ANALYSIS OF NYC'S SERVICE REQUEST CALLS TO 311

A report by:
Cedric Bhihe <cedric.bhihe@gmail.com>
Santi Calvo <s.calvo93@gmail.com>
Delivered on:
2018.06.26

Foreword

A few years ago, Ben Wellington published an article[i] about mapping New York City's noisiest neighborhoods, soon followed by another one producing results on the hidden circumstances behind New York City's traffic's permanent gridlock[ii]. Those two articles, published in the New Yorker, were meant for a wide somewhat upscale readership. However they revealed the fact that the author actually had used analytical and statistical methods based on a rich data base. That database (DB) is *NYC Open Data*[iii], a trove of information geographically and temporally more precise than census tract scale data made publicly available by the US Federal Government. We tapped it. This report describes why, how and how much.

Table of Contents

i https://www.newyorker.com/tech/elements/mapping-new-york-noise-complaints
ii https://www.newyorker.com/tech/elements/uber-isnt-causing-new-york-citys-traffic-slowdown
iii https://opendata.cityofnewyork.us/

1. Introduction

Since 2011, between 2,500 and 12,000 daily calls to 311 are recorded in New York City, NY (NYC). Those service request calls (SRCs) are logged with a slew of attributes (more than 50 fields are available per call), on the location of the incident, its nature (e.g. noise, public housing conditions, street potholes, stray animals, rodent sighting, ailing trees, barking dogs, unsanitary food establishments, etc.). SRCs' attributes include time and date , as well as geo-location of the incident, reason and object of the call. Simultaneously the NYPD, New York's Police Department, registers over 1000 daily felonies, misdemeanors and violations . This affords the curious analyst a rich overview on the type of issues being reported and about their frequency. It is also an invitation to scrutinize possible correlations between the statistics of geo-located 311 SRCs and other factors such as population density, type of criminality, median income and IRS declared jobless benefits in income tax returns. We will restrict our geographical reach to ZIP based neighborhoods in the 5 boroughs of NYC: Manhattan (1), Brooklyn (2), Queens (3), The Bronx(4), and Staten Island (5). All other ZIP codes are excluded.

In the end curiosity is what really subtends every human endeavor. More specifically in our case, the motivation to embark on this study was to ascertain how much insight can be gained from realistic multidimensional data using classical multi-variate analysis (MVA) exploratory tools. We present results based on Correspondence Analysis (CA), Principal Component Analysis (PCA), Clustering and Multiple Correspondence Analysis (MCA) to conduct data exploration, feature extraction and predictive modelling. Whenever suitable an effort is made to also offer a critical discussion of obtained results.

A less theoretically minded question is ultimately to reveal evolution patterns in the urban fabric of NYC. Our objective is to try to extract predictor-variables on the scale of a ZIP code[i] area. This is more precise than the census tract scale which may normally includes many ZIP codes. Possible applications are many:

- predict crime,
- link complaints about urban nuisance to certain neighborhoods and illustrate those neighborhoods in terms of social-economical categories,
- produce the basis reference model to help decide where to locate what business for maximum attractiveness to customers and return on investment for investors,
- optimize resources to better manage dense urban areas.

***Fig. 1**: NYC's five historical boroughs (source: Wikipedia)*

Although we provide a Table of Contents, a brief description of how this report is organized seems in order.

- In section 2, we present the protracted process of extracting data from various databases. This included cleaning it (in particular in terms of missing values) and modifying it from a time record format to a location oriented frequency table. Data cleaning, while not intrinsically or conceptually difficult, is a task laden with traps. It occupied over 170 hours of our time. This section sheds light on why and how. It can be skipped and the reader can jump directly to the analysis of Section 3.

- Section 3, encompasses the multivariate data analysis including CA and PCA on NYC311 SRCs, Clustering and MCA on 2 categorical variables and a total of 16 modalities, plus 1 (illustrative) supplementary variable and 2 quantitative variables. The analysis is performed on the April 2014 data-set, the which constitutes our training data. Our testing or validation data is the April 2015 data-set.

i ZIP or "Zone Improvement Plan" is a territorial mapping used by the US Postal Service (USPS) since 1963 to optimize mail delivery.

- Section 4, offers a general conclusion on obtained results and suggests new directions to pursue this work.

Due to external constraints imposed on this work, results produced in this report were obtained exclusively by relying on custom R scripts. Notwithstanding those constraints, we cannot but warmly advise interested coders, not to code with R during the data cleaning phase. R is quirky at times, and has either scant or too much documentation to wade through at other times. Being FOSS, it benefits from a community based ecosystem, and it is correct to say that the answer to many questions during development can be crowd-sourced. This however does not normally include extremely specific situations, where the coder is largely left to her own device. All in all data cleaning with R can be done, but is is at best irksome, grueling and slow depending on the exact nature of the task. Many R proponents will readily swear under oath that the same is true about any alternative to R, but heed our dispassionate advice: if you have the choice between R and Python for data cleaning, pick Python to go down the aisle and be forever thankful you did so.

All digital files (including input files, raw and processed data sets, scripts and result files) are made fully available to the reader, in a way which preserves the data structure and the files' hierarchical organization on any computing platform. Paths in adjoined scripts and occasionally in the body of this report are shown using Unix-like formats. However they can be transposed easily to any addressing format of the file system of your choice.

From the top containing folder "*NYC311*", the complete project's file tree is organized as follows. below means that we omit mention of some intermediate data files, obtained during the preliminary data processing phase. Those files are provided for the record. Their name usually starts with a time-stamp identifying the period to which they refer and ends with `__procXX.csv`, where XX is a double digit processing sequence identifier.

```
NYC311/
|___ Bibliography/
|___ Data/
|    |___ Geolocation/
|    |    |___ [7 shape files for NYC ZIP codes perimeter 2D drawing]
|    |___ 20140400_nyc311_raw.csv
|    |___ 20150400_nyc311_raw.csv
|    |___ 2014_zip-income_10-14_raw.csv
|    |___ 2015_zip-income_10-14_raw.csv
|    |___ 20140400_nyc-crime-map_raw.csv
|    |___ 20150400_nyc-crime-map_raw.csv
|    |___ 2014_zip-irs-exempt-unemp.csv
|    |___ 2015_zip-irs-exempt-unemp.csv
|    |___ [...]
|    |___ nyc_borough-zip.csv
|    |___ nyc311_00083-neighbors-common-border.csv # for ghost zip 00083 processing
|    |___ 20140400_nyc_whole-data-set.csv # April 2014 data-set at start of analysis
|    |___ 20150400_nyc_whole-data-set.csv # April 2015 data-set at start of analysis
|___ Report/
|___ Scripts/
     |__ 00_nyc311_input-parameters.R # defines basic period parameters and more
     |__ 01_nyc311_data-prep.R # clean up of raw data, serv. req. modalities reduction
     |__ 02_nyc311_missing-impute.R # NN-imputation or direct localization (GoogleMaps API)
     |__ 03_apportion-ghost-zip_prep.R # prepare ghost ZIPs' obs apportionment to neighbors
     |__ 04_nyc311_calls-by-zip.R # consolidates service request calls modalities per ZIP
     |__ 05_irs_median-inc-jobless.R # compute median income and joblessness per zip from IRS
     |__ 06_nypd_data-prep.R # clean up raw data, reduce crime modalities to 3
     |__ 07_nypd_crimes-by-zip.R # consolidates crime modalities per ZIP
     |__ 08_consolidate-by-zip.R # general consolidation
     |__ 09_apportion-ghost-zip_proc.R # apportion ghost ZIP's categorical variables' counts
```

2. Data-sets

2-1. Terms and conditions of use

All raw data-sets used in this project are public and accessible for free under the US Freedom of Information Act[i] (FOIA). Their use is regulated by the terms and conditions of use pertaining to each governing body responsible for their publication or production. Data dictionaries are generally made available in Appendix A, and the web pages harboring those terms are:

- http://www1.nyc.gov/home/terms-of-use.page for ZIP code centric and time-based NYC311 SRC data
- https://data.cityofnewyork.us/Business/Zip-Code-Boundaries/i8iw-xf4u for geometric ZIP code area boundary data
- https://www.irs.gov/statistics for ZIP code-centric income tax declaration data made available by the IRS
- https://www.census.gov/topics/income-poverty/income/data/tables/acs.html
  for ZIP code-centric unemployment benefit declared to the IR

2-2. Data scope and preparation

Data was generally available from various location on the web, from 2011 onward. We specialized our study to the months of April 2014 and April 2015 in order to be able to handle the corresponding volume of data. Raw files are available in ods and cvs formats at `NYC311/Data/`. Census data on population densities per ZIP code area was only available to us for the year 2016 and only for a limited number of ZIP code areas. We therefore do not include it in either one of our data-sets.

2-2-1. Duplicates, missings, and imputations

Every downloaded data-set was already fully labelled. A rapid inspection of raw data shows that "NA" (non-assigned / not-available) or erroneous values, referred to as "missings", exist, but in such proportion that dealing with them was tractable. As described below, we either imputed, re-imputed, suppressed or researched missings by cross-referencing them between DBs, with the goal of avoiding issues of data bias.

Service request calls (SRCs) to NYC 311:
The two data sets *`yyyy0400_nyc311_raw.csv`* contain the raw data of NYC SRCs for yyyy={2014,2015} as downloaded from *NYC Open Data*. That includes the call's object (description), date, time, ZIP codes and/or location (in several forms) of the reported matter and other less relevant information. We checked that data-sets contains SRCs (heretofore referred to as "dupes") from different callers with the same object. Tracking down dupes is inherently complex and we did not attempt it. More importantly, our study is concerned with people's spontaneous and independent tendency to call NYC 311 about aspects of their urban environment, which are important to them. In that sense dupes need not be eliminated; they are significant and represent a natural weighting for the data-set's observations. This will naturally influence observations' weights as represented later by marginals (row sums) in frequency tables.

Raw (unfiltered) data characteristics are shown in Table 1. Figures 2 and 3 below represent missings for the period April 2014.

Period	Raw data's obs number	Obs # with missing ZIP	Obs # missing all location info	Service requests' modalities #	Unique ZIP
April 2014	81645	3206	2740	170	278
April 2015	101890	4231	3069	178	260

***Table 1**: Summary table of salient missings and other characteristics for raw NYC 311 SRCs data sets (before data cleaning). SRCs' modalities are available in the 2 files:* `Report/yyyy0400_nyc311_proc01_modalities.csv` *with yyyy={2014,2015}*

As can be observed on Figure 2 below, during the April 2014 period, 2740 observations or 3.4% of all observations, and 85.5% of the 3206 observations missing a ZIP code have no other geographic locator. Those observations cannot be attributed to any ZIP code and are therefore useless. Figure 3 compares the service request calls' modality distributions for observations missing all location information (including a ZIP code and denoted "*loc-missing*") and the whole data set. It is readily apparent that simply eliminating "*Loc-missing*" observations would disrupt our analysis in terms of the *SocServ* modality, while for other modalities the effect would be negligible.

For that reason, we proceeded to impute a ZIP code to the 466 RFC observations missing it in 2014, but not included in the *Loc-missing* subset of missings. In practice those observations miss a ZIP code but are nevertheless endowed with some

i The FOIA is a companion to the US Privacy Act of 1974 (5 U.S.C. 552a). Under the FOIA, anyone residing legally in the USA can make a request for a Federal Agency record.

other geolocation information:

- an address, and/or
- 2 cross-streets in the form of (Xstreet_1,Xstreet_2), and/or
- an cross-road in the form of (Intersect_1,Intersect_2), and/or
- planar (Euclidian) coordinates (planeX, planeY), and/or
- GPS coordinates (latitude and longitude)

Imputation was done by fully implementing automated requests to GoogleMaps, through its API, in R, for each one of the aforementioned cases. As a result more than 97% of all 466 observations missing a ZIP code could be imputed for the April 2014 data-set. The rest including the *Loc-missing* subset of observations were given the bogus ZIP code "99999" to be uses later as a supplementary observation.

As there is no structural difference between the April 2014 and April 2015 data-sets, graphical analysis results for missings were only shown for April 2014. From Table 1, in April 2015, 3069 observations or 3.0% of all observations, and 72.5% of all observations missing a ZIP code have no other geographic locator. Here again we treat missings following the same pattern and with a similar success rate as before.

NYPD's crime reports:

Crimes are reported according to 3 general categories, which coincide with the crime modalities used in our analysis. In decreasing order of severity, they are: **felonies**, **misdemeanors**, and **violations**. . They are described and instances listed in Appendix B per the NYPD's DBs,

Data made publicly available by NYDP is completely devoid of ZIP information. However it does include planar localization and regular GPS coordinates. Because of the large amount of data involved in this study (close to 80,000 criminal observations) and of Google's imposed limitation on the number of queries (2500/day/account, as of 2018.04.30) , relying on our Google Maps API's implementation to impute a ZIP code to each crime was not deemed practical. We therefore developed two original algorithms to determine the ZIP code of each NYPD crime observation based on its planar (Cartesian) coordinates.

The first algorithm to be developed was based on nearest neighbor topological distance. It uses previously compiled ZIP code areas with planar and/or GPS coordinates for SRCs to NYC 311. The ZIP code of the 311 SRC closest in space to a crime's GPS or planar coordinates is imputed to the crime. This method is approximate and yield mixed results.

The second algorithm is exact and yields excellent results. It determines the ZIP code of every crime observation based on its planar coordinates and shape-formatted ZIP boundaries mapping data, downloaded from the *NYC Open Data* repository and made available to the reader under `Data/Geolocation/`.

The latter algorithm is general and is implemented in the form of a function, `whichBoxF()`, available at `Scripts/06_nypd_data-prep.R`. Its reaches its imputation target in more than 96% of all recorded observations. The rest, i.e. less than 4%, falls in the *missings* category and kept in supplementary observation with imputed bogus ZIP code "99999". Tables 2 below summarizes missing ZIP code "99999" imputation for crime data collected by NYPD in April 2014 and April 2015. A Chi square test of the NYPD crime data-sets' missings show that there is a significant association between missings and crime modalities. Simply suppressing missings would introduce a bias in the distribution.

2-2-2. SRCs' modality dimensional reduction

Service Request Calls' modality dimensional reduction was conducted by applying filters tailored to the semantics of the raw data's two columns: "`Complaint`", and "`Descriptor`".

The reduced modalities data-sets exhibit 13 modalities down from 170 and 178 (in Table 1, for April 2014 and April 2015 respectively) according to the description and distribution of Table 3. Noise related complaints remain the first reason for SRC to 311 with overall frequencies in noise related calls of 31.1% and 31.5% in 2014.and 2015 respectively. Table 3 is based on data after ZIP cleaning and missings imputation.

SRC modality ranking change show that the perceived (and perhaps also real) traffic noise related SRCs increased markedly between April 2014 and April 2015.

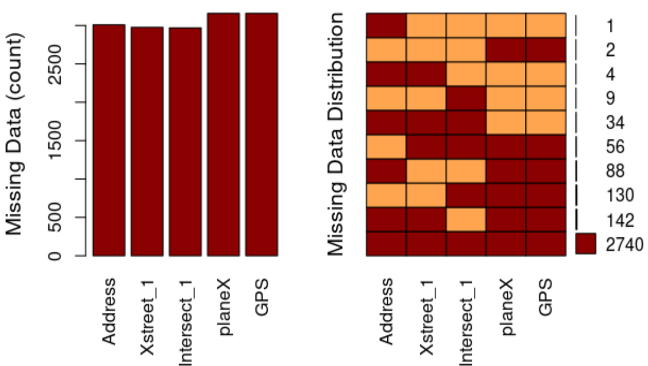

Fig. 2: *Analysis of missings for April 2014 on the 3206 raw observations missing a ZIP code.*

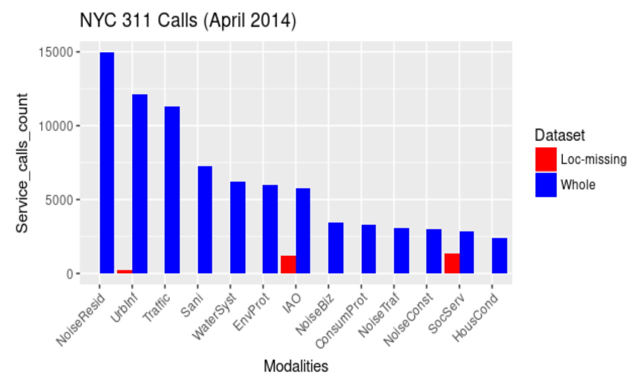

Fig. 3: *NYC 311 SRCs' modality distribution for the whole data-set (blue) and for observations missing all location information.*

April 2014	Felony	Misdemeanor	Violation	Total
non-missings	11,327	22,094	4,784	38,205
missings	481	985	64	1,530
Total	11,808	23,079	4,848	39,735

April 2015	Felony	Misdemeanor	Violation	Total
non-missings	11,669	22,080	5,010	38,759
missings	193	473	11	677
Total	11,862	22,553	5,021	39,436

Table 2*: Summary of misssings after imputation for the NYPD's crime datasets in NYC*

2-2-3. ZIP code cleaning

At this data preparation stage, our data consists of a mixture of correctly formed and ill-formed ZIP fields for each observation. An ill-formed ZIP code is a code that does not have 5 digits or does not exists officially or is otherwise not consistently found in federal US government DBs.

For our purposes, ill-formed ZIPs include ZIP+4 codes of the form 11355-1024, where the last four digits identify a geographic segment or a PO box within the five-digit ZIP delivery area. In those cases we simply suppress string characters ranging from position 6 to the end.

Other inadmissible ZIP codes are ghost ZIP codes. One of them appears in our DBs as "00083". The NYC 311 service request call data-set includes it along with surrounding and overlapping ZIP codes. So do the NYPD's crime DB, and the topological ZIP code area boundary DB also found in the *NYC Open Data* repository. Within the NYC area it designates the Central Park area in Manhattan. But because it overlaps with other official ZIP code areas surrounding it, observations identified by that ZIP code should be instead apportioned to neighboring ZIP code areas. Figure 4 (above) reveals the Zip mapping in that area, showing official ZIP code areas boundaries mapping Central Park in Manhattan. Surrounding ZIP codes are 10019, 10022, 10065, 10023, 10021, 10075, 10028, 10024, 10128, 10025, 10029, and 10026.

The use of ZIP code 00083 is
incompatible with IRS and
Census Agency DBs. To
overcome that difficulty we
calculate the common
boundaries between the 00083
ZIP code area boundary and
surrounding ZIP areas boundaries.
Our goal is to apportion
observations attributed to ZIP
code 00083 to surrounding ZIP
codes areas proportionally to
the lengths of the boundaries
they share, and in a way which
should remain modality agnostic.

Figure 5 represents the
Cartesian topological
mapping, and Table 3 shows the
computed proportion of
common boundary lengths
between Central Park's 00083
ghost ZIP and surrounding ZIP
code areas. The algorithm
developed can operate on
arbitrary sets of ZIP codes.

After correcting for ill-formed ZIP
codes, ghost ZIP codes, and ZIP

Service request calls' modalities	Modality description	Service request call frequencies		Change in rank from 2014 to 2015
		April 2014	April 2015	
NoiseResid	Residential Noise	19.00%	17.50%	–
UrbInf	Urban Infrastructure	15.00%	13.40%	↘
Traffic	Traffic related Issues	14.30%	17.20%	↗
Sani	Unsanitary Conditions	9.20%	10.50%	–
WaterSyst	Water Systems	7.80%	7.60%	–
EnvProt	Environmental Protection	7.60%	5.90%	–
IAO	Inspect, Audit, Order	5.80%	5.20%	↘
NoiseBiz	Commercial Noise	4.40%	4.90%	↘
ConsumProt	Comsumer Protection	4.20%	3.40%	↘
NoiseTraf	Traffic Noise	3.90%	5.40%	↗↗
NoiseConst	Construction Noise	3.80%	3.70%	↗
HousCond	Housing Conditions	3.10%	3.40%	–
SocServ	Social Services	1.90%	1.90%	–
	Total number of SRCs	78825	98649	↗↗

Table 3*: SRCs' consolidated modalities after dimensional reduction. The right most column indicates changes in modality ranking from 2014 to 2015.*

codes situated outside the 5-borough area, we observe respectively 208 and 206 unique observation ZIP codes in each one of our data sets: April 2014 and April 2015.

The starting point for our statistical analysis of section 3 are the two data-set located at:

```
    NYC311/
    |___ Data/
        |___ 20140400_nyc_whole-data-set.csv
        |___ 20150400_nyc_whole-data-set.csv
```

Inexplicably we note a few cases (fewer than 30 observations per data set) of missing response variables ("`medianInc`" and "`jlBenefit`") in each data set. This means that the Internal Revenue Service (IRS) chose not to make the corresponding ZIP code area's tax return statistical data public. This should not pose a problem for the coming clustering analysis (classification) , but obviously does so in any regression-like approach.

So the reader may comfortably relate ZIP codes to NYC boroughs, we provide a list of more than 200 ZIP codes and their corresponding boroughs in Appendix C.

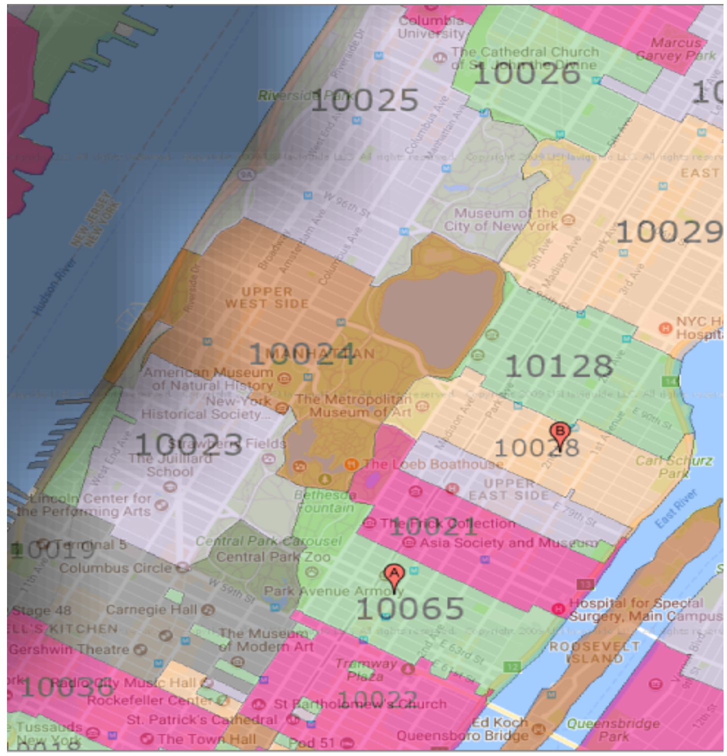

Fig. 4*: Detail of the ZIP code area map of Manhattan, showing how neighboring ZIP code areas pave Central Park piece-wise.*

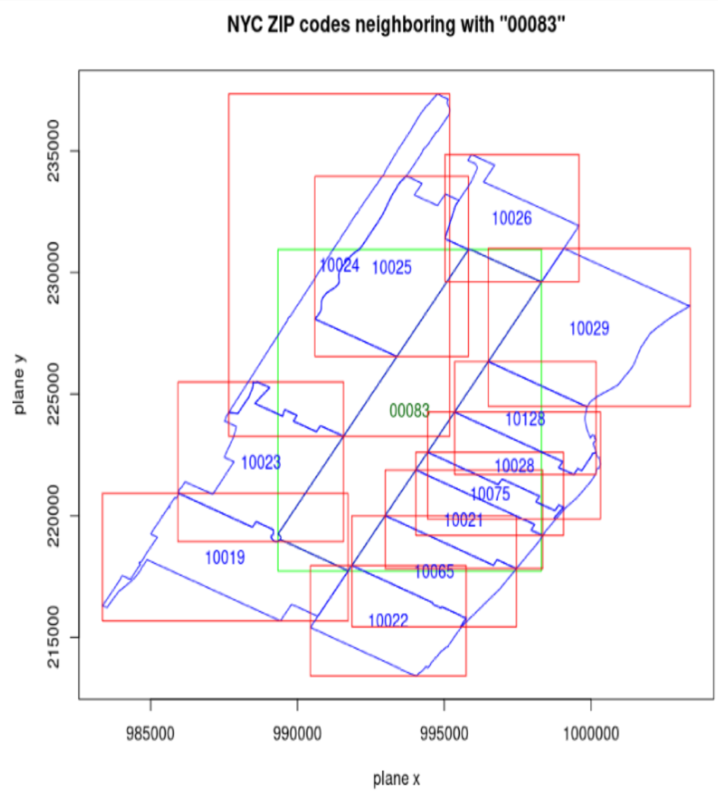

Fig.5*: New York City government ZIP code limits mapping Central Park in Manhattan and resorting to ghost ZIP code 00083 (at center in green).*

ZIP code	Common boundary length (ft)	Common boundary length proportion (%)
00083	32,710.8	100.0
10019	2,651.4	8.1
10022	259.2	0.8
10065	2,341.4	7.2
10023	4,864.4	14.9
10021	2,150.5	6.6
10075	833.0	2.5
10028	1,889.7	5.8
10024	3,748.7	11.5
10128	2,371.1	7.2
10025	5,031.3	15.4
10029	3,749.0	11.5
10026	2,821.2	8.6

Table 4*: 00083 ghost ZIP code area shared boundary analysis.*

3. Multi-Variate Analysis

Preliminary:
To approach our data, we first consider the contingency table made of the *NYC 311 service request calls* (SRC) categorical variable's 13 modalities and 208 zip codes seen as the modalities of a second categorical variable we name *Location*. Among the zip codes the last one, "99999", will be treated a supplementary variable.

We identify 33 zip codes with row marginals smaller than 5 10^{-4}, which we suppress from our contingency table. The resulting table for April 2014 is made of 174 zip codes (row labels, row index i) and 13 SRC modalities (column labels, column index j).

Next we identify table cells where low frequency and (simultaneously) high contributions to the χ^2-statistic value for the test of association of the two categorical variables may perturb the subsequent analysis. We define as low cell count or low frequency any contingency table cell count smaller than 5. There are 346 such cells. Based on the chi-square-test statistic:

$$\chi^2 = \sum_{i=1}^{195} \frac{(Count_{obs} - Count_{\mathrm{exp}})^2}{Count_{\mathrm{exp}}}$$

we calculated the contribution of every low frequency cell to the overall χ^2 statistic value and found that for low frequency cells: (i) no contribution exceeds 1%, and (ii) only 1 contributions exceed 0.1%, for a 2-sided χ^2 test statistics of 42,962. As a result the Pearson chi-square test for significant association (dependence) between row & column categories is deemed appropriate. It led to the clear rejection of the null hypothesis, with a p-value of the order of 10^{-4}:

H_0: *"In the population, the two categorical variables are independent."*

The above p-value was computed from Monte-Carlo simulations with 10,000 replicates.

Inspecting marginals, we see that SRC modalities with lowest weight across zip codes are:
"SocServ" ($f_{\cdot j} \approx 0.019$ for j=11), followed by HousCond ($f_{\cdot j} \approx 0.030$ for j=1), and "NoiseConst" ($f_{\cdot j} \approx 0.038$ for j=4).

3-1. Correspondence and Principal Components Analysis (CA, PCA)

CA was run from the FactoMineR package with row marginals as row weights to incorporate the χ^2 metric effect into the row-profile cloud projected on PC1-2, PC2-3 and PC1-3 – in Fig. 5.

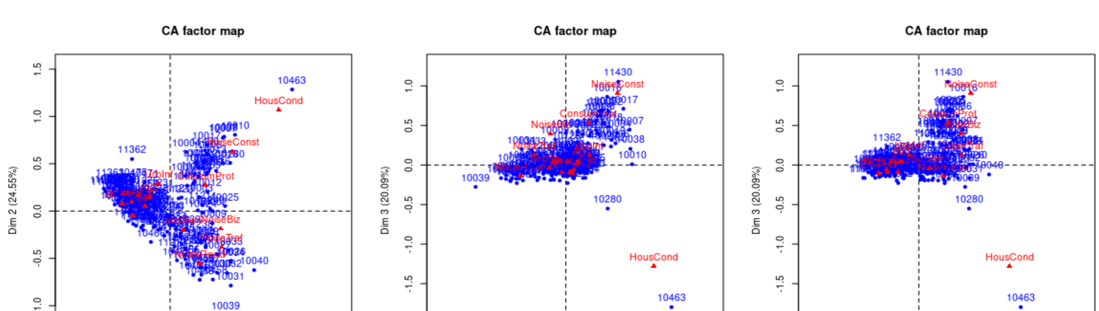

Fig. 5: *Scatter plots of row (blue) and column (red) factors in (from left to right) factorial planes: PC1-2, PC2-3 and PC1-3. The effect of the χ^2 metric is incorp-orated in the*

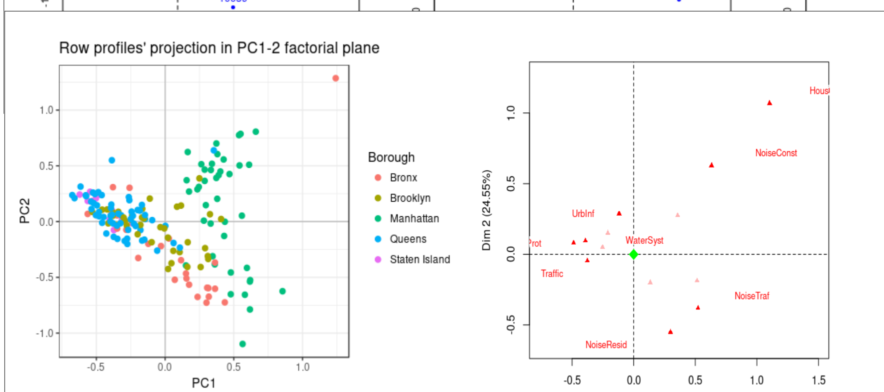

Fig. 6: *Separate projections on PC1-2 for: **(a)** using PCA **(left)** row profile color coded per NYC borough and **(b)** using CA **(right)** column profiles (labeled only for cos2>0.4.*

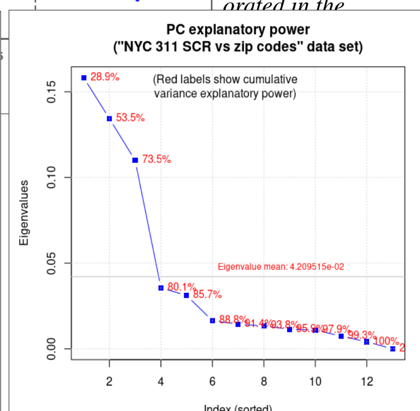

Fig. 7: *Screening plot, The first 4 eigenvalues are in decreasing order of importance 0.158, 0.134, 0.110, 0.036.*

In Fig. 5, **row (blue)** and **column (red)** profiles are printed together with distinct colors for easier differentiation. Distances *between same-colored points* are distances in the χ^2 sense to correct for the relative scarcity of factors. A red point (column profile) is a barycenter for the blue points (row profiles) expressing that column modality, weighted by said column, and vice versa[i]. To visualize projections of row and column profiles without the cloud deformation due to the incorporated χ^2 metric, we also plot separate projections for row and column profiles in Fig.6. Close and identically colored points have similar profiles.

We can retain either 3 or 4 significant dimensions. For 3 significant dimensions, the corresponding principal components (PCs) PC1, PC2, PC3 account for 73.5%. For 4 significant dimensions (adding PC4) a little over 80% of all inertia is accounted for – cf. Fig. 7. For ease of interpretation, we retain just 3 PCs.

Table 5 summarizes SRC modalities whose contributions to the construction of dimensions is greater than 10%. It also lists the corresponding quality of representation, cos2.

	Dim 1		**Dim 2**		**Dim 3**	
	ctr	***cos2***	***ctr***	***cos2***	***ctr***	***cos2***
HousCond	23.6	0.30	26.1	0.29	45.7	0.41
NoiseResid	10.7	0.20	42.9	0.69	-	-
NoiseConst	-	-	11.2	0.22	28.6	0.46
Traffic	12.8	0.50	-	-	-	-
ConsumProt	-	-	-	-	10.7	0.37
EnvProt	11.4	0.51	-	-	-	-

Table 5*: Contributions (shown for ctr > 10%) are dimension specific percentages, while quality of representation (cos2) are in [0,1]. Closer to 1 is better.*

▪ ZIP codes 10463 (the Bronx), and 10162 (Manhattan) are at similar distances from the centroid in all three factorial planes. They have similar intensity of correlations (correlations squared) with all three PCs.
▪ Due to a large number of SRCs in particular (but not only) under modality "HousCond", ZIP code 10463 (the Bronx) stands out as the biggest contributor to the construction of each of all 3 first dimensions with 17%, 22%, and a whopping 52% respectively.
▪ Altogether only 16 ZIP codes contributes more than 2% to the construction of at least one dimension. Dilution is the result of a large number of points in the row profiles' cloud.
▪ Table 5 with Fig. 6 further reveal that modalities:
- HousCond ("housing condition") and both NoiseResid ("residential noise") and NoiseTraf ("traffic noise") have near zero correlation in PC1-2, while other factorial plane cloud disposition are inconclusive.
- HousCond and NoiseConst ("construction noise") appear to be highly correlated, where their factorial plane representation is of good quality.

Figures 6 (a) and (b) give us interesting visual information on each borough's distribution on PC1-2, as well as on how predominant various significant modalities (cos2 > 0.4) of SRCs are among them:
▪ SRCs for the Bronx (orange dots)and Staten Island ZIP codes lie primarily along the second diagonal, close to the centroid, with the exception of ZIP code 10463, at approximate coordinates (+1.5, +1,5) on the plot, already discussed earlier. That ZIP code represents a one kilometer radius in the Bronx, known as Riverdale. The area has the highest population density in NYC with more than 30,000 housing units and more than 18,000 registered inhabitants per square kilometer. Understandably **HousCond** related calls to NYC 311 are disproportionately large in Riverdale, when compared to other NYC areas. Topologically neighboring ZIP areas have ZIP codes: 10467, 10468, 10471.
▪ From Figure 6, Queens and Staten Island are noted for SRCs focused on **Traffic** and **EnvProt** ("environmental protection").
▪ Manhattan meanwhile appears to be the center of **NoiseConst** ("construction noise") related complaints.
▪ **UrbInf** ("urban infrastructure") related SRCs appear common to Queens and Manhattan.
▪ Pending further examination, **WaterSyst** ("water systems") seems to be an important preocupation primarily in Both Brooklyn and Queens.

3-2. Clustering and Multiple Correspondence Analysis (MCA)

i *The reader should resist the temptation of interpreting row (blue) and column (red) profiles' proximity on the factorial planes in the χ^2 distance sense. Differently colored points may appear close, but no conclusion can be drawn from it.*

4. Conclusions

4-1. Section

As they stand, results presented in this document would certainly benefit from a comparison (perhaps at a later stage) with results obtained with the more diversified toolkit of Machine Learning.

Appendices

Appendix A: Data-set's variables' dictionaries

NYC 311 Service Request Calls – Raw Data Dictionary

Column Name	Description
Unique Key	Unique identifier of a Service Request (SR) in the open data set
Created Date	Date SR was created Date in format MM/DD/YY HH:MM:SS AM/PM
Closed Date	Date SR was closed by responding agency. Date in format MM/DD/YY HH:MM:SS AM/PM
Agency	Acronym of responding City Government Agency
Agency Name	Full Agency name of responding City Government Agency
Complaint Type	This is the fist level of a hierarchy identifying the topic of the incident or condition. Complaint Type may have a corresponding Descriptor (below) or may stand alone.
Descriptor	This is associated to the Complaint Type, and provides further detail on the incident or condition. Descriptor values are dependent on the Complaint Type, and are not always required in SR.
Status	Status of SR submitted: Assigned, Canceled, Closed, Pending, +… (Prior column indicates most frequent)
Due Date	Date when responding agency is expected to update the SR. This is based on the Complaint Type and internal SLAs. Date in format MM/DD/YY HH:MM:SS AM/PM
Resolution Action Updated Date	Date when responding agency last updated the SR. Date in format MM/DD/YY HH:MM:SS AM/PM
Resolution Description	Describes the last action taken on the SR by the responding agency. May describe next or future steps.
Location Type	Describes the type of location used in the address information
Incident Zip	Incident location zip code, provided by geo validation.
Incident Address	House number of incident address provided by submitter.
Street Name	Street name of incident address provided by the submitter
Cross Street 1	First Cross street based on the geo validated incident location
Cross Street 2	Second Cross Street based on the geo validated incident location
Intersection Street 1	First intersecting street based on geo validated incident location
Intersection Street 2	Second intersecting street based on geo validated incident location
Address Type	Type of incident location information available (Values: Address; Block face; Intersection; LatLong; Placename)
City	City of the incident location provided by geovalidation.

Landmark	If the incident location is identified as a Landmark the name of the landmark will display here
Facility Type	If available, this field describes the type of city facility associated to the SR
Community Board	Provided by geovalidation.
Borough	Provided by the submitter and confirmed by geovalidation.
X Coordinate (State Plane)	Geo validated, X coordinate of the incident location.
Y Coordinate (State Plane)	Geo validated, Y coordinate of the incident location.
Latitude	Geo based Lat of the incident location
Longitude	Geo based Long of the incident location
Location	Combination of the geo based lat & long of the incident location
Park Facility Name	If the incident location is a Parks Dept facility, the Name of the facility will appear here
Park Borough	The borough of incident if it is a Parks Dept facility
School Name	If the incident location is a Dept of Education school, the name of the school will appear in this field. If the incident is a Parks Dept facility its name will appear here.
School Number	If the incident location is a Dept of Education school, the Number of the school will appear in this field. This field is also used for Parks Dept Facilities.
School Region	If the incident location is a Dept of Education School, the school region number will be appear in this field.
School Code	If the incident location is a Dept of Education School, the school code number will be appear in this field.
School Phone Number	If the facility = Dept for the Aging or Parks Dept, the phone number will appear here. (note - Dept of Education facilities do not display phone number)
School Address	Address of facility of incident location, if the facility is associated with Dept of Education, Dept for the Aging or Parks Dept
School City	City of facilities incident location, if the facility is associated with Dept of Education, Dept for the Aging or Parks Dept
School State	State of facility incident location, if the facility is associated with Dept of Education, Dept for the Aging or Parks Dept NY
School Zip	Zip of facility incident location, if the facility is associated with Dept of Education, Dept for the Aging or Parks Dept
School Not Found	Y' in this field indicates the facility was not found (Y; N; BLANK)
School or Citywide Complaint	If the incident is about a Dept of Education facility, this field will indicate if the complaint is about a particualr school or a citywide issue. (Y; N; BLANK)
Vehicle Type	If the incident is a taxi, this field describes the type of TLC vehicle.
Taxi Company Borough	If the incident is identified as a taxi, this field will display the borough of the taxi company.
Taxi Pick Up Location	If the incident is identified as a taxi, this field displays the taxi pick up location
Bridge Highway Name	If the incident is identified as a Bridge/Highway, the name will be displayed here.
Bridge Highway Direction	If the incident is identified as a Bridge/Highway, the direction where the issue took place would be displayed here.
Road Ramp	If the incident location was Bridge/Highway this column differentiates if the issue was on the Road or the Ramp.

Bridge Highway Segment	Additional information on the section of the Bridge/Highway were the incident took place.
Garage Lot Name	Related to DOT Parking Meter SR, this field shows what garage lot the meter is located in
Ferry Direction	Used when the incident location is within a Ferry, this field indicates the direction of ferry
Ferry Terminal Name	Used when the incident location is Ferry, this field indicates the ferry terminal where the incident took place.

NYPD Crime Reports – Raw Data Dictionary

CMPLNT_NUM	Randomly generated persistent ID for each complaint
CMPLNT_FR_DT	Exact date of occurrence for the reported event (or starting date of occurrence if CMPLNT_TO_DT exists)
CMPLNT_FR_TM	Exact time of occurrence for the reported event (or starting time of occurrence if CMPLNT_TO_TM exists)
CMPLNT_TO_DT	Ending date of occurrence for the reported event if exact time of occurrence is unknown
CMPLNT_TO_TM	Ending time of occurrence for the reported event if exact time of occurrence is unknown
RPT_DT	Date event was reported to police
KY_CD	Three digit offense classification code
OFNS_DESC	Description of offense corresponding with key code (KY_CD)
PD_CD	Three digit internal classification code (more granular than Key Code)
PD_DESC	Description of internal classification corresponding with PD code; more granular than Offense Description (OFNS_DESC).
CRM_ATPT_CPTD_CD	Crime completion indicator (completed, attempted but failed, interrupted prematurely)
LAW_CAT_CD	Level of offense (felony, misdemeanor, violation)
JURIS_DESC	Jurisdiction responsible for incident. Either internal (Police, Transit, Housing) or external (Correction, Port Authority, etc."
BORO_NM	The name of the borough in which the incident occurred
ADDR_PCT_CD	The precinct in which the incident occurred
LOC_OF_OCCUR_DESC	"Specific location of occurrence in or around the premises (inside, opposite of, in front of, at the rear of)
PREM_TYP_DESC	Specific description of premises (grocery store, residence, street, etc.)
PARKS_NM	Name of NYC park, playground or greenspace of occurrence if applicable (state parks are not included)
HADEVELOPT	Name of NYCHA housing development of occurrence if applicable
X_COORD_CD	X-coordinate for New York State Plane Coordinate System, Long Island Zone (NAD 83) in units of feet (FIPS 3104)
Y_COORD_CD	"Y-coordinate for New York State Plane Coordinate System, Long Island Zone (NAD 83) in units of feet (FIPS 3104)
Latitude	"Latitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326) "
Longitude	"Longitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326)

IRS Statistics of Income per ZIP code– Raw Data Dictionary

IRS Documentation Guide (year 2014)

Contents

A. Overview

The Statistics of Income (SOI) division bases its ZIP code data on administrative records of individual income tax returns (Forms 1040) from the Internal Revenue Service (IRS) Individual Master File (IMF) system. Included in these data are returns filed during the 12-month period, January 1, 2015 to December 31, 2015. While the bulk of returns filed during the 12-month period are primarily for Tax Year 2014, the IRS received a limited number of returns for tax years before 2014 and these have been included within the ZIP code data.

B. Nature of Changes

The following changes have been made to the Tax Year 2014 ZIP Code data:

• Two new variables have been added for volunteer prepared returns: volunteered income tax assistance (VITA) and tax counseling for the elderly (TCE) prepared returns.

• Five new variables, related to the Affordable Care Act (ACA), have been added to the data: Excess advance premium tax credit repayment, Total premium tax credit, Advance premium tax credit, Health care individual responsibility payment, and Net premium tax credit. Please refer to section F for a complete list of variables and their corresponding names.

C. Population Definitions and Tax Return Addresses

• ZIP Code data are based on population data that was filed and processed by the IRS during the 2015 calendar year.

• State totals may not be comparable to State totals published elsewhere by SOI because of specific disclosure protection features in the ZIP code data.

• Data do not represent the full U.S. population because many individuals are not required to file an individual income tax return.

• The address shown on the tax return may differ from the taxpayer's actual residence.

• State codes were based on the ZIP code shown on the return.

• Excluded were tax returns filed without a ZIP code and returns filed with a ZIP code that did not match the State code shown on the return.

• Excluded were tax returns filed using Army Post Office (APO) and Fleet Post Office addresses, foreign addresses, and addresses in Puerto Rico, Guam, Virgin Islands, American Samoa, Marshall Islands, Northern Marianas, and Palau.

D. Disclosure Protection Procedures

SOI did not attempt to correct any ZIP codes on the returns; however, it did take the following precautions to avoid disclosing information about specific taxpayers:

• ZIP codes with less than 100 returns and those identified as a single building or nonresidential ZIP code were categorized as "other" (99999).

• Income and tax items with less than 20 returns for a particular AGI class were combined with another AGI class within the same ZIP Code. Collapsed AGI classes are identified with a double asterisk (**).

• All number of returns variables have been rounded to the nearest 10.

• Excluded from the data are items with less than 20 returns within a ZIP code.

• Excluded from the data are tax returns with a negative adjusted gross income.

• Excluded are tax returns representing a specified percentage of the total of any particular cell. For example, if one return represented 75 percent of the value of a given cell, the return was suppressed from the tabulation. The actual threshold percentage used cannot be released.

E. File Characteristics

The ZIP code data are available in three formats:

(1) Individual state excel files—14zp##xx.xls (## = 01-51; xx = AL-WY)

(2) A comma separated file (.csv) with AGI classes —14zpallagi.csv

(3) A comma separated filewithout AGI classes(The AGI_STUB variable has been set to zero for this file)—14zpallnoagi.csv

For all the files, the money amounts are reported in thousands of dollars.

F. Selected Income and Tax Items

STATEFIPS	The State Federal Information Processing System (FIPS) code
STATE	The State associated with the ZIP code
ZIPCODE	5-digit Zip code
AGI_STUB	Size of Adjusted Gross Income (AGI) 1 = $1 under $25,000 2 = $25,000 under $50,000 3 = $50,000 under $75,000 4 = $75,000 under $100,000 5 = $100,000 under $200,000 6 = $200,000 or more
N1	Number of returns
...	**...**

G. Endnotes:

For complete individual income tax tabulations at the State level, see the historic table posted to Tax Stats at http://www.irs.gov/uac/SOI-Tax-Stats---Historic-Table-2.

Does not include returns with adjusted gross deficit.

The "Number of volunteer prepared returns" shows counts of returns prepared by IRS-certified volunteers to taxpayers with limited income, persons with disabilities, limited English speaking taxpayers, current and former members of the military, and taxpayers who are 60 years of age and older.

“Qualified dividends” are ordinary dividends received in tax years beginning after 2002 that meet certain conditions and receive preferential tax rates. The maximum qualified dividends tax rate is 15%.

Includes the Alaskan permanent fund, reported by residents of Alaska on Forms 1040A and 1040EZ's.

This fund only applies to statistics in the totals, and the state of Alaska.

Earned income credit includes both the refundable and non-refundable portions. The non-refundable portion could reduce income tax and certain related taxes to zero. The earned income credit amounts in excess of total tax liability, or amounts when there was no tax liability at all, were refundable. See footnote 6 below for explanation of the refundable portion of the earned income credit.

The refundable portion of the earned income credit equals total income tax minus the earned income credit. If the result is negative, this amount is considered the refundable portion. No other refundable credits were taken into account for this calculation.

Income tax reflects the amount reported on Form 1040 line 56. It also includes data from Form 1040A and 1040EZ filers.

“Total tax liability” differs from “Income tax”, in that “Total tax liability” includes the taxes from recapture of certain prior-year credits, tax applicable to individual retirement arrangements (IRA's), social security taxes on self-employment income and on certain tip income, advanced earned income payments, household employment taxes, and certain other taxes listed in the Form 1040 instructions.

[10] Reflects payments to or with-holdings made to "Total tax liability". This is the amount the tax filer owes when the income tax return is filed.

[11] The amount of over-payments the tax filer requested to have refunded.

Appendix B: NYPD crime categorization

Crime modalities are: felony, misdemeanor, and violation.

FELONY is the most serious of offenses and gives rise to a more thorough classification. Felonies are lettered, with Class A being the most serious and Class E being the least serious. They are also divided into a smaller sub category; violent and non violent. In the state of NY, a non-violent, Class D felony would call for 1 to 4 years of probation. However, a violent Class D felony would automatically require a prison sentence of at least 2 years. What characterizes each felony as violent or non-violent is usually the presence of a weapon (possession of a firearm) or bodily harm to another person (aggravated assault/battery). A Class A Felony (e.g a 1st degree murder) is punishable by life in prison, with or without parole, depending on the circumstances.

MISDEMEANOR is the second type of criminal offenses, less severe than felonies but more serious than violations. Misdemeanors can carry up to a year in jail. In addition to jail time, a person convicted of a misdemeanor can also be subject to fines, probation, community service or restitution (victim compensation). A classic case of a misdemeanor would be simple assault, possession of a small amount of marijuana, or driving under the influence.

VIOLATION (also known as "infractions") is a minor offense. A speeding ticket, public intoxication, or jaywalking are some of the many petty offenses that could fall under the umbrella of violations. Violations are punishable by fines primarily, and do not result in jail or prison time.

In the subsequent listings, a number following a label within each category indicates the degree of the charge within that category, i.e. sub-categorization for judicial purposes.

Felonies

RAPE 1 *(means "1st degree rape", etc.)*
LARCENY,GRAND BY OPEN/COMPROMISE CELL PHONE ACCT
LARCENY,GRAND BY OPEN CREDIT CARD (NEW ACCT)
RAPE 3
FRAUD,UNCLASSIFIED-FELONY
LARCENY,GRAND BY DISHONEST EMP
BURGLARY,RESIDENCE,NIGHT
SEX CRIMES
RAPE 2
LARCENY,GRAND BY BANK ACCT COMPROMISE-REPRODUCED CHECK
SODOMY 1
LARCENY,GRAND BY THEFT OF CREDIT CARD
LARCENY,GRAND BY FALSE PROMISE-NOT IN PERSON CONTACT
LARCENY,GRAND FROM RESIDENCE, UNATTENDED
SEXUAL ABUSE
LARCENY,GRAND FROM BUILDING (NON-RESIDENCE) UNATTENDED
COERCION 1
PUBLIC ADMINISTRATION,UNCLASSI
COMPUTER TAMPER/TRESSPASS
LARCENY,GRAND FROM OPEN AREAS, UNATTENDED
LARCENY,GRAND BY IDENTITY THEFT-UNCLASSIFIED
BURGLARY,RESIDENCE,UNKNOWN TIM
BURGLARY,RESIDENCE,DAY
LARCENY,GRAND BY FALSE PROMISE-IN PERSON CONTACT
TAMPERING 1,CRIMINAL
RAPE 1,ATTEMPT
LARCENY,GRAND BY CREDIT CARD ACCT COMPROMISE-EXISTING ACCT

LARCENY,GRAND BY BANK ACCT COMPROMISE-TELLER
FORGERY,ETC.,UNCLASSIFIED-FELO
NY STATE LAWS,UNCLASSIFIED FEL
CRIMINAL CONTEMPT 1
LARCENY,GRAND BY BANK ACCT COMPROMISE-ATM TRANSACTION
LARCENY,GRAND BY ACQUIRING LOST CREDIT CARD
MISCHIEF,CRIMINAL,   UNCL 2ND
ARSON 2,3,4
RECKLESS ENDANGERMENT 1
MISCHIEF, CRIMINAL 3 & 2, OF M
LARCENY,GRAND OF VEHICULAR/MOTORCYCLE ACCESSORIES
LARCENY,GRAND FROM STORE-SHOPL
LARCENY,GRAND BY BANK ACCT COMPROMISE-UNCLASSIFIED
LARCENY,GRAND BY ACQUIRING LOS
LARCENY,GRAND FROM VEHICLE/MOTORCYCLE
LARCENY,GRAND OF AUTO
BURGLARY,COMMERCIAL,NIGHT
LARCENY,GRAND FROM RETAIL STORE, UNATTENDED
BURGLARY,COMMERCIAL,UNKNOWN TI
LARCENY,GRAND FROM PERSON,PICK
LARCENY,GRAND OF MOTORCYCLE
LARCENY,GRAND BY EXTORTION
WEAPONS POSSESSION 3
FORGERY,DRIVERS LICENSE
LARCENY,GRAND FROM PERSON,PERSONAL ELECTRONIC DEVICE(SNATCH)
ROBBERY,OPEN AREA UNCLASSIFIED
LARCENY,GRAND FROM NIGHT CLUB, UNATTENDED
CONTROLLED SUBSTANCE,INTENT TO
ASSAULT 2,1,UNCLASSIFIED
CONTROLLED SUBSTANCE,POSSESS.
ROBBERY,DWELLING
IMPRISONMENT 1,UNLAWFUL
STRANGULATION 1ST
LARCENY,GRAND FROM EATERY, UNATTENDED
STOLEN PROPERTY 2,1,POSSESSION
LARCENY, GRAND OF AUTO - ATTEM
BURGLARY,TRUCK NIGHT
ROBBERY,PERSONAL ELECTRONIC DEVICE
BURGLARY,UNCLASSIFIED,NIGHT
LARCENY,GRAND OF BICYCLE
ARSON, MOTOR VEHICLE 1 2 3 & 4
WEAPONS POSSESSION 1 & 2
CONTROLLED SUBSTANCE, SALE 5
FORGERY,M.V. REGISTRATION
ASSAULT 2,1,PEACE OFFICER
ROBBERY,COMMERCIAL UNCLASSIFIED
FORGERY-ILLEGAL POSSESSION,VEH
ROBBERY,RESIDENTIAL COMMON AREA
LARCENY,GRAND FROM PERSON, BAG OPEN/DIP
CONTROLLED SUBSTANCE,SALE 1
BRIBERY,PUBLIC ADMINISTRATION
IMPERSONATION 1, POLICE OFFICER
MARIJUANA, SALE 1, 2 & 3
ROBBERY,PUBLIC PLACE INSIDE

MENACING 1ST DEGREE (VICT NOT
CRIMINAL MIS 2 & 3
ROBBERY, PAYROLL
ROBBERY,HOME INVASION
CONTROLLED SUBSTANCE,SALE 3
LARCENY,GRAND FROM PERSON,PURS
THEFT,RELATED OFFENSES,UNCLASS
LARCENY,GRAND FROM PERSON,UNCL
ROBBERY,CAR JACKING
AGGRAVATED HARASSMENT 1
BURGLARY,COMMERCIAL,DAY
LARCENY,GRAND BY BANK ACCT COMPROMISE-UNAUTHORIZED PURCHASE
ROBBERY,POCKETBOOK/CARRIED BAG
CONTROLLED SUBSTANCE, POSSESSI
UNAUTHORIZED USE VEHICLE 2
CONTROLLED SUBSTANCE, INTENT T
BURGLARY,TRUCK DAY
MARIJUANA, POSSESSION 1, 2 & 3
ROBBERY,OF TRUCK DRIVER
CRIMINAL DISPOSAL FIREARM 1 &
CONTROLLED SUBSTANCE,SALE 2
LARCENY,GRAND BY OPEN BANK ACCT
BURGLARY,UNCLASSIFIED,UNKNOWN
FORGERY,PRESCRIPTION
SODOMY 2
GAMBLING 1,PROMOTING,BOOKMAKIN
AGGRAVATED CRIMINAL CONTEMPT
ROBBERY, CHAIN STORE
FALSE REPORT 1,FIRE
ROBBERY,PHARMACY
ROBBERY,LICENSED MEDALLION CAB
STOLEN PROPERTY-MOTOR VEH 2ND,
LARCENY,GRAND OF TRUCK
ROBBERY,LIQUOR STORE
LARCENY,GRAND FROM PERSON,LUSH WORKER(SLEEPING/UNCON VICTIM)
BRIBERY, POLICE OFFICER
ARSON 1
TRESPASS 1,CRIMINAL
ROBBERY,UNLICENSED FOR HIRE VEHICLE
CONTROLLED SUBSTANCE, SALE 4
ROBBERY,BICYCLE
OBSCENE MATERIAL - UNDER 17 YE
ROBBERY,BANK
ROBBERY,NECKCHAIN/JEWELRY
LARCENY,GRAND PERSON,NECK CHAI
ROBBERY,BODEGA/CONVENIENCE STORE
DRUG PARAPHERNALIA,   POSSESSE
CUSTODIAL INTERFERENCE 1
ESCAPE 2,1
PROMOTING A SEXUAL PERFORMANCE
BURGLARY,UNCLASSIFIED,DAY
ROBBERY,GAS STATION
MENACING 1ST DEGREE (VICT PEAC
USE OF A CHILD IN A SEXUAL PER

CONSPIRACY 2, 1
SEX TRAFFICKING
INCOMPETENT PERSON,KNOWINGLY ENDANGERING
TAX LAW
MANUFACTURE UNAUTHORIZED RECOR
MISCHIEF, CRIMINAL 3&2, BY FIR
ROBBERY,ON BUS/ OR BUS DRIVER
ROBBERY,ATM LOCATION
LARCENY,GRAND FROM TRUCK, UNATTENDED
OBSCENITY 1
CHILD ABANDONMENT
INTOXICATED DRIVING,ALCOHOL
HOMICIDE, NEGLIGENT, VEHICLE,
MAKING TERRORISTIC THREAT
BURGLARY,UNKNOWN TIME
KIDNAPPING 2
BAIL JUMPING 1 & 2
FACILITATION 3,2,1, CRIMINAL
SOLICITATION 3,2,1, CRIMINAL
END WELFARE VULNERABLE ELDERLY PERSON
AGGRAVATED SEXUAL ASBUSE
LARCENY,GRAND FROM PIER, UNATTENDED
ROBBERY,BAR/RESTAURANT
SODOMY 3
SUPP. ACT TERR 2ND
LARCENY, GRAND OF MOPED
LARCENY,GRAND FROM BOAT, UNATTENDED
SALE SCHOOL GROUNDS 4
KIDNAPPING 1
ROBBERY,CHECK CASHING BUSINESS

Misdemeanors

ASSAULT 3
LARCENY,PETIT FROM BUILDING,UN
FRAUD,UNCLASSIFIED-MISDEMEANOR
AGGRAVATED HARASSMENT 2
SEXUAL ABUSE 3,2
CRIMINAL MISCHIEF 4TH, GRAFFIT
SEXUAL MISCONDUCT,INTERCOURSE
CRIMINAL MISCHIEF,UNCLASSIFIED 4
MISCHIEF, CRIMINAL 4, BY FIRE
MISCHIEF, CRIMINAL 4, OF MOTOR
LARCENY,PETIT OF LICENSE PLATE
CHILD, ENDANGERING WELFARE
UNAUTHORIZED USE VEHICLE 3
VIOLATION OF ORDER OF PROTECTI
PUBLIC ADMINISTATION,UNCLASS M
LARCENY,PETIT BY CREDIT CARD U
CUSTODIAL INTERFERENCE 2
LARCENY,PETIT FROM OPEN AREAS,
NY STATE LAWS,UNCLASSIFIED MIS
LARCENY,PETIT FROM STORE-SHOPL

FORGERY,ETC.-MISD.
LARCENY,PETIT FROM AUTO
STOLEN PROPERTY 3,POSSESSION
LARCENY,PETIT BY FALSE PROMISE
CONTEMPT,CRIMINAL
LARCENY,PETIT BY CHECK USE
BRIBERY,COMMERCIAL
MENACING,UNCLASSIFIED
OBSTR BREATH/CIRCUL
ADM.CODE,UNCLASSIFIED MISDEMEA
LARCENY,PETIT OF VEHICLE ACCES
LEWDNESS,PUBLIC
CONTROLLED SUBSTANCE, POSSESSI
MARIJUANA, POSSESSION 4 & 5
WEAPONS, POSSESSION, ETC
INTOXICATED DRIVING,ALCOHOL
TRESPASS 2, CRIMINAL
THEFT,RELATED OFFENSES,UNCLASS
ACCOSTING,FRAUDULENT
MARIJUANA, SALE 4 & 5
LARCENY,PETIT OF MOTORCYCLE
LARCENY,PETIT OF BICYCLE
RECKLESS ENDANGERMENT 2
LEAVING SCENE-ACCIDENT-PERSONA
IMPERSONATION 2, PUBLIC SERVAN
RESISTING ARREST
TRAFFIC,UNCLASSIFIED MISDEMEAN
LARCENY,PETIT BY ACQUIRING LOS
TRESPASS 3, CRIMINAL
LARCENY,PETIT FROM TRUCK
IMPRISONMENT 2,UNLAWFUL
BURGLARS TOOLS,UNCLASSIFIED
THEFT OF SERVICES, UNCLASSIFIE
LARCENY,PETIT FROM BOAT
LARCENY,PETIT BY DISHONEST EMP
RECKLESS ENDANGERMENT OF PROPE
TAX LAW
UNAUTH. SALE OF TRANS. SERVICE
PETIT LARCENY-CHECK FROM MAILB
IMPAIRED DRIVING,DRUG
ASSEMBLY,UNLAWFUL
BAIL JUMPING 3
FALSE REPORT UNCLASSIFIED
RECORDS,FALSIFY-TAMPER
SEXUAL MISCONDUCT,DEVIATE
PROSTITUTION, PATRONIZING 4, 3
SALE OF UNAUTHORIZED RECORDING
DRUG PARAPHERNALIA,   POSSESSE
CHILD,ALCOHOL SALE TO
GAMBLING 2,PROMOTING,UNCLASSIF
CHECK,BAD
FALSE REPORT BOMB
LARCENY, PETIT OF AUTO - ATTEM
RECKLESS DRIVING

AGRICULTURE & MARKETS LAW,UNCL
TAMPERING 3,2, CRIMINAL
PROSTITUTION 4,PROMOTING&SECUR
GENERAL BUSINESS LAW,TICKET SP
LARCENY,PETIT OF BOAT
POSSESSION HYPODERMIC INSTRUME
ALCOHOLIC BEVERAGE CONTROL LAW
GAMBLING, DEVICE, POSSESSION
STOLEN PROP-MOTOR VEHICLE 3RD,
CHILD,OFFENSES AGAINST,UNCLASS
LARCENY,PETIT OF AUTO
PUBLIC SAFETY,UNCLASSIFIED MIS
LARCENY, PETIT OF MOPED
DOG STEALING
DIS. CON.,AGGRAVATED
RIOT 2/INCITING
MENACING,PEACE OFFICER
JOSTLING
PERJURY 3,ETC.
ESCAPE 3
PUBLIC HEALTH LAW,UNCLASSIFIED
COMPUTER UNAUTH. USE/TAMPER
FALSE ALARM FIRE
NUISANCE,CRIMINAL,UNCLASSIFIED
WOUNDS,REPORTING OF
LARCENY, PETIT FROM COIN MACHINE

Violations

HARASSMENT,SUBD 3,4,5
HARASSMENT,SUBD 1,CIVILIAN
MARIJUANA, POSSESSION
ALCOHOLIC BEVERAGES,PUBLIC CON
THEFT OF SERVICES- CABLE TV SE
POSSES OR CARRY A KNIFE
ADM.CODE,UNCLASSIFIED VIOLATIO
PEDDLING,UNLAWFUL
TRESPASS 4,CRIMINAL SUB 2
DISORDERLY CONDUCT
IMITATION PISTOL/AIR RIFLE
PARKR&R,UNCLASSIFIED VIOLATION
NY STATE LAWS,UNCLASSIFIED VIO
APPEARANCE TICKET FAIL TO RESP
IMITATION PISTOL/AIR RIFLE
TRAFFIC,UNCLASSIFIED INFRACTION
LOITERING,GAMBLING,OTHER
ENVIRONMENTAL CONTROL BOARD
INAPPROPIATE SHELTER DOG LEFT
EXPOSURE OF A PERSON
UNDER THE INFLUENCE OF DRUGS

Appendix C: Index of ZIP codes and New York city boroughs

ZIP	Borough
10001	Manhattan
10002	Manhattan
10003	Manhattan
10004	Manhattan
10005	Manhattan
10006	Manhattan
10007	Manhattan
10009	Manhattan
10010	Manhattan
10011	Manhattan
10012	Manhattan
10013	Manhattan
10014	Manhattan
10016	Manhattan
10017	Manhattan
10018	Manhattan
10019	Manhattan
10020	Manhattan
10021	Manhattan
10022	Manhattan
10023	Manhattan
10024	Manhattan
10025	Manhattan
10026	Manhattan
10027	Manhattan
10028	Manhattan
10029	Manhattan
10030	Manhattan
10031	Manhattan
10032	Manhattan
10033	Manhattan
10034	Manhattan
10035	Manhattan
10036	Manhattan
10037	Manhattan
10038	Manhattan
10039	Manhattan
10040	Manhattan
10041	Manhattan
10044	Manhattan
10045	Manhattan
10048	Manhattan
10065	Manhattan
10069	Manhattan
10075	Manhattan
10103	Manhattan
10107	Manhattan
10111	Manhattan
10112	Manhattan
10118	Manhattan
10119	Manhattan
10121	Manhattan
10128	Manhattan
10129	Manhattan
10162	Manhattan
10163	Manhattan
10167	Manhattan
10170	Manhattan
10172	Manhattan
10178	Manhattan
10203	Manhattan
10259	Manhattan
10278	Manhattan
10280	Manhattan
10281	Manhattan
10282	Manhattan
10301	Staten Island
10302	Staten Island
10303	Staten Island
10304	Staten Island
10305	Staten Island
10306	Staten Island
10307	Staten Island
10308	Staten Island
10309	Staten Island
10310	Staten Island
10312	Staten Island
10314	Staten Island
10451	Bronx
10452	Bronx
10453	Bronx
10454	Bronx
10455	Bronx
10456	Bronx
10457	Bronx
10458	Bronx
10459	Bronx
10460	Bronx
10461	Bronx
10462	Bronx
10463	Bronx
10464	Bronx
10465	Bronx
10466	Bronx
10467	Bronx
10468	Bronx
10469	Bronx
10470	Bronx
10471	Bronx
10472	Bronx
10473	Bronx
10474	Bronx
10475	Bronx
11004	Queens
11101	Queens
11102	Queens
11103	Queens
11104	Queens
11105	Queens
11106	Queens
11109	Queens
11201	Brooklyn
11202	Brooklyn
11203	Brooklyn
11204	Brooklyn
11205	Brooklyn
11206	Brooklyn
11207	Brooklyn
11208	Brooklyn
11209	Brooklyn
11210	Brooklyn
11211	Brooklyn
11212	Brooklyn
11213	Brooklyn
11214	Brooklyn
11215	Brooklyn
11216	Brooklyn
11217	Brooklyn
11218	Brooklyn
11219	Brooklyn
11220	Brooklyn
11221	Brooklyn
11222	Brooklyn
11223	Brooklyn
11224	Brooklyn
11225	Brooklyn
11226	Brooklyn
11228	Brooklyn
11229	Brooklyn
11230	Brooklyn
11231	Brooklyn
11232	Brooklyn
11233	Brooklyn
11234	Brooklyn
11235	Brooklyn
11236	Brooklyn
11237	Brooklyn
11238	Brooklyn
11239	Brooklyn
11249	Brooklyn
11251	Brooklyn
11354	Queens
11355	Queens
11356	Queens
11357	Queens
11358	Queens
11359	Queens
11360	Queens
11361	Queens
11362	Queens
11363	Queens
11364	Queens
11365	Queens
11366	Queens
11367	Queens
11368	Queens
11369	Queens
11370	Queens
11371	Queens
11372	Queens
11373	Queens
11374	Queens
11375	Queens
11377	Queens
11378	Queens
11379	Queens
11385	Queens
11411	Queens
11412	Queens
11413	Queens
11414	Queens
11415	Queens
11416	Queens
11417	Queens
11418	Queens
11419	Queens
11420	Queens
11421	Queens
11422	Queens
11423	Queens
11426	Queens
11427	Queens
11428	Queens
11429	Queens
11430	Queens
11432	Queens
11433	Queens
11434	Queens
11435	Queens
11436	Queens
11451	Queens
11691	Queens
11692	Queens
11693	Queens
11694	Queens
11695	Queens
11697	Queens
99999	bogus ZIP