Project Title: Crime Analysis with a Focus on Crime-Related Trends in Brooklyn and Weather-Related Anomaly Detection

Team Name: Big Data CSI

Authors: Julie Helmers (jch609), Zeynep Doganata (zd507), Kanishk Dugar (kd1783)

Abstract: We have chosen to use the NYPD crime incidents dataset. For Part 1 of the project, we explored the data using PySpark and Matplotlib. We found that the dataset is relatively clean overall but does have some null and/or invalid values. We also generated a number of visualizations for different ways of "slicing" the data, to help us spot useful trends. For Part 2 of the project, we incorporated three additional datasets – housing prices, restaurant permits, and weather data – and mined relationships between the features from those datasets and crime rates.

Introduction: Our primary dataset is the set of felony, misdemeanor, and violation crimes reported to the New York City Police Department (NYPD) from 2006 to 2016. The dataset is available for download at https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i. The downloadable file (NYPD_Complaint_Data_Historic.csv) is 1.3 GB and contains 5,101,232 lines. Its sheer size motivated our use of big data tools for this project, including PySpark.

Part 1: Data Summary and Data Quality Issues

The historic crime dataset contains 23 columns, which are detailed in the table below. This table is provided by the NYPD and/or NYC OpenData and is available for download (NYPD_Incident_Level_Data_Column_Descriptions.csv) at the same link as the complete dataset.

Table 1: Column Descriptions from NYC OpenData

Column	Description
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
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)
CRM_ATPT_CPTD_CD	Indicator of whether crime was successfully completed or attempted, but failed or was interrupted prematurely
LAW_CAT_CD	Level of offense: felony, misdemeanor, violation
JURIS_DESC	Jurisdiction responsible for incident. Either internal, like Police, Transit, and Housing; or external, like 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, front of, rear of
PREM_TYP_DESC	Specific description of premises; grocery store, residence, street, etc.
PARKS_NM	Name of NYC park, playground or green space 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, units feet (FIPS 3104)

Y_COORD_CD	Y-coordinate for New York State Plane Coordinate System, Long Island Zone, NAD 83, units 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)

To begin our data exploration, we wrote and ran PySpark scripts for each category of offense: felony, misdemeanor, and violation. Each script takes a different "slice" of the data, for example, the counts of crime incidents by borough, by year, by both borough and year, etc. A complete list of the "slices" we took for each type of offense can be found in the tables below, along with where the script can be found in our GitHub.

Table 2: Summary of PySpark Data Exploration Scripts for Part 1

Script Description	**Felony**	**Misdemeanor**	**Violation**
Count by borough	felonies/source/felony_count_by_borough.py	misdemeanors/source/misdemeanor_count_by_borough.py	violations/source/violation_count_by_borough.py
Count by year	felonies/source/felony_count_by_year.py	misdemeanors/source/misdemeanor_count_by_year.py	violations/source/violation_count_by_year.py
Count by borough and year	felonies/source/felony_count_by_borough_year.py	misdemeanors/source/misdemeanor_count_by_borough_year.py	violations/source/violation_count_by_borough_year.py
Count by preposition for location ("front of", "inside", etc.)	felonies/source/felony_count_by_preposition.py	misdemeanors/source/misdemeanor_count_by_preposition.py	violations/source/violation_count_by_preposition.py
Count by premise and preposition	felonies/source/felony_count_by_loc_and_prep.py	misdemeanors/source/misdemeanor_count_by_loc_and_prep.py	violations/source/violation_count_by_loc_and_prep.py
Count by month	felonies/source/felony_count_by_month.py	misdemeanors/source/misdemeanor_count_by_month.py	violations/source/violation_count_by_month.py
Count by month and	felonies/source/felon	misdemeanors/source	violations/source/viol

year	y_count_by_month_year.py	/misdemeanor_count_by_month_year.py	ation_count_by_month_year.py
Count by season (by binning months)	felonies/source/felony_count_by_season.py	misdemeanors/source/misdemeanor_count_by_season.py	violations/source/violation_count_by_season.py
Count by season and preposition	felonies/source/felony_count_by_prep_season.py	misdemeanors/source/misdemeanor_count_by_prep_season.py	violations/source/violation_count_by_prep_season.py
Count by season and year	felonies/source/felony_count_by_season_year.py	misdemeanors/source/misdemeanor_count_by_season_year.py	violations/source/violation_count_by_season_year.py
Count by hour	felonies/source/felony_count_by_hour.py	misdemeanors/source/misdemeanor_count_by_hour.py	violations/source/violation_count_by_hour.py
Count by time of day (by binning hours)	felonies/source/felony_count_by_tod.py	misdemeanors/source/misdemeanor_count_by_tod.py	violations/source/violation_count_by_tod.py
Count by description	felonies/source/felony_count_by_desc_lvl1.py	misdemeanors/source/misdemeanor_count_by_desc_lvl1.py	violations/source/violation_count_by_desc_lvl1.py
Count by description (more granular)	felonies/source/felony_count_by_desc_lvl2.py	misdemeanors/source/misdemeanor_count_by_desc_lvl2.py	violations/source/violation_count_by_desc_lvl2.py
Count by description and year	felonies/source/felony_count_by_desc_lvl1_year.py	misdemeanors/source/misdemeanor_count_by_desc_lvl1_year.py	violations/source/violation_count_by_desc_lvl1_year.py

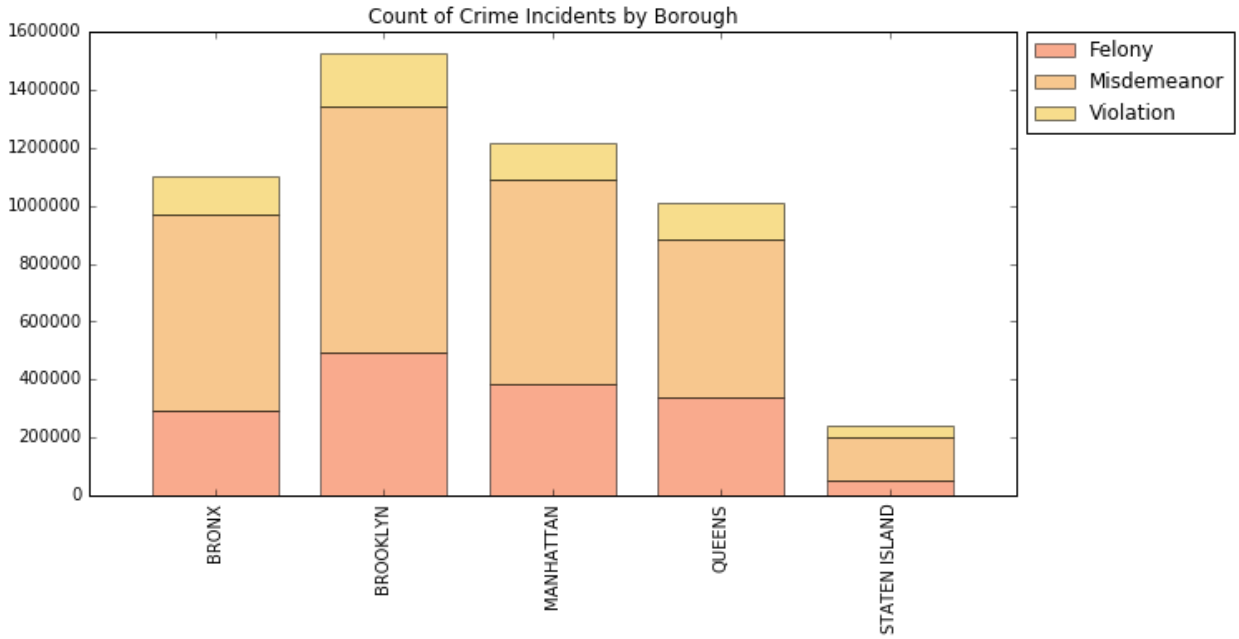
We then used the output of these scripts to visualize several “slices” of the data to help us identify any patterns and/or anomalies. The visualizations were created using Jupyter Notebook, Pandas, and Matplotlib and can be re-created by running the notebook visualizations/initial_visualizations.ipynb in our GitHub. For this initial stage of data exploration, we were most interested in patterns in crime incidents over time and by location, so we focused on visualizing the “slices” related to borough, year, hour, and season. We were also interested in how the different types of crime (felony, misdemeanor, and violation) might vary differently across time and space, so we visualized the counts of each type of crime separately

instead of visualizing the total count of all crimes. The offense types are color-coded by severity, from yellow (violation, least serious) to red (felony, most serious).

The visualizations by borough reveal that the crime rates in New York are very uneven across boroughs. They do seem to vary roughly with population: Staten Island is much less populated than all the other boroughs and has many fewer crime incidents from 2006-2016; Brooklyn is the most populated borough and also has the highest number of crime incidents for all three offense types. However, Queens is the next most populated borough, while Manhattan has the next highest number of crimes.

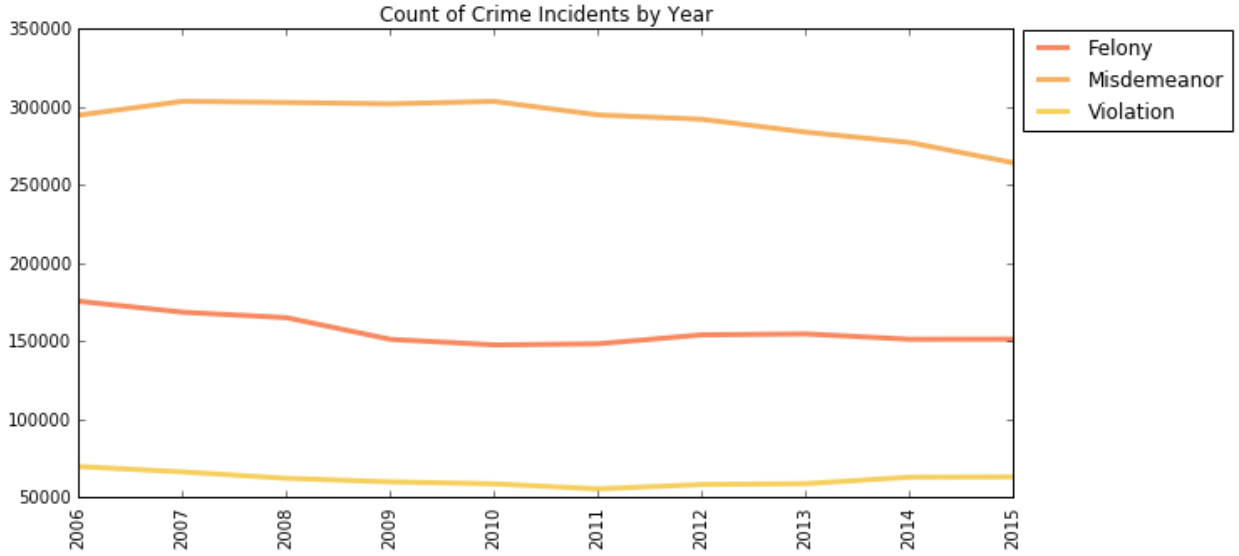
One aspect of the data that is relatively consistent across boroughs is the breakdown of the number of crimes into felonies, misdemeanors, and violations. There is not a borough that appears to have a disproportionate number of felonies, for example.

Figure 1: Count by Borough

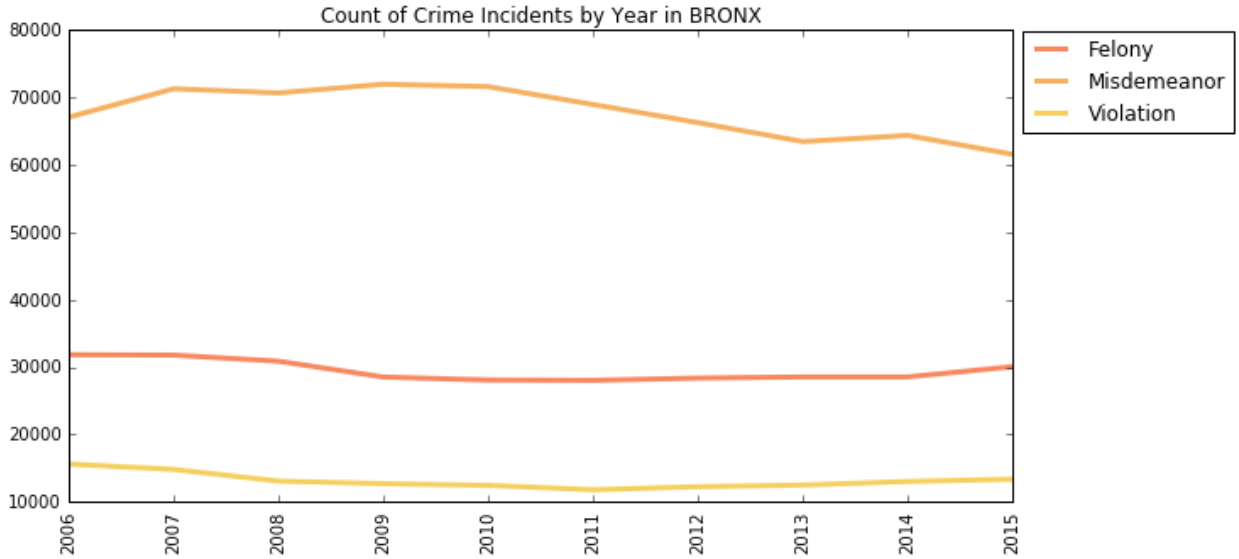


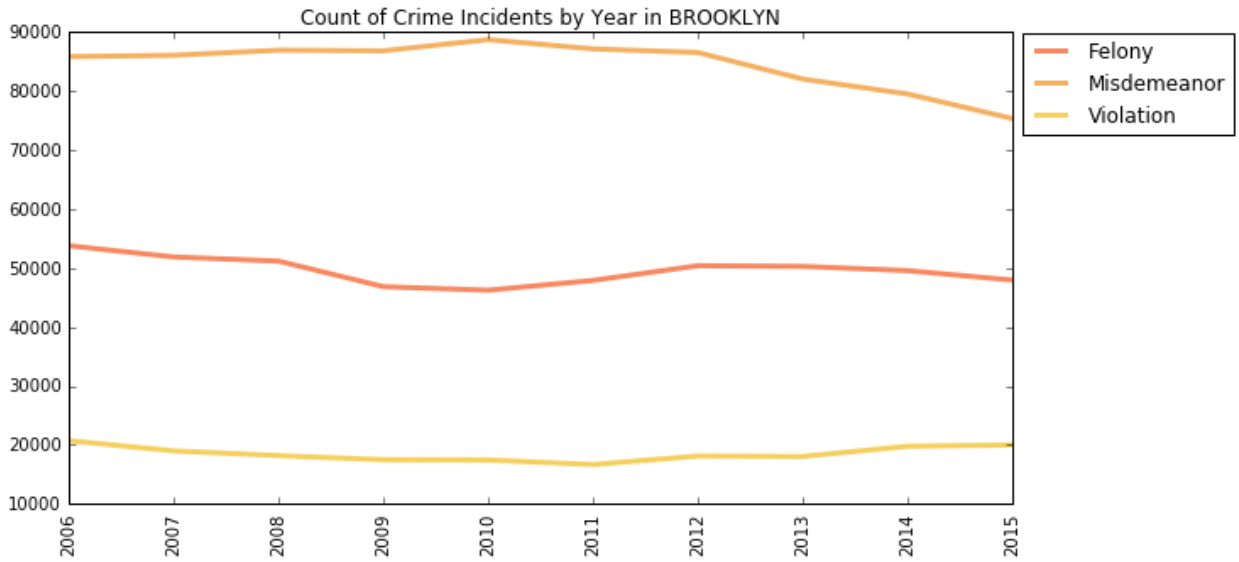
The patterns across the years from 2006-2016 also seem to be relatively consistent across boroughs, with all boroughs showing a slight downward trend in numbers of crimes, especially for misdemeanors, but the shapes of the curves are different across boroughs.

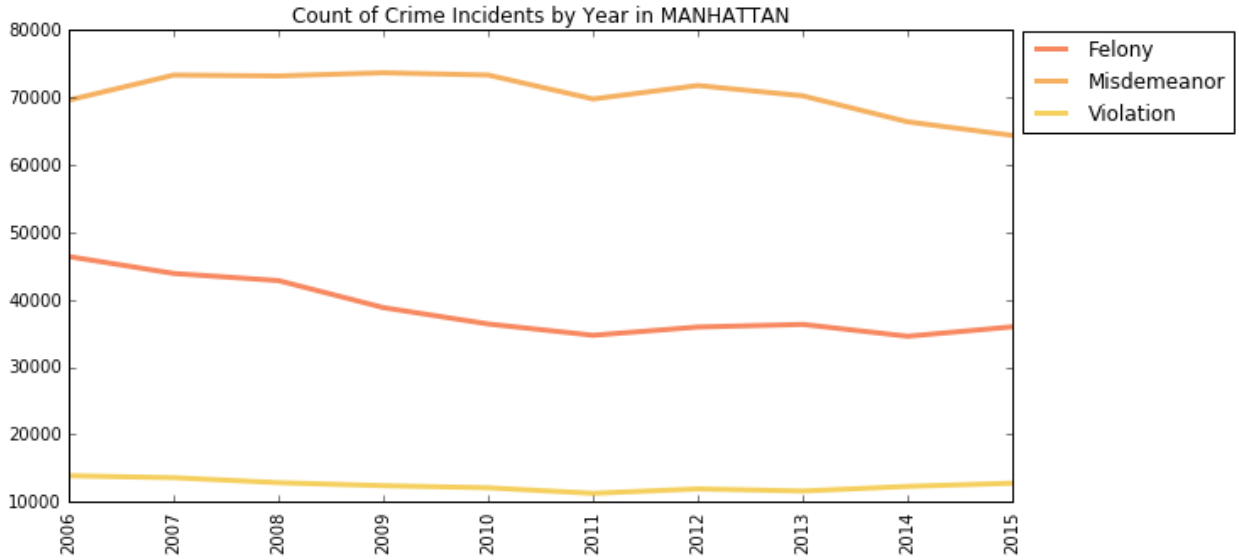
Figure 2: Count by Year

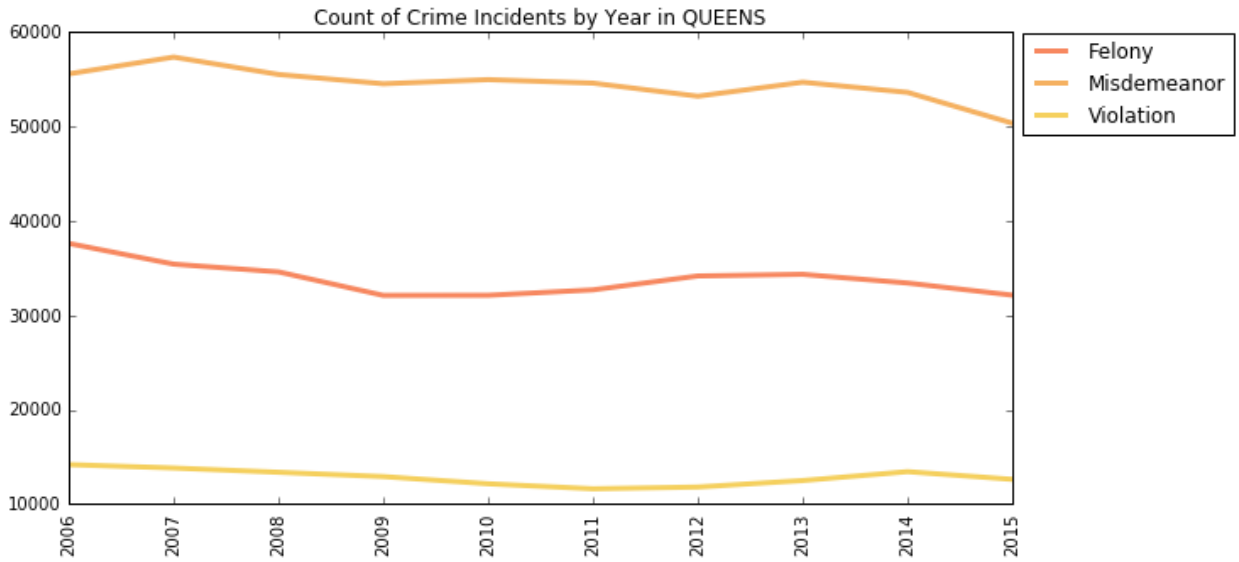


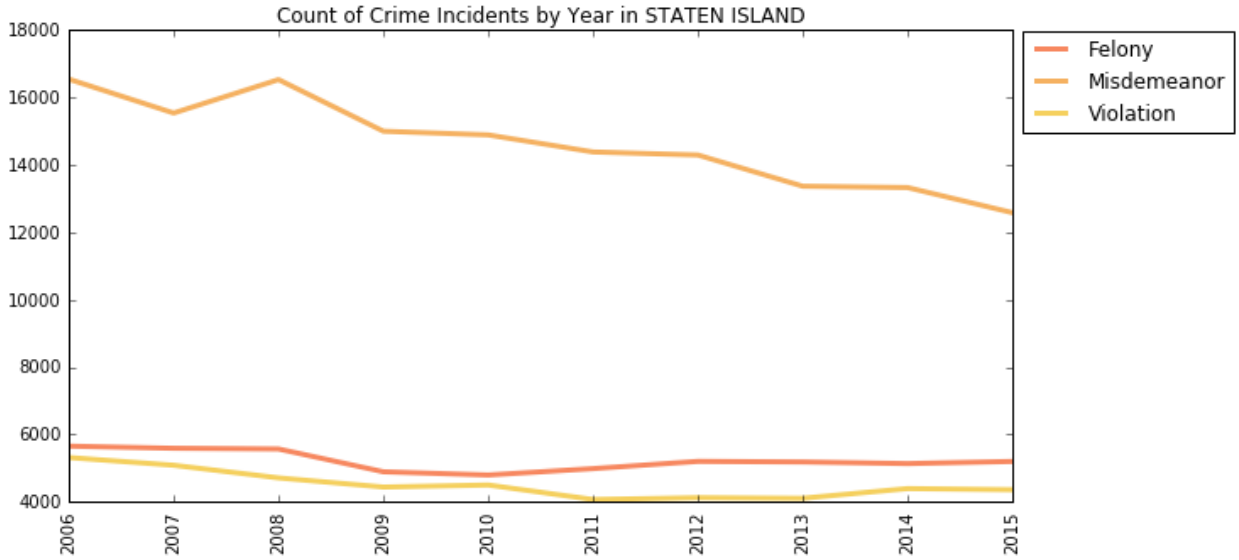
Figures 3-7: Count by Borough and Year




Count of Crime Incidents by Year in BROOKLYN
90000
80000
70000
60000
50000
40000
30000
20000
10000
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
Felony
Misdemeanor
Violation

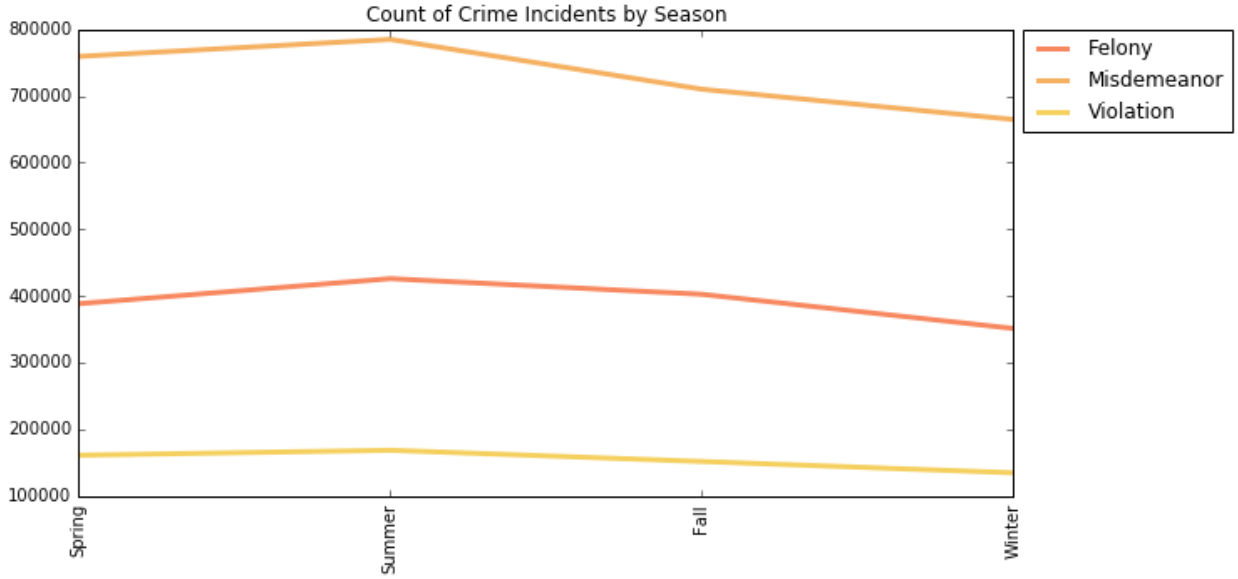

Count of Crime Incidents by Year in MANHATTAN
80000
70000
60000
50000
40000
30000
20000
10000
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
Felony
Misdemeanor
Violation





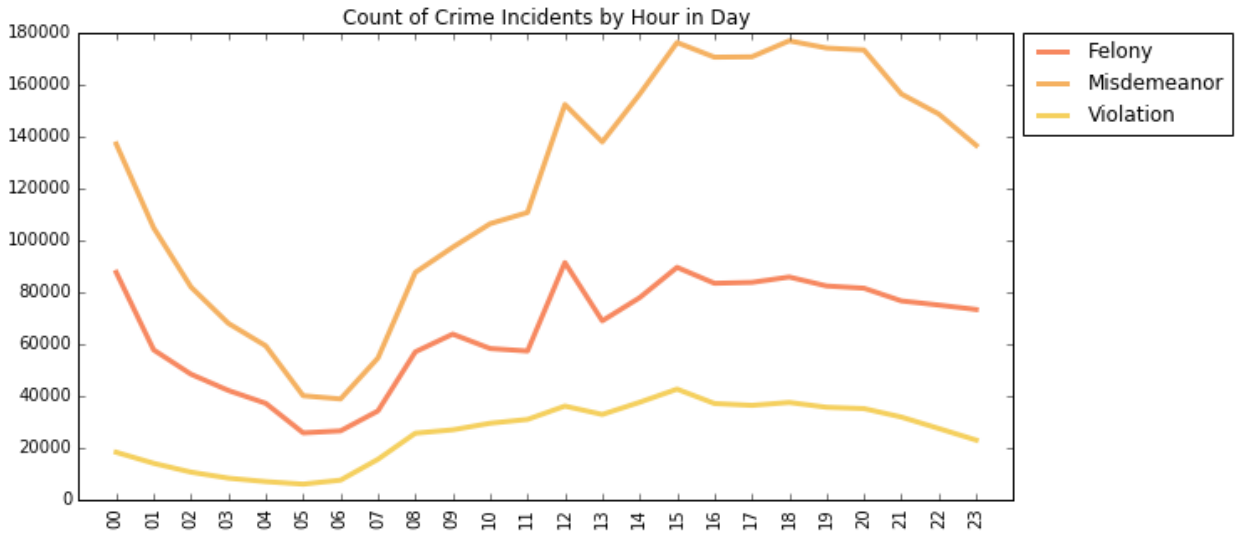
Before we looked at the data, we had predicted that we would find more crimes during the summer months than in the winter months, because there are more people out and about the city − and often on vacation from work or school − during the warmer parts of the year. In several of our "slices" of the dataset, we binned consecutive months to examine crime rates by season. The resulting visualization supports our hypothesis about summer, which we will test more rigorously moving forward.

Figure 8: Count by Season



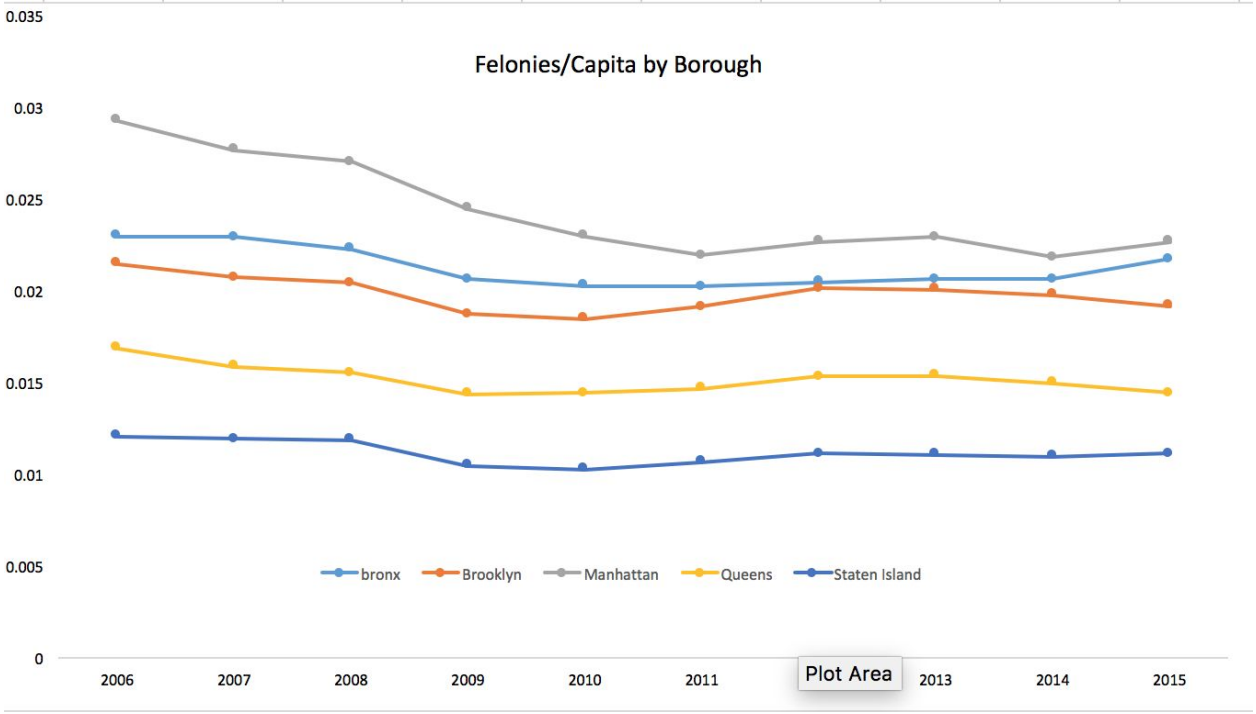
One of the most striking visualizations we have made so far is the chart of crime rates versus hour in the day. We had expected there to be more crimes at night than during broad daylight. The visualization suggests that this may be partially true; the numbers of misdemeanors and violations are highest from 6pm until midnight. However, the number of felonies is slightly higher between noon and 6pm, and all three offense types show a surprising spike in count around noon as well as a steep drop-off during the early morning hours. Perhaps the time of noon was entered in the dataset for any and all incidents that were missing reliable time information.

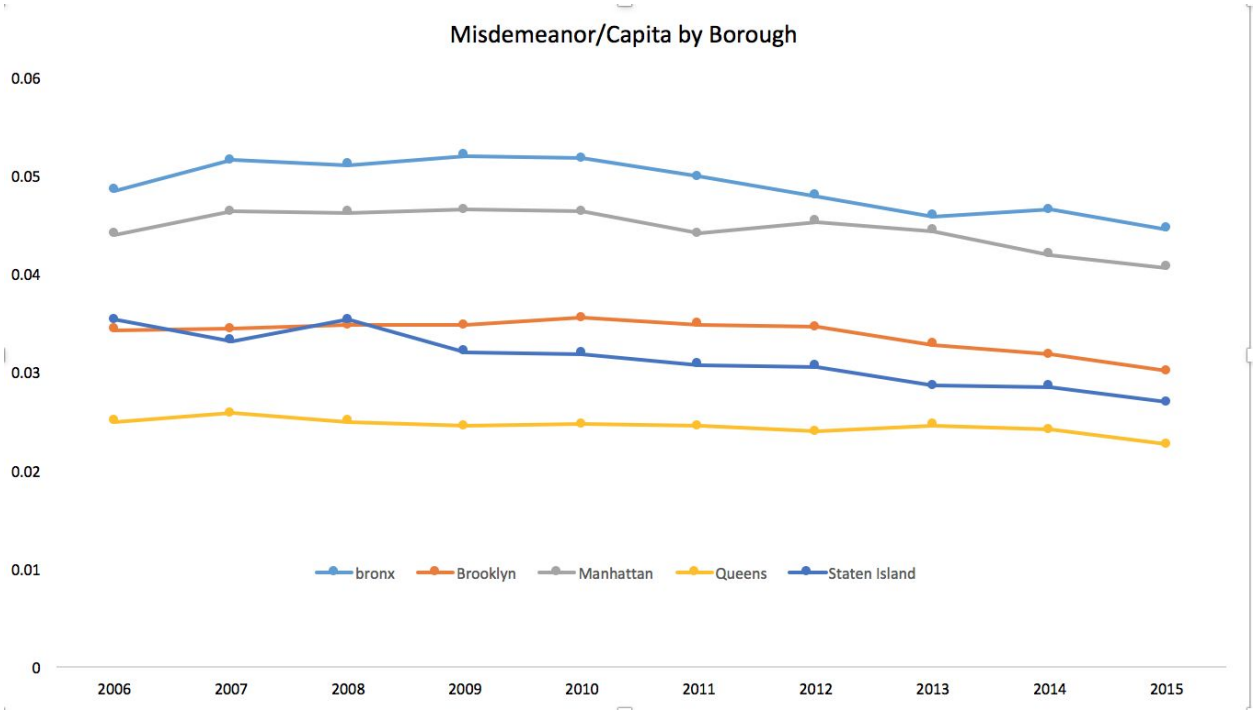
Figure 9: Count by Hour

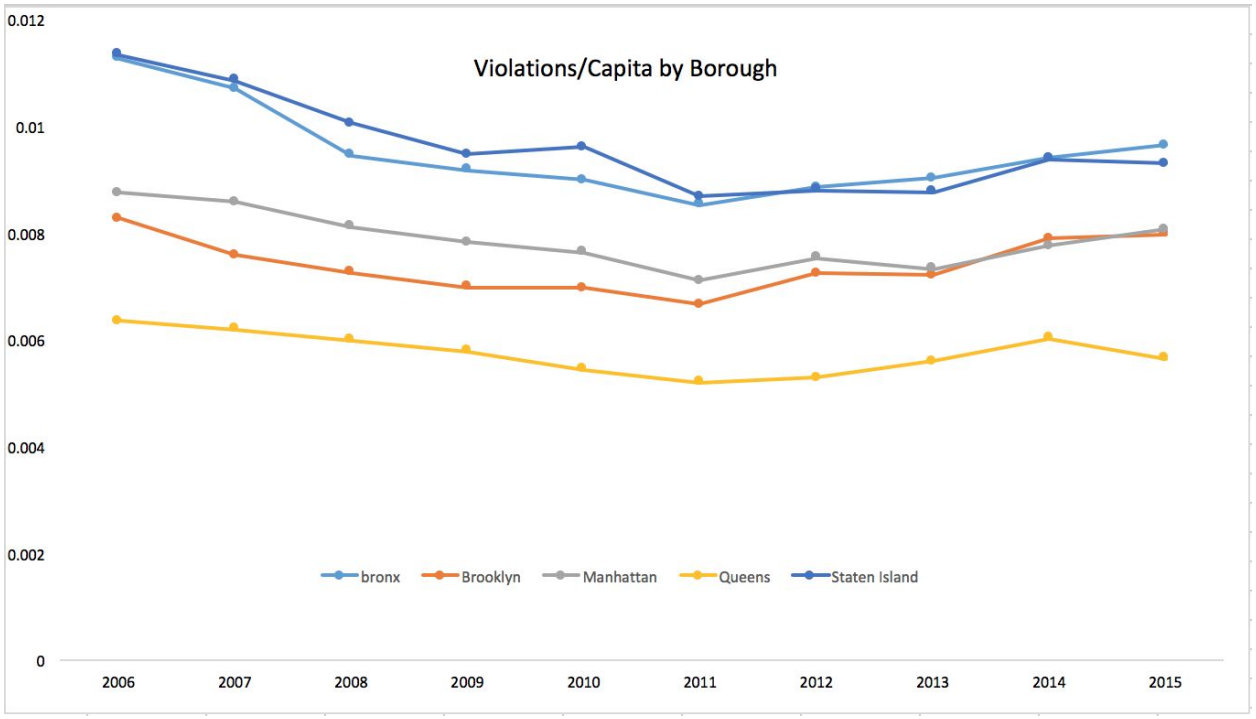


Given the varying populations of the 5 boroughs, we also calculated the complaint type by capita for each borough to give us more perspective using the census data recorded in 2010.

Figures 9-11: Count per Capita by Borough







Finally, as requested, we wrote a PySpark script (crimedata.py) that takes each of the columns in the dataset and outputs a base type, a semantic data type, and a label of "NULL," "VALID," or "INVALID" for each value. The following column-specific observations were made based on the output of the script:

Complaint from date

Invalid	7	Incorrect entry of year as 1015, possibly instead of 2015
Null	655	No data entry

Complaint from time

Invalid	903	Incorrect entry of hour as 24:00:00 possibly instead of 00:00:00
Null	48	No data entry

Complaint to date

Invalid	35	Last occurrence date falling before the first occurrence
Null	1391478	Not necessarily a data quality concern as 'to date' is optional

Complaint to time

Invalid	1376	Incorrect entry of hour as 24:00:00 possibly instead of 00:00:00
Null	1387785	Optional, not a data quality concern

Complaint report date

Invalid	2	Report date before the first occurrence of crime

Offense Code

No null or invalid entry

Offense Description

Null	18840	No data entry

PD Internal Code

Null	4574	No data entry

PD Internal Description

Null	4574	No data entry

Crime Completed or Attempted

Null	7	No data entry

Offense Level

No null or invalid entry

Jurisdiction
No null or invalid entry

Borough
Null 463 No data entry

Precinct
Null 390 No data entry

Specific Location
Null 1127341 No data entry, specific location not available

Premises
Null 33279 No data entry, specific premise not available

Park
Null 5093632 Optional entry, not a data quality concern

Housing Development Authority
Null 4848026 No data entry

X-coordinate for New York State Plane Coordinate System
Null 188146 No data entry

Y-coordinate for New York State Plane Coordinate System
Null 188146 No data entry

Latitude
Null 188146 No data entry

Longitude
Null 188146 No data entry

Most of the invalid entries can be corrected by simple adjustment; for example, the year entry 1015 in the 'Complaint from date' is probably 2015. This can be confirmed by checking the corresponding ‘Complaint to date’ and ‘Report Date’. Also, the time 24:00:00 can be replaced by 00:00:00. Apart from the date and time, there are no invalid entries in the whole database. The overall quality of data is quite reliable. There are many blanks in the database. However, some of

those entries like “Parks,” “Complaint to time,” and “Complaint to date” are situation-specific and optional and therefore are not quality concerns.

Part 2: Crime Analysis with a Focus on Crime-Related Trends in Brooklyn and Weather-Related Anomaly Detection

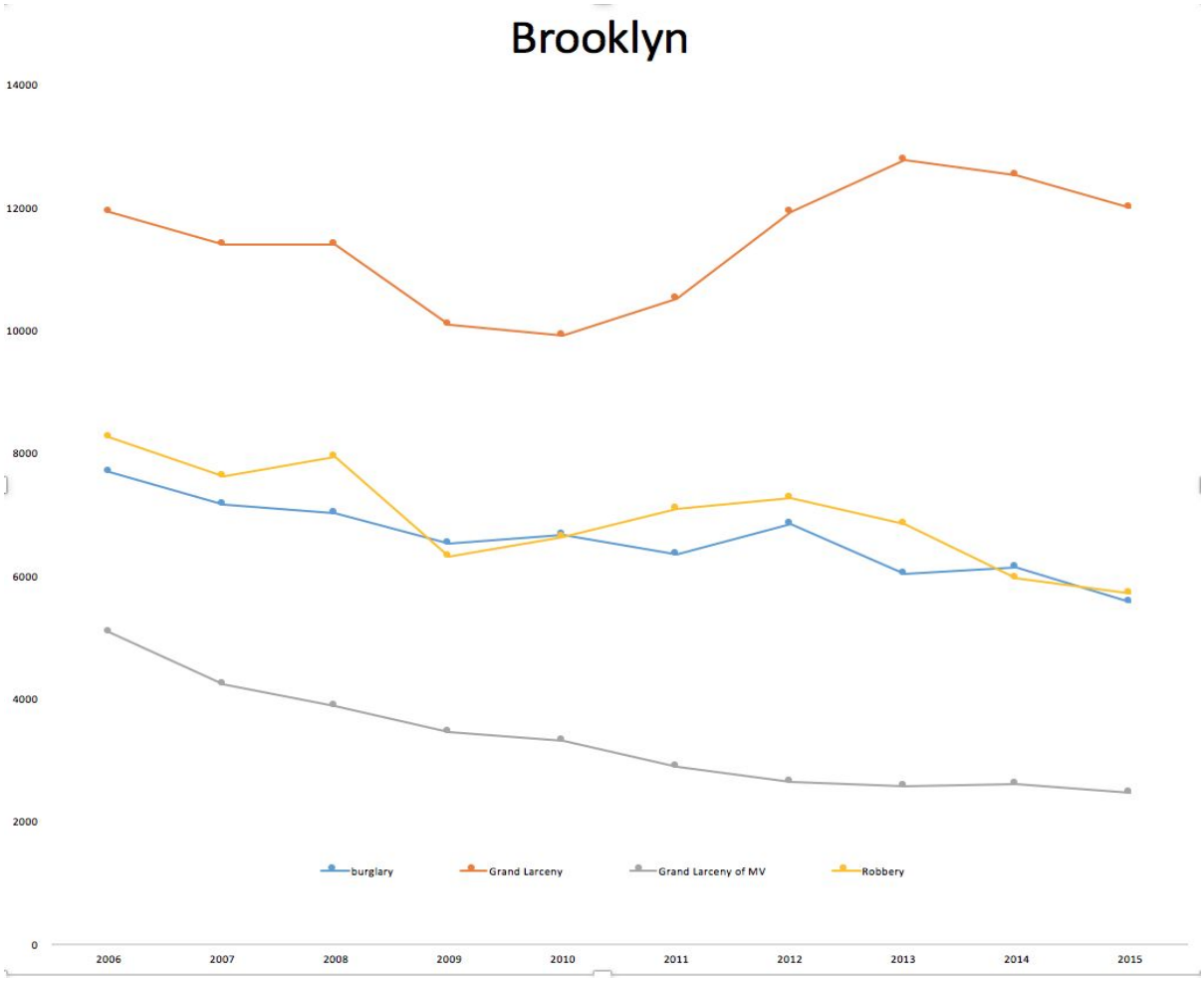
For Part 2 of this project, we investigated three hypotheses that we had formed from exploring and thinking about the data: (1) that crime rates in a given area should decrease as housing prices increase, (2) that crime rates in a given area should decrease as the number of restaurants increases, and (3) that crime rates overall should vary with weather, specifically that they should increase with higher temperatures and decrease with higher precipitation and lower visibility. We analyzed an external dataset for each hypothesis.

Hypothesis 1: Exploring The Relationship Between Crime Complaint Patterns and Real-Estate Demand

After taking a look at different visualizations of borough-level slices of crime complaints over the last 10 years, we decided to drill down to another level of detail. Since the categories of felonies, misdemeanors and violations are quite broad, we used the description field to cluster types of complaints to give us a more solid understanding of the changes in types of complaints. For example, under the umbrella of felonies, we clustered robbery, burglary, grand larceny and grand larceny of motor vehicles and found a steady decrease for all of these categories in Brooklyn (accept for grand larceny which seems to follow a steep incline after 2010). As residents of NYC with friends flocking to the now “hipster” neighborhoods of Brooklyn, it seemed likely that an increase in the safety of valuable items in a given area would go hand in hand with an increase in demand for real-estate in the area.

The following plot shows the count of complaints across all of Brooklyn for felonies in the categories of robbery, burglary, grand larceny and grand larceny of motor vehicles.

Figure 12: Number of Felonies in Brooklyn by Category

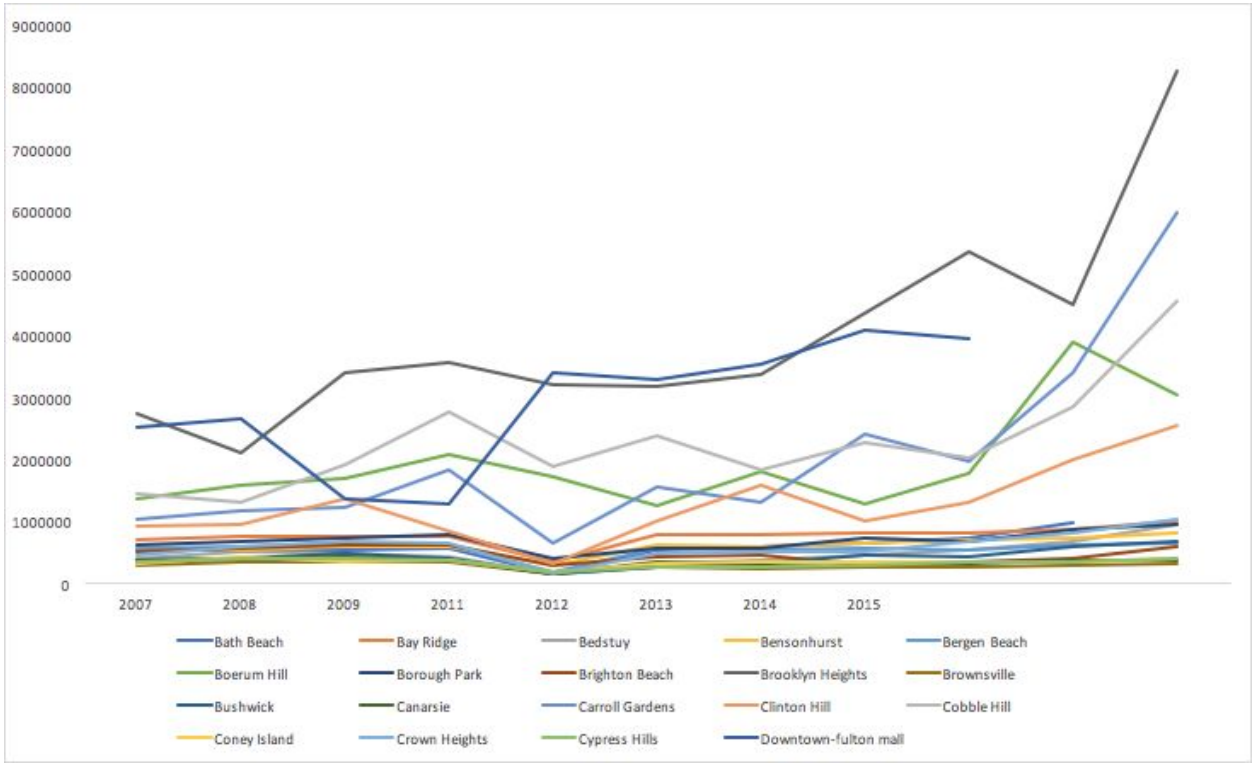


For data on real-estate interest and cost shifts, which we thought should have an impact on crime rates, we turned to the ACRIS NYC archives, which break down prices by borough for varying types of sales. Sure enough, from the chart below we saw there was a steep increase in the average cost of 1-family, 2-family and 3-family homes.

Figure 13: Real Estate Prices in Brooklyn

We also examined housing prices at the neighborhood-level. The chart below shows the breakdown in real-estate purchases for 1-family, 2-family and 3-family homes for a sample of Brooklyn neighborhoods. It was interesting to see the infamous dent in 2012 of real-estate values!

Figure 14: Real Estate Prices in Brooklyn by Neighborhood



As evident from the neighborhood breakdown above, the variance can make it seem like Brooklyn contains many cities within itself. We wanted to take a closer look and pinpoint different areas in Brooklyn for crime and real-estate costs. We chose Bedford-Stuyvesant and Park Slope to investigate. This involved using the geopandas library to map shapefiles, available from https://www1.nyc.gov/site/planning/data-maps/open-data/bytes-archive.page, and neighborhood polygons to the city X-Y coordinates to enhance our crime dataset with a neighborhood column. Since the geopandas library wasn't available on dumbo, we broke the files into smaller slices (which can be found in github directory under data/bk_slice) and enhanced them locally before using PySpark to recalculate our counts, this time incorporating neighborhood into our key.

While understanding our new housing price dataset, we also learned that the definition of a "home" in NYC sales mean that the purchase is a stand-alone building, not an apartment, which made us seek a more comprehensive dataset from ACRIS that included all real-estate purchase from rentals, condos, to coops and homes. The following analysis includes the latter categorization of sales.

Table 3: Bedford-Stuyvesant - Average Price of all Real-Estate Sales and Count of Burglary, Robbery, Grand Larceny and Grand Larceny of Motor Vehicles (BRGL)

Bedford-Stuyvesant
10 Year Increase in Sale Cost: **61%**
10 Year Decrease in BRGL: **15%**

Year	Avg Sale Cost	Count of BRGL
2006	$605,831.34	2,053
2007	$599,073.63	2,133
2008	$536,218.40	1,957
2009	$414,480.32	1,781
2010	$373,402.03	1,915
2011	$419,655.64	1,978
2012	$481,524.01	2,086
2013	$626,258.19	1,812

2014	$804,190.59	1,898
2015	$ 978,138.33	1,750

Table 4: Park Slope/Park Slope South - Average Price of all Real-Estate Sales and Count of Burglary, Robbery, Grand Larceny and Grand Larceny of Motor Vehicles (BRGL)

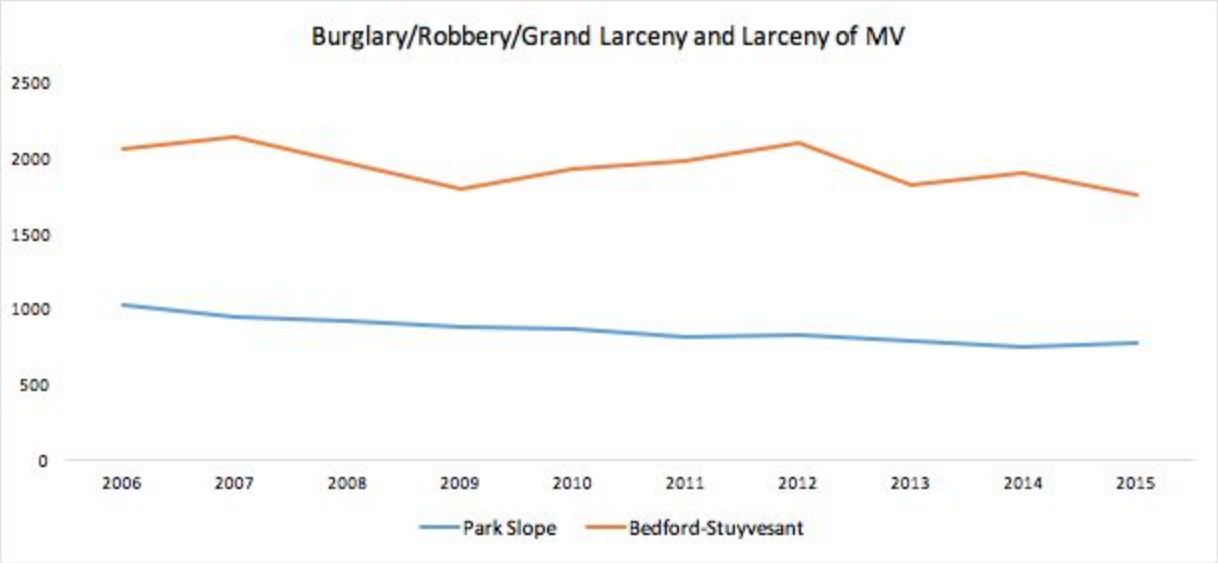
Park Slope

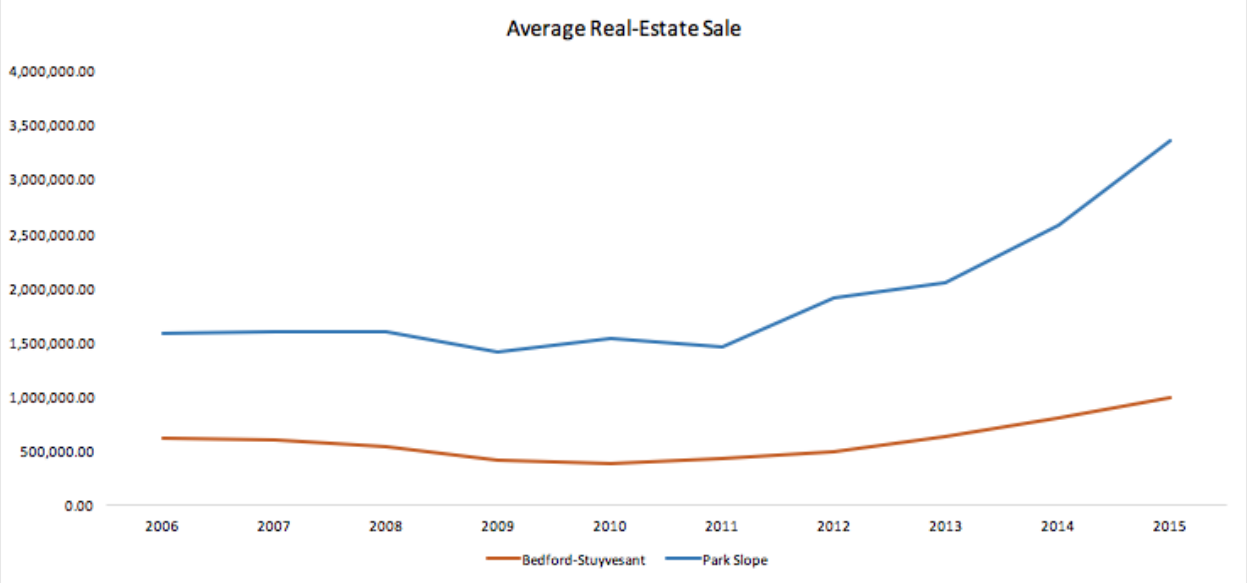
10 Year Increase in Sale Cost: **112%**

10 Year Decrease in BRGL: **25%**

Year	Avg Sale Cost	Count of BRGL
2006	$1,578,509.14	1,018
2007	$1,588,772.97	944
2008	$1,590,645.56	910
2009	$1,411,799.96	875
2010	$1,533,640.79	855
2011	$1,446,689.03	801
2012	$1,909,316.35	817
2013	$2,044,284.97	777
2014	$2,575,820.84	743
2015	$3,343,544.29	765

Figures 15 and 16: Park Slope and Bedford Stuyvesant - Comparison of Average Price of all Real-Estate Sales and Count of Burglary, Robbery, Grand Larceny and Grand Larceny of Motor Vehicles (BRGL)





A simple Pearson correlation between real estate prices and the rates of burglary, robbery, and larceny resulted in -.3345 for Bedstuy with a p-value of .3446, while the results for Park Slope were -.6161 and and p-value of .058. As we might have guessed, the up-and-coming neighborhood of Park Slope has seen an increase in real-estate costs as the worries of burglary, robbery, and larceny have decreased over the last decade. Bedford-Stuyvesant, on the other hand, shows this trend less drastically or perhaps hasn't yet joined the bandwagon of rapidly evolving

Brooklyn neighborhoods. We should note that this is a very naive correlation result as the sample set we drilled down to is too small for statistical significance. This was a lesson learned and side-effect of taking on the complexity of exploring many-to-many relationships within sub-categories of crimes and neighborhood-level slicing of boroughs. Regardless, we discovered the richness of the available urban dataset to test our assumptions about the city we live in, and how to leverage Spark to decimate the effort involved.

Hypothesis 2: Investigating Gentrification via New Restaurants

Returning to the borough-level instead of focusing on neighborhoods in Brooklyn, we hypothesized that the differences between boroughs in how violations, misdemeanors, and felonies varied over time could be due to differences in rates of gentrification. We thought that higher levels of gentrification might lead to fewer crimes. To investigate, we turned to one potential metric of gentrification: the number of new restaurants opening in an area. We downloaded the Department of Health and Mental Hygiene's dataset of food safety violations found in New York restaurants, filtered to include only records corresponding to pre-permit inspections. We hoped that the number of restaurants receiving pre-permit inspections would be a reliable metric of the number of new restaurants. The complete DOHMH food safety dataset can be downloaded from https://data.cityofnewyork.us/Health/DOHMH-New-York-City-Restaurant-Inspection-Results/43nn-pn8j/data, and the filtered view that we created using the NYC OpenData interactive website can be downloaded from https://data.cityofnewyork.us/Health/Pre-Permit-Restaurant-Inspections/jzz4-5r78/data.

We had originally wanted to look at the relationship between crime and new restaurants on a neighborhood level, but unfortunately the food safety dataset did not have X and Y city coordinates (or latitude and longitude), so we could not easily map the restaurants to neighborhoods as we'd done with our crime data. Instead, we explored the data on a borough-level. We used PySpark to eliminate duplicate restaurants, as identified by their unique CAMIS ids, and then count them by borough and year (count_restaurants_by_boro_year.py). Then we matched the counts with our counts of felonies, misdemeanors, and violations by borough and year, which we had produced for Part 1. The food safety dataset only included data from 2013 on (after we eliminated years with counts of only 1 new restaurant), so we did not have very many data points. The tables and plots below, generated by the notebook Correlation with Restaurants.ipynb, show each of the three crime types plotted versus the number of new restaurants in a given borough. Each square represents a different year.

Table 5 and Figure 17: Felonies and New Restaurants by Borough

	Year	Boro	Felonies	Restaurants
0	2013	BRONX	28541	47
3	2013	BROOKLYN	50268	145
6	2013	MANHATTAN	36334	233
9	2013	QUEENS	34325	89
12	2013	STATEN ISLAND	5186	19
1	2014	BRONX	28538	231
4	2014	BROOKLYN	49554	604
7	2014	MANHATTAN	34575	905
10	2014	QUEENS	33405	616
13	2014	STATEN ISLAND	5137	77
2	2015	BRONX	30063	256
5	2015	BROOKLYN	47942	741
8	2015	MANHATTAN	35972	1043
11	2015	QUEENS	32118	648
14	2015	STATEN ISLAND	5201	93

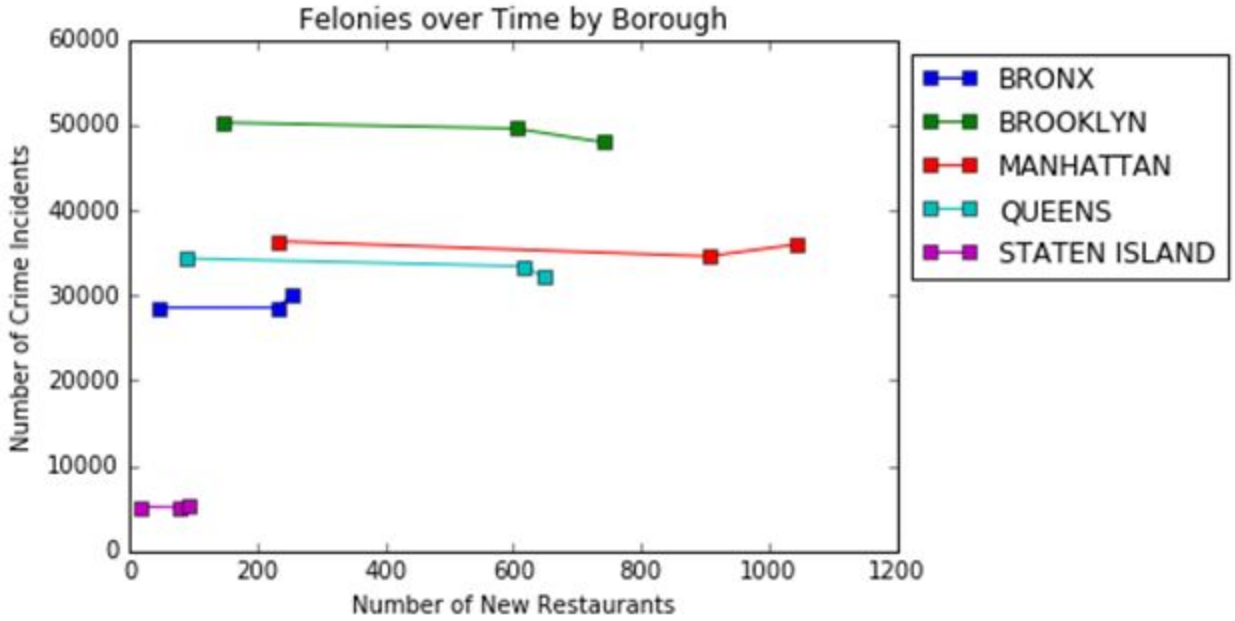


Table 6 and Figure 18: Misdemeanors and New Restaurants by Borough

	Year	Boro	Misdemeanors	Restaurants
0	2013	BRONX	63452	47
3	2013	BROOKLYN	82029	145
6	2013	MANHATTAN	70238	233
9	2013	QUEENS	54682	89
12	2013	STATEN ISLAND	13366	19
1	2014	BRONX	64367	231
4	2014	BROOKLYN	79503	604
7	2014	MANHATTAN	66385	905
10	2014	QUEENS	53604	616
13	2014	STATEN ISLAND	13327	77
2	2015	BRONX	61570	256
5	2015	BROOKLYN	75330	741
8	2015	MANHATTAN	64365	1043
11	2015	QUEENS	50333	648
14	2015	STATEN ISLAND	12581	93

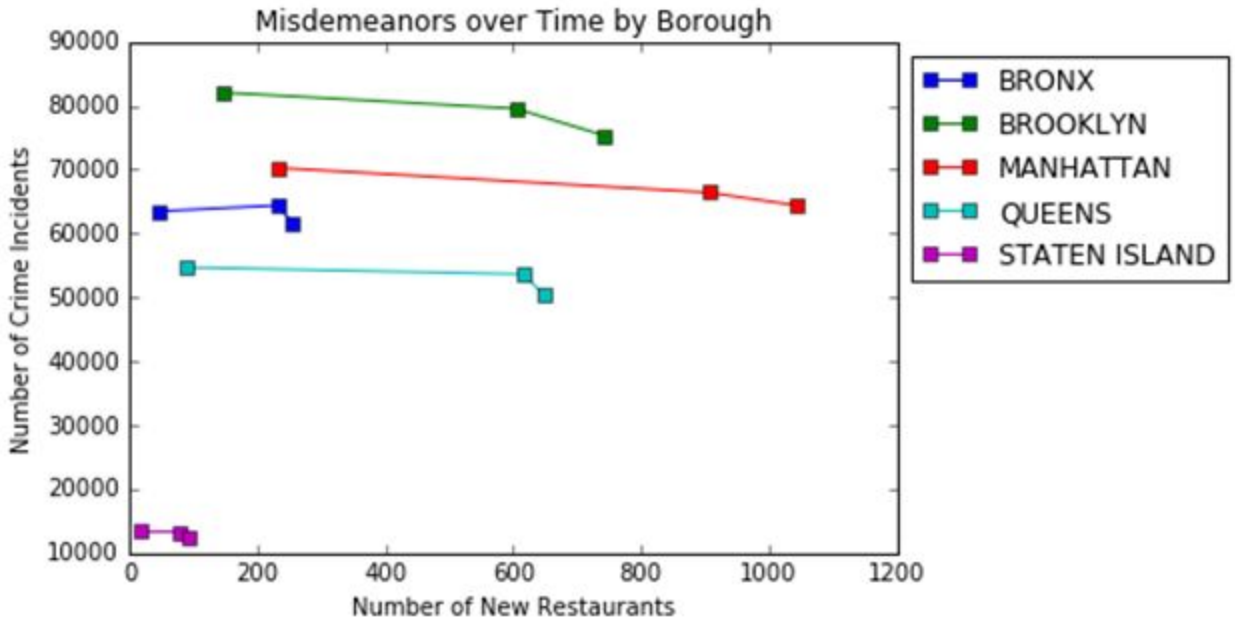
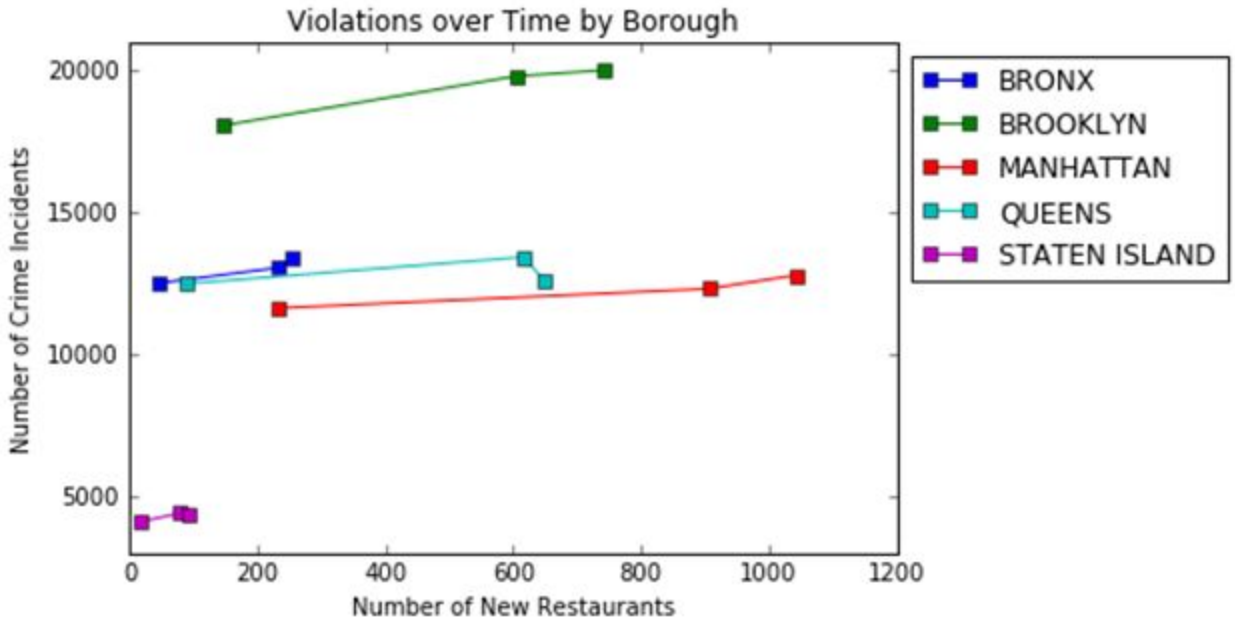


Table 7 and Figure 19: Violations and New Restaurants by Borough

	Year	Boro	Violations	Restaurants
0	2013	BRONX	12504	47
3	2013	BROOKLYN	18040	145
6	2013	MANHATTAN	11618	233
9	2013	QUEENS	12477	89
12	2013	STATEN ISLAND	4108	19
1	2014	BRONX	13032	231
4	2014	BROOKLYN	19780	604
7	2014	MANHATTAN	12300	905
10	2014	QUEENS	13416	616
13	2014	STATEN ISLAND	4396	77
2	2015	BRONX	13366	256
5	2015	BROOKLYN	19990	741
8	2015	MANHATTAN	12781	1043
11	2015	QUEENS	12604	648
14	2015	STATEN ISLAND	4362	93



From the graphs, it appears that felonies are negatively related to the number of new restaurants, as expected, in two boroughs, Brooklyn and Queens, but positively related in the other three boroughs. It also appears that violations may be positively related to the number of new restaurants in all five boroughs. Only misdemeanors seem to be negatively related to the number of new restaurants in all five boroughs.

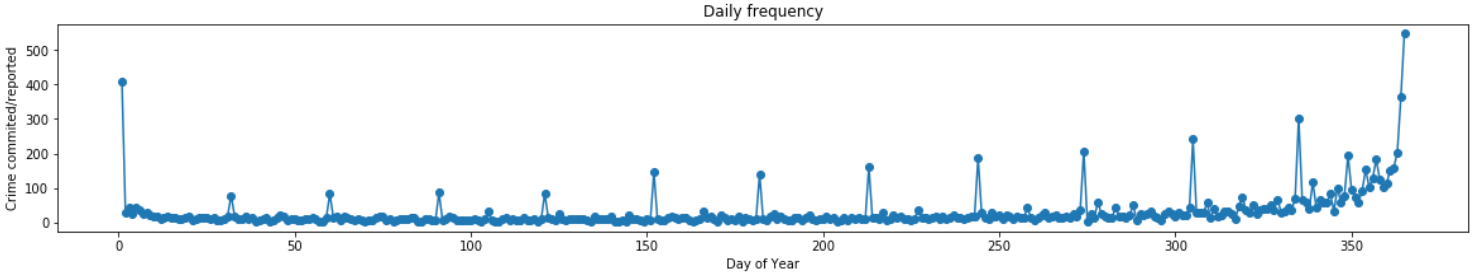
Because our sample sizes were too small for formal correlation tests like Pearson correlation, we decided to use linear regression to try to understand the effects of new restaurants on crime. For each type of crime, we built an OLS model that included borough (categorical, dummy-encoded), year (categorical, dummy-encoded), and number of restaurants (numerical, centered) as predictors for the number of crimes of that type. We examined the size of the coefficients learned by the model to gauge the relationship between new restaurants and crime, controlling for borough and year. The coefficient for restaurants was negative for felonies but not significantly (coef -2.8585, p-value 0.175). It was positive for violations but not significantly (coef 1.1986, p-value 0.223). Finally, it was significantly negative for misdemeanors (coef -7.2786, p-value 0.004), in keeping with what we'd observed from the graphs.

These results suggest that, if the number of new restaurants is a good measure of gentrification, gentrification could possibly decrease misdemeanors but not have a significant effect on felonies (more severe than misdemeanors) or violations (less severe). However, it should be stressed that we are working with very small samples sizes because of the availability of the restaurant data, are only examining one aspect of gentrification, and are looking at correlation without any causal evidence, so we can't draw broad conclusions.

Hypothesis 3: Weather-Related Anomaly Detection

From our exploration of the data in Part 1, we suspected there was a relationship between weather and crime. We explored this hypothesis further in Part 2. To begin, we counted the number of crimes per day over the days of the year (DailyTrendByYear.py), using PySpark. To do so, we had to decide which date to use, since the dataset includes three date fields, the "from" date, the "to" date, and the report date. We prioritized them in that order, so that the "to" date was used if the "from" date was invalid or null, and the report date was used if both the "from" and "to" dates were invalid or null. After obtaining the counts per day, we plotted the number of crimes per day for every year from 2005-2015 using the notebook DaysbyYearsPlot.ipynb. The figure is included below. For these analyses, we collapsed across all types of crime.

Figure 20: Count by Day of the Year



The plot shows very clear spikes on certain days of every year, which we wanted to investigate. The days whose crime rates fell outside of 3 standard deviations from the mean were noted using the notebook SpikesInYear.ipynb.

We noticed that the year 2005 was an outlier with very low numbers of crimes, inconsistent with the rest of the data, and hence was excluded from subsequent analyses. The annual mean and standard deviation of the number of crimes per year are consistent for the rest of the years. The number of incidents in a year, however, has gradually decreased with time.

All of the highest spikes occur on the first of a month. New Year's has the highest number of incidents. As suggested in the class discussion on Part 1, this appears to be related to an internal factor, prima facie, rather than an external one. Therefore, this observation did not stimulate immediate interest for finding a correlation.

The lower spikes, however, seem to be mainly related to external factors. Most of them lie either on a holiday or on a day of bad weather. The number of incidents was low on and around Christmas and Thanksgiving. Other "low" days seem to be weather-related, for example:

- Feb 02, 2012: Snow storm with 15 inches of snow fall.
- Mar 02, 2009: Mercury plunges to 16°, 6.5 inches of snow fall.
- Aug 28, 2011: Hurricane Irene
- Oct 29, 2012: Hurricane Sandy
- Nov 07, 2012: Snow storm, record snow for this day of year.
- Feb 09, 2013: North American Blizzard

These observations motivated us to obtain a score for the correlation between number of crime incidents on a day and weather factors. We randomly chose year 2013 to analyze more closely. We obtained the weather data for this year from Weather Underground and the National Centers for Environmental Information. Using the PreciptationCrimeCorrelation.ipynb, TemperatureCrimeCorrelation.ipynb, and VisibilityCrimeCorrelation.ipynb notebooks, we

computed the correlation between the number of incidents and the following weather components:

- Average Daily Temperature: This was the most satisfying revelation of this analysis. A positive Pearson's correlation coefficient of 0.54 was obtained, which means the number of incidents somewhat decreased when the temperature was low, like in winters, and increased when the temperature was warm. In a climate like New York's, the temperature range falls towards the lower end of the spectrum, i.e. low temperatures are generally unpleasant and harsh, while higher temperatures are pleasant and stimulate more outdoor activity – and by extension, more criminal activity. Therefore, this correlation makes intuitive sense in the case of New York.
- Precipitation: It is important to note that precipitation is not a regular phenomenon. It will only bear its effect on our observations when rain occurs, i.e. precipitation > 0, and any variation in the number of crime incidents when precipitation = 0 would be irrelevant. Therefore, obtaining a high correlation coefficient is not possible, if we consider the whole dataset. However, it is intuitive to expect a negative correlation, as a rainy day would tend to inhibit outdoor activities and possibly incidents of crime. The correlation coefficient obtained was -0.07, which was not very surprising.
- Visibility: Just like precipitation, visibility is not a highly fluctuating value. Therefore, the magnitude of coefficient should not be high. But again, it could be interesting to observe the nature of the correlation. The correlation coefficient obtained was positive (0.09), although very low, which means the number of incidents somewhat decreased, but definitely did not increase, when the visibility was low.

Summary

In Part 2 of the project, we followed up on several hypotheses we had formed about the crime dataset, including that crime would be related to housing price, the number of new restaurants, holidays, and the weather. We found interesting results in all of these areas, as summarized below.

Table 8: Data Relationships

Attribute	**Metric**
Housing Price to Burglary/Robbery	Pearson correlation: -0.33463, p = 0.345 for Bedford Stuyvesant -0.6161, p = 0.0580 for Park Slope
New Restaurants	OLS coefficient and p-value -7.72786, p = 0.004 for misdemeanors

	-2.8585, p = 0.0175 for felonies 1.1986, p = 0.223 for violations
Temperature	Pearson correlation 0.54
Precipitation	Pearson correlation -0.07
Visibility	Pearson correlation 0.09

Individual Contributions

All team members contributed to brainstorming ideas, analyzing data, interpreting results, and writing. For Part 1, Zeynep wrote the PySpark "slice" scripts, Julie created the visualizations, and Kanishk analyzed the data quality. For Part 2, Zeynep generated additional "slices" and analyzed housing price data, Kanishk analyzed weather and day-by-trends, and Julie mapped crimes to neighborhoods and analyzed restaurant data.

References (Data Sources)

Crime data:

- "NYPD Complaint Data Historic" from https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i

Real-estate:

- "City of New York ACRIS Real Property" from https://data.cityofnewyork.us/City-Government/ACRIS-Real-Property-Master/bnx9-e6tj/data

Restaurant data:

- "DOHMH New York City Restaurant Inspection Results" from https://data.cityofnewyork.us/Health/DOHMH-New-York-City-Restaurant-Inspection-Results/43nn-pn8j/data

Neighborhood shapefiles:

- "Neighborhood Tabulation Areas" from https://www1.nyc.gov/site/planning/data-maps/open-data/bytes-archive.page

Weather data:

- Weather Underground, https://www.wunderground.com
- National Centers for Environmental Information, https://www.ncdc.noaa.gov

NYC Census Data:

- "NYC Population Estimates" NYC Planning, http://www1.nyc.gov/site/planning/data-maps/nyc-population/current-future-populations.page

Acknowledgements

We'd like to thank Professor Juliana Freire, Dr. Erin Carson, Dr. Nick Knight, and NYC OpenData.