New York City Arrests Analysis

Célia Detrez

Abstract— This project analyses New York City arrests from January 1, 2018 to December 31, 2018. Data visual analysis techniques are used to provide an insight into the characteristics of NYC arrests and law enforcement policy efficiency. Additional data with points of interest, schools and population census information has been used to analyse NYC arrests. Firstly, missing values and outliers are studied. Missing values are highly dependent on the day and the location of the arrests. Interactive visualisation is used to study the time of the year correlation. Correlation with the number of arrests is investigated to identify the relevant features. A seasonal classification model is fitted to highlight what influences the type of arrest. The significant features are the temperature (DEP) and the day of the week. Spatial variation and correlation allow exposing regional variation and geographical relationships using geographically-weighted statistics at the district levels. Ethnicity, the number of social and religious points of interest, the number of schools significantly influence the number of arrests. Brooklyn's district five has a particularly high number of arrests and points of interest. Three clusters are obtained and highlight NYC's economic and social disparities.

1 Problem Statement

NYC is the most populous city in the US with a population of about 8,398,700 residents distributed over 302.6 square miles **[12,17]**. Working in the most densely populated, the NYC police must work with the latest technologies available to prevent crimes. Indeed, in the 1990s the crime rates were very high compared to other cities in the US whereas NYC is now one of the safest. To face the crimes, NYC has adopted some revolutionary techniques such as the "broken window" policy and a management tool CompStat **[19]**. Efficiency of such measures can be assessed by visual inspection. This project will answer the following questions:

- What are the hotspots of NYC and did they change over the year 2018?
- Is there a socio-economic correlation with the number of arrests?
- Are there patterns in NYC arrests at the district level?

To do this analysis, the arrests data provided by the NYPD **[7]**, points of interest, schools and population census information **[6,7,8,9,10,11]** have been collected. Indeed, there is a strong relationship between the crime rate and low socio-economic backgrounds **[18]**. One sample of data is associated with a specific date of arrest. The data has the latitude and longitude characteristics for each arrest that allow multi-scale study and to define an appropriate scale. Thus, arrests data are spatiotemporal data that enable to answer both geographically- and temporal-related questions. In addition, the data is unstructured which not only allows to understand what happened but also why an event happened.

2 State of the Art

Historical crime records are used by Feng et al., Xia and Zhou, and Towers et al. to analyse crime issues **[1,2,3]**. Feng et al. and Towers et al. focus on Chicago crime analysis to predict the probability of arrests and improve the decision-making process using exogenous data such as the unemployment rate, weather and holiday information. Besides, Feng et al. compare the analysis of Chicago with Philadelphia crime analysis. Alternatively, Xia and Zhou assess the neighbourhood risk around user-provided trajectories in Detroit City. Their dataset is composed of a range of 5 to 14 years of arrests with more than one million records.

The main features studied by the authors are the year of arrest, the month, the day, the hour, the minute and the crime category. The authors also state the importance of the scale of the study. Feng et al. replace missing coordinates by the mean of non-missing values. They discretise the time to interpret the data as time series. As the NYPD arrests data only have the day of arrest, the time series analysis cannot be applied to this project.

Pie charts are plotted to interpret the distribution of type of crime. When Feng et al. use a box plot to highlight the evolution of the number of arrests across the month in each city, Xia and Zhou, and Towers et al. plot the number of arrests versus the day of the year. Towers et al. play on the scale of the x-axis and the type of crime to identify patterns in the arrests. They also use a box plot to highlight the variance of the number of arrests across the years for each month. Xia and Zhou compute an entropy with the probability of incidents based on the historical number of crimes and a reference number of crimes. The hotspots are shown along the route using a heatmap. A heatmap showing the number of arrests by month and day of the week highlights the most dangerous days for this route. Similarly, Feng et al. use a marker cluster algorithm with different levels of clusters. Alternatively, Towers et al. implement linear regression methods to analyse the temporal relationship between the exogenous variables and the number of arrests. As the features are heteroskedastic, a robust model is fitted, and the model is assessed on a test set. The weather features are used to adjust prediction models. As the NYPD data does not have the arrest time, this project focuses on the influence of the weather of the day on the arrests. Hence, the Towers et al. analysis is adapted.

These papers have highlighted the importance of the choice of the scale of the study and the necessity of interactivity in the analysis. The authors state that crime data is highly dependent on geographical information and the time of the year. Moreover, Xia and Zhou assumed that visual analytics is made to present and interpret multiscale, unstructured, and temporal information such as crime data.

3 Properties of the Data

The data is collected from US public databases NYC Open Data, Cityofnewyork.us, Weather.gov and NYC.gov websites **[5,6,7,8,9,10,11]**.

The main table is Arrests and has one record for each arrest. The population for each district in 2019 is computed using a linear regression of the three population values available. The census features are normalized by dividing them by the 2016 population.

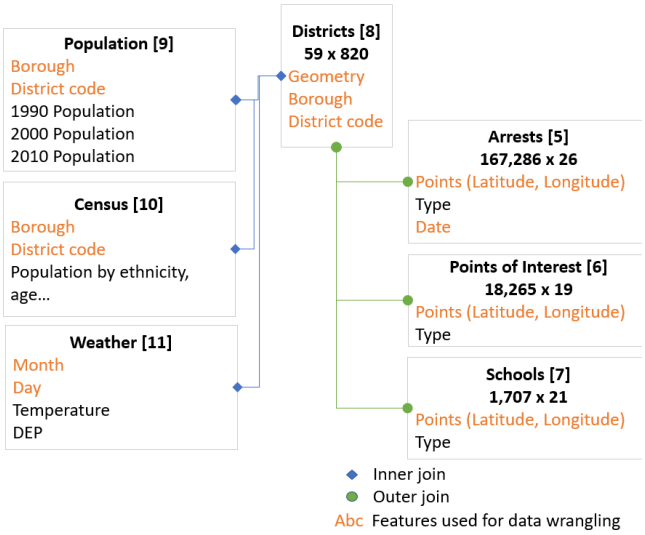

Fig. 1. Database description

The weather is supposed to be the same across NYC and to be dependent only of the day. A new table is created by counting the number of arrests, points of interest and schools in each district and dividing it by the total population of the corresponding district. The final values are the numbers per 100 inhabitants.

The data covers the arrests from January 1, 2018 to December 31, 2018 in 59 out of 77 community districts of NYC. The arrests, points of interest and schools tables have the longitude and the latitude features that enable defining the scale of the study. The GeoPandas geometry object allows defining if a point is in an area. After different analyses, the district scale has been selected. Regarding the time scale, only the arrest day is available. After joining the features, there are 820 features in the main table Districts with the 59 community districts records with 2 categorical features corresponding to the borough name and the district number, 1 for the geometry and 817 numerical. The Arrests table has 167,286 records and 26 features describing the characteristics of the arrest with 12 categorical and 14 numerical, the Points of Interest 18,265 and 19 describing the type of points of interest with 8 categorical and 11 numerical and the Schools table 1,707 and 21 describing the type of school with 13 categorical and 8 numerical.

There are some missing values of the population in the population table or with the wrong community district code. These community districts have been ignored. It introduces bias as for example a community district in the middle of the Bronx will not be studied. There are 4 features in the Arrests table with missing values:

- 0.6% of PD_CD: internal classification code
- 0.07% of PD_DESC: description of internal classification corresponding to the PD_CD
- 0.07% of KY_CD: category of PD_CD
- 0.07% of OFNS_DESC: description linked to the PD_DESC
- 0.02% of LAW_CAT_CD: level of offense: felony F, misdemeanour M, violation V, incivility I.

The first one is dependent on others. Thus, when one of the others is missing PD_CD is missing. The graph below highlights that most of the missing values are related to arrests in Manhattan.

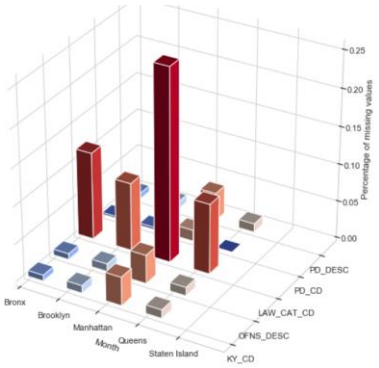

Fig. 2. Percentage of missing values per type and month of arrests

Besides, there are fewer missing values during the weekends and no missing values in Staten Island and for July arrests. It highlights the differences between the precinct regarding administrative records. As the number of missing values is very low, concerned samples will be ignored if these features need to be analysed.

4 Analysis

4.1 Approach

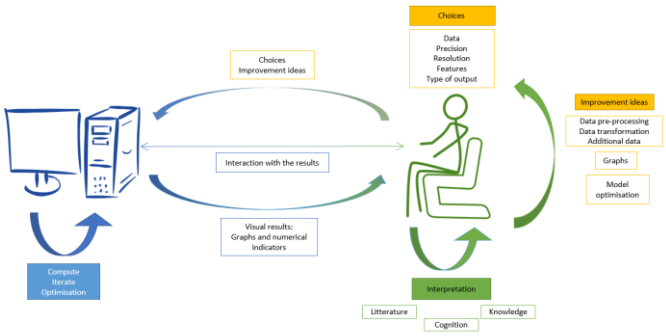

Fig. 3. Analysis workflow

The analysis workflow of this project is organised as follows: the human decides the data needed to answer its questions and what he wants to see. Firstly, the computer outputs some insights of the dataset as the missing values or some visual mapping of the features. Then, the human uses its cognition, knowledge or reading of the literature to decide the data pre-processing and transformation needed. The computer received its choices and compute them. Some outputs are fully interactive between the human and computer to allow computer-machine interaction. Iteratively, data wrangling,

choice of the scale of the analysis, data collection is performed. Once the data is judged ready by the human, the computer can compute the model and further visualisation. Iteratively, the human decides to go further in the analysis or to change the data.

For this project, the visualisation used are bar charts, interactive 3D plots, pie charts, interactive 2D plots and line graphs, interactive 2D plot of maps with a color bar representing quantitative values and heatmaps. Indeed, those plots are defined as suitable for this type of data **[1,2,3,13,14,15]**. Firstly, the time of the year correlation is explored. Then, as read in Towers et al. article, the weather dependence is significant. Seasonal classification models using a random forest classifier has been implemented to analyse the importance of the features in each season for the prediction of the type of arrests. The class imbalanced has been considered for the analysis. Then, the spatial variation and correlation have been explored and some features have been highly correlated with the number of arrests. A model is defined based on the spatial correlation. The assumptions of linear regression are verified with the visualisation of a heatmap of the correlation. The regional variation is exposed with a two-dimensional plot with a color bar for the quantitative values. The spatially-varying relationships are explored using geographically-weighted statistics. An agglomeration clustering is computed using correlation coefficients previously standardized.

Throughout this study, the first steps of data collection, data transformation, and data wrangling had to be repeated until the results were judged successful. New data based on the knowledge and previous studies have been included. The main table has been designed to improve the efficiency of the computation and make the visualisation more user-friendly. The derivation of new features, the representation of the features and the identification of independent and dependent features needed a lot of effort. Different configurations have been tested and implemented to visualise the effect on the models. The scale of the study of the arrests has been challenging as the precinct intervention areas are different from the districts. As the computer is not able to choose automatically the scale of the study and the results of the final clustering are very sensitive to the resolution, the human must choose carefully the parameter of its visualisation and its models.

4.2 Process

Firstly, the time of year correlation is explored. For this purpose, the arrest table is studied. The distribution of the arrests across the month of years is investigated by plotting a pie chart by type of arrests. Interactive visualisation is used to select a borough. Felony and misdemeanour arrests are equally distributed around the year regardless the borough. In Manhattan, the incivility and violation arrests are higher during summer. In the Queens and the Bronx, the number of incivility arrests is higher in spring whereas in Staten Island and Brooklyn the number of arrests is higher in January. The number of violation arrests has decreased during the summer in the Queens, Staten Island and the Bronx and stays the same across the years in Brooklyn. It may be the result of the law enforcement policy in New York City for violation arrests. It highlights similarities between the boroughs. Then, an interactive heatmap (Fig. 4.) at the district level is used to visualise the hotspots of arrests **[13,14]**. The location of the hotspots evolves during the year, but some patterns can be identified. That statement raises the interest of clustering the geographical areas.

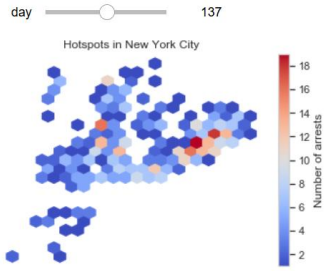

Fig. 4. Hotspots in New York City

To dig further into the evolution of the arrests, the evolution of the number of arrests is analysed with an interactive line-plot that allows choosing the area of the study. The arrests table is used and the number of arrests per month is computed **[2]**.

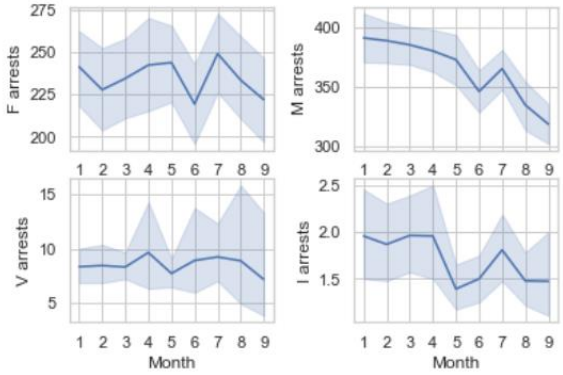

Fig. 5. Number of arrests per month in NYC

At the New York City scale, the number of arrests per day of the year has decreased on average during the year regardless the type of arrest. The miscellaneous arrests have been decreasing from January to September in NYC borough except in the Queens where the number of arrests has started to decrease sharply only from July. The felony arrests have been decreasing in Staten Island and the Bronx, slowly in Brooklyn, from July in Manhattan and stagnate in Staten Island. The violation arrests have increased in Manhattan, increased in Spring and decreased after in Brooklyn, decreased from May to June in the Queens, decreased in the Bronx, and stagnate in Staten Island. The incivility arrests have decreased from June in the Queens, decreased sharply between January and February in Brooklyn, increased from June to September in Manhattan, increased in the spring and decreased after in the Bronx, remains constant in Staten Island. Those observations strengthen the idea of dependency between the time and the number of arrests. To use a similar line plot to visualise the number of arrests per weekday, the number of arrests by weekday is counted for each type of arrest. The felony and miscellaneous arrests are higher during working days in every borough. However, the number of incivility arrests is higher on Wednesday in Manhattan, on

Friday in the Queens, on Tuesday and Sunday in the Bronx, on Saturday in Brooklyn and is the same regardless the day in Staten Island. The number of violation arrests is higher on Sunday in Manhattan and the Queens, on Friday in Staten Island, on Tuesday in the Bronx and on Saturday in Brooklyn.

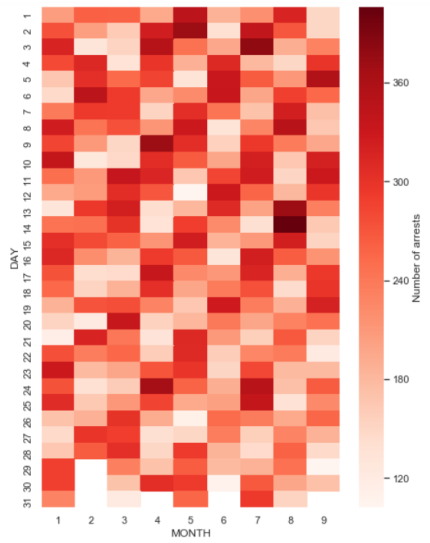

Fig. 6. Heatmap of the number of arrests per day and month

The days, weekdays and months correlation with number of arrests can be quickly analysed with a heatmap for each type of arrests **[2, 15]**. The figure above is an example. Most of the felony arrests happened between Tuesday and Friday. The miscellaneous arrests are mostly on Tuesday and Sunday. The violation arrests are mostly on the 14th of September, the 10th of August, the 15th of July, the 17th of April, on Saturday during the summer, on Wednesday in May and Monday in July. The number of incivility arrests is lower during the summer. It raises the following question: Is the number of arrests of each category influenced by the seasons?

To answer this question, random forest classifier models are fitted to predict the category of arrests based on the other features. As in Towers et al. article **[3]**, one model for each season is fitted. The dataset classes are imbalanced with the number of felonies and miscellaneous arrests ten or one hundred times higher than the number of incivility and violation arrests.

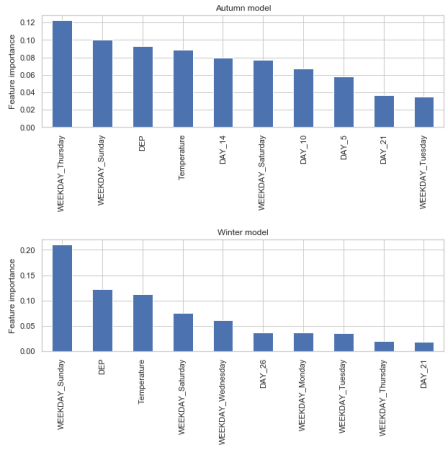

Fig. 7. Feature importance bar plots for the random forest classifier fitted for data of autumn and winter

The feature importance bar plots highlight that regardless the season, the DEP, the temperature the weekdays Sundays and Saturdays are the most significant features to predict the type of arrest. The days of the week's importance differ from a model to another. It can be linked with the previous statement made with the heatmaps (Fig 6.).

The analysis is carried on with the exploration of the correlation with the number of arrests. Interactive visualisation enables to observe the patterns. The number of female and male arrests follows the same trends with more miscellaneous and felony arrests. The districts are ranked by descending number of arrests as follows: Manhattan, Bronx, Brooklyn, and Staten Island. The ethnicity ranked by number of arrests is: black, white and black Hispanic arrests. In descending number of arrests people arrested is aged of 24-44, 18-24, 45-64, <18. It highlights the correlation of ethnicity and age with the number of arrests.

The spatial variation and correlation are explored with a visualisation like the ones used for the analysis for the Brexit referendum **[16]**. 3D bar plots with a map (Fig. 8), 2D map highlights the positive correlation of the number of arrests with the number of schools, points of interest per 100 inhabitants, the black and white density of population and the density of population. An interactive tool has been created to choose the borough, the type of arrest, the age, the ethnicity, and the feature to colour the map.

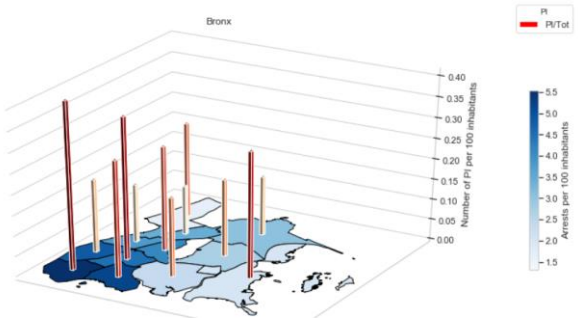

Fig. 8. Points of interest and number of arrests correlation in the Bronx

Then based on these observations and a heatmap of the feature's correlation with the number of arrests, the spatial correlation is assessed. The number of arrests is highly positively correlated to the number of high schools, the cultural, recreational, residential, education, social and water points of interest. 2D plots with points size linked to the population of the districts and one colour for each district are used. They highlight that for the number of Navajo tribal grouping population, two groups are different:

- the Bronx, Manhattan and Brooklyn.
- the Queens and Staten Island.

Then, the colour of dots shows the number of arrests. It put under the spotlight that district five of Brooklyn is different from the others with a high number of arrests, points of interest and schools. Those statements reinforce the need to cluster the data to understand better the distribution of the number of arrests.

First, a linear regression is performed to define the features for the clustering. The assumptions of linear regression are checked with a heatmap of the mutual correlation of the

features. It highlights that the total number of interest points is highly correlated with the number of points of interest of a specific type. The same statement has been observed for the schools. A first linear regression has been fitted on the total numbers whereas the second one has been fitted on counts by type of schools and interest points. The variance inflation factors (VIF) are in the range of 3 and 5 in the first model and 4 and 70 for the second model. Hence, the model variables considered in the next steps are the features representing the total numbers. Moreover, with a 2D plot with the residuals as colour of the districts, it can be observed that the first model underestimates and overestimates more the number of arrests than the second one. The differences between the areas are more pronounced. For instance, the model underestimates the number of arrests in all districts of Staten Island. The choice of the features is stopped when the variance inflation factors of the model features are lower than 10, the effect of removing the explanatory variable on VIF scores and model fit is explored iteratively. Using the features without the total number of schools and points of interest has been more successful for the next steps of the analysis and the results coincide with the knowledge of the different areas of NYC.

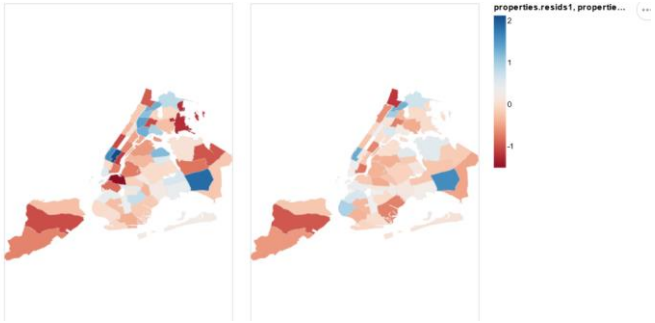

Fig. 9. Residuals spatial variation

The spatially-varying relationships are explored using geographically-weighted statistics. Centroids are defined and normalised correlations for each feature with the number of arrests are computed. The results are visualised with a map. North and South of NYC are different. Agglomerative clustering is used, and the silhouette scores allow to select the best number of clusters. The analysis is stopped once the silhouette of the clusters is judged satisfying.

4.3 Results

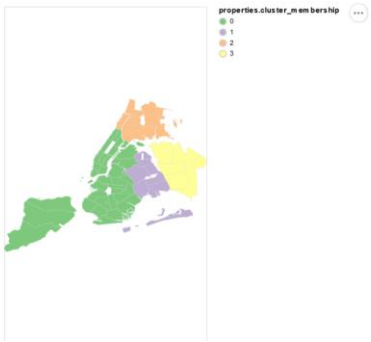

Fig. 10. New York City clusters

Firstly, this study has highlighted that the number of arrests has been decreasing on average in every borough. In addition, the number of arrests is correlated with the population density even though there is some disparities at the district level. The four clusters highlighted by the agglomerative clustering correspond to the four different areas of New York City. Cluster 2 stands for the Bronx and East Manhattan. This region was inaccessible for a long time by the policemen because of the high crime rate. Cluster 0 stands for the wealthier and safer districts of New York City. Cluster 1 for the area between Brooklyn and the Queens with the high-rise apartment buildings. Finally, cluster 3 stands for East Queens with suburban neighbourhoods. The analysis at the district level enables to overcome borough borders. The characteristics of the borough can be interpreted with the 2D interactive line-plots.

This study can be of great help for the NYPD to target their actions to the most sensitive areas and assess the quality of their new policy. The number of arrests keeping falling is a good indicator of its success **[4]**.

5 Critical Reflection

The analysis process is highly dependent on thinking for taking further steps. The choice of the features, of additional data, is made through the interpretation of visual results. Even the choice of graphs is focused on the purpose of the analysis and the results expected. As an example, the famous “broken windows” policy was expected to be effective. Results show that it is indeed the case. The visual representations have helped to understand better the data, identify the patterns, the correlations between the features and to decide the further steps.

This approach is challenged when the assumptions for the linear regression are checked. The model without the features of the total numbers of arrests and points of interest works better based on knowledge for the clustering but do not meet the linear assumptions.

This process can be altered by using the precinct areas instead of the district areas. Indeed, the precinct areas are different and reflect the policemen's work. Normalising the number of arrests by the number of policemen on duty could help identify the vulnerable areas. But the availability of data is an issue. Moreover, playing on the scale of the study could enable having a more detailed study. Also, including the social-economic background instead of points of interest and the schools could improve significantly the results.

This analysis can be potentially applicable to the analysis of every crime data, arrests data to other cities or even countries. The choice of the scale allows some flexibility in the study. It can even be applied to the different spatial-temporal events such as the fire brigade interventions.

This project has put under the spotlight some significant points to remember when using visual analysis: the data preparation and the iterative steps. Indeed, data preparation is the most demanding part of the project. It is essential to prepare the data correctly, to choose how to store it and how to represent it. The high resolution of arrest events makes it hard to work with Tableau on a regular laptop. Python select by default only a subset of points and enables fast visualisation.

I would recommend analysts who have to deal with the same kind of data to interpret step by step the data. Taking time to analyse the features, understand their meaning, analysing if they are co-varying is primordial to further the analysis. It is easier if the data provided is already clean and has all the features needed. The data wrangling can be more challenging if the data comes from various data sources. The compatibility of the data sets must be considered. The project would be successful if the model assumptions are checked and challenged throughout the project by using the resources available: literature review, cognition, and knowledge.

Table of word counts

Problem statement	249
State of the art	495
Properties of the data	498
Analysis: Approach	480
Analysis: Process	1468
Analysis: Results	193
Critical reflection	456

REFERENCES

[1] M. Feng et al., "Big Data Analytics and Mining for Effective Visualization and Trends Forecasting of Crime Data," IEEE Access, vol. 7, pp. 106111–106123, 2019.

[2] J. Xiao and X. Zhou, "Crime Exposure along My Way Home: Estimating Crime Risk along Personal Trajectory by Visual Analytics," Geographical Analysis, Jan. 2019.

[3] S. Towers, S. Chen, A. Malik, and D. Ebert, "Factors influencing temporal patterns in crime in a large American city: A predictive analytics perspective," PLOS ONE, vol. 13, no. 10, p. e0205151, Oct. 2018.

[4] S. Khurshid, "Why Does Crime Keep Falling in New York City?," Gotham Gazette, 2019. [Online]. Available: https://www.gothamgazette.com/city/7410-why-does-crime-keep-falling-in-new-york-city. [Accessed: 05-Dec-2019].

[5] Police Department (NYPD), "NYPD Arrest Data (Year to Date)," Cityofnewyork.us, 05-Jun-2018. [Online]. Available: http://data.cityofnewyork.us/Public-Safety/NYPD-Arrest-Data-Year-to-Date-/uip8-fykc. [Accessed: 05-Dec-2019].

[6] Open Calgary, "Points Of Interest | NYC Open Data," NYC Open Data, 2018. [Online]. Available: http://data.cityofnewyork.us/City-Government/Points-Of-Interest/rxuy-2muj. [Accessed: 05-Dec-2019].

[7] Department of Education (DOE, "School Point Locations," Cityofnewyork.us, 22-Sep-2011. [Online]. Available: http://data.cityofnewyork.us/Education/School-Point-Locations/jfju-ynrr. [Accessed: 05-Dec-2019].

[8] Open Calgary, "Community Districts | NYC Open Data," NYC Open Data, 2018. [Online]. Available: http://data.cityofnewyork.us/City-Government/Community-Districts/yfnk-k7r4. [Accessed: 05-Dec-2019].

[9] Department of City Planning (DCP, "New York City Population By Community Districts," Cityofnewyork.us, 20-Feb-2013. [Online]. Available: http://data.cityofnewyork.us/City-Government/New-York-City-Population-By-Community-Districts/xi7c-iiu2. [Accessed: 05-Dec-2019].

[10] Department of City Planning (DCP, "Demographic and Housing Profiles by Borough," Cityofnewyork.us, 09-Aug-2011. [Online]. Available: http://data.cityofnewyork.us/City-Government/Demographic-and-Housing-Profiles-by-Borough/cu9u-3r5e. [Accessed: 05-Dec-2019].

[11] National Weather Service Corporate Image Web Team, "National Weather Service Climate," Weather.gov, 2019. [Online]. Available: https://w2.weather.gov/climate/index.php?wfo=okx.

[12] "Planning-Population-NYC Population Facts - DCP," Nyc.gov, 2015. [Online]. Available: https://www1.nyc.gov/site/planning/planning-level/nyc-population/population-facts.page.

[13] J. Chae et al., "Visual analytics of heterogeneous data for criminal event analysis VAST challenge 2015: Grand challenge," 2015 IEEE Conference on Visual Analytics Science and Technology (VAST), Oct. 2015.

[14] A. Malik, R. Maciejewski, T. F. Collins, and D. Ebert, "Visual Analytics Law Enforcement Toolkit," 2010 IEEE International Conference on Technologies for Homeland Security (HST), Nov. 2010.

[15] G. Gorczynski, "Chicago Crime Scene | Tableau Picasso," Tableaupicasso.com, 2019. [Online]. Available: http://tableaupicasso.com/chicago-crime-scene/. [Accessed: 05-Dec-2019].

[16] R. Henkin, "VA_brexit_practical_w7," INM433 Visual Analytics (PRD1 A 2019/20), 2019.

[17] Worldpopulationreview.com, 2019. [Online]. Available: http://worldpopulationreview.com/us-cities/. [Accessed: 12-Dec- 2019].

[18] R. TARLING and R. DENNIS, "Socio-Economic Determinants of Crime Rates: Modelling Local Area Police-Recorded Crime," *The Howard Journal of Crime and Justice*, vol. 55, no. 1–2, pp. 207–225, Feb. 2016.

[19] C. Smith, "The Controversial Crime-Fighting Program That Changed Big-City Policing Forever," Intelligencer, 02-Mar-2018. [Online]. Available: http://nymag.com/intelligencer/2018/03/the-crime-fighting-program-that-changed-new-york-forever.html. [Accessed: 17-Dec-2019].