Data Integration and Processing

London's Boroughs Crimes - Crimes Origin

Diogo Azevedo
Informatics Department
Faculdade de Ciências da
Universidade de Lisboa
Lisboa, Campo Grande
fc55773@fc.ul.pt

Gonçalo Silva
Informatics Department
Faculdade de Ciências da
Universidade de Lisboa
Lisboa Portugal
fc50038@fc.ul.pt

Brian Marques
Informatics Department
Faculdade de Ciências da
Universidade de Lisboa
Lisboa, Campo Grande
fc51020@alunos.fc.ul.pt

José David Domingues
Informatics Department
Faculdade de Ciências da
Universidade de Lisboa
Lisboa, Campo Grande
fc51084@alunos.fc.ul.pt

ABSTRACT

In this project, London's boroughs demographics and its crimes will be analysed trying to find some correlation between both.

All datasets were cleaned, analysed and merged into the desired relations for analysis and after various iterations of processing the data and generating reports of datasets attributes, some conclusions were taken.

This first part focuses more on collecting, cleaning and merging the datasets, which are contemplated in the notebooks attached to this paper, using tools like pandas and pandas profiling.

In the second part, the work focuses more on constructing the data warehouse, using a star schema, and answering the defined questions via SQL queries or using software tools like Tableau.

In the sections below, all main processes of the project are described and explained.

CCS CONCEPTS

• Data Cleaning • Data Integration • London • London's Boroughs • Crimes • Data Processing • Data Analysis • Tableau • SQL • Data Warehouse • Star Schema • Postgres

KEYWORDS

Data Integration, Data Cleaning, Data Model, DataSet, London, Crimes, Data Analysis

1 Dataset Definition and Questions

In the first approach of the project, the process was inverted starting first by searching for interesting datasets in platforms such as Kaggle and then defining the questions based on the chosen datasets. For this study, datasets relative to crimes, criminals and demographics in London were chosen to try to identify possible causes and what can be done to prevent such crimes. After the first iteration the questions were redesigned due to the fact the gathered information revealed to be interesting in many other aspects, other than the questions proposed. As so, some of the questions that are interesting to see answered are:

- Is there any relation between the socio-economic factors and the number of crimes in each borough?
- Does the school level have any impact on the number of crimes?
- The number of crimes has any impact on the migration of each borough?

2 Datasets Description

All the datasets presented below were described using pandas dataframes method Info(). This was done to check both the type of the column and also for null values. Notice also that in an a posteriori analysis of the columns data, some of them were converted to other types that made more sense to the project.

London Boroughs House Price, Crime, and Population[1]

In this full dataset only the Borough Level Crime dataset was used. The dataset is composed of 27 columns, where most of them are represented by a year and a month where a certain crime occurred a certain number of times. The other remaining columns are used to identify the borough where those crimes happened and to categorize them. Regarding data quality, this dataset brought more info about more recent crimes, adding more data to crimes that were previously gathered in the London Crime Dataset. In this dataset the columns respective to the dates remained integer values and for the others they remained string values.

	Borough	Major Category	Minor Category	201609	201610	201611	201612	201701	201702	201703	...	201711	201712	201801	201802	201803	2018
0	Barking and Dagenham	Burglary	Burglary - Business and Community	45	41	24	19	25	47	44	...	27	21	38	33	38	
1	Barking and Dagenham	Burglary	Burglary - Residential	49	60	73	100	118	124	93	...	88	124	143	134	122	
2	Barking and Dagenham	Criminal Damage	Arson	15	3	1	5	5	5	2	...	7	4	2	3	6	
3	Barking and Dagenham	Criminal Damage	Criminal Damage To M/V	61	68	67	59	65	62	61	...	48	57	60	51	53	
4	Barking and Dagenham	Criminal Damage	Criminal Damage To Other Bldg	11	12	16	8	10	6	13	...	11	10	10	5	6	

5 rows × 27 columns

London Crime Data[2]

This dataset was the "main" dataset in terms of crime data. It's originally composed of 7 columns describing the category of the crime (minor and major category), as well as the date, number of times and the borough where the certain crime occurred. This dataset due to size reasons was filtered to only grab a certain part of data (rows where the number of times a crime happened bigger than 0) and then the dataset above merged with this one. In terms of data quality, it's possible to ensure that data is trustworthy and that the dataset above complement this one, since it adds more recent data. As for the types of the columns they were left as they are, not converting for example the year and month column to the type date.

	lsoa_code	borough	major_category	minor_category	value	year	month
0	E01000002	City of London	Violence Against the Person	Harassment	0	2011	9
1	E01000005	City of London	Violence Against the Person	Harassment	0	2014	7
2	E01000001	City of London	Violence Against the Person	Harassment	0	2008	3
3	E01032739	City of London	Violence Against the Person	Harassment	0	2016	3
4	E01000001	City of London	Violence Against the Person	Harassment	0	2011	10

Borough Demographics[3]

As for this dataset, it presents us many columns about statistics of a certain borough for a certain year (specified in each column name) which can help identify possible justifications for the crime rate in that borough and boroughs in general. The dataset in terms of quality gives a lot of useful information and additional rows for comparison measures. Cleaning had to be done to only gather the information about the London boroughs and the other rows were set aside for later comparison. In terms of cleaning the following was applied to the boroughs dataset (by order): remotion of the '£' sign; remotion of the ',' separator and use the '.' separator; convert the possible columns to be float type; and in the remaining columns that have null values (annual pay) these values were replaced by the column mean. Adding to this a simple map was done to substitute the 'Inner/Outer London' string to be float type, being 1 if the borough is Inner London, and 0 if Outer. As for the statistics in each column, they were left as they were since the pre-processing required to extract each date from a column name and refactoring the data model to assign some statistics to each year proved to be not worth it.

	Area name	Inner/ Outer London	GLA Population Estimate 2016	GLA Household Estimate 2016	Inland Area (Hectares)	Population density (per hectare) 2016	Average Age, 2016	Proportion of population aged 0-15, 2016	Proportion of population of working-age, 2016	Proportion of population aged 65 and over, 2016	...	Male life expectancy, (2012-14)	Female life expectancy, (2012-14)	Anxiety score 2011-14 (out of 10)
	Barking and Dagenham	Outer London	205,773	76,841	3,610.8	57.3	32.9	21.0	86.1	13.9	...	77.6	82.1	3.05
	Barnet	Outer London	385,108	149,147	8,674.8	44.5	37.2	21.0	83.3	16.7	...	82.1	85.1	2.75

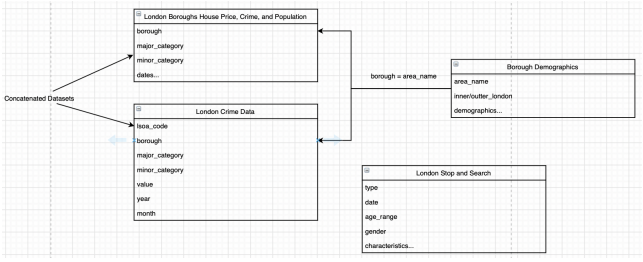
London Stop and Search[4]

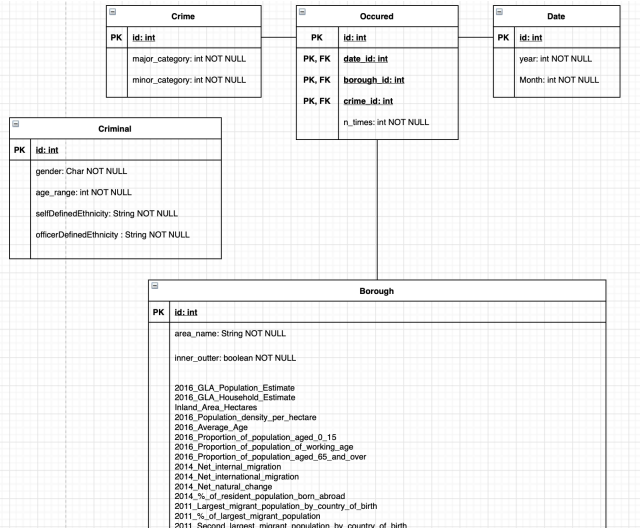
Last but not least there is the stop-and-search dataset that gets the information about criminals in London. Notice that these criminals were not matched to a certain crime that happened, and in further progress were abandoned due to the fact it didn't represent really interesting data for analysis or any kind of integration with the other datasets. The objective here was to simply have the criminal's characteristics. The original dataset had 15 columns, most of them describing what happened and the outcome of a situation but only the characteristics of the criminals were interesting. As so, the criminals table was abandoned and the posterior analysis won't be touching it.

	Type	Date	Part of a policing operation	Policing operation	Latitude	Longitude	Gender	Age range	Self-defined ethnicity	Officer-defined ethnicity	Legislation	Object of search	Outcome	Outcome linked to object of search
0	Person search	2015-03-02T16:40:00+00:00	NaN	NaN	NaN	NaN	Male	25-34	Asian or Asian British - Bangladeshi (A3)	Asian	Police and Criminal Evidence Act 1984 (section 1)	Stolen goods	Suspect arrested	True
1	Person search	2015-03-02T16:40:00+00:00	NaN	NaN	NaN	NaN	Male	25-34	Asian or Asian British - Bangladeshi (A3)	Asian	Police and Criminal Evidence Act 1984 (section 1)	Stolen goods	Suspect arrested	False
2	Person search	2015-03-02T18:45:00+00:00	NaN	NaN	NaN	NaN	Male	25-34	White - Any other White ethnic background (W9)	White	Police and Criminal Evidence Act 1984 (section 1)	NaN	Suspect arrested	True
3	Person search	2015-03-02T19:15:00+00:00	NaN	NaN	NaN	NaN	Male	over 34	White - White British (W1)	White	Police and Criminal Evidence Act 1984 (section 1)	Stolen goods	Suspect arrested	False
4	Person and Vehicle search	2015-03-03T15:50:00+00:00	NaN	NaN	NaN	NaN	Male	25-34	White - White British (W1)	White	Police and Criminal Evidence Act 1984 (section 1)	Stolen goods	Suspect arrested	True

3 Integrated Data Model

In the images below is possible to see the correspondences between the original datasets, and after that the fully integrated data model that was used for the dataset integration.





The criminals as it's possible to see are not in any relation with the other part of the model since they would be just used for analysis and insights about the characteristics of the criminals. Notice, however, that the number of columns is reduced like mentioned before since the columns related to other kinds of information that are not related to the people were discarded. As for the rest of the model, the merge and construction of it are straight-forward defining that a certain crime, that has a major and minor category, occurred a certain number of times, on a certain date, at a certain borough. The preprocessing required to build this model involved slicing the crimes dataset and creating an entity *Crime* that is composed of categories, and also extracting the date that it occurred into an entity to facilitate queries. As for the boroughs, everything was left from the original except the remotion of the columns of the codes and also some other columns that with previous analysis using pandas profiling didn't prove to be interesting enough to be kept.

4 Data Extraction and Integration

In the data extraction process, the panda's python library was used along with pandas profiling. For each dataset used in the project a report about each one was built with pandas profiling to understand each attribute, how are the values range, what is their type, if there are some outliers, among others, and after doing the first iteration of this analysis, some changes were made to each dataset, like mentioned before, to extract only the most useful and interesting data. After the second iteration of the profiling on the improved and arranged datasets, some conclusions were made about some attributes values, mostly about the boroughs demographics, and then prepared data to be exported to a CSV format being able to be imported to a database. These CSV's were generated by preparing each dataset using pandas with the required changes. A primary key was given to each entity to make it uniquely identifiable and after understanding how the integrated data model would be mapped into code, the process involved gathering the respective entity by its key and adding it to the desired table, to then export to CSV. Notice that besides this, the CSV columns and values had to match the SQL script to create the tables, otherwise, it would give an error. The imports from CSV were preferred over the insert statements due to the enormous file size that would generate. Annexed to this report there are all the notebooks done during data extraction and integration.

5 String Similarity - Cases

As for the string similarity, the datasets of the study didn't represent much necessity of using these tools for entity matching since they just add data and complement each other adding more rows or more columns. However, due to demonstration purposes a string similarity was used to show the similarity between the London boroughs and also the categories of the crimes.

For this string similarity the team splitted the categories and boroughs by buckets representing the length of the word(s). After this each of these buckets were converted to lowercase, ordered and compared, and since the number of cases is low a manually verification was done.

Here is an illustrative example:

Dataset 1 - ['Harrow', 'Brent', 'Richmond Upon Thames'].
Dataset 2 - ['brent', 'H@rrow', 'Richmond Upon Thames'].

Dataset 1, Bucket 1 - ['brent', 'harrow'].
Dataset 1, Bucket 2 - ['Richmond Upon Thames'].

Dataset 2, Bucket 1 - ['brent', 'h@rrow'].
Dataset 2, Bucket 2 - ['Richmond Upon Thames'].

In this case a mismatch would be identified in Harrow and fixed.

If this had to be scaled what would've been done is to simply sort the borough names and compare them using a certain metric, for example, the Hamming Distance.

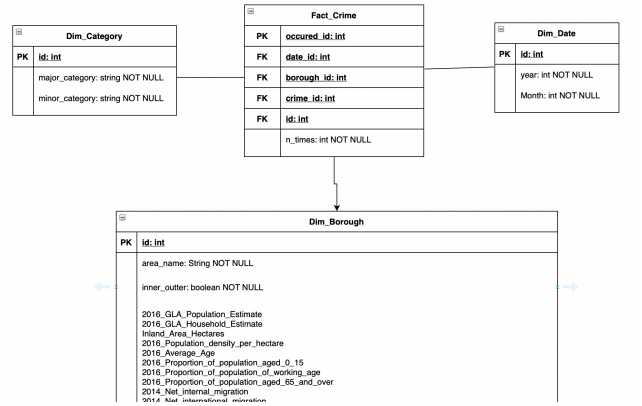
6 Macro Analysis

As for the macro analysis done in this first phase of the project, it's possible to see some outliers that stand out from others, and also some distinguishable values that are very different from the standards of England.

For most cases, there are some boroughs that stand out from others in many aspects (when just considering London itself), but the most common outliers are Richmond upon Thames, Kingston upon Thames, Westminster, Camden, and 'Kensington and Chelsea'. While comparing to England stats, it's possible to understand that some major differences that can be found are in terms of the proportion of the population of working age, in terms of migration, overseas nationals entering the UK, the gross annual pay for men, the household median income, the crime rates per thousand people, the median house price, total carbon emissions and pupils whom the first language is not English.

7 Data Warehouse - Design

As for the data warehouse, the process involved in the construction of the star schema was quite simple since the defined questions were quite open. In terms of fact tables there is only one main fact table, the crime fact table, and the rest of the tables are the dimensions that are used in different ways for the queries. The fact crime is composed by the foreign keys of the other dimensions plus the number of times a certain crime occured. It also has a primary key even though in this case it wouldn't have been necessary. As for the dimension tables their core was maintained the same as presented before with a small change, the major and minor category are now part of the category dimension that characterizes a crime instead of being in a particular table, avoiding snowflaking. This was maintained this way since the queries use distinct information and don't "compare" a certain property for all the other dimensions, and so no other property from a dimension table was added to the fact table. Even though the more open analysis used specific columns and compare them, since the queries were more specific for this case and in the Tableau software the connection between the tables is easier to do then the data warehouse was designed as follows:



The two main queries that were chosen by the team were:

- For each borough, what's the major crime category that occurs the most.
- What are the top 5 boroughs to live? Those that have less number of crimes, bigger employment rate, and bigger salary.

For this case, the first query will use all the tables. However, the columns for the demographics dimension will be completely unused. As for the second query it will use also all of the tables but it won't consider either a specific date or the type of the crime and will focus exclusively on the borough dimension. In the next section, is shown the data warehouse construction and next the queries and their results along with some insights.

8 Data Warehouse - Construction

When constructing the data warehouse from the definition of the star schema the process was quite straightforward. Initially the team started from the creation of the actual database of the information. This was done simply exporting the initial files retrieved from Kaggle and cleaned to csv and then importing them to PostgreSQL via PgAdmin.

After this, an ETL workflow was defined, and with the use of psycopg2 library for python a notebook with the required operations was coded to retrieve data from the database, clean it and transform it, and then insert it on the data warehouse previously created (using also this notebook). The example notebook is annexed to this report for consulting.

9 Queries Results & Optimizations

Starting analysing the result of the queries it was possible to observe some really interesting insights. Note also that the queries file are annexed to this report for consulting. As for the optimization part due to the performance and the way the queries were written the team didn't feel the necessity to add any kind of indexing for better performance but instead focused more on getting interesting insights using Tableau on the section ahead.

9.1 Query 1

Corresponding to the query "For each borough, what's the major crime category that occurs the most." Some insights were possible to be taken in terms of what major crimes affect the most each borough.

The query to solve this question and its results are the following:

```
SELECT A1.*
FROM (SELECT SUM(fc.value) as n_crimes, dc.major_category_id as major, b.area_name as bid
      FROM fact_crime fc, dim_category dc, dim_dates dd, dim_boroughs b
      WHERE fc.date_id = dd.date_id
      AND fc.borough_id = b.id
      AND fc.crime_id = dc.crime_id
      GROUP BY dc.major_category_id, b.area_name
      ORDER BY 3, 1 DESC, 2) AS A1
LEFT JOIN (SELECT SUM(fc.value) as n_crimes, dc.major_category_id as major, b.area_name as bid
           FROM fact_crime fc, dim_category dc, dim_dates dd, dim_boroughs b
           WHERE fc.date_id = dd.date_id
           AND fc.borough_id = b.id
           AND fc.crime_id = dc.crime_id
           GROUP BY dc.major_category_id, b.area_name
           ORDER BY 3, 1 DESC, 2) AS A2
    ON A1.bid = A2.bid AND A1.n_crimes < A2.n_crimes
WHERE A2.n_crimes IS NULL
```

	n_crimes bigint	major character varying (200)	bid character varying (100)
1	12575	Violence Against the Person	Barking and Dagenham
2	21676	Theft and Handling	Barnet
3	10055	Violence Against the Person	Bexley
4	20119	Violence Against the Person	Brent
5	16748	Theft and Handling	Bromley
6	36517	Theft and Handling	Camden
7	110434	Theft and Handling	Croydon
8	113754	Theft and Handling	Ealing
9	19315	Theft and Handling	Enfield
10	18036	Violence Against the Person	Greenwich
11	25121	Theft and Handling	Hackney
12	18660	Theft and Handling	Hammersmith and Fulham
13	23451	Theft and Handling	Haringey
14	9533	Violence Against the Person	Harrow

As it's possible to see the major crime for most boroughs during the collected data period (2008-2018) is Theft and Handling, being Violence Against the Person the second most crime. All the other major crimes are completely irrelevant in terms of occurrences compared to these ones. Further ahead, when analysing the data on Tableau was possible to confirm these results and also notice that the Violence Against The Person is mostly cases of 'Harassment' and the Theft and Handling is mostly cases of 'Other Thefts'.

9.2 Query 2

Corresponding to the query "What are the top 5 boroughs to live? Those that have less number of crimes, bigger employment rate, and bigger salary.." Some insights were possible to be taken in terms of what major crimes affect the most each borough.

The query to solve this question is the following

```
SELECT
        b.area_name
        ,b.employment_rate_pct_2015
        ,b.gross_annual_pay_2015
        ,sum(fc.value) as n_crimes
        ,((b.employment_rate_pct_2015 * b.gross_annual_pay_2015) / sum(fc.value)) as statistic

FROM
        fact_crime fc
        ,dim_boroughs b

WHERE
        fc.borough_id = b.id

GROUP BY
         b.area_name
        ,b.employment_rate_pct_2015
        ,b.gross_annual_pay_2015

ORDER BY
        statistic DESC
```

	area_name character varying (100)	employment_rate_pct_2015 double precision	gross_annual_pay_2015 double precision	n_crimes bigint	statistic double precision
1	Harrow	73.9	32529	28090	85.57825204699182
2	Bexley	75.1	32040	29479	81.62434275246785
3	Hammersmith and Fulham	77.5	38029	43034	68.48648742854488
4	Havering	76.5	32274	37076	66.59189232926961
5	Wandsworth	78.8	39562	51146	60.952676665115552

As it's possible to see the best boroughs to live according to query and the established parameters are the boroughs presented above. The formula used to calculate the statistic column simply combines the employment rate and the gross annual pay and divides it by the number of crimes in that specific borough.

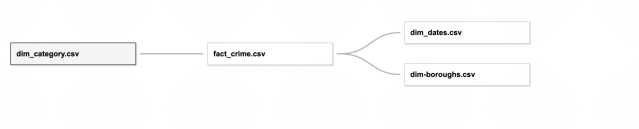
10 Open Analysis - Tableau

When using the data from the data warehouse in the Tableau software to analyse data in a more open way it is possible to generate maps that help us to understand if some phenomenon occurs due to others, and this is why this was a major part of the actual findings of the project.

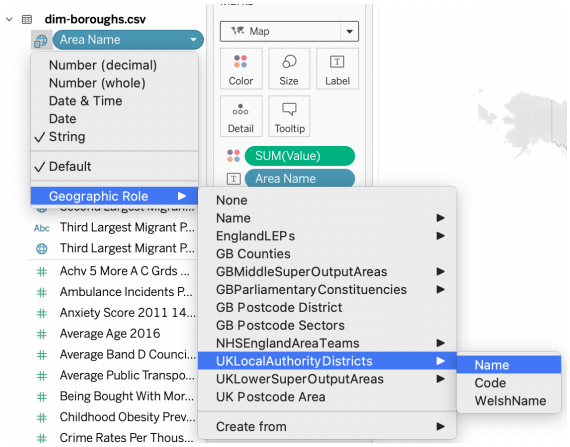
10.1 Dashboard Construction

Before diving into the actual analysis it's relevant to explain how these dashboards are done. Through this process the example to produce the crimes dashboard will be done, but the process is similar for the others.

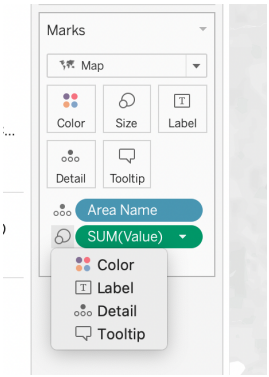
In the first place Tableau should be installed, and the annex files should be downloaded. After this, download here the required metadata that represents many features about London. After this, unzip the folder and copy the folder "Local Data" to the "My Tableau Repository" (created when Tableau was installed). With these tools the dashboards can now be built. Start by importing the fact and dimensions csv files to Tableau ("Text File"), and constructing the relation between the tables as follows.



Now create a new worksheet, and name it "crimes_map". When starting to build the dashboard change the property of the area name column on the demographics dimension.



After this simply add the area name to be plotted by double clicking on the feature. The number of crimes should also be added by double clicking on the "value" feature in the fact crime csv. By default, Tableau plots the values automatically but we want to plot a map, so change the option for map. Besides this, ensure the values are plotted according to a color scale.



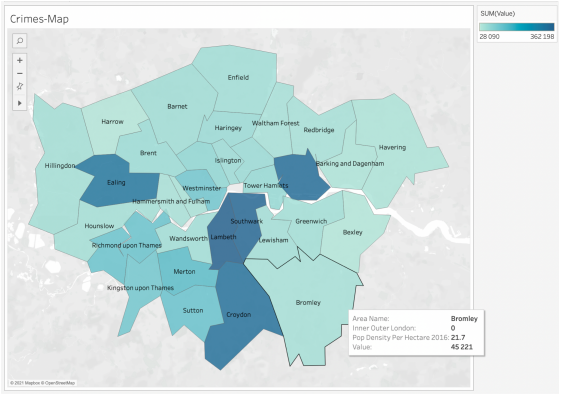
By now you should see a map plot of the boroughs and the colors varying according to the number of crimes. More features could

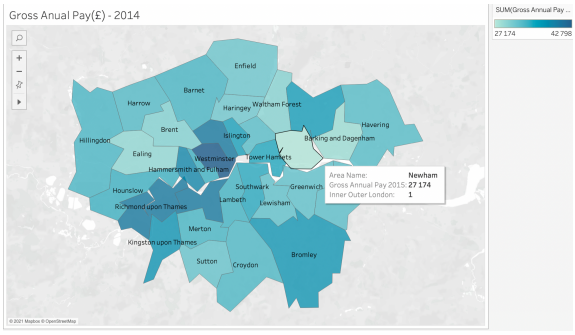
also be analysed and it's a matter of adding them or creating more worksheets and repeating the process.

10.2 Analysis

The team started looking at the happiness, anxiety and worthwhileness scores on the borough demographics trying to identify something interesting but indeed to see that the values are really close to each other and don't variate a lot. Plus, they don't really relate to other characteristics to have possible justifications to certain values.

After this observation, the team decided to look through the possibly most interesting demographics and plot the values to have insights that could possibly justify some values. In the first place, a plot of the number of crimes for each borough was done (2008-2018) using the tableau map plot. Since this approach seemed really good due to the fact it organized the map with colors according to the value passed and other values could also be obtained, the same plot technique was used to plot multiple other information. Among others tested, the core information plotted and maintained were: crimes per borough; employment per borough; minor category per borough; major category per borough; crimes per year per borough; net migration per borough; homes owned per borough; education level per borough; and finally the annual payment per borough. From all these sheets the maps were analysed and some really interesting conclusions were found and also related to the queries before.





One of the major conclusions was the fact that socioeconomic and educational factors might have impact on the number of crimes of a borough. When analysing the maps it was noticed on boroughs with high number of crimes like Ealing that the gross annual pay is below most values and when looking to the education the number of A*-C grades were inferior to most boroughs. For this case a deeper analysis could be done and found if actually this influinciates the number of crimes.

One other important observation is the fact that in the boroughs that have a high number of crimes the net migration value is negative low, which indicates there are more people leaving the borough than entering, probably because of these crimes.

Last but not least, there is the crimes per category that shows most of the crimes practiced throughout the years are Theft and Handling (related to the "Other Theft" minor category) and also Violence Against the Person (related to the "Harassment" minor category). If something could be done to prevent these crimes and since the majority of these types of crimes are really focused on more certain boroughs than it would improve a lot the security of the boroughs.

All these graphics and dashboards are available for consulting here , on the London-Crimes dashboard, and also annexed to the project delivery.

11 Future Improvements & Development

Regarding data improvements, some things could boost the knowledge taken from these datasets that unfortunately the group wasn't able to do. Having more data, retrieved from the Census for example, could boost a lot some insights since a correlation matrix would be plotted to see if some characteristics are related and then plotting some charts to see their relation.

Going deeper in the case it would be interesting to find out how the formula used for the second query works and exactly what value outputs. The annual pay, employment and number of crimes are related, but in what way? And what does this formula represent? Having this information and continuing this study, more useful insights would've been gathered.

Besides this since the demographics about the boroughs were a lot, the team focused on the most interesting ones. If the study continued the team would go deeper in each feature and plot more graphs for more interesting insights.

12 Final Conclusions

Taking in consideration the whole process of the project it's possible to draw some interesting insights not only from the data collected but also in terms of data modelation, cleaning and loading.

The major observation from the data retrieved is essentially that the socio-economic factors may have an impact on the number of crimes that happen and also on the net migration value. The types of crimes that happen the most were also identified and from there

some actions from the different security forces in London could be done to reduce this situation.

As for the data modelation, cleaning and loading it was important in terms defining a workflow for those operations, and also how data correlates and how to establish a well built data warehouse. One thing to notice also is that the data model was tried to fit the gathered data but also good for developing from the current state adding more information that complements the one gathered.

ACKNOWLEDGMENTS

Kaggle, for the datasets
Jupyter notebook
Pandas library
Pandas profiling
Bigquery for obtaining Kaggle datasets
Tableau
PostgreSQL

REFERENCES

https://www.kaggle.com/LondonDataStore/london-crime
https://www.kaggle.com/sabihxh/london-boroughs?select=MPS+Borough+Level+Crime.csv

https://public.tableau.com/profile/jos.domingues#!/

Download UK Geocoding Pack for Tableau

https://www.kaggle.com/marshald/london-boroughs
https://www.kaggle.com/sohier/london-police-records