Analysis on NYC illegal Revenue

Abul Hasan Fazulullah (*Author*)
Department of Computer Science and Engineering
University of Bridgeport
Bridgeport, CT
afazulul@my.bridgeport.edu

Siddarth (Co- *Author*)
Department of Computer Science and Engineering
University of Bridgeport
Bridgeport, CT
e-mail address goes here

***Abstract*— New York City has always been the framed picture in terms of quality and wealthy lifestyle. Though being the small state in United States, the Cost of living still ranks third most expensive state. Comparing to Connecticut, New York is 65% more expensive. Forty out of Fifty people would love to settle themselves in the city of New York irrespective of the expenses. But they have no idea about the real living in New York, in fact, the people who already live in NYC have very few idea about expenses in NYC. Also, NYC has a top-notch finance department for themselves and they know how to make money to run the city. To be specific, they know how to make miscellaneous money out of people to run the city. We have thoroughly worked on variety of datasets from Department of Finance and various open information that is available to public. After our deep research, we could attain at a point that NYC is making approximately $860M a year with Parking Violation. (*Abstract*)**

Keywords—open information; miscellaneous money; datasets; parking violation (key words)

I. INTRODUCTION

Every one of us would love to have a better lifestyle than what we actually have now, no matter which point they stand at the moment. There are few couple factors which demonstrates the life of any individual. These factors includes Time, Money, health and etcetera. Out of which, the most primary factor that can control any persons' life at a given point of time would be money. In that way money plays an important role as a building block in building their own life. In United States, to have a better life, people move towards the state with high quality of lifestyle. Of which, New York being the small state stands third in terms of cost of living. This is because this state has a lot of potential wide variety of opportunities to support everyone's life. Comparing to Connecticut, New York is 65% more expensive. Forty out of Fifty people would love to settle themselves in the city of New York irrespective of the expenses[1]. But they have no idea about the real living in New York, in fact, the people who already live in NYC have very few idea about expenses in NYC. We started off with this project to find an answer for how New York makes money to run the city though being the small and densely populated city. Since we deal with the economic healthiness of that state, we started our research with overviewing the open information about New York and most specifically the information from the Department of Finance of NYC[2]. Coming before all other depictions, we focused on income and expense estimation of the state which would give us a rough numbers on what we actually going to work. Figure 1 represents the various income sources for New York City in the fiscal year 2017 while Figure 2 depicts the expenses that the state has to make in the Fiscal year 2017.

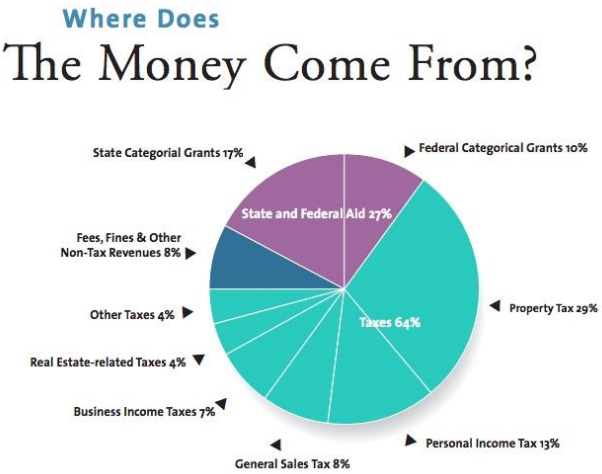

Fig 1. Income Stats Fiscal year 2017

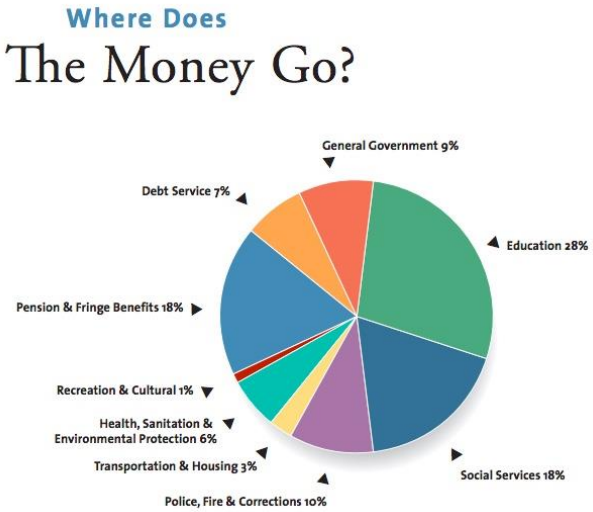

Fig 2. Expense stats Fiscal Year 2017

The expense chart is perfectly crafted to show the complete depiction of all of the New York City's expenses. To brief a little about the income stats, the city is making money majorly from three different categories. Of that, taxes comes to 64% and Federal Aid comes to 27%. Fees, Fines and Other Non-tax revenues accounts to only 8%. The only way the city can make money is either of these three ways.

We started working on each of those categories while assessing the expense reports. According to the New York budget guide, The rough revenue of NYC is $85B a year and it goes up exponentially every two years. The same report says that the expense would hit approximately $85B a year but the Mayor's Management Report(MMR) which is released from Mayor's Office of Operations twice a year has the actual expense reports which accounts more than $85B, coming to a conclusion for our findings that NYC is making unaccounted extra revenue out of the files[3][4]. Also, while making an expense, the city has to prioritize based on the current need. Figure 3 gives information on what expenses that the city face in any given year.

In New York City's budget $10M could have funded any one of the following:	
Child Care	1,377 child care vouchers (out of a total of 67,420)
Education	132 new teachers (the city employs about 75,000 teachers)
Environment	9.7 billion gallons of wastewater treated (about 2 percent of wastewater treated annually)
Fire	7 ladder trucks
Health	100 additional school based nurses
Homeless Services	213 homeless family shelter units for a year (about 1 percent of the average annual cost of sheltering families)
Parks	952 summer pool and beach season lifeguards (68 percent of seasonal lifeguards)
Police	72 police officers per year (the city employs about 23,800 personnel at the rank of police officer)
Jails	1.7 days of incarcerating the average daily population of 9,790 inmates in city jails
Public Assistance	Annual Safety Net Assistance grants for 1,936 recipients (there are 145,000 individuals receiving Safety Net Assistance)
Sanitation	9 days of disposal of residential garbage
Seniors	1.2 million home-delivered meals (4.5 million meals are delivered annually)
Street Resurfacing	67 lane-miles of city streets (about 5 percent of total lane miles resurfaced each year)
Small Business Services	13,558 job placements (48 percent of the workers hired through the Workforce1 Career Centers)
Tax Relief	$2.68 personal income tax savings per city taxpayer

SOURCE: IBO

NOTES: All numbers are approximations based on information available at the time of publication.
Personnel costs include salary and fringe benefits.

Fig 3. Mayor's Management Report

When getting deep into analyzing the revenue and the expenses, we also would be able to find that the city have a lot of debts previously for which they are paying off the payments every other time. This also adds up on top of the city's expenses and would account to even more money than what they actually spend. This also includes the minor expenses that are unaccounted.

Though the expenses are allocated to each department, the government also prioritize their complete expenses. Figure 3 show the expenses list for the first $10 Million that they care to spend.

The first document that is released by the Mayor is known as Preliminary Mayor's Management report. This report gives an approximate expense that is about to be spent for that particular year. Later that year, another report is published which is a refined version of the first report. This report will have the updated expenses that the city has to make in the remaining months of the year. Usually the Preliminary MMR is released in the second month of an year. i.e. February. This issuance of the reports is taken care by the acting director after collaborating with 45 city agencies and organizations. Behind the scenes, the staff of the Mayor's office of operations monitors this complete progress throughout the year which would help them to plan the next following year. Figure 4 gives an information from the Fiscal year 2016 to Fiscal Year 2018 and how the expenses change between the Preliminary MMR and the MMR.

SELECTED PERFORMANCE INDICATORS	Actual		4-month Actual		Target	
	FY16	FY17	FY17	FY18	FY18	FY19
Change the Culture						
Individuals trained in Mental Health First Aid (DOHMH)	4,771	21,796	3,813	11,020	72,000	72,000
Act Early						
Individuals (staff and parents) who receive mental health consultation in early care and education programs (DOHMH)	NA	1,584	108	1,629	1,457	1,457
Schools served by the school Mental Health Consultant Program (DOE)	206	930	738	899	950	950
Eligible families residing in DHS shelters who have been successfully visited by the Newborn Home Visiting Program (DOHMH)	448	1,252	466	697	1,100	1,100
Close Treatment Gaps						
Naloxone kits distributed from DOHMH to Opioid Overdose Prevention Programs (DOHMH)	10,110	30,671	6,325	31,296	48,500	48,500
NYC Well: Direct callers/texters/ chatters (non-service providers) who report that they are accessing mental health care for the first time (DOHMH) (%)	NA	16%	NA	14%	*	*
NYC Well: Inbound calls, texts, and chats answered within 30 seconds or less (DOHMH) (%)	NA	88%	84%	87%	90%	90%
Runaway and homeless youth served (DYCD)	1,835	2,408	608	928	2,600	2,600
Partner with Communities						
Mental Health Service Corps members placed in primary care practices, mental health clinics, and substance use disorder programs (DOHMH)	NA	128	126	121	130	130
Staff trained through Connections to Care (DOHMH, Mayor's Office for Economic Opportunity)	99	848	699	109	241	241

Fig 4. Expenses report FY2016 to FY2019

The MMR provides narrative and statistical information on the activities of city departments and agencies. Figure 5 is the budget value published in the MMR for the current fiscal year.

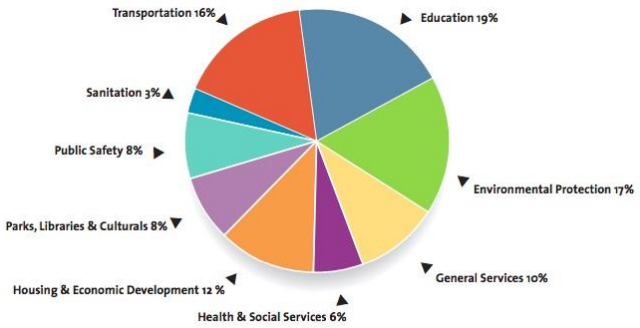

Fig 5. MMR Budget

As you can see, the budget they publish is not as same as the expenses happened in that particular Fiscal Year. For example, the budget says, 19% for education but the actual expenses says, 28% for education. But this is not the issue, it can usually happen in any firm.

To overcome this problem, Budget modifications occur year-round, with more emphasis placed on modifications associated with each quarterly Financial Plan. With this budget modification, the NYC is able to run that particular year. Also, The city's constitutional debt limit, from the statement of debt affordability, is $90.2 billion in 2017 and $97.2 billion in 2018. When we take the previous year's debts, we were able to find that the debt increases every year. Comparing to the early 2000s, the debt right now is very high. The city has to figure out some solutions to pay off all the debts. But they use the PAYGO method to pay off the debts which will continue for more years. There are few smaller debts that would be paid off in near future which also adds up to that current year's budget. Somehow the government has to find a solution to match the expenses and debts as well.

The question arises on how the government were able to match with the expenses. Based on the revenue report, the sources of revenue for NYC are taxes and fee collections. Fees also includes Fines as well which includes the Traffic Tickets, Parking Tickets and other fines. We wanted to explore over the taxes and the fines[5]. When we researched over the tax reports, we found that the taxes has been cut down in the past year. This is what they call as "Pre-trump Tax Plan". Below tables and charts will give more information on how the taxes has been reduced. The weakening in tax collections this year has particularly worried council budget officials and government watchdogs because of the sharp growth of the city's budget since Mr. de Blasio, a Democrat, took office in 2014. The budget has grown to about $82 billion from about $70 billion in 2014.

Your 2017 Income Taxes (Pre-Trump Tax Plan)

Tax Type	Marginal Tax Rate	Effective Tax Rate	2017 Taxes	2018 Trump Taxes*
Federal	25.00%	17.01%	$14,610	$12,194
FICA	7.65%	7.65%	$6,570	$6,570
State	6.45%	5.46%	$4,692	$4,675
Local	0.00%	0.00%	$0	$0
Total Income Taxes			$25,873	$23,440
Income After Taxes			$60,013	$62,446

* These will be the taxes owed for the 2018 - 2019 filing season.

Your 2017 Tax Breakdown

Income Tax	$25,873
Sales Tax	$1,812
Fuel Tax	$237
Property Tax	$5,771
Total Estimated Tax Burden	$33,692

Percent of income to taxes = 39%

Total Estimated Tax Burden

$33,692

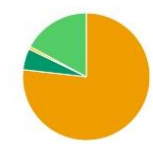

Fig. 6 2017 Income Taxes(Pre-Trump Tax plan)

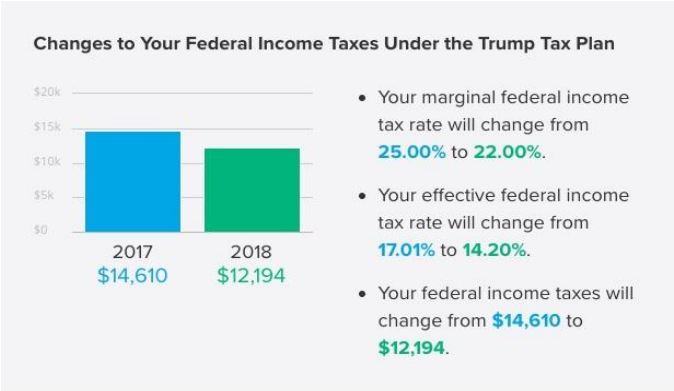

Fig. 7 Federal Income Taxes under Trump Tax Plan

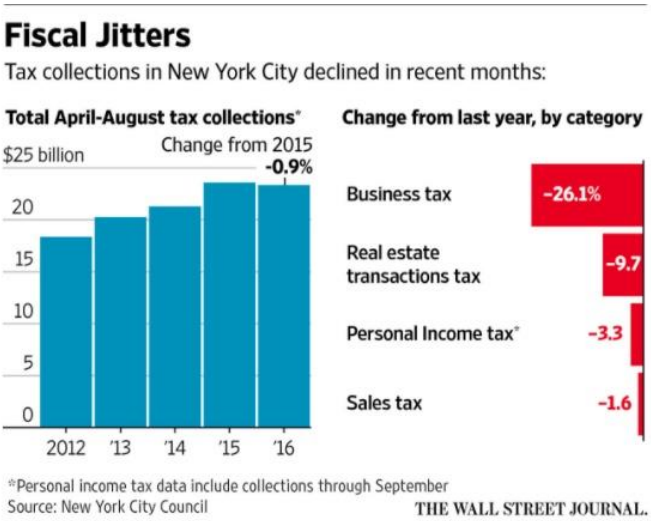

Fig. 8 Fiscal Jitters

Surprisingly, NYC also have to cover the tax differences to run the government. From all the above researches, we were able to conclude that NYC is making extra revenue from fines, and most specifically, from parking tickets.

The revenue they make can be viewed in two perspectives, the visible money and the dark money. We were very much concerned about that the dark money that NYC is making out of New Yorkers. The primary aim of this project is to find the illegal revenue made by New York City and give a rough approximation of how much money they make. We would like to support our statement with the references that would help the reader to follow-up throughout this document.

When we started narrowing down our research on the dark money the government makes, we were able to infer that, the NYC is making illegal money from the Parking tickets. Below paragraphs would suffice to prove our theory.

II. Problem Defintion

The aim of this project is to analyze and calculate the illegal money that the New York City is making from parking tickets. There are more than 10 million vehicle parking violation tickets issued every year in New York City. (Estimated $765 million has been charged for parking violation in 2016, before the late fees and excluding towing charges.) But a lot of it comes from illegal parking tickets.

This illegal issuing of tickets started after 2009 change in traffic laws, it's now legal to park in front of pedestrian ramps not connected to crosswalks in New York, but according to a big data analysis at the blog I Quant NY, police officers have still been issuing tickets, with some individual spots racking up fees worth $50,000 over just more than two years.

So, the problem is to find the number of illegal parking tickets issued and how much extra money that the government makes out of it.

III. Dataset

This project deals with a variety of datasets with variable velocity and veracity. The two primary dataset that we use for this project is Parking Tickets Dataset and the dataset from the department of finance. We would be more working on the Parking Ticket dataset and finally we use the other dataset to arrive at the solution. Also we use a little of information from the NYC Laws database and NYPD Database which is available at NYC OpenData repository. The dataset includes the dataset from Fiscal year 2013 to Fiscal Year 2018 which comprises to approximately more than 8 Gigabytes of Textual Data.

Being specific about the parking tickets dataset, it has approximately 54.6M+ Records and 44 Attributes (After processing the Dataset)(FY2013 to FY2018) and similar for the other set of datasets from Department of Finance. NYC openData repository is the major source of information for our project. We also collect quite a little information from internet i.e. Not exactly from the NYC government databases. We use this anonymous information to do a cross reference to our findings which eventually helped us in proceeding the project in the right direction.

Talking about the parking tickets datasets, it is a realtime data which will automatically downloaded every last date of the month. We enabled a feature that can pull up information from the NYC OpenData repository and store in the local storage system in CSV file format which will then be easy for the data cleaning and processing.

A. Data Cleaning

- Since the data is dynamic, it requires data cleaning process all the time to correlate with the previously existing dataset.
- This Data Cleaning includes removing unused attributes, replacing empty spaces with NULL, and make the attributes identical to the previous datasets. This Data Cleaning process also includes the Data Normalization.
- This normalization of data is usually done manually by checking across the other datasets. Once all these granular updates are done, the nuclear datasets are now combined into a single molecular dataset.
- By doing this process, we have one single huge dataset which would further reduce the time and space complexity as well.

B. Data Processing

- Processing the dataset is one important operation that has to be performed before using them in the application.
- This process includes removing unnecessary data, version controlling and filling up the empty spaces and missing information.
- One major operation of the data processing is to combine the relevant information into a single dataset and removing the duplications.

C. Generating Realtime Data

- Generating real-time data is usually done using Web Crawling and by running a CRON Job at the desired repository.
- This both functionalities runs a periodic check on the database and look for any new information added.
- To reduce the complexity of this operation, the CRON job is also setup in a way that it checks only with the specified naming conventions which will narrow down the search operation.
- Once it finds a new data, it appends the data with the existing dataset in the local system and then to the HDFS.

IV. Setting up the Environment

Setting up the environment to run these jobs takes a little time. Considering into the configurations, number of machines, latency time, memory, routes are the primary things that needs to be taken care. We used four separate machines to work on this project. Since we rely on Hadoop File System, We need one Namenode which takes care of all other nodes. We had a setup of three Datanodes. These nodes run Linux Operating System for themselves are can communicate with each other using Secure Shell(SSH). This secure shell enables these nodes to easily distribute the data across them without any intervention from the public traffic[6][7][8][9].

For our project, we created an environment with total of 60 Gigabytes storage space and 16 Gigabytes of memory. This configuration is decent enough to process a very large dataset and process a query at a decent total time. To avoid the latency

time, we enabled the passwordless SSH which is easy for the nodes to establish the communication[10].

V. Modelling the System

After collecting and cleaning the dataset, it is also important on how to approach the problem. To start with that, it is more important to model our system. The important modules that we need to look into is, MapReduce module and the Spark module. At its core, Hadoop is a distributed data store that provides a platform for implementing powerful parallel processing frameworks. The reliability of this data store when it comes to storing massive volumes of data, coupled with its flexibility in running multiple processing frameworks makes it an ideal choice for our data hub[11]. This characteristic of Hadoop means that we can store any type of data as is, without placing any constraints on how that data is processed.

*(input) <k1, v1> -> **map** -> <k2, v2> -> **combine** -> <k2, v2> -> **reduce** -<k3, v3> (output)*

A. Mapper Function

MapReduce provides an effective environment to attain automatic parallelism. The MapReduce runtime system handles these internal low-level details to attain parallelism. This ease of use feature of MapReduce framework enables us to focus more on the application logic rather than dealing with low-level parallelization details. In our MapReduce framework, we are only concerned with expressing our problem or computation in a functional programming model[12]. Once these functions are defined, the runtime system takes the responsibility of automatically parallelizing it on the given hardware architecture.

Map Tasks are used to split the input dataset into independent values. Each Cluster-node has a Job Tracker and a Task Tracker to take care of the data. Since the MapReduce operates exclusively on <Key, Value> pairs, we give the input in the same format. This Key-value classes have to be serializable by the framework and hence we implemented the writable interface. The Key class have to implement WritableComparable interface to enable the sorting feature.

Apache Organization suggests 82,000 maps for an input dataset with 10TB in size and 128MB block size. Also, the optimum level of parallelism for maps lies between 10 and 100. Thus number of maps that we used for this project is 11.

B. Reducer Function

Reducer includes three primary phases: Shuffle, Sort and Reduce. We use the following method to perform the reducing operation[13].

reduce(WritableComparable, Iterator, Output Collector, Reporter). The output is written in the File system using Output Collector . collect (Writable Comparable, Writable). The right number of reduces seems to be 0.95 or 1.75 multiplied by (<*no. of nodes* > * mapred . tasktracker . reduce . tasks . maximum).

Increasing the number of reduces increases the framework overhead, but increases load balancing and lowers the cost of failures.

The scaling factors above are slightly less than whole numbers to reserve a few reduce slots in the framework for speculative-tasks and failed tasks.

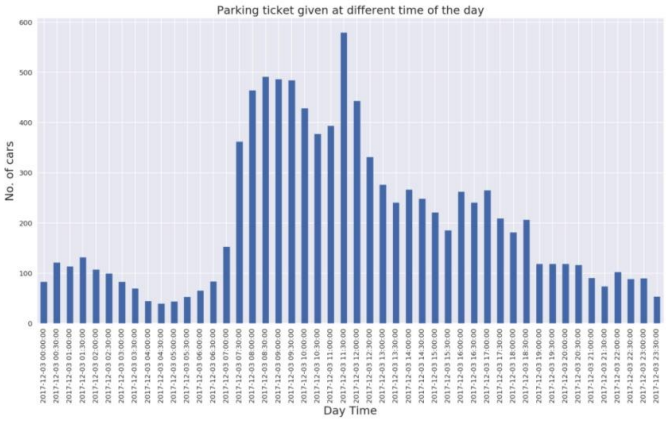

Fig. 9 Number of Cars Vs Part of the Day

C. Spark Functions

We use the core Spark APIs to operate on the data. We use the RDD API to process on the dataset. This API performs two different operations, Transformation and actions. We also use DataFrame API[14]. Spark is being the next generation of Hadoop which can process on realtime dataset but Hadoop lacks in processing the realtime dataset instead it can only process on static dataset and moreover to a fixed size. If the dataset is larger than the limit of Hadoop, then it performs the batch operation and combines the result at the end.

Fig. 10 Violation that occurs the most

D. PigLatin

We used PigLatin to work on few operations because Hadoop can do only batch operations. To run a Pig script, we need to copy the files from local system to HDFS[15]. If the dataset is in some repository, we can import the dataset into the system using the "wget" which copies the dataset from a location using a static link. Then we need to start up the grunt interactive shell using the command "pig -x mapreduce". You can either run the entire Pig script using the exec function or input each line into the grunt interactive shell to analyze the JSON file.

Fig. 11 Visualizing the Violations spots across New York

E. Google BigQuery

We used Google BigQuery to query on our datasets and do an analysis to find the way of approach for the project. For Example, We did a query to find the number of vehicles from each state that has been ticketed so that we were able to find the state that are being more concentrated on issuing the tickets. Google Bigquery allows to process millions of data in seconds.

In our case, it took approximately 50 to 60 seconds to process each query. We used one query and that is the final query to be executed to calculate the revenue generated[16]. This particular query compares the value and updates in the dataset.

The dataset can be updated using the update query which uses the template of the dataset from the JSON File. The following query will make this update action.

bq update –schema UpdateSchema.json -t Project:Dataset.table

We used the Google BigQuery API for Python to perform this operation. This piece of code provides an API for retrieving and inserting BigQuery data by wrapping Google's low-level API library. It also provides facilities that make it convenient to access data that is tied to an App Engine appspot, such as request logs[17].

VI. Performance Measures

Apart from developing the application, we also concentrated on the security and performance of the system. Since we use the Hadoop environment, we created a VPN environment that keeps all the nodes inside a single private network. All the nodes are managed from gateway and Https enabled Domain Name Server(DNS). This service is provided by CloudFlare. We make use of the CloudFlare system and integrated with our machines that are hosted in Google Cloud Platform and Amazon Web Services(AWS). We experienced an optimal latency time for every operation that we do on the dataset. Performance of the system went down when multiple operations are done simultaneously. For Example, running a job on HDFS while running another CRON Job at the backend and meanwhile running a Query on Google BigQuery using Google BigQuery API for Python from the Namenode which also acts as a Nginx webserver drops the machine performance drastically but as far as we checked, it never stopped because of heavy load. It still runs it job for a very long time and attain at a result.

VII. Conclusion

Many Data Analysts are working on this issue and they somehow try to find the exact revenue that the government is making from this parking tickets. In our project we were able to find the total amount the government is making from the parking tickets and other such violations. From this value we were able to roughly calculate the illegal revenue the government makes out of people which we call it as a miscellaneous revenue or dark money. As per our calculation, the government is making approximately $993M in the year 2016 and it has been increased to $1.9B in the year 2017. Also, this number is expected to increase every Fiscal Year. Out of this amount, we were able to find that $17M that is being collected from parking tickets are illegal. Extending more on this research would give the exact numbers but the numbers we mentioned are approximately close to the right value.

References

[1] Violation, Codes, Fines, Rules and Regulations from Inform NYC, http://www1.nyc.gov/site/finance/vehicles/services-violation-codes.page

[2] Issue on Illegal Parking Tickets issued by New York City, http://www .inverse.com/article/15564-how-new-york-city-s-open-data-revealed-the-nypd-was-issuing-illegal-parking-tickets

[3] Parking Ticket Services, http://www .nyc.gov/parkingservices

[4] The Guardian article on How an open data blogger proved NYPD issued wrong tickets, http://www.theguardian.com/cities/2016/jul/26/open-data-blogger-parking-tickets-new-york-nypd

[5] IquantNY, http://iquantny.tumblr.com/post/144197004989/the-nypd

[6] M. K.Kakhani, S. Kakhani and S. R.Biradar, Research issues in big data analytics, International Journal of Application or Innovation in Engineering & Management, 2(8) (2015), pp.228-232.

[7] A. Gandomi and M. Haider, Beyond the hype: Big data concepts, methods, and analytics, International Journal of Information Management, 35(2) (2015), pp.137-144.

[8] MH. Kuo, T. Sahama, A. W. Kushniruk, E. M. Borycki and D. K. Grunwell, Health big data analytics: current perspectives, challenges and potential solutions, International Journal of Big Data Intelligence, 1 (2014), pp.114-126.

[9] R. Nambiar, A. Sethi, R. Bhardwaj and R. Vargheese, A look at challenges and opportunities of big data analytics in healthcare, IEEE International Conference on Big Data, 2013, pp.17-22.

[10] T. K. Das and P. M. Kumar, Big data analytics: A framework for unstructured data analysis, International Journal of Engineering and Technology, 5(1) (2013), pp.153-156.

[11] T. K. Das, D. P. Acharjya and M. R. Patra, Opinion mining about a product by analyzing public tweets in twitter, International Conference on Computer Communication and Informatics, 2014.

[12] L. A. Zadeh, Fuzzy sets, Information and Control, 8 (1965), pp.338- 353. [15] Z. Pawlak, Rough sets, International Journal of Computer Information Science, 11 (1982), pp.341-356.

[13] D. Molodtsov, Soft set theory first results, Computers and Mathematics with Aplications, 37(4/5) (1999), pp.19-31.

[14] J. F.Peters, Near sets. General theory about nearness of objects, Applied Mathematical Sciences, 1(53) (2007), pp.2609-2629.

[15] R. Wille, Formal concept analysis as mathematical theory of concept and concept hierarchies, Lecture Notes in Artificial Intelligence, 3626 (2005), pp.1-33.

[16] O. Y. Al-Jarrah, P. D. Yoo, S. Muhaidat, G. K. Karagiannidis and K. Taha, Efficient machine learning for big data: A review, Big Data Research, 2(3) (2015), pp.87-93.

[17] Changwon. Y, Luis. Ramirez and Juan. Liuzzi, Big data analysis using modern statistical and machine learning methods in medicine, International Neurourology Journal, 18 (2014), pp.50-57.