**Data Management & Data Visualization**

**Densità di popolazione all’interno del comune di Milano**

Emanuele Rebesco matricola 864006

Gianluca Simionato matricola 864045

Kevin Tranchina matricola 861785

Anno Accademico 2019-2020

**Densità di popolazione**

Lo studio dei sistemi socio-territoriali è stato rivoluzionato dall’introduzione di una quantità di dati digitali senza precedenti costantemente prodotti dalla continua attività dei nostri smartphone.  
L’obiettivo del presente elaborato è la stima della densità urbana del comune di Milano sfruttando i dati di telefonia mobile.  
Per il raggiungimento dello scopo abbiamo utilizzato parte del dataset sull’attività telefonica raccolto durante l’evento di “Telecom Italia Big Data Challenge 2014”, composto da un settimana di Call Details Records (CDRs) sulla città di Milano, uniti con i dati geografici prelevati direttamente da OpenStreetMap.  
La stima sulla densità effettuata è stata poi messa in comparazione con i dati ISTAT, per evidenziarne differenze e somiglianze. Per l’appunto sono state realizzate 3 visualizzazioni, entrambe con lo scopo di mettere in luce l’ ‘infosfera’ del comune milanese al variare dei giorni della settimana.  
Le visualizzazioni realizzate sono state infine sottoposte al giudizio di un gruppo di 20 persone attraverso dei questionari con moduli Google.

Indice

**1 Data Management1**

* 1. Introduzioni2
  2. Descrizione del dataset2

1.2 Processo e metodi3

**2 Data Visualitation4**

2.1 Prima Visualizzazione5

2.2 Seconda Visualizzazione6

2.2 Terza Visualizzazione6

* 1. Questionari Psicometrici e Valutazione degli utenti6

**3 Conclusioni e Bibliografia4**

1. **Data Management**

**1.1 Introduzione**

In questo studio affrontiamo il problema di sfruttare i dati basati sulla rete di telefonia mobile con lo scopo stimare la densità abitativa all’interno del comune di Milano.

**1.2 Descrizione dataset**

**Origine dataset**

Il dataset preso in esame è una parte della “Telecom Italia Big Data Challenge 2014”. Il dataset originario infatti è stato creato da Telecom Italia in associazione con EIT ICT Labs, SpazioDati, MIT Media Lab, l’Università di Northeastern, il Politecnico di Milano, Fondazione Bruno Kessler, Università del Trentino e Trento e contiene una grande varietà di dati sulle telecomunicazioni, clima, notizie, dati energetici, social networks presi dal comune di Milano e di Trento.

**Call detail records(CDR)**

I Call Detail Records sono forniti da Telecom Italia.   
Ogni volta che un utente impegna un'interazione di telecomunicazione, una stazione base radio (RBS) viene assegnata dall'operatore a quell’utente e fornisce la comunicazione attraverso la rete. Quindi, viene creato un nuovo CDR che registra il tempo dell'interazione e l'RBS che l'ha gestita.   
Dalla RBS è possibile ottenere un'indicazione della posizione geografica dell'utente, grazie alle mappe di copertura che associa ciascuna RBS alla porzione di territorio che serve.

Nella nostra analisi abbiamo preso solo una porzione dei dati relativi solamente al comune di Milano, per un totale di 1,43 gigabit. Di seguito i dati di nostro interesse:

**sms-call-internet-mi-2013-11-0(1:7)**

* **Datetime** : datetime della creazione della CDR
* **CellID** :identificativo della cella 270x270 ,all’interno del comune di Milano
* **Countrycode** :il codice identificativo della nazione
* **Smsin** : un CDR è generato ogni volta che un utente riceve un SMS
* **Smsout** : un CDR è generato ogni volta che un utente riceve un SMS
* **Callin** : un CDR è generato ogni volta che riceve una chiamata
* **Callout** : un CDR è generato ogni volta che un utente effettua una chiamata
* **Internet** : viene generato un CDR ogni volta che un utente avvia una connessione internet o termina una connessione internet. Se la stessa connessione dura più di 15 minuti o se l’utente trasferisce più di 5 MB viene generato un nuovo CDR.

7 tabelle in formato csv dove sono contenute le informazioni sopra indicate, una tabella per ogni giorno dal primo al 7 novembre 2013.

**milano-grid.geojson**

Griglia in formato geojson.  
Sono presenti 1000 celle (235m\*235m) comprendenti il comune di milano.Come id ritroviamo il cellID.  
Le coordinate sono espresse in WGS84 standard.

**Istat**Densità abitativa del comune di Milano nel 2012.

**1.3 Processo e metodi**

**2 Data Visualization**

Per lo sviluppo della parte di visualizzazione è stato utilizzato Qlik Sense, un piattaforma proprietaria per le analytics. E’ stata utilizzata nella sua forma gratuita per una prova complessiva di 30 giorni.

* 1. **Prima Visualizzazione**

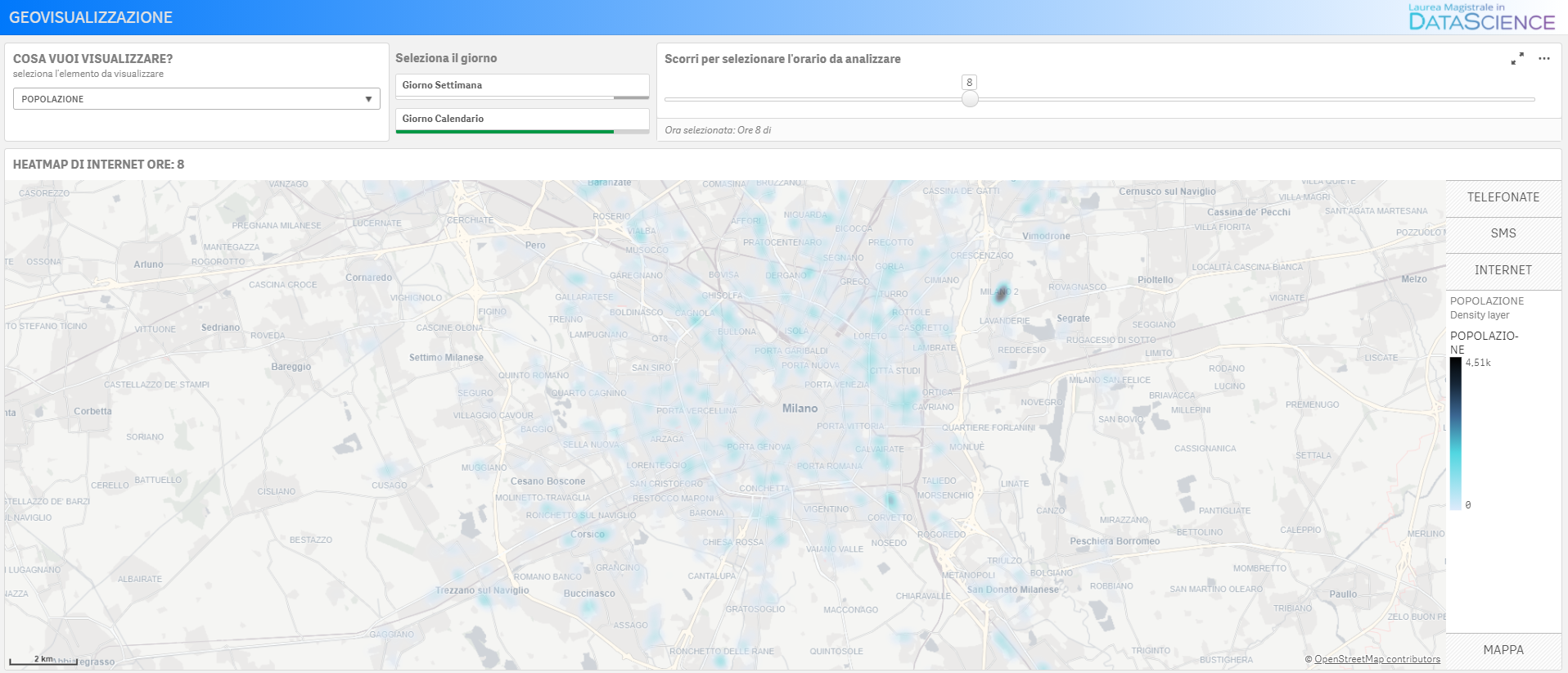
La prima visualizzazione è una heatmap.

Partiamo dai dati mostrati.  
E’ possibile vedere all’interno della mappa, filtrandoli tramite menù a scelta a destra del grafico i seguenti dati ad una specifica ora:

* *“internet”,* colonna internet di una cella
* *“sms“*, somma degli sms in ingresso e in uscita all’interno di una cella
* *“telefonate“,* somma delle telefonate in ingresso e in uscita all’interno di una cella
* *“popolazione“,* stima della densità abitativa elaborata nel parte di DataManagement

Per selezionare l’ora di interesse è stata inserita una linea di selezione a scorrimento che va dalle ore 0 alle 23.

E’ possibile anche scegliere il giorno in cui si vogliono osservare i dati sopra citati, dal 1-novembre-2013 al 7-novembre-2013, come si evince in figura.



La scopo della heatmap realizzata è quello di evidenziare a colpo d’occhi il traffico internet, sms, chiamate in ingresso e in uscita nelle diverse ore e nei diversi giorni.

Questo ci porta subito a capire senza analizzare i dati ma osservando immediatamente la mappa, se ci sono delle differenze , per esempio tra il giorno e la notte, tra il lunedì mattina e il sabato sera, per smentire o confermare delle movimentazione che ci aspetteremmo.

**Analisi migliorie prima visualizzazione**

Quello che è emerso dalle valutazione del lavoro da parte di persone esterne allo sviluppo   
sono state interessanti correttive e modifiche anche per sviluppi futuri.

Quasi tutti gli utenti che si interfacciavano con la visualizzazione erano portati ad osservare come primo elemento con cui poi interagire l’elemento in alto a sinistra,  
probabilmente per come, noi europei, siamo abituati a leggere un testo.  
Pertanto in quella posizione abbiamo preferito mettere come prima scelta il tipo di dato da visualizzare. Come naturale prosecuzione subito sulla destra abbiamo aggiunto le ore. E poi la possibilità di scegliere il giorno. In questo modo l’utente posto difronte la nostra visualizzazione ha avuto in generale un comportamento più in linea con le nostre aspettative.

Le label sono state modificate dopo un iniziale fase di sviluppo. Al posto del nome del componente grafico inserito in fase iniziale è stata inserita l’azione che l’utente doveva svolgere. Questo per cercare di parlare anche con gli utenti meno avvezzi all’utilizzo di componenti grafici.

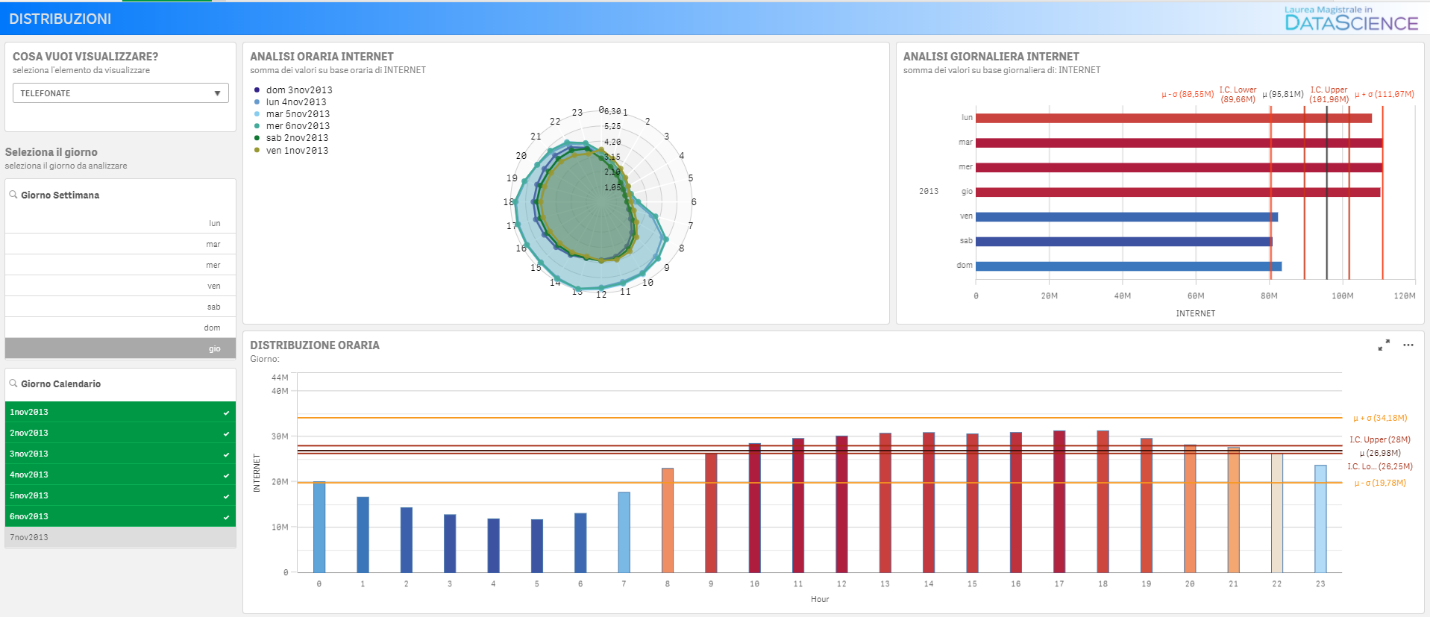
Nel corso dell’analisi è emerso anche qualche spunto di sviluppo futuro che segnaliamo come nota, ma sui quali il tool di sviluppo non permetteva una rapida modifica:

-la possibilità di modificare la linea orizzontale di selezione degli orari attraverso un orologio piuttosto che una linea dritta.

-aggiungere un + o – per riuscire subito ad ingrandire la mappa.

**2.2 Seconda Visualizzazione**

La seconda visualizzazione ha lo scopo di comparare il traffico internet, sms, telefonate in ingresso e in uscita dal territorio milanese, popolazione nei diversi giorni o nelle diverse ore della giornata



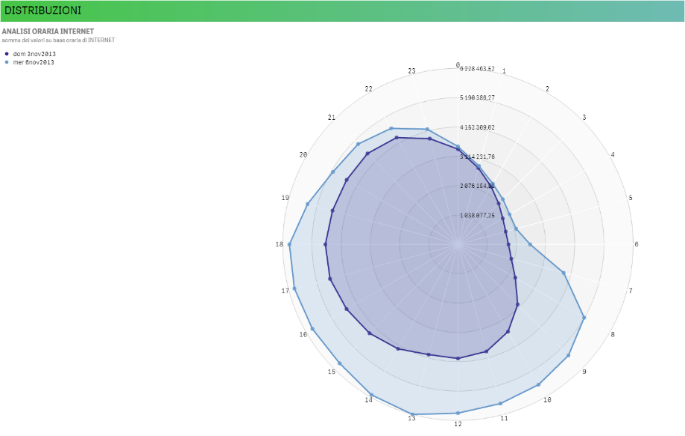
**Panoramica dei grafici:**

1. *radar chart*, risulta immediatamente chiara la differenza tra le ore notturne e diurne, come ci aspettavamo, e dopo qualche scelta appropriata delle giornate risulta anche chiara a colpo d’occhio la differenza tra l’utilizzo della grandezza selezionata nel fine settimana è maggiore che durante la settimana.
2. *Grafico a barre con i giorni della settimana,* inizialmente per questo grafico era stato pensato ad un radar chart, tuttavia la somiglianza tra i vari giorni intrasettimanali e la differenza tra giorni weekend-intrasettimanali non era apprezzabile graficamente. Con l’utilizzo del grafico a barre risultavano più leggibili i valori e più chiare le differenze
3. *Grafico della distribuzione dell’utilizzo della variabile selezionata,* sono stati inseriti i valori anche i valori di deviazione standard e mediana in modo da avere una panoramica dei dati.

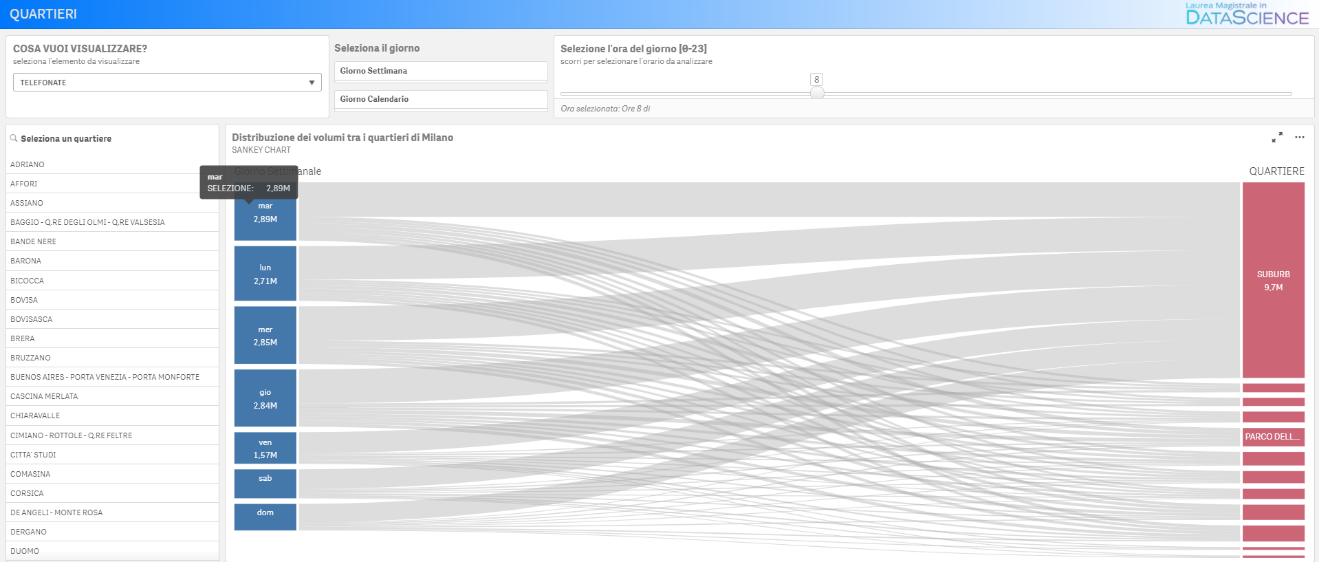
**Analisi migliorie seconda visualizzazione**

Così come nella prima, l’utente come prima scelta si trova la selezione dei dati che desidera visualizzare in alto a sinistra. Subito sotto la scelta del giorno. Fatta l’operazione noterà le infografie modificarsi in base alla selezione fatta.

Questa seconda visualizzazione perde la parte grafica della mappa che piaceva molto nella prima, quindi risulta, per meno esperti lettori di visualizzazioni, di non immediata comprensione.

Il grafico risulta immediatamente comunicativo a colpo d’occhio delle differenze tra le ore notturne e diurne, e con un oppurtuno scelta di giorni le differenze tra i giorni intrasettimanali e il weekend.

* 1. **Terza visualizzazione**

Per questa terza visualizzazione abbiamo scelto una suddivisione del traffico telefonico in quartieri.  
Dai dati delle coordinate delle celle telefoniche incrociati con i dati che identificavano i vari quartieri di Milano è nato il sankey chart che domani centrale sulla visualizzazione, il resto sono solo filtri che per coerenza con la prima visualizzazione sono stati mantenuti nella stessa posizione, in cima al grafico.

Aprendo di default la visualizzazione troviamo il volume smistato 10 quartieri su cui è presenta la maggiore distribuzione in ordine descrescente della credenza “telefonate”.

E’ possibile scegliere in questo caso, oltre che come sempre ora e giorno, anche i quartieri che si vogliono visualizzare.

Uno degli utilizzi maggiormente interessanti del grafico è la selezione di 2 o 3 quartieri d’interesse in modo da non avere un grafico troppo confusionario e modificare l’orario per capire come varia il flusso al variare delle ore.

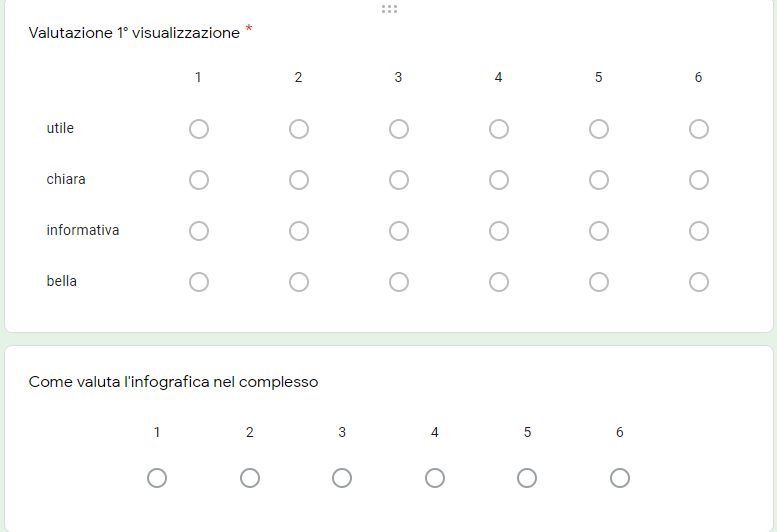
**Analisi migliorie terza visualizzazione**:

Inizialmente non è possibile vedere i valori numerici associati alla grandezza se durante quella giornata o in quello specifico quartiere non sono presenti grandi traffici. Questo perché semplicemente il tool di sviluppo elimina le informazioni d’interesse se il componente risulta troppo piccolo. Per questo motivo è stato inserito un Tooltip come è visibile nello screen sopra.

Agli occhi degli utenti il grafico è risultato molto bello ed intuitivo nel complesso. Non subito chiaro invece l’uso che ne può fare, al valore aggiunto che porta. Selezionando però pochi quartieri d’interesse coglie subito delle differenze o somigliare confrontando anche i diversi giorni tra loro.

* 1. **Questionari Psicometrici e Valutazione Utenti**

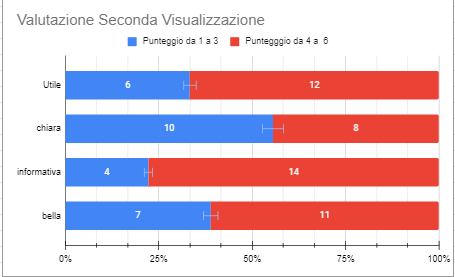
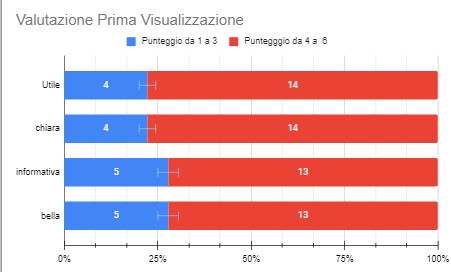
Per la realizzazione del questionario sono stati sottoposti ad un gruppo di venti persone tre questionari Moduli Google, ognuno riferito ad una specifica visualizzazione, compilabili al seguente link: [Valutazione visualizzazione](https://forms.gle/6LxSxydkJXXDk1XB6).

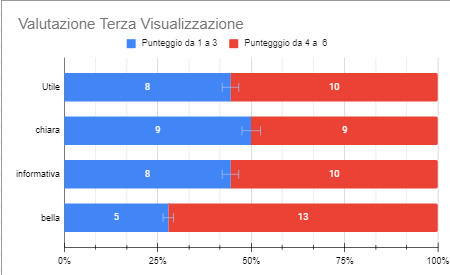


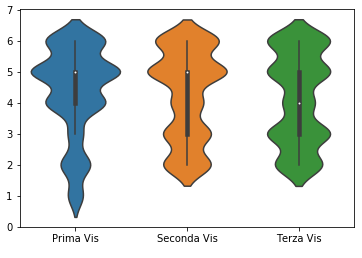
Per ogni visualizzazione l’utente che ha esaminato la visualizzazione può esprimere un giudizio sui 5 aggettivi elencati in figura con votazione da 1 a 6 (1 aggettivo completamente distante con la visualizzazione e 6 aggettivo perfettamente calzante con la descrizione).

Successivamente sono stati realizzati due tipologie di grafico:

1. **Grafico a barre**, realizzato all’interno dello stesso Google Sheet dove venivano salvati i risultati delle risposte.

****



1. **Violin Plot**, comparativo delle 3 visualizzazioni, complessivo di tutte le valutazioni e realizzato tramite api google invocata da codice python che va direttamente a prelevare i risultati dal Google sheet associato al questinario. La libreria grafica usata è seaborn.

E’ possibile trovare il codice nel github in bibliografia.

**Conclusioni**

**Bibliografia**

Fonte dati e informazioni progetto Telecom Italia Big Data Challenge **<https://www.nature.com/articles/sdata201555>**

Dataset Kaggle **<https://www.kaggle.com/marcodena/mobile-phone-activity?select=sms-call-internet-mi-2013-11-01.csv>**

Software di visualizzazione **<https://www.qlik.com/it-it/trial/download-qlik-sense-desktop>**

**Progetto git :**