Chicago Crimes Story

Candidati:

Giulia Sanfilippo

Antonio Manlio D'Agostino

Chiara Suraci

Docente:

Domenico Ursino

INDICE

Descrizione della realtà

In questa tesina ciò che ci si propone è un'analisi dei dati relativi alle attività criminali nella città di Chicago, USA.

Prima di intraprendere la nostra analisi, è fondamentale comprendere come è organizzata la città di Chicago. Essa è suddivisa in quelle che prendono il nome di *Community Area,* sono ben 77 e sono distribuite come segue:

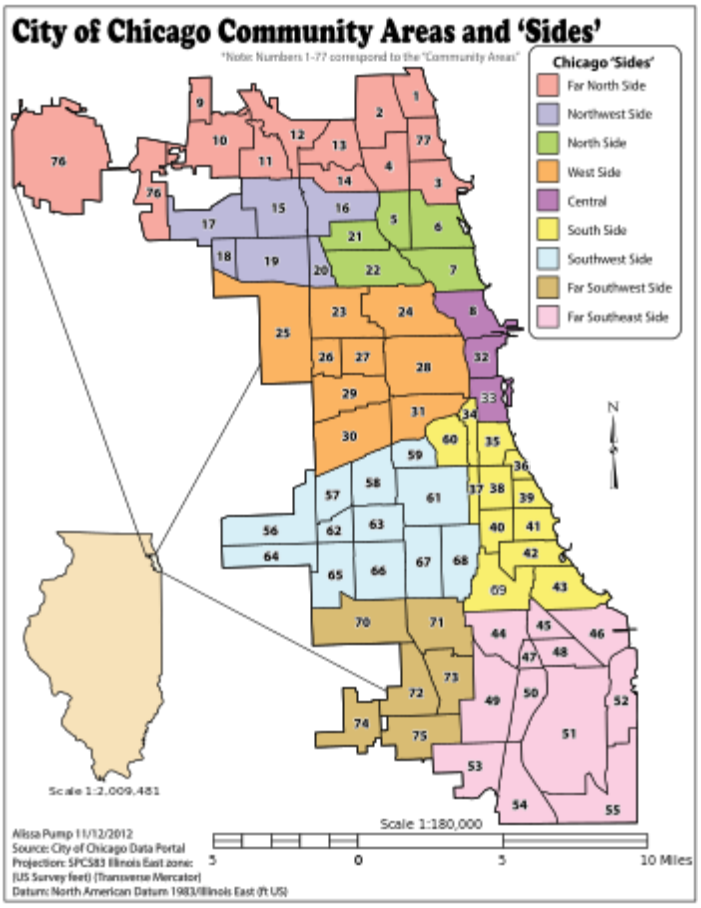
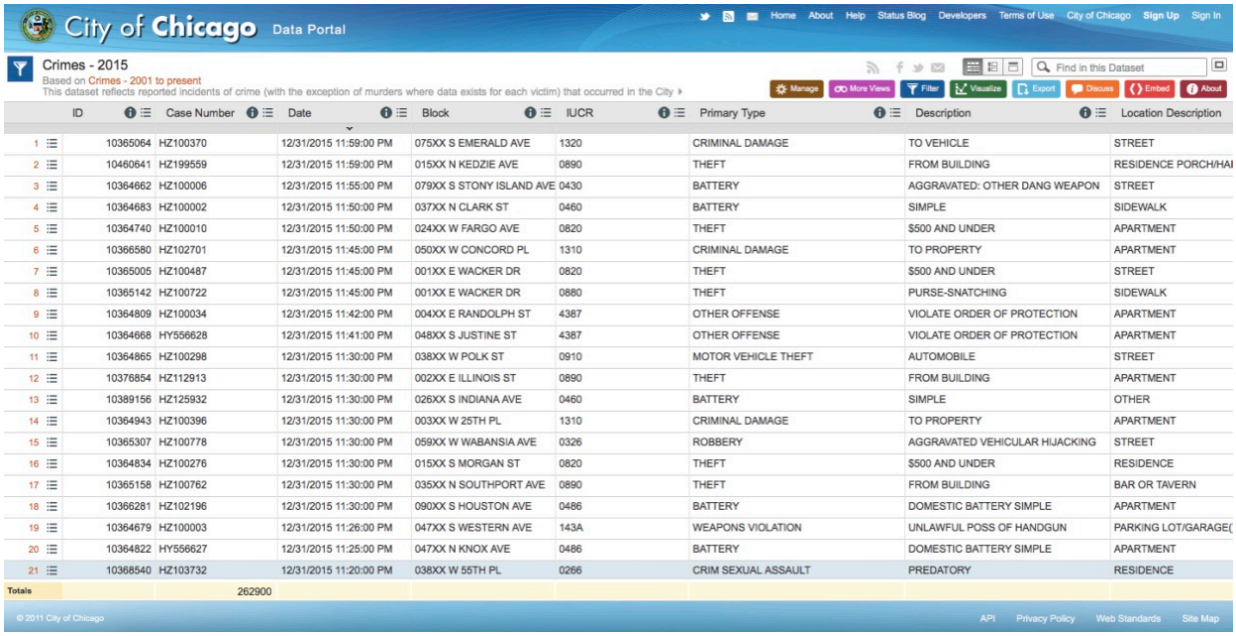


Figura 1 *Community Areas;*

Come già anticipato, ciò che faremo sarà effettuare delle analisi relativamente ai crimini commessi a Chicago, in particolare furti, omicidi, crimini di natura sessuale, rapine, etc.

I dati analizzati sono stati prelevati da https://data.cityofchicago.org. Nella seguente figura mostreremo alcuni dei campi che saranno molto utili durante le nostre analisi



City of Chicago Data Portal

Home About Help Status Blog Developers Terms of Use City of Chicago Sign Up Sign In

Crimes - 2015

Based on Crimes - 2001 to present

This dataset reflects reported incidents of crime (with the exception of murders where data exists for each victim) that occurred in the City

Find in this Dataset

Manage More Views Filter Visualize Export Discuss Embed About

	ID	Case Number	Date	Block	IUCR	Primary Type	Description	Location Description
1	10365064	HZ100370	12/31/2015 11:59:00 PM	075XX S EMERALD AVE	1320	CRIMINAL DAMAGE	TO VEHICLE	STREET
2	10460641	HZ199559	12/31/2015 11:59:00 PM	015XX N KEDZIE AVE	0890	THEFT	FROM BUILDING	RESIDENCE PORCH/HAI
3	10364662	HZ100006	12/31/2015 11:55:00 PM	079XX S STONY ISLAND AVE	0430	BATTERY	AGGRAVATED: OTHER DANG WEAPON	STREET
4	10364683	HZ100002	12/31/2015 11:50:00 PM	037XX N CLARK ST	0460	BATTERY	SIMPLE	SIDEWALK
5	10364740	HZ100010	12/31/2015 11:50:00 PM	024XX W FARGO AVE	0820	THEFT	$500 AND UNDER	APARTMENT
6	10366580	HZ102701	12/31/2015 11:45:00 PM	050XX W CONCORD PL	1310	CRIMINAL DAMAGE	TO PROPERTY	APARTMENT
7	10365005	HZ100487	12/31/2015 11:45:00 PM	001XX E WACKER DR	0820	THEFT	$500 AND UNDER	STREET
8	10365142	HZ100722	12/31/2015 11:45:00 PM	001XX E WACKER DR	0880	THEFT	PURSE-SNATCHING	SIDEWALK
9	10364809	HZ100034	12/31/2015 11:42:00 PM	004XX E RANDOLPH ST	4387	OTHER OFFENSE	VIOLATE ORDER OF PROTECTION	APARTMENT
10	10364668	HY556628	12/31/2015 11:41:00 PM	048XX S JUSTINE ST	4387	OTHER OFFENSE	VIOLATE ORDER OF PROTECTION	APARTMENT
11	10364865	HZ100298	12/31/2015 11:30:00 PM	038XX W POLK ST	0910	MOTOR VEHICLE THEFT	AUTOMOBILE	STREET
12	10376854	HZ112913	12/31/2015 11:30:00 PM	002XX E ILLINOIS ST	0890	THEFT	FROM BUILDING	APARTMENT
13	10389156	HZ125932	12/31/2015 11:30:00 PM	026XX S INDIANA AVE	0460	BATTERY	SIMPLE	OTHER
14	10364943	HZ100396	12/31/2015 11:30:00 PM	003XX W 25TH PL	1310	CRIMINAL DAMAGE	TO PROPERTY	APARTMENT
15	10365307	HZ100778	12/31/2015 11:30:00 PM	059XX W WABANSIA AVE	0326	ROBBERY	AGGRAVATED VEHICULAR HIJACKING	STREET
16	10364834	HZ100276	12/31/2015 11:30:00 PM	015XX S MORGAN ST	0820	THEFT	$500 AND UNDER	RESIDENCE
17	10365158	HZ100762	12/31/2015 11:30:00 PM	035XX N SOUTHPORT AVE	0890	THEFT	FROM BUILDING	BAR OR TAVERN
18	10366281	HZ102196	12/31/2015 11:30:00 PM	090XX S HOUSTON AVE	0486	BATTERY	DOMESTIC BATTERY SIMPLE	APARTMENT
19	10364679	HZ100003	12/31/2015 11:26:00 PM	047XX S WESTERN AVE	143A	WEAPONS VIOLATION	UNLAWFUL POSS OF HANDGUN	PARKING LOT/GARAGE(
20	10364822	HY556627	12/31/2015 11:25:00 PM	047XX N KNOX AVE	0486	BATTERY	DOMESTIC BATTERY SIMPLE	APARTMENT
21	10368540	HZ103732	12/31/2015 11:20:00 PM	038XX W 55TH PL	0266	CRIM SEXUAL ASSAULT	PREDATORY	RESIDENCE
Totals		262900						

© 2011 City of Chicago API Privacy Policy Web Standards Site Map

Figura 2 *Schermata sorgente dei dati;*

Sono state effettuate operazioni di ETL durante lo studio, alcuni dei campi risultavano essere ridondanti e poco utili per i nostri fini, ci siamo occupati di elaborare la data, cercando di portarla in una forma semplice da comprendere: mm/dd/yyyy, e cercando di separarla dall'ora. Saranno di nostro interesse i campi:

- Case Number: identificativo del singolo crimine;
- Date;
- Hour;
- Primary Type: tipo di crimine commesso;
- Description: descrizione del contesto in cui è stato commesso il crimine, strada, marciapiede, etc.;
- Location Description: luogo in cui è avvenuto il crimine;
- Arrest: valore booleano che indica se l'arresto per quel particolare crimine è avvnuto o meno;
- Domestic: valore booleano che indica se il crimine in questione è o meno di natura domestica;
- District: numero del distretto di polizia che si occupa del crimine in questione;
- Community Area: suddivisione geografica della città;
- Latitude;
- Longitude;

- Location: insieme di latitudine-longitudine.

Dimensioni e misure

Abbiamo individuato un fatto: "*Crime".* La granularità prescelta è il singolo crimine. Sono state inoltre identificate le seguenti **dimensioni:**

1. **Time:** che si compone di
    - Date
    - Hour
    - Day
    - Month
    - Years
2. **Type:** che si compone di
    - Case Number
    - Description
    - Community Area
    - Location Description
    - Arrest
    - Domestic
    - Primary Type
    - Latitude
    - Longitude
    - Location
3. **District:** che si compone di
    - District
    - ward

E le seguenti **misure**

1. **Arrest;**
2. **Crimes;**
3. **Domestic;**

Schema a stella

Lo schema a stella ottenuto è il seguente:

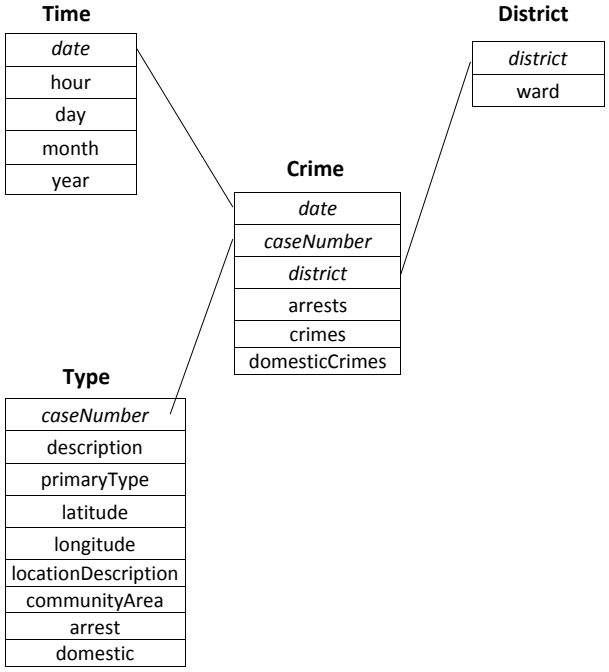


Figura 3 *Schema a stella;*

Dopo queste analisi preliminari, essenziali per comprendere al meglio il lavoro che faremo, possiamo passare ai tool che ci hanno consentito di ottenere i risultati che ci eravamo prefissati:

- **Qlik sense**
- **Tableau**
- **MongoDB**

Qlik sense

Qlik sense è una piattaforma per l'analisi dei dati. Con Qlik sense è possibile effettuare delle rilevazioni e ottenere i risultati con una velocità elevatissima. Questo tool, difatti, genera immediatamente viste di informazioni: ogni volta che si fa click, i grafici si aggiornano con dei nuovi set di dati calcolati sul momento.

Per quanto riguarda le nostre analisi, dopo aver opportunamente inserito le dimensioni e le misure prestabilite, abbiamo ottenuto i seguenti risultati:

- Numero di crimini e di arresti nell'arco 2013-2015;
- Numero di crimini per *Primary Type* nel 2014;
- Numero di crimini e di arresti per mese;
- Numero di crimini domestici negli anni considerati;
- Numero di arresti e crimini per ora;
- Numero di crimini e di arresti per *Community Area* e *Location Description*;

Risultati ottenuti

Per l'utilizzo dei dati in questione su Qlik sense, è stata fatta fatta un'unica operazione di ETL (oltre le iniziali precedentemente illustrate). Difatti, abbiamo modificato il formato dell'ora in base alla compatibilità di Qlik sense.

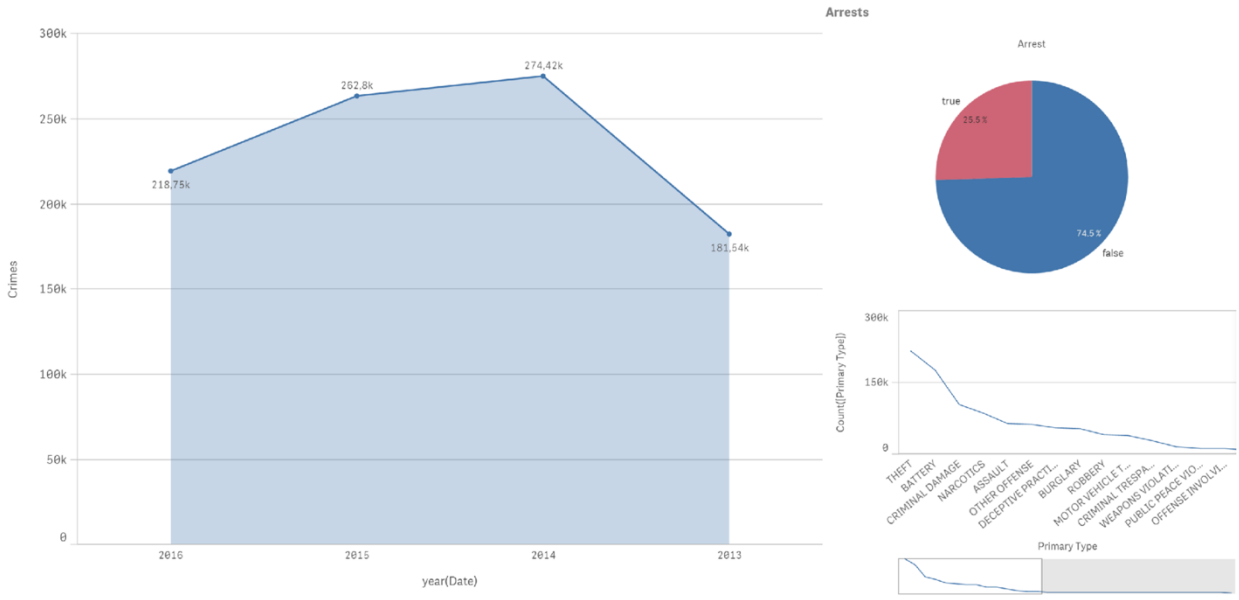


Figura 4 Numero di crimini per anno e crimine maggiormente commesso;

In questo grafico viene mostrato l'andamento dei crimini in numero e degli arresti in percentuale nei vari anni presi in considerazione per le nostre analisi. Si può notare come il 2014 sia l'anno in cui si è verificato il maggior numero di crimini. Si è inoltre ricercato il tipo di crimine più commesso: risulta il furto.

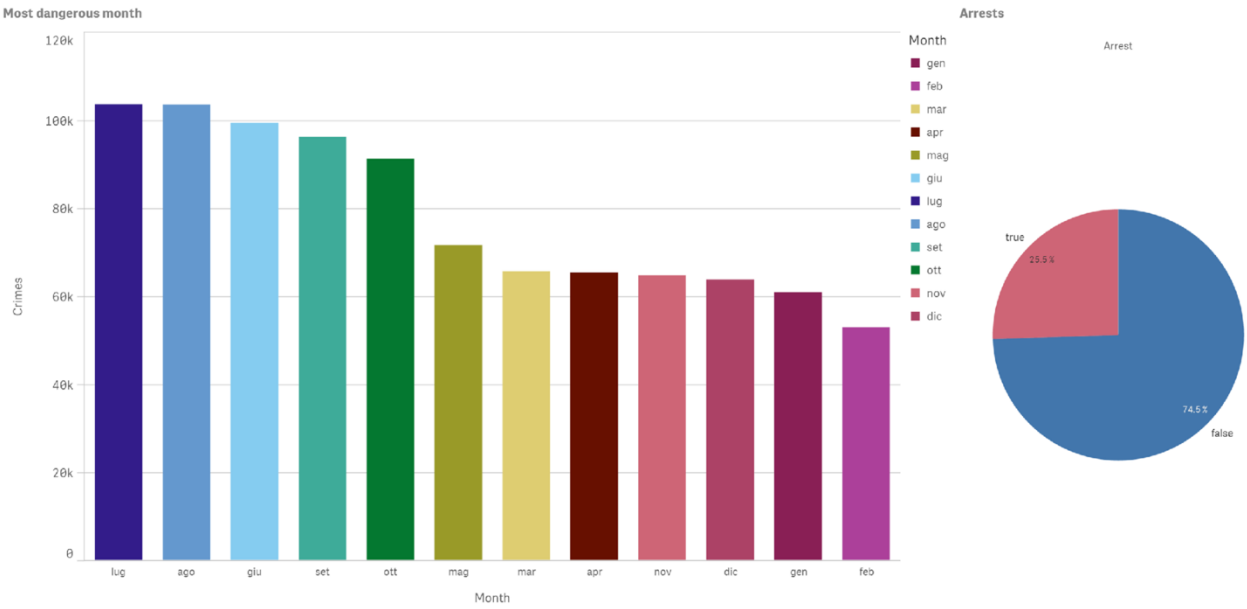


Figura 5 Numero di crimini per mese;

Il mese più pericoloso risulta il mese di luglio, con oltre 100.000 crimini all'attivo. In ogni grafico abbiamo rappresentato la percentuale di arresti fatti in base al numero di crimini.

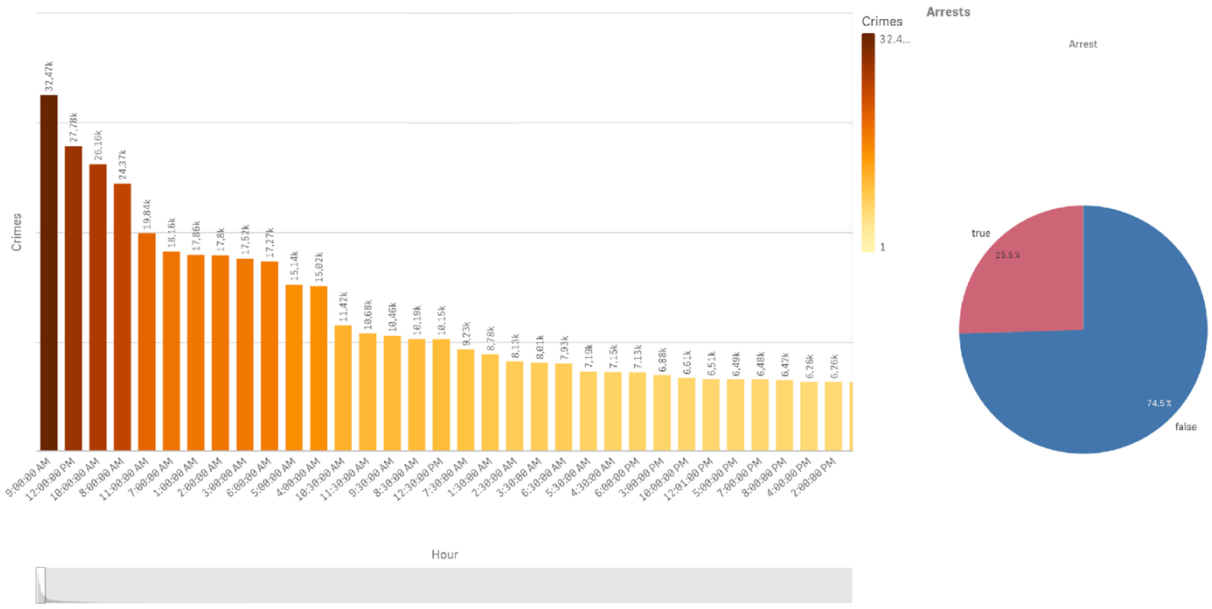


Figura 6 Ora con il maggior numero di crimini;

Sono due gli orari più pericolosi: troviamo le 9.00 AM di mattina e le 12.00 PM, quindi mezzanotte. Risultano molti più crimini nella prima parte della giornata, tra le 00.00 AM e le 12.00 AM.

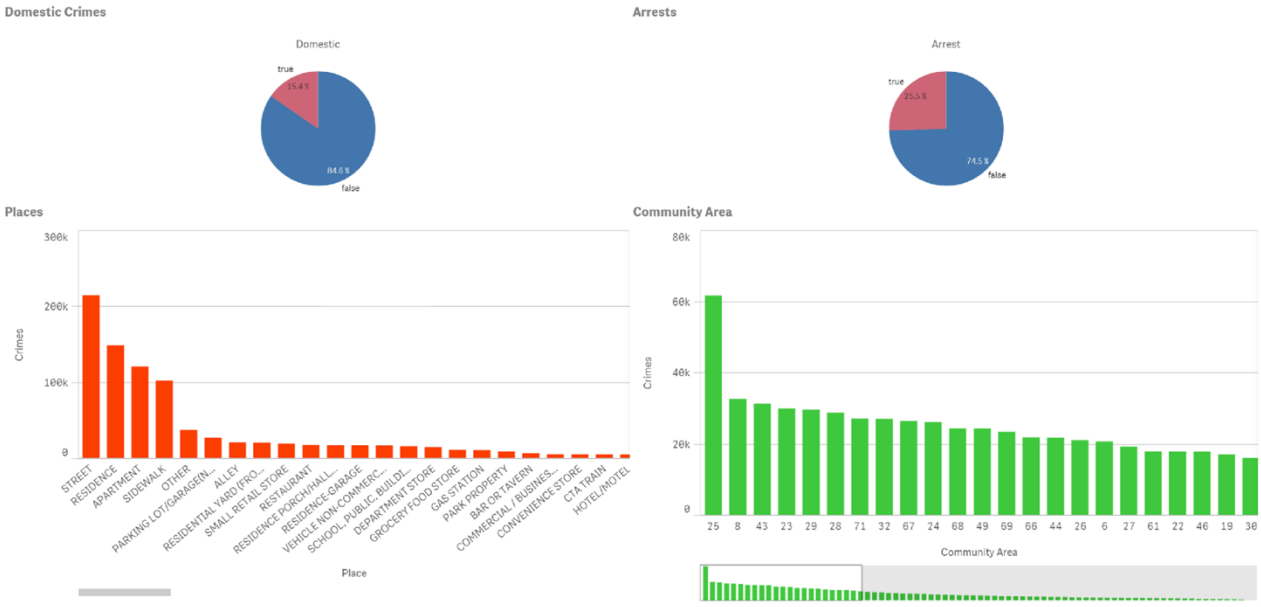


Figura 7 Luoghi e Community Area con il maggior numero di crimini;

Nell'ultimo sheet abbiamo analizzato i luoghi in cui avvengono più crimini, facendo sempre un confronto con i crimini domestici: questi ultimi risultano meno numerosi rispetto ai crimini di strada. La community area più criminosa è la 25.


tableau

Tableau

Tableau è anch'essa una piattaforma per l'analisi dei dati. Seppur entrambe le piattaforme perseguono lo stesso obiettivo, le possibilità offerte dai due tool sono abbastanza differenti. Grazie alla nostra esperienza di utilizzo abbiamo avuto modo di constatare che Qlik sense, seppur molto più usabile di Tableau, ci offrisse qualcosa in meno, in particolare relativamente al trattamento dei dati geografici e nella presentazione dei risultati ottenuti. Per quanto riguarda le nostre analisi, dopo aver opportunamente inserito le dimensioni e le misure prestabilite, abbiamo ottenuto i seguenti risultati:

- Numero di crimini e di arresti nell'arco 2013-2015;
- Numero di crimini e di arresti per mese;
- Numero di crimini domestici negli anni considerati;
- Numero di arresti, per *Description* ed ora;
- Numero di crimini per *Location Description*;
- Numero di crimini per *Community Area;*
- Numero di arresti per *District*;
- Numero di arresti per *District* e *Primary Type*;
- Numero di crimini per ora;
- Numero di crimini e di arresti per *Community Area* e *Location Description*;
- Numero di crimini per *Primary Type* nel 2014;
- Confronto tra arresti true e false per *Primary Type*;

Il piano di lavoro sarà sostanzialmente quello di analizzare e commentare, figura per figura, i risultati ottenuti.

Risutati ottenuti

Figura 8 Numero di crimini e di arresti per mese;

In questo grafico viene invece mostrato l'andamento dei crimini e degli arresti (sempre in numero) per tutti i mesi dell'anno, in tutti gli anni considerati. Da tale grafico si nota quanto il mese più colpito dai crimini sia luglio (in media, negli anni considerati), al secondo posto agosto e al terzo giugno.

Per quanto riguarda invece gli arresti troviamo al primo posto il mese di luglio, seguono i mesi di giugno e agosto. Ciò combacia esattamente con le analisi ricavate da Qlik sense.

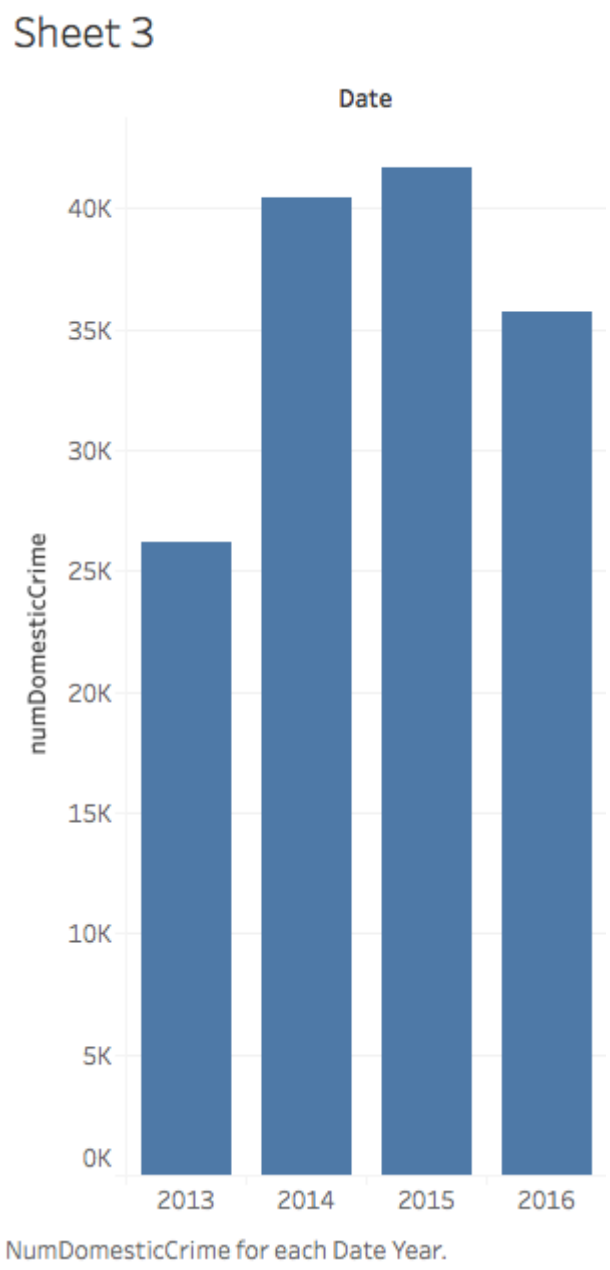


Figura 9 Numero di crimini domestici negli anni considerati;

In questo grafico viene mostrato il numero di crimini di tipo “domestico” negli anni considerati per le nostre analisi. Il risultato maggiore si trova per l’anno 2015, con ben 41.662 crimini domestici. Seguono gli anni 2014 e 2016. Ciò combacia esattamente con le analisi ricavate da Qlik sense.

Figura 10 Numero di crimini per Community Area;

In questo grafico viene mostrato il numero di crimini per “Community Area”, in particolare la più colpita è la numero 25 con 61.587 crimini, seguono la Community Area 8 con 32.545 crimini e la 43 con 31.208 crimini. Ciò combacia esattamente con le analisi ricavate da Qlik sense.

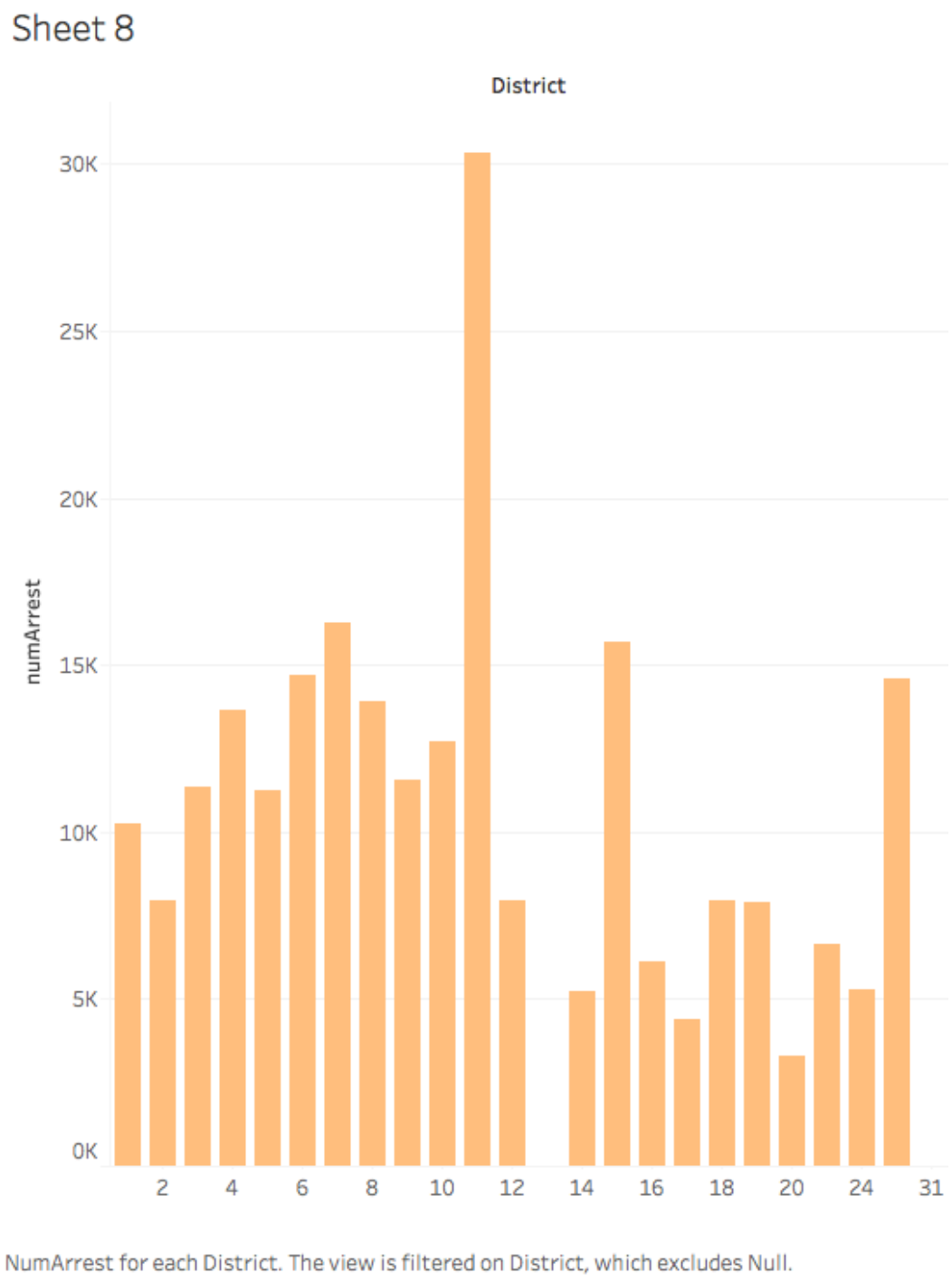


Figura 11 Numero di arresti per District;

Questo grafico mostra l'andamento del numero di arresti per i vari distretti di polizia considerati. In particolare possiamo notare come il distretto numero 11 sia quello con il maggior numero di arresti.

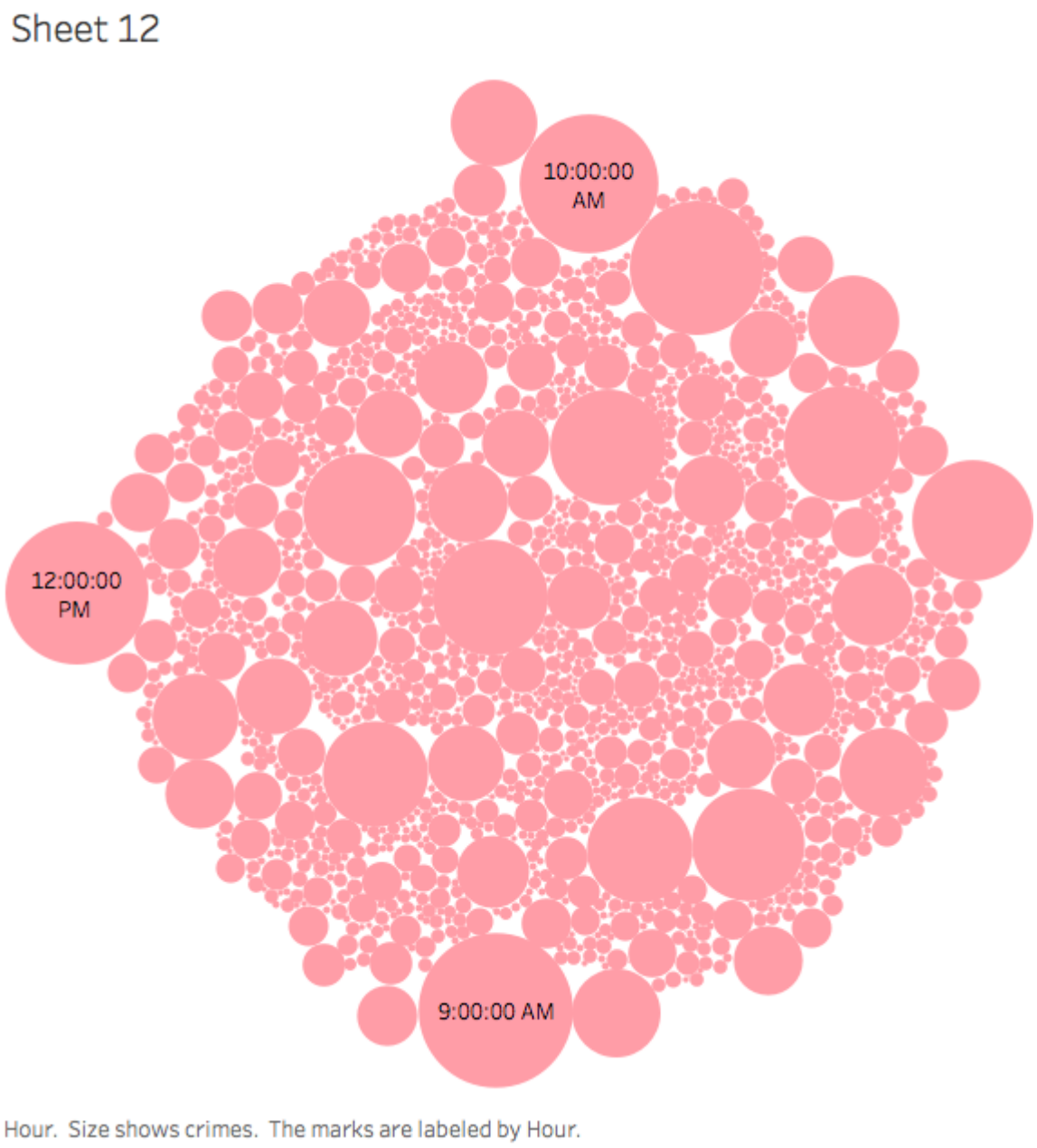


Figura 12 Numero di crimini per ora;

In questa figura vengono invece mostrate le ore maggiormente colpite dai crimini. Possiamo notare come le ore 9:00:00 AM siano le maggiormente colpite dai crimini, in particolare 32.446. Seguono poi le ore 12:00:00 PM e le ore 10:00:00 AM, rispettivamente con 27.778 e 26.155 crimini. Ciò combacia esattamente con le analisi ricavate da Qlik sense.

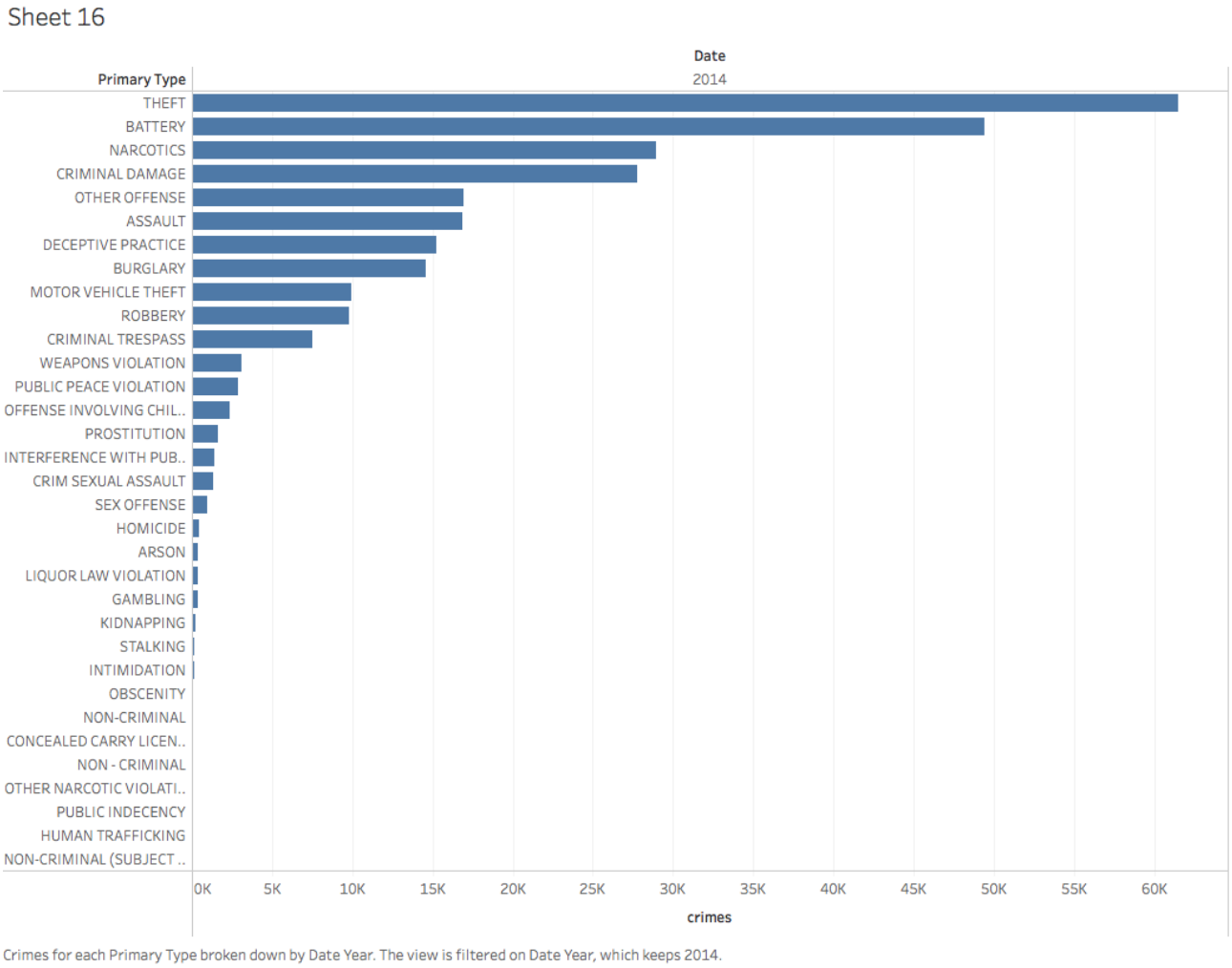


Figura 13 Numero di crimini per Primary Type nel 2014;

Il seguente grafico mostra il numero di crimini, nell'anno 2014, categorizzati per "Primary Type". Si nota come il crimine maggiormente commesso sia il furto (theft), 61.523, seguono poi il furto con scasso (battery) e lo spaccio di droga (narcotics), rispettivamente 49.443 e 21.951 crimini. Ciò combacia esattamente con le analisi ricavate da Qlik sense.

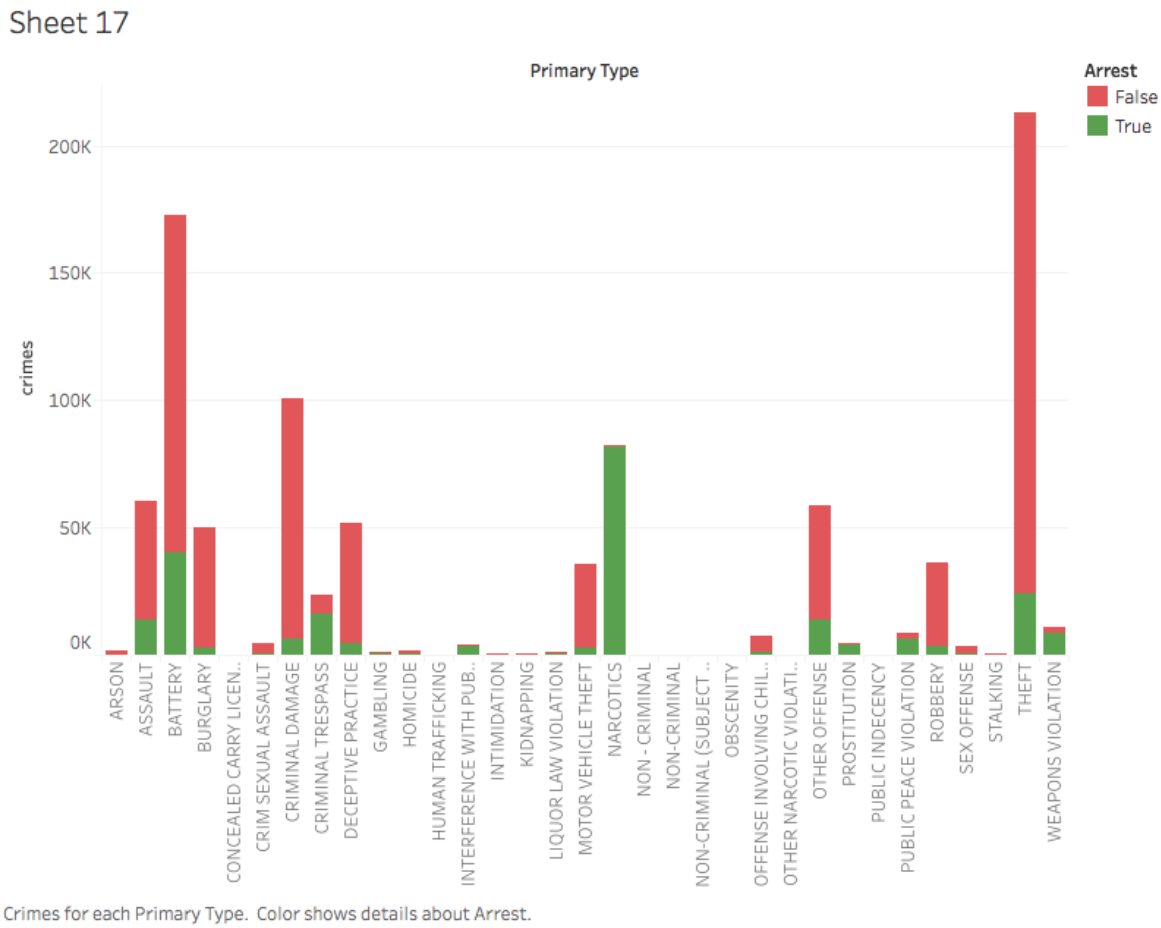


Figura 14 Confronto tra arresti true e false per Primary Type;

In questo grafico vengono mostrati invece i valori “true” e “false” degli arresti. Per quanto riguarda i valori “false”, troviamo al primo posto il furto con 189.231 arresti mancati, seguono poi il furto con scasso e “Criminal damage”, rispettivamente 132.801 e 94.018 arresti mancati. Per quanto riguarda i valori “true”, molto minori rispetto ai valori false, notiamo come il maggior numero di arresti si sia ottenuto con il crimine “spaccio di droga”, 81.759 arresti.


mongoDB

MongoDb

MongoDB è uno tra i database NoSQL open source più diffusi e utilizzati al momento. I database NoSQL si adattano soprattutto a situazioni in cui si ha a che fare con grandi quantità di dati e con sistemi real-time. MongoDB è una vera e propria suite, dal momento che offre anche alcuni tools legati al database stesso, che presenta una shell quale strumento di amministrazione ed alcuni driver per l'interfacciamento con i più noti linguaggi di programmazione. Per lavorare con MongoDB abbiamo utilizzato l'interfaccia utente MongoChef, che ci ha consentito il caricamento e l'elaborazione stessa dei dati.

Installazione e avvio

Dopo aver eseguito l'installazione di MongoDB, è necessario avviare il server da terminale:

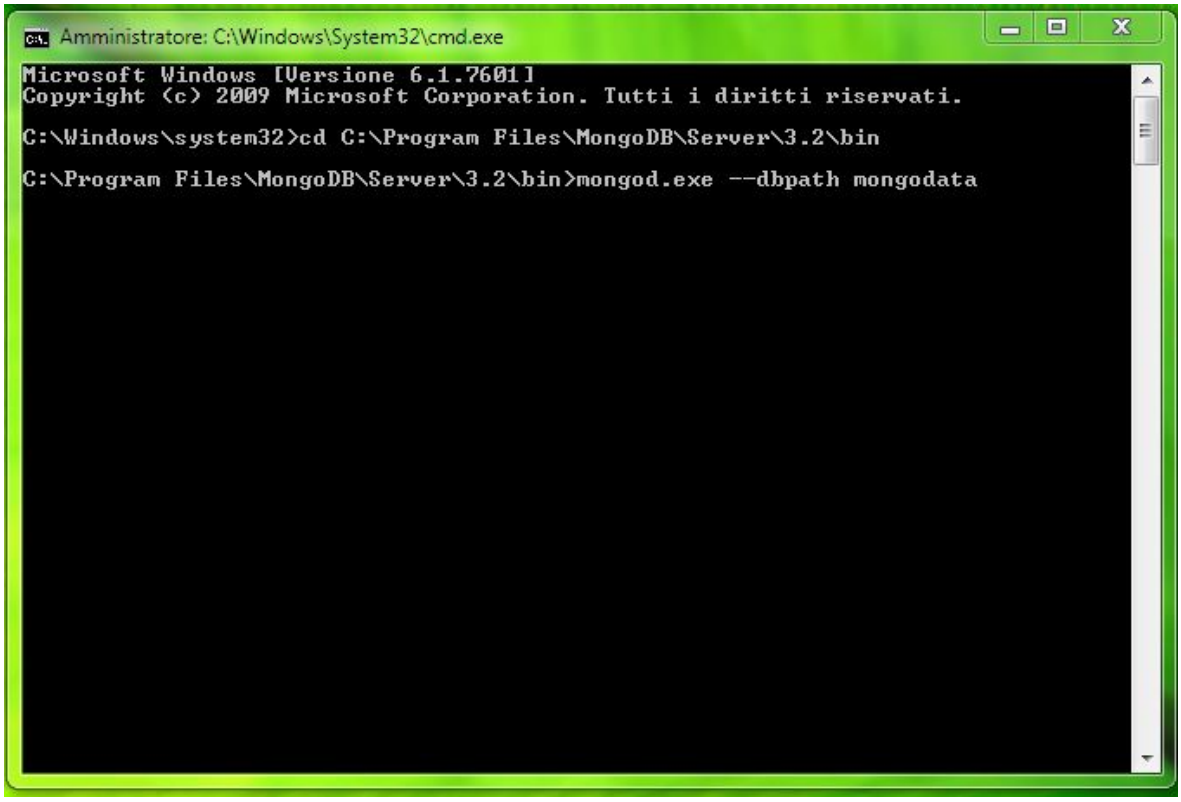


Figura 15 Avvio server;

In seguito è stato possibile effettuare il caricamento dei dati tramite MongoChef.

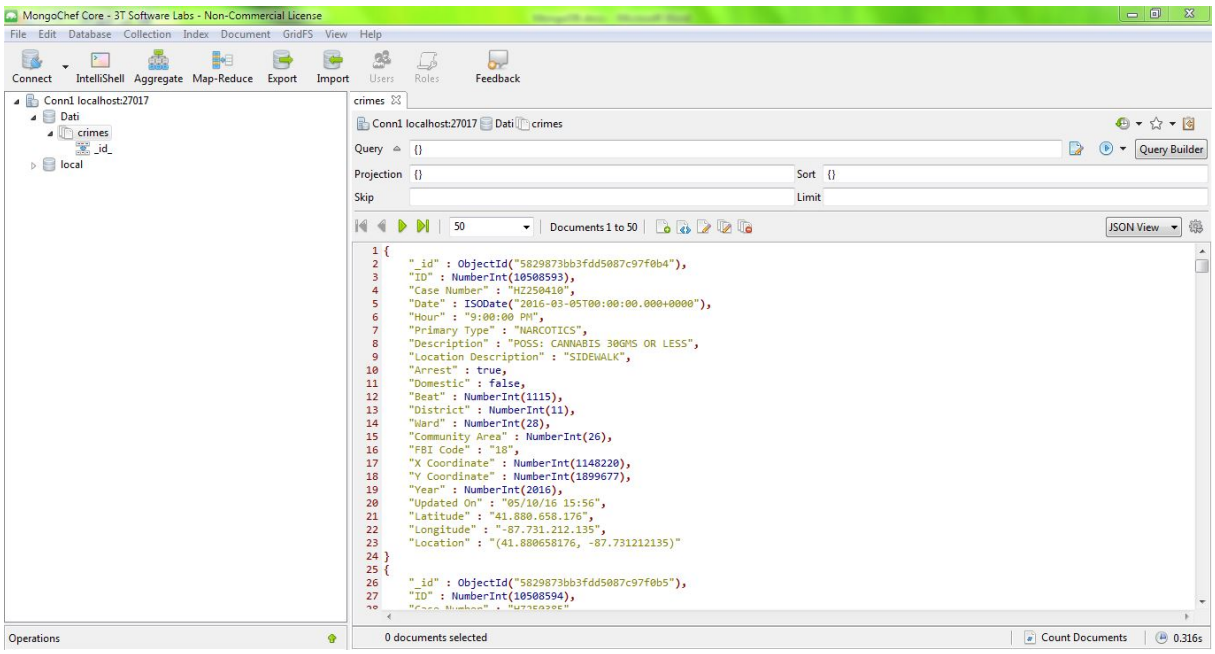


Figura 16 Aspetto dei dati caricati in formato JSON;

Risultati ottenuti

Grazie alla nostra analisi siamo riusciti ad ottenere informazioni su:

- Anno in cui è stato commesso il maggior numero di crimini;
- Ora del giorno in cui è stato commesso il maggior numero di crimini;
- Tipo di crimine più commesso;
- Community Area in cui è stato commesso il maggior numero di crimini;
- Distretto che ha effettuato il maggior numero di arresti;

Si riportano di seguito i risultati ottenuti grazie alla scrittura di script specifici:

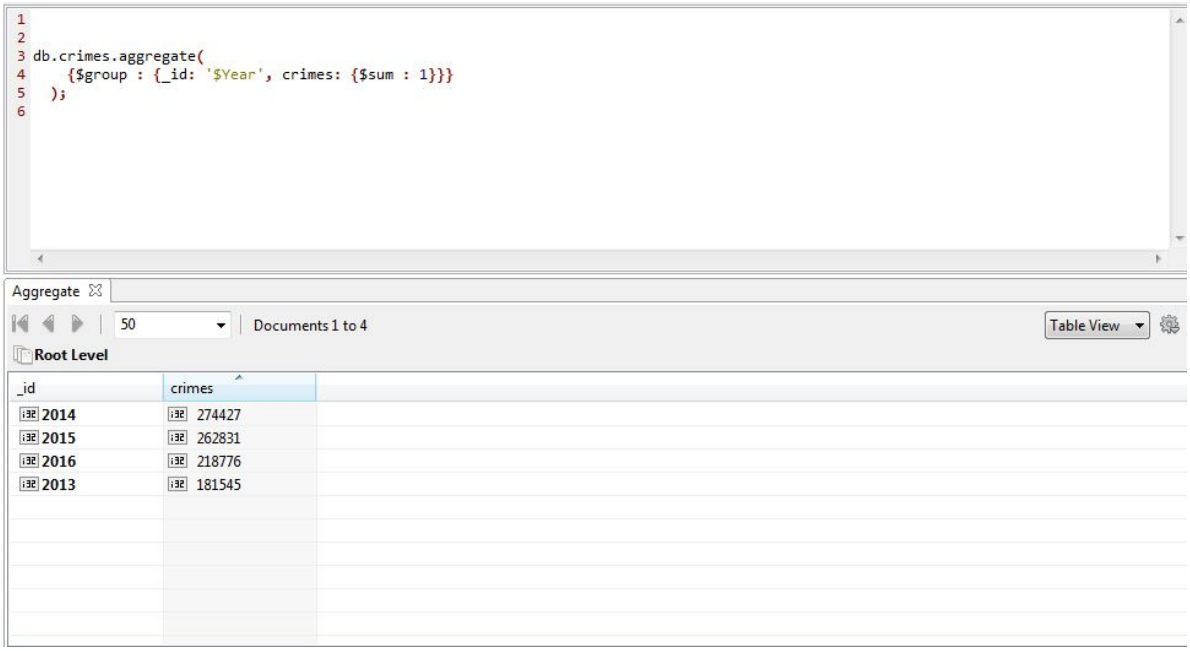


Figura 17 Anno con maggior numero di crimini;

Il 2014 è risultato l'anno in cui sono stati commessi più crimini.

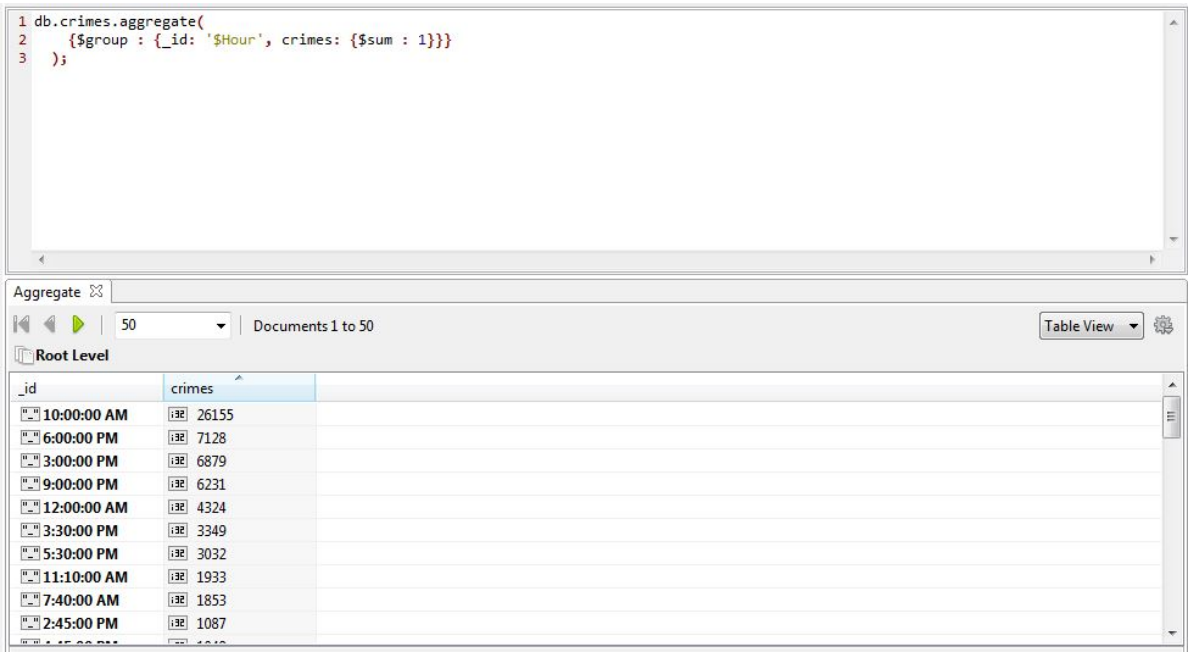


Figura 18 Ora con maggior numero di crimini;

L'ora del giorno in cui si è commesso il maggior numero di crimini è le 10 del mattino, seguono le 18:00 e le 15:00. I valori numerici combaciano con quelli ottenuti dai

precedenti tool, ma i risultati in termini di classifica decrescente differiscono in quanto MongoDB visualizza solo un certo campione di risultati, in particolare 50.

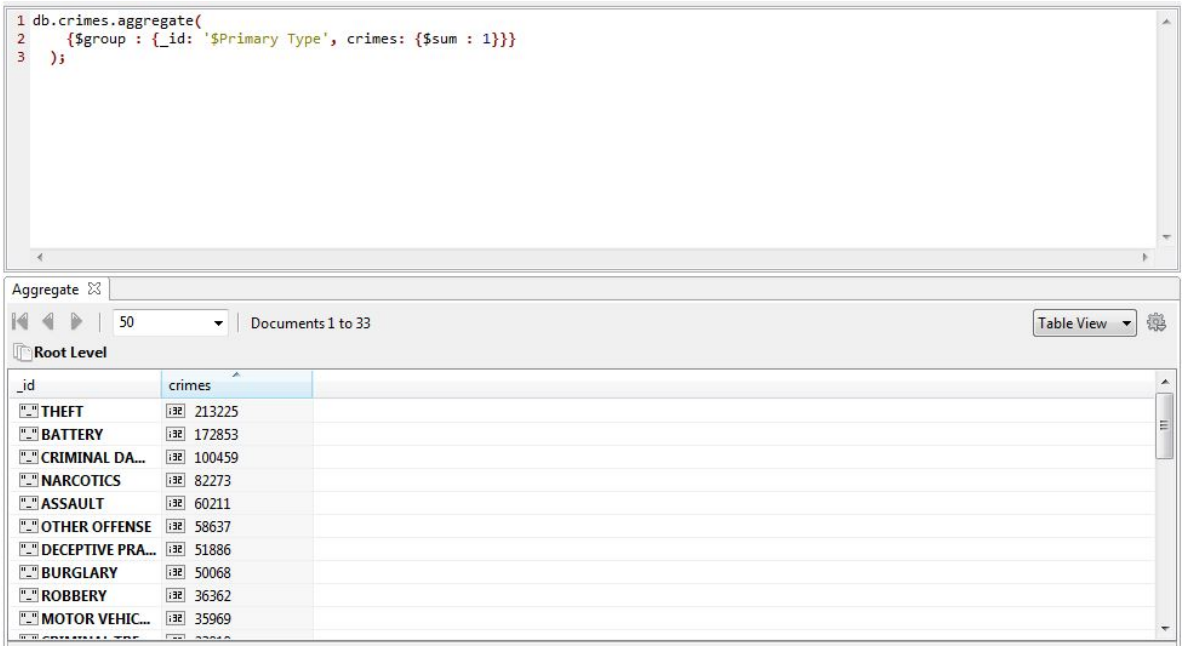


Figura 19 Tipo di crimine più commesso;

Il furto è risultato essere il crimine maggiormente commesso, segue il reato di percosse.

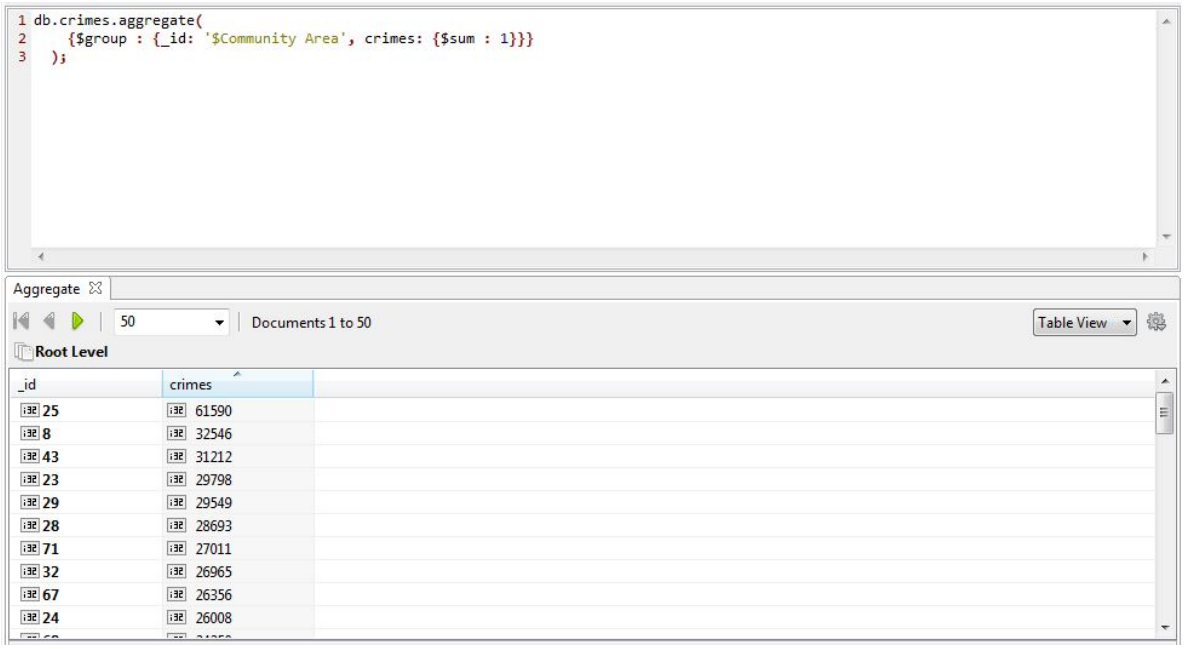


Figura 20 CA con maggior numero di crimini;

L'area più pericolosa di Chicago è senza dubbio la numero 25, a seguire ci sono la 8 e la 43.

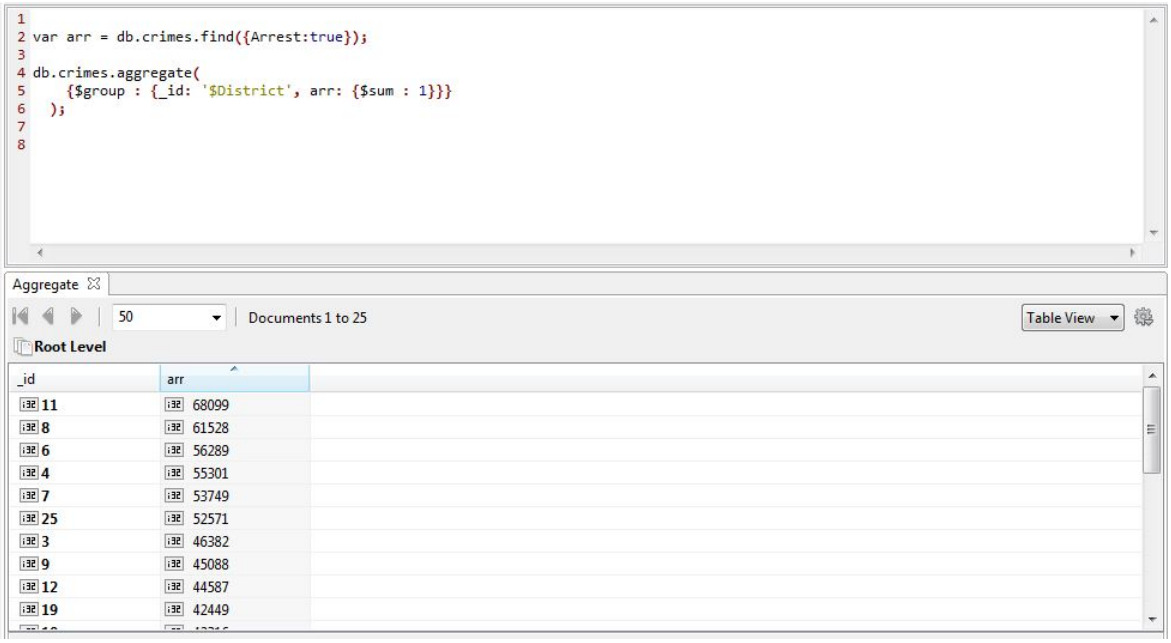


Figura 21 Distretto con maggior numero di arresti;

Il distretto numero 11 è quello che ha effettuato il maggior numero di arresti.

Conclusioni

La nostra esperienza con Qlik sense, Tableau e MongoDB è stata molto formativa. Abbiamo imparato molto relativamente all'analisi dei dati, questi tool ci hanno dato infatti l'opportunità di confrontarci con un problema di frontiera e di grande interesse aziendale.

Qlike sense è stato sicuramente il tool più intuitivo da utilizzare, l'interfaccia e la creazione delle Dashboard è molto semplice e immediata, con una piacevole visualizzazione dei risultati intuitiva anche per un utente medio.

Tableau è stato invece meno intuitivo. Anche se il loro fine è sostanzialmente lo stesso possiamo notare come i due tool abbiano approcci e strumenti differenti, in particolare relativamente alla creazione dei grafici e la visualizzazione dei risultati.

L’esperienza con MongoDB è stata sicuramente la meno immediata, abbiamo avuto modo di interfacciarci ad un mondo nuovo, il mondo dei database non relazionali. Tutto ciò ci ha dato l’occasione di affrontare un unico problema sotto punti di vista differenti, di crescere e acquisire nuove capacità che sicuramente ci saranno di grande aiuto nel mondo del lavoro.