2014 Trends in NYC Marijuana Arrests

A data-driven story by
Lylla Younes, Christopher Traver, and Christopher Yoo

Introduction

For this project, we were interested in looking at trends in marijuana related arrests. In order to solve for confounding variables, we decided to narrow down our two our datasets by focusing on a single location and year - New York City in 2014. In doing so, we were able to not only track the price of marijuana and the number of related arrests but also explore other possible trends related to marijuana, such as the number of arrests per month, the location of arrests within the the five boroughs of New York City, and the demographics of those that were arrested. We detail our approach and findings below.

Data Description

We utilized two data sets for this project: NYC Law Enforcement data from 2014 and the price of marijuana in NYC throughout 2014. Details of both of these datasets are outlined below.

The NYC Law Enforcement Data from 2014

Data source and accuracy. We were able to download this dataset directly from the nyc.gov website (link). Since this data set is directly compiled by the NYC government and the law enforcement officers involved in the arrests, we believe that this data is extremely reliable.

Criteria for selection. The dataset is extremely nuanced and provides lots of detailed information on different characteristics of each of the arrests, including things like the arrest location and different characteristics of person(s) arrested, among other things. With such a detailed dataset like this, we were forced to choose variables that we believed would tell the best story. We were specifically interested in variables that could tell us more about the demographics of those arrested and the location of the arrests. Therefore, we utilized the documentation of the dataset (link) to use the crime code description (detailcm) to filter the data to only include marijuana related arrests (codes 27 & 29). Using the filtered marijuana related arrests data, we then focused on the date (datestop), sex (sex), race (race), age (age), and address of the arrest (addrnum & stname). We then cut out unnecessary variables such as hair color (haircolor), and binary variables such as stopped and frisked (yes/no) because they strayed from the goals we had for the story we wanted to tell for this project.

Pertinent variables. Ultimately, we focused on the sheer number of arrests in 2014, the location of the arrests, and demographic attributes related to those that were arrested, such as race and gender. Each of these are detailed in some manner throughout our visualizations. As mentioned above, as a result we focused on the following variable names in the NYPD arrest dataset: crime code description (detailcm) to filter the data to only include marijuana related arrests (codes 27 & 29), date of the arrest (datestop), sex of the offender (sex), race of the offender (race), age of the offender (age), and the address of the arrest (addrnum & stname).

Parsing/Formatting Data. The original NYPD arrest dataset came with over 20,000 data points, each of which contained nearly 100 variables. The sheer size of this dataset not only made it difficult for the browser to handle when rendering visualizations, but it also made it difficult for us as a team to comprehend in order to develop the best visualizations. Therefore, we had to filter out most of the unnecessary variables such as eye color and decided to focus only on the location and date of the arrests. This was simple to do in Excel, and we deleted the corresponding columns for variables that we were not interested in. The dataset also came with an address where the arrest took place. This was indispensable in plotting the locations of the arrests. However, in order to plot the data on the map that we would be using in our visualization, the plaintext street addresses in the dataset had to be parsed into latitude and longitude coordinates. We accomplished this conversion by utilizing a Google Maps API batch geocoding tool (link) to input street coordinates and get the associated latitude and longitude coordinates of that address. Those coordinates were then manually inputted into our CSV dataset and used to plot each arrest on our projected map. In addition, we also had to re-format the date field provided by the original dataset into a format that d3 would understand. Dates in the original dataset were formatted using only numbers (ex. mmddyyy), however we needed the dates to be in the format mm/dd/yy. We utilized Excel transform the date into the format “d-m”.

Marijuana Price Dataset

Data source and accuracy. We downloaded this dataset from www.priceofweed.com, which is an open-source dataset that tracks the price fluctuations of marijuana globally. This website allows users to anonymously log marijuana purchases, inputting how much they bought, of what quality, at what price, and in what location. Although the title and URL of the website seems questionable at first glance, under further examination we found that the data is regularly updated by a large number of users from around the world. Therefore, we concluded that because this was the case it was most likely the most accurate dataset to use of all of the available options currently.

Criteria for selection. This price dataset was simpler than the aforementioned NYPD dataset. The original dataset came with date information, the price for that date, and location. We were interested only in the price of marijuana in New York throughout 2014, therefore we filtered the dataset accordingly so that it listed only the street price of marijuana for each day in 2014

in NYC. One slight complication was deciding which quality of marijuana we were interested in as the original dataset included fields for 3 different types of marijuana, ranging from LowQ (low quality) to MedQ (medium quality) to HighQ (high quality). For our visualizations, we decided to focus on the medium-quality Marijuana (MedQ).

Pertinent variables. As mentioned above, we extracted the street price of medium-quality (MedQ) marijuana throughout 2014 in New York City. The original dataset contained data for every day of the year (not just a monthly average), and we utilized this data to compare with the NYPD dataset of arrests.

Parsing/Formatting Data. We filtered the data by date to only include data points for 2014. We then further filtered the data to only include data from NYC. The resulting filtered dataset yielded information on the street price of MedQ marijuana for each day in 2014. We then standardized the date format into mm/dd/yy order to integrate it with the NYPD dataset.

Dataset Integration

We integrated the two datasets by narrowing them down in the same way and utilizing a common field between the two. To do this, we utilized the common date field in each of the datasets to link the price of marijuana on a certain day to each of the arrests in the NYPD dataset. In order to link the two together programmatically, we utilized the VLOOKUP function in Microsoft Excel to search for a common field in each of the datasets and map the two together into a single worksheet. The result left us with a CSV file that contained all of the arrest data from the NYPD and the corresponding street price of marijuana for the days of each of the arrests.

Additional Data

In order to develop our visualization for the New York City map, we relied on a JSON shape file of coordinates for each of the New York City boroughs created by Phil Pedruco (source link) in order to build our map since the world.json file that we had been using in class did not contain enough detail to display the New York City boroughs in enough detail.

Mapping Data to Visual Elements Description

Our data-driven story includes a number of visualizations, each of which has different corresponding mappings of data to visualization elements. We outline the details of each below.

NYC Arrests Map. The map of arrests throughout New York City is one of our main visualizations, and we hope it drives the story that we are trying to tell. Each dot on the map

represents a marijuana-related arrest in 2014. The color of the dot represents the price of marijuana at the time of arrest. The darkest green color expresses the highest price in the dataset, and the lightest green expresses the lowest price. A scale on the side of the map shows the corresponding colors for prices using a color gradient. We feel that this map helps tells the viewer several things clearly. Firstly, it illustrated the locations in which marijuana-related arrests are prevalent. Secondly, it shows that types of marijuana (expensive or cheap) are most common in a given burough of New York City. What this visualization does not graph is the number of arrests per month. The time aspect is excluded from this graph, however we do display data on this topic in other visualizations of the project.

Demographic Data. These visualizations focus on conveying a better understanding of the demographics of people that are often arrested on marijuana related charges. Our visualizations for this second include a breakdown of male vs. female arrests, as well as an additional breakdown of ages and race in the form of histograms. The gender breakdown includes symbolic icons of male and female (as seen on bathroom doors) and the appropriate percentage breakdown of arrests. The race histogram uses an ordinal scale for the x-axis to group by race, and a linear scale for the y-axis to display the frequency of arrests for each race group. Finally, the age graph uses a linear scale to increment the radiuses of the circles, with with each circle representing a group of ages and the size of the circle representing the frequency of arrests in 2014 for each age group.

Price over time. This area graph exists to depict the change over marijuana street prices for one ounce over the course of 2014 within all of New York City. An area graph was chosen over a scatterplot and line graph because we felt that these would not show the decreasing trend in price as well as an area graph would. The graph uses a time related scale for the x-axis by first parsing the data and converting it into d3's time format and a linear scale for the y-axis to display the price of marijuana for each date.

Arrests per month. This line graph illustrates the general trends of marijuana-related arrests per month over the course of 2014 within all of New York City. The line graph was chosen to see the various peaks for each month. Although a per day interval was considered, the graph quickly became cluttered and confusing, so this scale of measurement was avoided. The graph uses a time related scale for the x-axis by parsing the dates and converting them to the month integers (i.e. November → 11) and a linear scale for the y-axis to display the number of arrests per month.

Story

The purpose of this project is to provide visualizations about trends in marijuana related arrests in New York City during 2014. We chose to focus on 2014 specifically for our visualizations as this year yielded the most accurate data and we decided that there would be too much data if we incorporated multiple years. We hope that our visualizations shed light

on interesting trends about marijuana in New York City. Specifically, the map communicates the following information: (1) the areas where marijuana activity is most dense in New York City, (2) how the price of marijuana impacts where it is commonly used and sold throughout the City, and (3) whether the price influences the density of arrests in a location. In addition, the demographic data that we provide about the arrests is intended to shed light on the types of people that are often being arrested for marijuana related offenses. The price of marijuana over time graph shows how the street price of marijuana has been steadily decreasing throughout 2014. We use this graph in conjunction with the graph showing marijuana related arrests over time in order to show if a correlation exists between price and arrests. Overall, there were several very interesting and surprising findings that came to light after looking at our visualizations.

It seems marijuana arrests occur in several specific clusters throughout the five New York City boroughs. We see this within the map visualization where various places have a significant number of stacked data points. With this data, we can consider underlying causes of these marijuana arrests. For example, the density of arrests could be related to population. Notice that the northern area of Staten Island is the only location in the borough with dense number of arrests. This is because this area is marked by the cities West New Brighton and St. George, which are the most populated areas of the borough. The density of arrests could also deal with safety and demographics of a neighborhood. For example, high marijuana arrests are seen in Northern Manhattan (Harlem) and the South Bronx, which are known for higher minority populations and lower wealth standards. Despite these hypotheses, it is difficult to pinpoint the reasons for certain marijuana prices in various areas; for instance, since Manhattan is particularly known for higher standards of living, one would expect it to have consistently higher prices of marijuana; however, Queens has the most uniformly high prices of the five boroughs.

It was also surprising to see how little the price of marijuana fluctuated over the course of an entire year. Specifically, we can see that the price of marijuana followed a steady decreasing trend over the twelve months. We also see the same trend for arrests over time. However, although there certainly is a correlation, we cannot necessarily extrapolate to conclude that a decreasing trend of prices *causes* a decreasing trend in arrests. For example, it is possible that arrests in November and December are lower simply because of holiday season. It is also possible that prices later in the year (post-May) are lower simply because the prime marijuana growing season is in late March, thus increasing supply and driving price down when it becomes available.

Finally, it is important to note the skew in race and age demographics. Certainly, the skew of marijuana arrests towards people <24 years of age exists because drug usage in general appeals more to this age group. In relation to the race demographics, marijuana arrests are likely higher among minority populations (Hispanic and Black) because general arrests are higher among minorities as well. However, another consideration is that marijuana sales and,

thus, arrests are more likely to occur in underdeveloped neighborhoods, which are more likely to be populated by minorities.