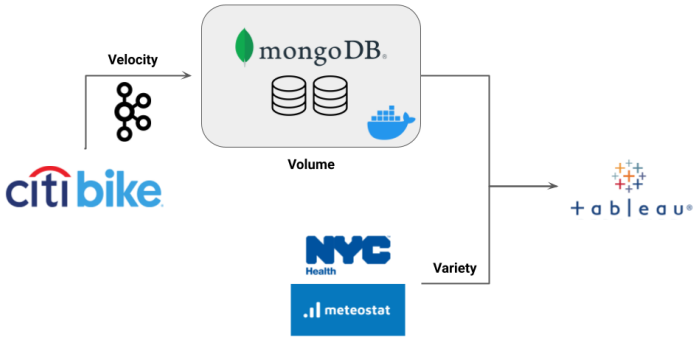


Figura 1: Workflow

3.1 Volume

L'ingente mole di dati a disposizione (quasi 8 gigabyte) si è posta come un problema da affrontare fin da subito. Per gestire in modo adeguato il volume è stato quindi necessario valutare non solo un DBMS opportuno agli scopi prefissati ma anche un'architettura scalabile in grado di ottimizzare le prestazioni.
Dopo alcune valutazioni, si è scelto di utilizzare MongoDB, un DMBS di tipo non relazionale in cui i dati vengono archiviati in formato BSON (Binary JSON) e letti tramite indici. MongoDB è particolarmente noto per la sua scalabilità orizzontale ottenibile aumentando il numero di nodi. Infatti, esso lascia la possibilità di creare dei cluster per permettere sia di aumentare le capacità di storage sia di aumentare le performance. Questa procedura viene comunemente chiamata sharding: con questa espressione si intende la tecnica per la creazione di un database distribuito in cui ogni nodo contiene una porzione del database [9].
Per realizzare un cluster in sharding sono necessarie 3 componenti, di cui almeno una per tipo:

- *router mongos:* nodo che funge da interfaccia alle applicazioni client.
- *shard:* partizione del database che sfrutta la tecnica della replicazione attraverso due o più nodi in modo tale da evitare perdite permanenti di dati nel caso di fallimento di un nodo.
- *config server:* l'entità che conosce la distribuzione dei dati nel cluster e quindi come reperirli o a quale nodo inviare le operazioni da effettuare.

L'architettura è stata implementata su una macchina locale tramite Docker con questa configurazione: 2 shard (ciascuno con 3 nodi replica), 1 config server (con 3 nodi replica) e 1 router mongos [10]. Per disporre in modo equilibrato i dati sui 2 shard, si è utilizzata una chiave di tipo *hashed* che viene generata automaticamente da mongoDB. Si è scelta questa configurazione al fine di effettuare una simulazione semplice ma quanto più realistica.
Una volta abilitato lo sharding sul database creato, sono stati caricati progressivamente i dati. La risultante distribuzione dei dati negli shard è riassunta in Figura 2.

```
mongos> db.NewYork.getShardDistribution()

Shard shard1 at shard1/mongoshard11:27017,mongoshard12:27017,mongoshard13:27017
 data : 1.98GiB docs : 19680975 chunks : 42
 estimated data per chunk : 48.34MiB
 estimated docs per chunk : 468594

Shard shard2 at shard2/mongoshard21:27017,mongoshard22:27017,mongoshard23:27017
 data : 2.05GiB docs : 20377579 chunks : 41
 estimated data per chunk : 51.27MiB
 estimated docs per chunk : 497014

Totals
 data : 4.03GiB docs : 40058554 chunks : 83
 Shard shard1 contains 49.13% data, 49.13% docs in cluster, avg obj size on shard : 108B
 Shard shard2 contains 50.86% data, 50.86% docs in cluster, avg obj size on shard : 108B
```

Figura 2: Distribuzione dei dati tramite sharding

3.2 Velocità

I dati utilizzati nel corso del progetto sono composti dallo storico dei viaggi avvenuti negli anni 2019 e 2020. Per loro natura però, i dati derivanti dall'attività di bike sharing possono considerarsi "veloci" in quanto essi vengono generati ad altissima frequenza e quindi possono essere acquisiti in tempo reale. Per questo motivo si è deciso di effettuare una simulazione della reale acquisizione e memorizzazione dei log dei noleggi utilizzando Apache Kafka, una piattaforma ben nota per il data streaming.
Kafka permette di implementare un meccanismo che garantisce la gestione della velocità dei dati tramite un approccio di *data ingestion* con un sistema a coda caratterizzato da due fasi fondamentali: *publish & subscribe*. Nella prima fase una componente *producer* si occupa di "catturare" il flusso di dati in tempo reale e di pubblicarli in una coda kafka senza curarsi che vengano successivamente letti da qualcuno.
Nella seconda fase una componente *consumer* potrà leggere i dati processati dal primo all'ultimo arrivato nella coda e memorizzarli dove necessario. In questo modo i problemi di lettura e di memorizzazione dei dati, quando sono "veloci", sono separati e gestiti da entità distinte.
Nello specifico sono stati utilizzati per la simulazione i dati del mese di Dicembre 2020. La velocità è stata simulata tramite l'utilizzo congiunto delle funzioni `time.sleep(random.randint(0,5))` rispettivamente delle librerie time [11] e random [12] di Python. In questo modo il producer aggiunge un'osservazione alla coda Kafka in un tempo variabile tra 0 e 5 secondi. Il consumer poi legge i singoli dati dalla coda e li carica nella collezione creata di MongoDB, predisposta per la memorizzazione di tutti i dati dello storico necessari allo svolgimento del progetto.

3.3 Varietà

Al fine di rispondere alla domanda di ricerca nel modo più completo possibile, si è reso necessario un lavoro di integrazione dei dati.
In una prima fase si è recuperato un file JSON, proveniente dalla stessa fonte dei log dei viaggi, contenente molteplici informazioni riguardanti le singole stazioni di noleggio. Le più rilevanti allo scopo si sono rilevate nome, capacità e, in coppia, latitudine e longitudine. Difatti queste ultime hanno permesso la creazione e lo sviluppo di grafici geo-referenziati nella parte di data visualization. L'ID della stazione ha permesso il collegamento con i dati principali.
Successivamente, sono state utilizzate altre due fonti. La prima è il dipartimento della salute di New York [13], dalla quale si sono ricavate alcune variabili circa la diffusione della pandemia nella città, permettendo quindi un'analisi locale più accurata sulle variazioni di utilizzo del servizio di noleggio in relazione al fenomeno in corso.
La seconda invece è Meteostat, una banca dati meteorologici e climatici con record meteo dettagliati e statistiche climatiche per migliaia di stazioni meteorologiche e località in tutto il mondo. Questo perché si è voluto tenere conto anche del fattore meteorologico data la natura stagionale dell'utilizzo del servizio. Tuttavia, New York ha un'estensione piuttosto ampia e le stazioni climatiche sono molteplici. Per risolvere quindi il problema si è calcolato il baricentro geografico delle stazioni di noleggio per poi interpolare i valori delle diverse stazioni ed ottenere un valore unico giornaliero per i seguenti parametri: temperatura (C°), precipitazioni di pioggia (mm), precipitazioni di neve (mm) e velocità del vento (km/h). Entrambe le fonti sono state integrate temporalmente ai dati principali aggregati a livello giornaliero tramite Python.

4 Estrazione ed esplorazione dei dati

4.1 Creazione datasets

Una volta immagazzinati i dati, sono state create delle queries mongo ad hoc e dei notebook Python sia per effettuare una prima esplorazione sia per sviluppare dei dataset opportuni al fine di creare delle infografiche su Tableau.
Tra i diversi set di dati creati ne possiamo identificare 3 principali, sfruttati per la creazione delle infografiche finali:

- *dataset 1:* sono stati recuperati i dati aggregati a livello giornaliero da mongoDB (ex. numero di viaggi totali, durata media, età media, etc.) e sono stati integrati successivamente con i dati COVID e i dati meteo. Infine, per praticità, si è voluto trasformare il dataset da formato long a formato wide ottenendo tutte le variabili di entrambi gli anni su un'unica riga per ciascun giorno dell'anno.
- *dataset 2:* questo dataset si è sviluppato in modo analogo al precedente senza effettuare integrazioni ma considerando una variabile aggiuntiva: il numero di viaggi di piacere, *leisure* in inglese. In particolare si è preferito, in questo caso, tenere i dati in formato long.
- *dataset 3:* al fine di performare un confronto della mobilità giornaliera in modo più dettagliato tra i due anni è stato necessario scegliere un periodo di riferimento ridotto. Si è scelto il mese di Settembre poiché, sia a livello epidemiologico che meteorologico, è il mese caratterizzato da differenze minori tra i due anni. Ciò garantisce una minimizzazione del bias e quindi eventuali distorsioni nel confronto. Quindi, sono stati recuperati i dati a livello di tupla (stazione/ora/giorno della settimana) ed è stata calcolata la media delle variabili: viaggi in partenza, in arrivo, totali, durata, etc.

4.2 Primi risultati

Dopo aver reperito i dati, averli memorizzati in una collezione MongoDB ed averli integrati con le informazioni aggiuntive riguardanti il meteo e la diffusione di COVID-19, il lavoro è proceduto in una fase esplorativa di visualizzazione dei dati grazie all'utilizzo del software Tableau. In particolar modo sono stati utilizzati i dataset ottenuti al punto 3.3 per la creazione di visualizzazioni che permettessero di scoprire ed evidenziare alcune delle caratteristiche ed eventuali modifiche nelle abitudini di utilizzo del servizio di bike sharing tra 2019 e 2020. Il primo grafico in figura 3, rappresenta le serie storiche del numero di viaggi totali per settimana per gli anni 2019 e 2020 a confronto. Si legge dal grafico l'andamento stagionale proprio del fenomeno, inoltre è evidente come nei mesi primaverili di Aprile e Maggio ci sia stato un crollo nell'utilizzo del servizio, dovuto al diffondersi della pandemia e alle misure di restrizione messe in atto.

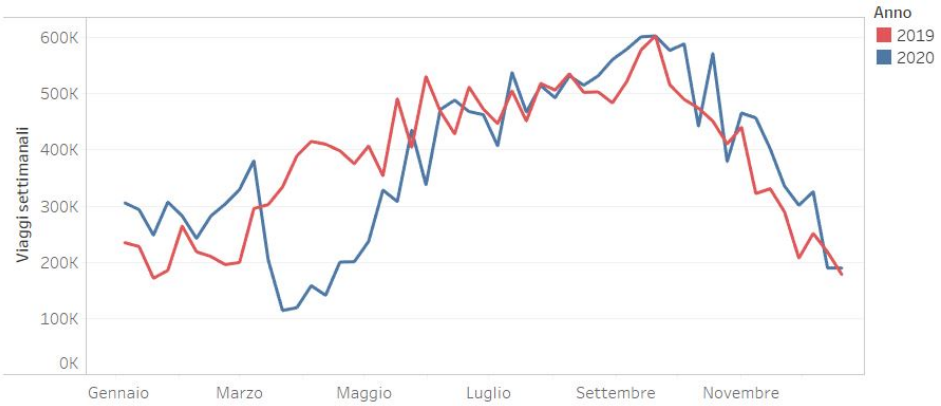


Figura 3: Serie storica del numero di viaggi totali per settimana del 2019 e 2020

I grafici in figura 4, mostrano a livello intuitivo gli incrementi e decrementi di utilizzo per giorno della settimana e per tipologia di utente. Si può notare come a livello generale siano diminuiti i viaggi infra-settimanali, a favore dei viaggi nei giorni di Sabato e Domenica. Questo andamento non è però generalizzabile a tutti gli utenti. É evidente come il numero di viaggi effettuati da utenti di genere maschile nel 2019 fossero per lo più concentrati nei giorni lavorativi, mentre nel 2020 si nota una generale diminuzione del numero di viaggi effettuati da questi utenti e una bassa differenziazione tra giorni della settimana. Un pattern simile si verifica nel comportamento degli utenti che possiedono un abbonamento annuale al servizio. Diverso è il caso degli utenti di genere femminile e degli utenti "occasionali". In questi due casi si nota un notevole aumento del numero di viaggi, principalmente concentrati nei giorni di Sabato e Domenica.

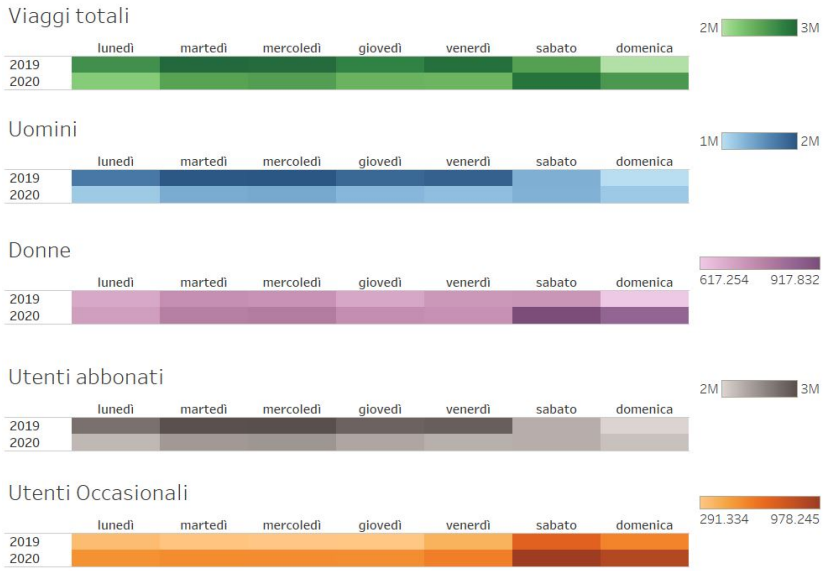


Figura 4: Numero di viaggi all'anno per tipologia di utente

Un'ulteriore evidenza è mostrata dal grafico in Figura 5, ottenuto dal dataset contenente il traffico orario per i mesi di Settembre 2019 e 2020. É possibile confrontare il numero di viaggi totali effettuati nei due anni per ciascuna ora del giorno. Si nota che il traffico nell'anno 2019 era caratterizzato da due incrementi nell'arco della giornata corrispondenti all'inizio e fine del tipico orario lavorativo, rispettivamente uno la mattina tra le 7 e le 9 ed uno nel tardo pomeriggio, nelle ore 17 e 18. Nel 2020 si può osservare una distribuzione dei viaggi differente. L'aumento di traffico nelle prime ore del mattino è pressoché scomparso, mentre si nota un generale incremento del numero di viaggi in tutte le ore pomeridiane. Questo fenomeno potrebbe essere giustificato in parte dal fatto che tra le misure introdotte per il contenimento della pandemia c'è quella del lavoro da casa.
È evidente come già questa prima fase esplorativa confermi il fatto che la pandemia ha avuto un impatto sulla mobilità cittadina e sull'utilizzo del servizio di bike-shairng.

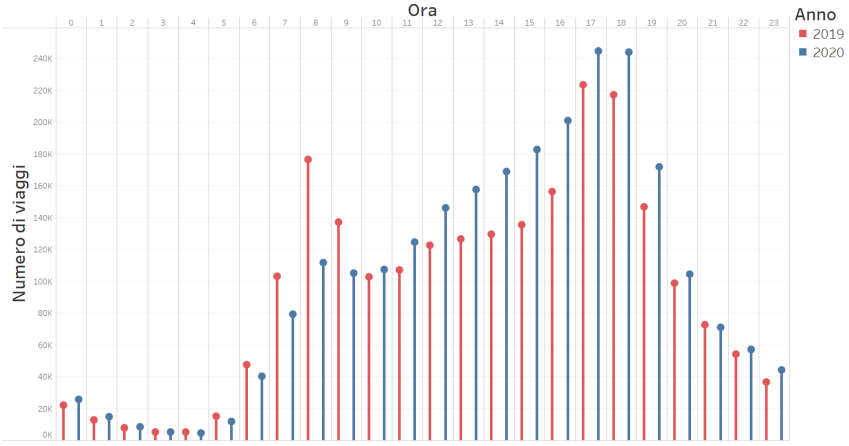


Figura 5: Numero di viaggi nel mese di Settembre per ora del giorno

5 Data Visualization

Nella seguente sezione ci si pone come obiettivo di avvicinare gli utenti all'interazione con i dati in studio e di rispondere alle domande di ricerca tramite la realizzazione di infografiche interattive che possano comunicare in maniera efficace i risultati ottenuti. L'attività di visualizzazione dei dati infatti, oltre ad essere uno strumento utile nelle fasi di esplorazione, è anche il modo migliore per portare i risultati ottenuti da un'analisi alla coscienza di molti. L'interattività delle visualizzazioni permette anche di lasciare all'utente stesso la possibilità

di esplorare e scoprire i dati in autonomia.
Al fine di rendere le infografiche realizzate accessibili ad un più ampio insieme di utenti possibile, si è cercato di coinvolgere nelle fasi di valutazione descritte nella sezione 5.2 un campione di utenti eterogeneo. In particolare, sono state coinvolte persone di entrambi i sessi di età tra i 15 e i 70 anni. Inoltre, sempre al fine di aumentare l'accessibilità alle infografiche, sono state utilizzate palette di colori che tenessero conto di eventuali forme di daltonismo dove necessario.
Le visualizzazioni finali possono essere consultate all'indirizzo [14]. Queste sono state ottenute, ciascuna partendo da un dataset diverso tra quelli descritti al paragrafo 4.1:

- Infografica 1 con il dataset 1
- Infografica 2 con il dataset 2
- Infografica 3 con il dataset 3

5.1 Infografiche

5.1.1 Infografica 1: L'impatto della pandemia sul noleggio delle biciclette a New York

La prima infografica rappresenta due serie storiche che confrontano in parallelo la differenza percentuale di viaggi tra il 2020 e il 2019 e i casi confermati di COVID-19. Sono affiancate verticalmente tenendo in comune l'asse delle ascisse sul quale c'è la variabile tempo rappresentata dall'anno 2020.
Per procedere con la creazione della prima infografica i dati sono stati estratti da MongoDB e aggregati a livello giornaliero garantendo, tra le tante, le seguenti variabili:

- *Date:* data di rilevazione con il formato gg/mm/aaaa
- *Trips:* numero di viaggi totali
- *Rainfall:* mm di precipitazione acquosa
- *Snow:* mm di neve caduta
- *Case Count:* numero di casi COVID-19

Dopo aver importato i dati sul software Tableau, sono stati calcolati due ulteriori campi:

- *Differenza percentuale dei viaggi:* variazione percentuale dei viaggi tra il 2020 e 2019

$$\frac{T_{20} - T_{19}}{T_{19}} * 100$$

  dove T_i è il valore di *Trips* nell'anno i.

- *Differenza assoluta delle precipitazioni:* differenza di precipitazioni totali tra il 2020 e il 2019

$$(R_{20} + \frac{S_{20}}{10}) - (R_{19} + \frac{S_{19}}{10})$$

  dove R_i indica la somma di mm di pioggia nell'anno i e S_i la somma di mm di neve nell'anno i.

Approssimativamente si considera che 1 cm di neve caduta equivalga a 1 mm di precipitazione acquosa, per questo nella formula della *Differenza assoluta delle precipitazioni* la neve caduta espressa in mm viene convertita in cm.
Il grafico è suddiviso in una parte superiore e una inferiore: la parte superiore mostra la variabile *Differenza percentuale dei viaggi* aggregata livello settimanale utilizzando barre colorate rispetto alla variabile *Differenza assoluta delle precipitazioni* mentre nella parte inferiore si riporta la curva dei contagi da COVID-19 a New York, più precisamente la somma di casi confermati a livello settimanale. In particolare, il colore delle barre del grafico superiore è determinato dalla variazione assoluta delle precipitazioni nel 2020 rispetto all'anno precedente: più il colore è intenso verso il blu, più le precipitazioni saranno state maggiori nel 2020, al contrario se il colore è più intenso verso il rosso, le precipitazioni saranno state maggiori nel 2019. L'infografica punta quindi ad avere una visione complessiva della variazione dei viaggi tra i due anni tenendo in considerazione tutti i fattori in questione.
Dal grafico riportato è evidente come ci sia una tendenza negativa con una variazione fino al -70% di viaggi in meno nel 2020 rispetto al 2019 corrispondente al picco dei contagi fino a maggio; al contrario, nei primi mesi del 2020 si nota una tendenza positiva fino al 60% di viaggi in più. Le barre riferite a quelle settimane hanno quasi tutte colori molto caldi, a testimonianza del fatto che l'inizio dell'anno 2020 è stato caratterizzato da giornate meno piovose rispetto al 2019. Prendendo in considerazione la settimana del 3 e 4 marzo, ad esempio, si riscontra un aumento del 59.8% di viaggi giustificato dal fatto che nel 2019 in quel periodo c'è stata una bufera di neve, mentre nel 2020 le precipitazioni sono state scarse (variazione assoluta di precipitazioni totali: -47.7 mm). Si nota un'altra settimana dal colore intenso nella scala del blu: il 10 luglio 2020 c'è stato un nubifragio nella città, infatti si ha un'ampia variazione assoluta di precipitazioni totali (100.7 mm) che accompagna una differenza percentuale di viaggi del -11.7%.

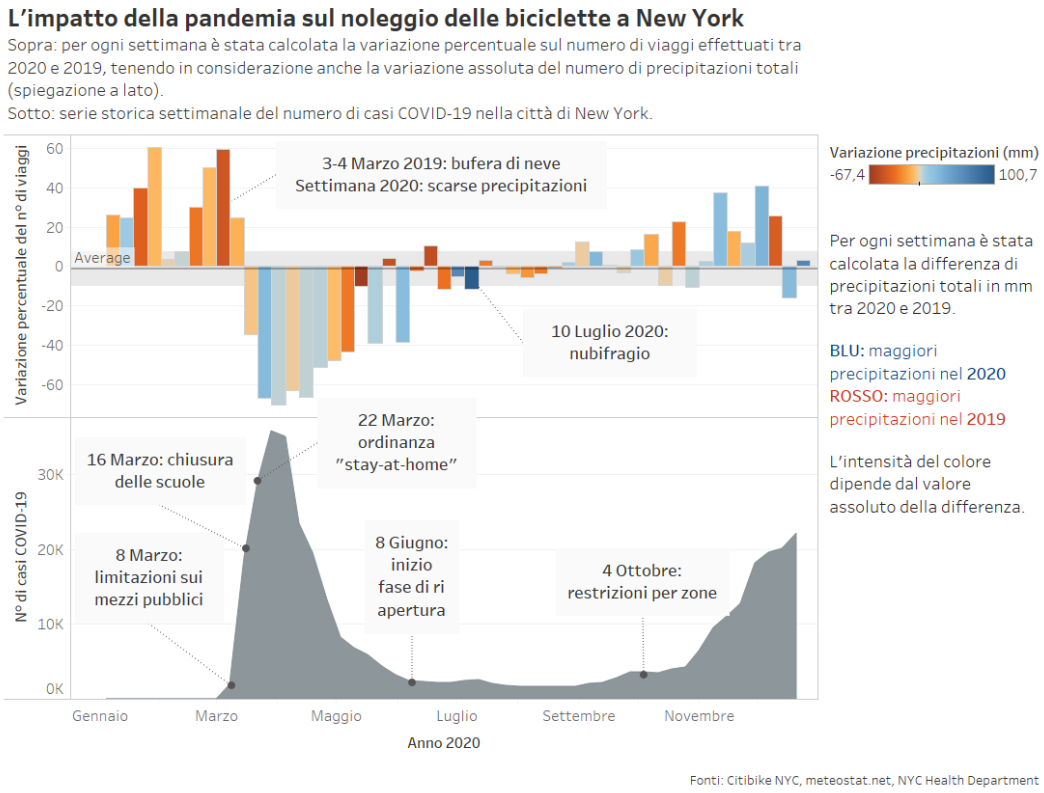


Figura 6: Infografica 1

5.1.2 Infografica 2: Viaggi totali vs Viaggi di piacere: una nuova tendenza

La seconda infografica è un grafico di dispersione con i *Viaggi di piacere* sull'asse delle ascisse e sull'asse delle ordinate i *Viaggi totali*. Con viaggi di piacere si fa riferimento a quei viaggi le cui la stazioni di partenza e arrivo coincidono. Sono definiti "di piacere" poiché si immagina che un simile comportamento derivi dalla volontà dell'utente di noleggiare una bicicletta per un semplice giro ricreativo, anziché utilizzare il servizio per recarsi in un luogo specifico.
Gli elementi del grafico sono i mesi del 2019 e 2020, congiunti da un segmento in modo tale da osservare la variazione a livello mensile da un anno a quello successivo. Inoltre le osservazioni sono colorate in base alla stagione meteorologica del mese di rilevazione.
È dunque utile soffermarsi sul campo stagione che si riferisce alle stagioni meteorologiche e non astronomiche. Infatti, le prime tengono conto esclusivamente di cambiamenti climatici, quindi più utili per le nostre analisi. Le stagioni meteorologiche sono definite nel seguente modo:

Stagioni Meteorologiche		
Nome	*Inizio*	*Fine*
Inverno	1 dicembre	28/29 febbraio
Primavera	1 marzo	31 maggio
Estate	1 giugno	31 agosto
Autunno	1 settembre	30 novembre

I segmenti con cui si presenta il grafico uniscono il punto che rappresenta il numero di viaggi di piacere e viaggi totali per il 2019 con lo stesso punto riferito al 2020. In tale maniera si riesce a evidenziare qual è stata la variazione del trend nei diversi mesi delle stagioni meteorologiche tra i due anni.
Essendo i segmenti raggruppati per stagioni, è possibile stabilire un andamento per ciascuna. Nei mesi di Gennaio e Febbraio 2020, precedenti il diffondersi della pandemia, si può osservare un leggero aumento proporzionale sia dei viaggi totali che di quelli di piacere. Nei mesi primaverili si assiste ad un crollo dei viaggi totali. In particolare nel mese di aprile, durante la prima ondata, sebbene il numero di viaggi totali risulti più che dimezzato, il numero dei viaggi di piacere è comunque aumentato. In estate, dopo le prime riaperture di giugno 2020, la lunghezza dei segmenti evidenzia un notevole aumento dei viaggi di piacere a parità del numero

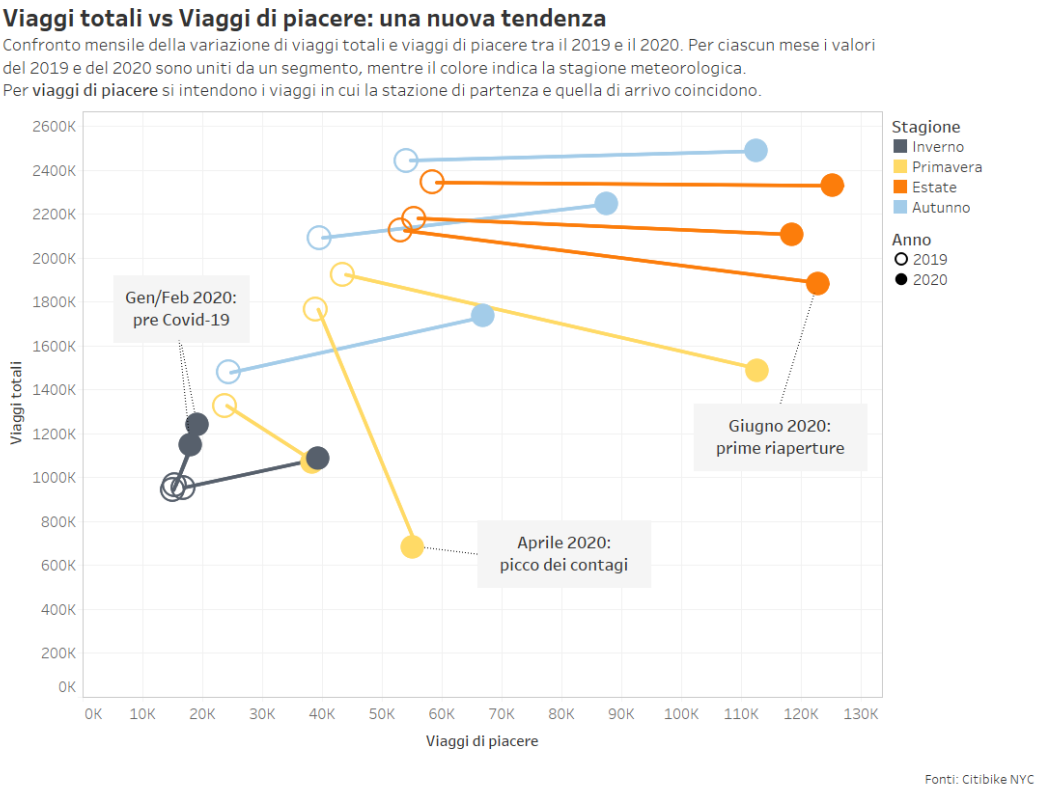


Figura 7: Infografica 2

di viaggi totali rispetto all'anno precedente. Infine, in autunno, si verifica un aumento sia del numero dei viaggi di piacere che del numero di viaggi totali, a testimoniare il fatto che questa abitudine si sia consolidata tra gli utenti anche dopo la fine del periodo estivo. Lo stesso può essere affermato per quanto riguarda il mese di Dicembre 2020.

5.1.3 Infografica 3: Mappa interattiva

La terza infografica riporta la mappa della zona di New York con la posizione di tutte le stazioni considerate nello studio rispetto ad alcune aree note della città come i parchi più importanti (Central Park e Prospect Park), il centro e le zone residenziali. Ogni stazione è rappresentata da un punto: la grandezza è proporzionale al traffico medio nel 2019 mentre il colore dipende dalla variazione percentuale di traffico medio tra i due anni. La tinta è più calda se il traffico è stato maggiore nel 2019 e più fredda se il traffico è stato maggiore nel 2020. La variazione percentuale viene calcolata definendo la variabile "traffico" come la somma dei viaggi in arrivo e in partenza da una determinata stazione.
Differenza percentuale traffico: variazione percentuale del traffico medio tra il 2020 e 2019

$$D_s = \frac{T_{s20} - T_{s19}}{T_{s19}} * 100$$

dove T_{si} è il valore di *Traffico* nell'anno i per la stazione s.
Trattandosi di una mappa interattiva, è possibile filtrare per fascia oraria e giorno della settimana in modo tale da esplorare i cambiamenti con diverse configurazioni.
In generale, si nota un forte calo dei viaggi nel 2020 rispetto al 2019 soprattutto nella zona centrale di New York, come nel caso della stazione di Pershing Square North che ha registrato un calo del -55.9% dei viaggi, passando da una media di circa 50 viaggi all'ora per il 2019 a una media di 21.69 nel 2020. Al contrario, nella zona settentrionale di Manhattan e nelle aree periferiche al di là del fiume Hudson, le stazioni hanno registrato un aumento notevole di viaggi. Ciò suggerisce una tendenza da parte degli utenti a preferire le aree meno affollate della città, lontane dal centro.

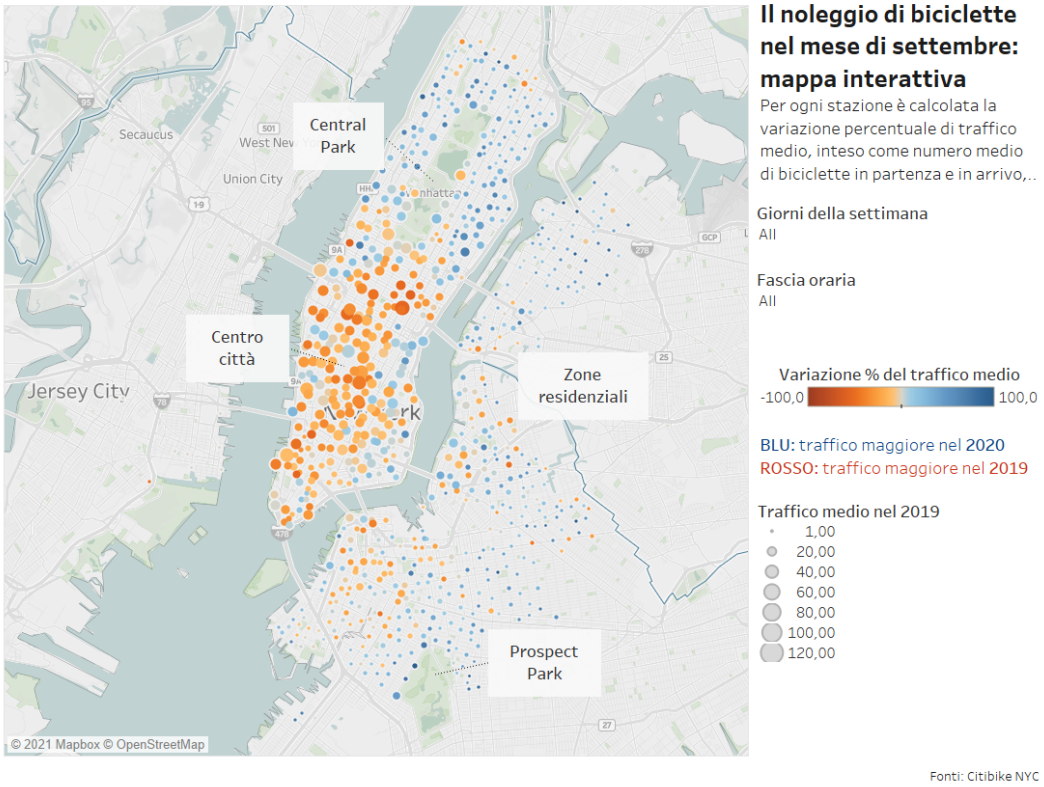


Figura 8: Infografica 3

5.2 Quality assessment

Nella realizzazione di un'infografica di successo è importante procedere in modo iterativo, alternando fasi di progettazione e creazione a fasi di verifica e correzione al fine di garantire un graduale miglioramento. A tale scopo campioni di utenti sono stati coinvolti in vari modi nelle diverse fasi di sviluppo con l'obiettivo di ottenere feedback utili a indirizzare il lavoro al miglior risultato possibile per gli utenti finali.

5.2.1 Valutazione euristica

In un primo momento è stato utilizzato il protocollo *think aloud*. Sono state coinvolte 4 persone a cui è stato chiesto di pensare ad alta voce mentre osservano le infografiche per la prima volta, senza alcuna indicazione specifica di lettura. Questo ha così permesso di raccogliere informazioni utili ad identificare ed eliminare eventuali ostacoli alla comprensione delle visualizzazioni e garantire una fruizione più *user friendly*.
Il primo problema emerso riguardava la comprensione del significato dei colori utilizzati nella serie storica della prima infografica in Figura 6. La semplice legenda "Variazione precipitazioni (mm)" si è rivelata non essere sufficiente, così si è deciso di aggiungere la dicitura:

"BLU: maggiori precipitazioni nel 2020
ROSSO: maggiori precipitazioni nel 2019".

Analogamente nella seconda infografica è sorto un altro problema di comprensione. Così è stata aggiunta nella descrizione la definizione di "viaggio di piacere", prima mancante e non facilmente intuibile per gli utenti.
L'infografica che a seguito della valutazione euristica ha subito maggiori modifiche è stata la terza. Inizialmente infatti la mappa interattiva si presentava come in Figura 9. Le stazioni potevano essere selezionata sia dalla cartina che dallo scatterplot, così che i grafici si aggiornassero automaticamente a vicenda. La bisettrice del grafico a destra era pensata per agevolare il confronto tra il traffico medio 2019 e 2020 per ogni stazione. La visualizzazione così strutturata però ha generato non poca confusione tra gli utenti intervistati. Si è deciso, per la versione finale, di eliminare lo scatterplot e di aggiungere alla cartina l'informazione riguardo al traffico medio dell'anno 2019 sfruttando la dimensione dei punti sulla mappa. In questo modo le stazioni che nel 2019 ricoprivano un ruolo maggiore spiccano nella mappa, mentre per il confronto tra gli anni è sempre possibile ricondursi ai colori. Inoltre, ora emerge molto più chiaramente che le stazioni che hanno subito un aumento di traffico medio nel 2020 sono generalmente quelle che nel 2019 erano meno frequentate, mentre in Figura

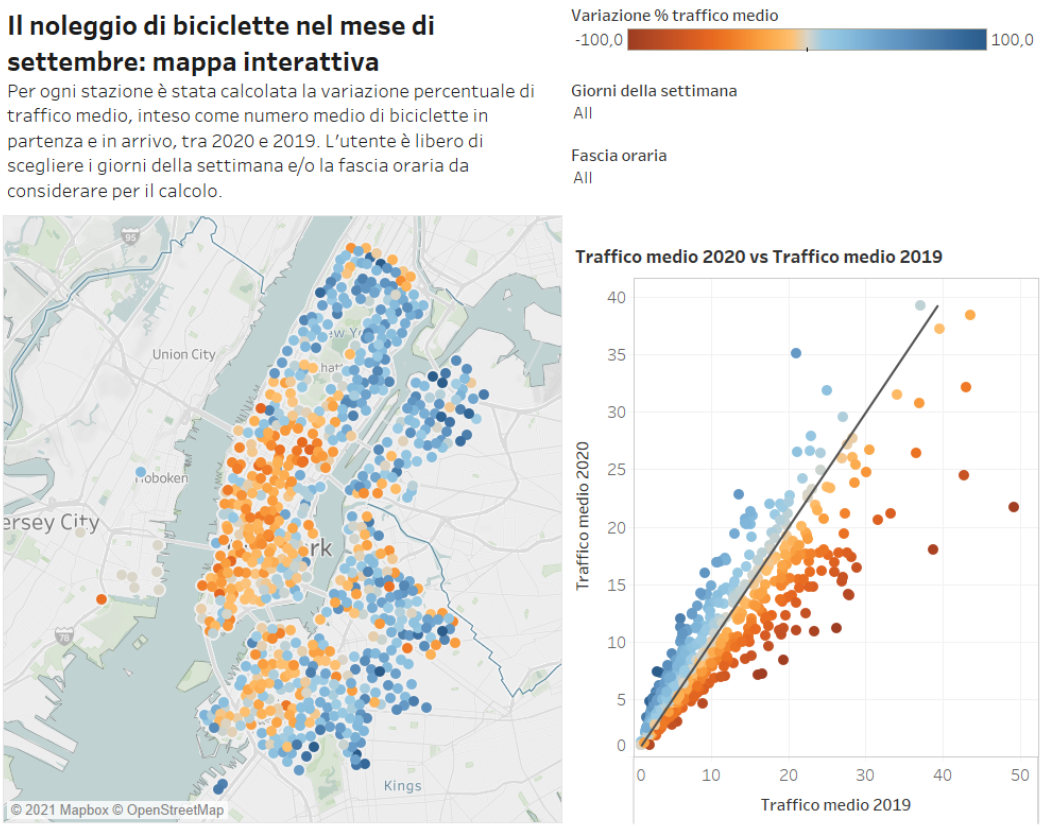


Figura 9: Mappa interattiva, versione non definitiva.

9 venivano tutte rappresentate indistintamente. Come per per la prima infografica è stato aggiunto nella descrizione:

"BLU: traffico maggiore nel 2020
ROSSO: traffico maggiore nel 2019".

5.2.2 User Test

Dopo le prime correzioni, per valutare quantitativamente le visualizzazioni sono state coinvolte 12 persone alle quali è stato chiesto di rispondere a tre domande per verificare la loro comprensione delle infografiche. Le domande sono state studiate in modo che risultassero dei task "complessi" e richiedessero all'utente di ottenere una piena comprensione dell'infografica prima di poter rispondere. I tempi di risposta sono stati registrati manualmente per ogni utente e successivamente confrontati con i tempi di risposta ottimali ipotizzati in precedenza. Di seguito sono riportati i tre task (uno per ogni infografica) e le relative opzioni di risposta.

1. Domanda relativa alla prima infografica: Nella settimana del 22 marzo 2020, a quanto corrispondono rispettivamente i casi di COVID-19, la variazione percentuale del numero di viaggi rispetto al 2019 e la variazione assoluta delle precipitazioni?
   (a) 35878 casi, -40% di viaggi, +4 mm di precipitazioni.
   (b) 29167 casi, -66.9% di viaggi, +23.9 mm di precipitazioni.
   (c) 35069 casi, +35% di viaggi, +24 mm di precipitazioni.
   (d) 29167 casi, +70% di viaggi, +4 mm di precipitazioni.
2. Domanda relativa alla seconda infografica: Selezionare l'affermazione corretta. Nella stagione estiva 2020, rispetto al 2019:
   (a) Il numero di viaggi totali è rimasto pressoché invariato mentre il numero di viaggi di piacere è raddoppiato.
   (b) Sono notevolmente aumentati sia il numero di viaggi totali che il numero di viaggi di piacere.
   (c) Non hanno subito una particolare variazione né il numero di viaggi totali né di viaggi di piacere.

(d) È diminuito il numero di viaggi in cui la stazione di partenza e quella di arrivo coincidono.

3. Domanda relativa alla terza infografica: Considerando i giorni lavorativi e la fascia oraria 03-06, si notano in particolare due stazioni: quali sono e cosa le caratterizza?

   (a) Park Ave & E 124 St, North Moore St & Greenwich St: hanno entrambe avuto un incremento di traffico nel 2020.

   (b) S8 Ave & W 31 St, North Moore St & Greenwich St: hanno entrambe registrato una differenza percentuale di -50%.

   (c) Park Ave & E 124 St, W 42 St & Dyer Ave: hanno entrambe avuto un traffico medio minore di 10 nel 2020.

   (d) 8 Ave & W 31 St, North Moore St & Greenwich St: partendo da un traffico medio elevato nel 2019, hanno avuto una variazione negativa significativa nel 2020.

I risultati, riportati nei grafici seguenti, mostrano che i tempi di risposta registrati si sono rivelati mediamente superiori ai tempi di esecuzione ottimali previsti, ma la maggior parte degli utenti intervistati ha saputo rispondere correttamente a tutte e tre le domande.

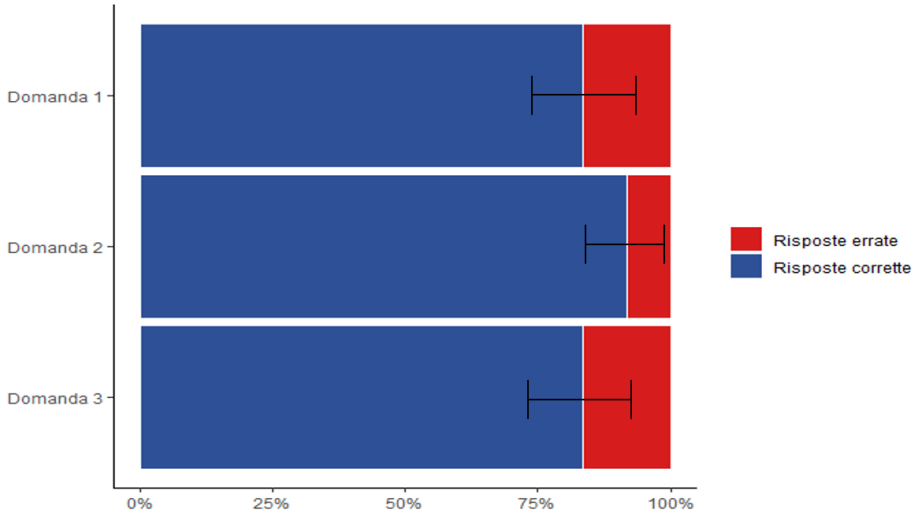


Figura 10: Risposte alle domande con intervallo di convidenza al 95%

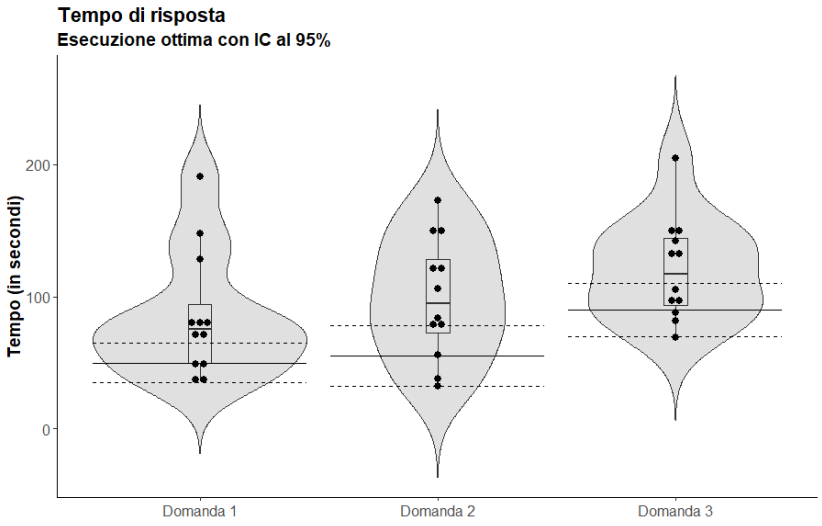


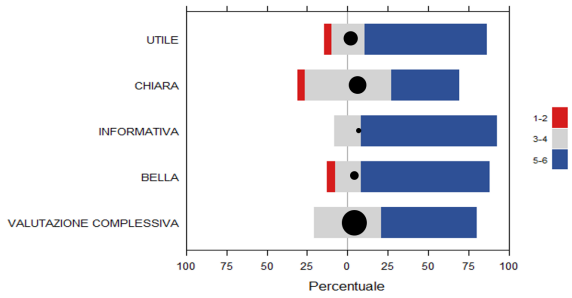
Figura 11: Violin plot con tempi di risposta

5.2.3 Questionario psicometrico

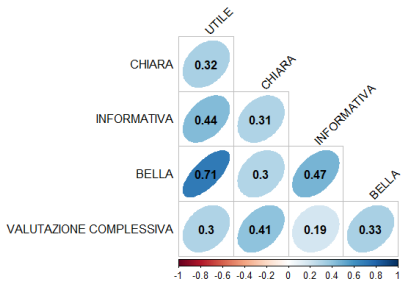
Infine, a 27 utenti è stato chiesto di rispondere al questionario psicometrico Cabitza-Locoro[15] per la misurazione di diverse dimensioni di qualità delle infografiche: Utilità, Chiarezza, Capacità informativa e Bellezza. Il questionario richiede in modo molto semplice di esprimere una valutazione da 1 a 6: Figura 12. I risultati sono riportati nei grafici seguenti. Per tutte e tre le infografiche si è ottenuta una maggioranza di giudizi positivi (5-6). Si è poi verificata la presenza di correlazione tra le risposte date ai questionari. Per la prima infografica si nota che le diverse misure di qualità non sono particolarmente correlate, ad eccezione della coppia bellezza ed utilità. È interessante perché si intuisce che, sebbene la valutazione complessiva sia positiva, le diverse dimensioni contribuiscono in modo diverso alla qualità totale dell'infografica. Invece, per quanto riguarda le valutazioni date alla seconda infografica, i valori maggiori si riscontrano per le coppie di dimensioni: chiarezza e capacità informativa, chiarezza e bellezza, utilità e capacità informativa. Queste dimensioni testimoniano insieme il fatto che una visualizzazione risulta generalmente più informativa quando riesce ad essere chiara e al tempo stessa bella esteticamente. Infine, le valutazioni ottenute dalla terza infografica risultano le più alte fra le tre, oltre che essere tutte correlate tra loro.

Figura 12: Questionario psicometrico Cabitza-Locoro

1) Prima infografica



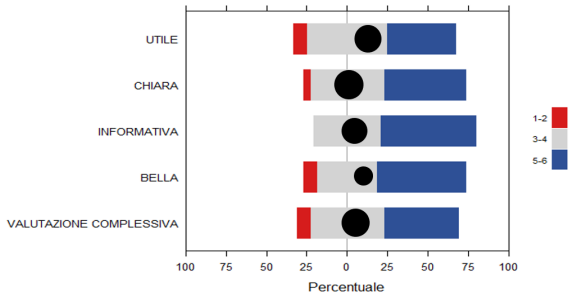
(a) Back-to-back stacked bar chart col valore del trend dei valori "incerti" (IC al 95%)



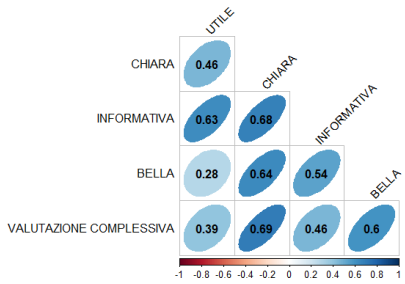
(b) Correlogramma

Figura 13: Serie storica

2) Seconda infografica



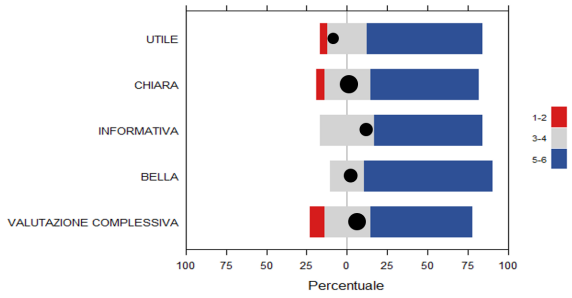
(a) Back-to-back stacked bar chart col valore del trend dei valori "incerti" (IC al 95%)



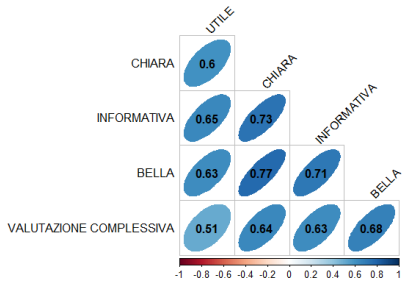
(b) Correlogramma

Figura 14: Scatterplot

3) Terza infografica



(a) Back-to-back stacked bar chart col valore del trend dei valori "incerti" (IC al 95%)



(b) Correlogramma

Figura 15: Mappa interattiva

6 Conclusione

Attraverso la manipolazione dei dati che ha permesso di produrre delle infografiche finali adeguate, è stato possibile rispondere alle domande di ricerca poste inizialmente. Innanzitutto, si evince da ogni infografica che nel 2020 la pandemia ha messo in atto una variazione della mobilità dei viaggi in bicicletta nella città di New York. Infatti, si è notato che gli abbonati a CitiBike hanno aumentato i viaggi in bici per i cosiddetti viaggi di piacere durante il 2020, modificando la tendenza rispetto all'anno precedente, prediligendo le zone meno affollate ed evitando così il centro della città. A tal riguardo, si è visto più nello specifico che alcune stazioni solite ad avere un traffico medio elevato, si sono ritrovate l'anno seguente a dimezzare la frequenza dei viaggi in media. Spostando l'attenzione su chi ha usufruito del servizio di noleggio, si è vista una diminuzione generale per gli utenti abbonati maschi nel corso della settimana, mentre un incremento nel fine settimana per le donne. Si è inoltre osservato come gli utenti occasionali siano aumentati, motivo in più che spinge ad affermare come ci sia una differenza sostanziale nelle abitudini di utilizzo tra i due anni.
Uno sviluppo ulteriore dello studio potrebbe riguardare l'approfondimento di queste tendenze, per poter confermare o meno le peculiarità dell'anno 2020 dovute alla pandemia. Quindi, raccogliendo i dati del prossimo futuro, si potrebbe monitorare il fenomeno della mobilità per poter valutare il consolidamento dei pattern emersi nell'anno in cui la situazione sanitaria ha scosso i ritmi di tutto il mondo.

Riferimenti bibliografici

[1] Citibike. https://www.citibikenyc.com/homepage.

[2] Meteostat. https://dev.meteostat.net.

[3] Kenneth Reitz et al. Requests: Http for humans. https://docs.python-requests.org.

[4] Mongodb. https://www.mongodb.com.

[5] Docker. https://www.docker.com/.

[6] Apache kafka. https://kafka.apache.org.

[7] MongoDB. Pymongo: the python driver for mongodb. https://github.com/mongodb/mongo-python-driver/.

[8] Tableau desktop. https://www.tableau.com.

[9] Onofrio Panzarino. Scalabilità: lo sharding. *<html>.it*, 2015.

[10] Marco Distrutti. Mongodb sharding with docker. https://github.com/kayne87/mongodb-sharding-docker, 2020.

[11] Time: Time access and conversions. https://github.com/python/cpython/blob/3.9/Doc/library/time.rst.

[12] Random: Generate pseudo-random numbers. https://github.com/python/cpython/blob/3.9/Doc/library/random.rst#id3.

[13] Nyc health. https://www1.nyc.gov/site/doh/index.page.

[14] Infografiche tableau. https://public.tableau.com/app/profile/guglielmo.muoio2364/viz/NYCCitibike_16234158643060/Storia.

[15] Federico Cabitza and Angela Locoro. Questionnaires in the design and evaluation of community-oriented technologies. *International journal of web based communities*, 2017.