# The Association Between Felonies in NYC and Weather and Temporal Conditions

Brian Weinstein

*Columbia University*
*STAT W4201: Advanced Data Analysis*

May 2, 2016

**Abstract**

UPDATE THE ABSTRACT

**Background:** The New York City Police Department recently released incident-level felony data to the New York City Open Data portal. The dataset includes timestamped information for all felonies committed in NYC.

We first examine the association between the daily number of felonies committed in NYC in 2015 and temperature, presence of precipitation, day of week, federal and New York holidays, and school days. Second, we examine the association between changes in the number of felonies (as compared to the previous day) and large increases in temperature ($> 8°$F from the previous day).

**Methods and Results:** We initially test for a difference between the number of felonies on warmer and cooler days ($\geq 51.98°$F and $< 51.98°$F, respectively — the NYC 2015 median), finding overwhelming evidence of a difference, with there being 44 more felonies, on average, on warmer days than on cooler day (95% CI 38 to 51 felonies; two sided p-value $< 0.000001$ from a two-sample t-test).

Remove this.

After accounting for presence of precipitation, holidays, school days, and day of week, the data provides overwhelming evidence that for every 1°F increase in temperature there are, on average, 1.4 additional felonies per day (95% CI 1.3 to 1.6 felonies; two-sided p-value $< 0.000001$ for a test that the linear regression coefficient is 0).

Explain that the results change after removing observations with large studentized residuals.

We next find that, after accounting for day of week, there is little evidence to suggest that increases in felonies from the previous day are associated with large increases in temperature (two-sided p-value 0.1001 for a test that the linear regression coefficient is 0).

**Conclusions:** There is a clear association between warmer temperatures and an increased number of felonies. There is no evidence that large increases in temperature from the previous day are associated with increases in the number of felonies.

# 1 Introduction

After many years of pressure, the New York City Police Department (NYPD) recently released incident-level felony data to the NYC Open Data portal, as part of their initiative to improve their accessibility, transparency, and accountability. Prior to this release, felony data had only been provided in an aggregated format (by week and police precinct), and was done so only in PDF and Excel files on a weekly and quarterly basis.

In this paper, we use the newly-released data to examine the association between the daily number of felonies committed in New York City (NYC) and: day of week, outside air temperature, precipitation, federal and New York (NY) holidays, and public school days.

## 1.1 Questions of Interest

In this paper we study three main questions of interest:

1. Are felonies associated with temperature? After taking temperature into account, is felonies associated with precipitation, school days, holidays, and day of week?
2. Although theres no causal relationship, for a given set of these conditions, how many felonies can the NYPD reasonably expect?
3. Are changes in the number of felonies, as compared to the previous day, associated with large ($> 8°$F) increases in temperature?

## 1.2 Dataset

### 1.2.1 Data Schema and Sources

The class, description, and source for each variable in the dataset is outlined below. Only those variables used in the analyses are included here — redundant and untranformed variables that were removed during exploratory analysis are not described below. The dataset contains 365 observations, one for each date in 2015.

- `felonies` (integer) is a count of the number of felonies committed on each day in NYC in 2015. The values are derived counts from the "NYPD 7 Major Felony Incidents" dataset in the NYC Open Data Portal [data.cityofnewyork.us/d/hyij-8hr7]. The felonies included in the dataset that contribute to the overall daily count are burglary, felony assault, grand larceny, grand larceny of motor vehicle, murder and non-negligent manslaughter, rape, and robbery.
- `temp_min_degF` (numeric) is the minimum daily temperature on the given date (in degrees Fahrenheit), as reported by the New York Central Park Belvedere Tower weather station. The data was requested via the National Centers for Environmental Information [ncdc.noaa.gov/cdo-web/search].
- `is_warm` (factor) is an indicator variable, taking value "1" if `temp_min_degF` is $\geq 51.98°$F (the median for 2015) on the given date, and "0" otherwise.
- `any_precip` (factor) is an indicator variable, taking value "1" if there was any precipitation on the given date, and "0" otherwise. See `temp_min_degF` for source information.
- `is_holiday` (factor) is an indicator variable, taking value "1" if the given date is a NY or federal holiday, or value "0" otherwise. The NY and federal holidays were defined using the lists provided by the NY State Department of Civil Service [cs.ny.gov/attendance_-leave/2015_legal_holidays.cfm] and U.S. Office of Personnel Management [opm.gov/policy-data-oversight/snow-dismissal-procedures/federal-holidays/#url=2015], respectively.

- `is_school_day` (factor) is an indicator variable, taking value "1" if NYC Public Schools were open and in session on the given date, or value "0" otherwise. Although the NYC Department of Education publishes this data to the NYC Open Data Portal, the historical data is only retained there for the current school year. Instead we scrape the attendance data from XML files in Aaron Schumacher's "NYCattends" Github repository [github.com/ajschumacher/NYCattends/tree/master/xml].
- `day_of_week` (factor) is a categorical variable indicating the day of week (Sunday="1", Monday="2" , . . . , Saturday="7").
- `felonies_diff` (numeric) indicates for a given date the difference in the number of felonies as compared to the previous day. On Jan. 3, 2015, for example, `felonies_diff` = 6, since there were 6 more felonies committed on Jan. 3 than on Jan. 2.
- `temp_min_degF_diff` (numeric) indicates for a given date the difference in `temp_min_degF` as compared to the previous day. On Jan. 3, 2015, for example, `temp_min_degF_diff` = -1.98, since the daily minimum temperature (`temp_min_degF`) was 1.98°F lower on Jan. 3 than on Jan. 2.
- `temp_jump` (factor) is an indicator variable, taking value "1" if `temp_min_degF_diff` $> 8$, or value "0" otherwise. An increase of $> 8$°F puts the day of interest in the top 10% of temperature increases in 2015.

### 1.3 Report Overview

In Section 2 we present some exploratory analysis and plots, in Sections **??** and **??** we walkthrough the model setup and statistical analysis, in Section **??** we do some model checking, and in Section 5 we present our conclusions. update this

## 2 Exploratory Analysis and Data Cleaning

We first examine pairwise scatterplots of some of the numeric variables in the raw dataset (`felonies`, `temp_min_degF`, `temp_max_degF`, and `school_attendance_pct`), as shown in Figure 1. From this figure we first notice that there are approximately linear relationships between `felonies` and `temp_min_degF`, and between `felonies` and `temp_max_degF`. There is also strong collinearity between `temp_min_degF` and `temp_max_degF` (correlation: 0.969), however, so we remove one of these variables (`temp_max_degF`) from the covariates that will be used in the regression model.

Figure 1 also shows that there is no linear relationship between `felonies` and `school_attendance_pct` (the percent of students present in school on a given day). Any non-school-day has 0% attendance, so instead of using this as a numeric variable, we convert it to the indicator variable `is_school_day`, taking value "1" if NYC Public Schools were open and in session on the given date (i.e., if `school_attendance_pct` is $> 0$), or value "0" otherwise.

Faceted boxplots of the categorical variables are shown in Figure 2.

## 3 Modeling the Number of Felonies Per Day

In this section we use simple and multiple linear regression to model the number of felonies per day (the first and second questions of interest).

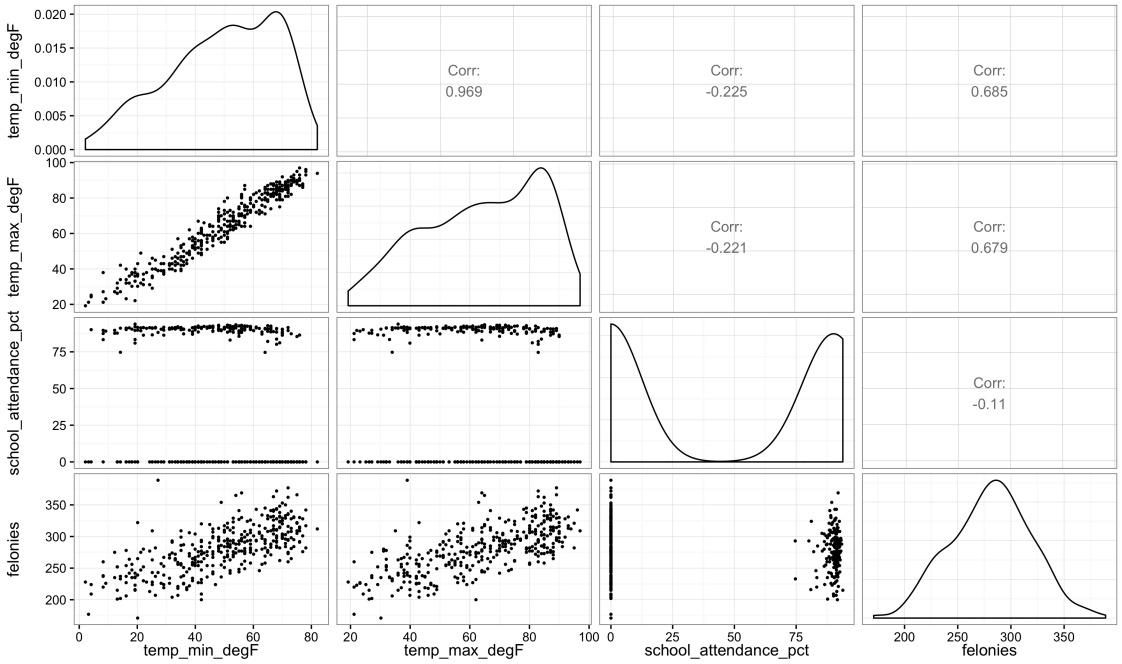


Figure 1: Pairwise scatterplots of some of the numeric variables in the raw dataset.

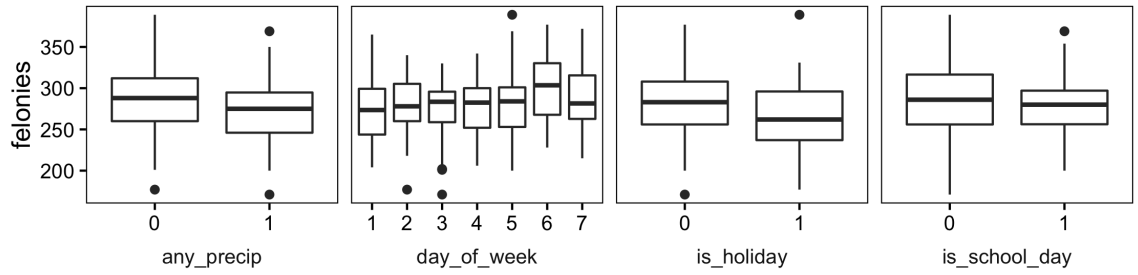


Figure 2: Faceted boxplots of the categorical variables in the dataset.

### 3.1 Assumptions

We first assume that each occurrence of a felony is an independent Bernoulli event with very low probability $p$. The sum of these Bernoulli events is the number of felonies that occur on a given day. It follows a binomial distribution where $n$ is the number of opportunities for a felony to occur — as a rough approximation this might be on the order of the population of NYC ($\sim 8.4$ million). Since $n$ is large enough, we can approximate this binomial distribution with a normal distribution.

We also assume the four assumptions of linear regression:

- Linearity: `felonies` can be expressed as linear combination of the independent variables
- Homoscedasticity (constant variance): $\text{Var}(Y|X_1, \ldots, X_p)$ is the same at all values of $X_1, \ldots, X_p$
- Normality: residuals of the fit are normally distributed
- Independence: residuals of the fit are independent

## 3.2 Exploratory Models

As an exploratory step, we initially perform a simple linear regression of felonies on temperature. The regression coefficients, standard errors, and p-values are shown below:

Might not even need to show these here, as long as I keep the summary in the next paragraph.

```
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   213.11913    4.07082   52.35   <2e-16 ***
temp_min_degF   1.38097    0.07709   17.91   <2e-16 ***
```

There is overwhelming evidence of an association between felonies and temperature. For every 1 °F increase in temperature, there are 1.38 additional felonies per day (95% CI 1.22938 to 1.532569, two sided p-value $< 2 \times 10^{-16}$ for a test that the coefficient is 0).

What's most interesting here, however, are the observations with high residuals, as shown in the residual plot in Figure 3. Some days with large residuals aren't modeled well by temperature alone — Jan. 1, 2015, for example, a federal holiday, had many more felonies than expected, given the temperature on that day. Including other covariates in the model, like `is_holiday` (an indicator as to whether the day is a NY/federal holiday), might help to account for some of this behavior. We next add these additional variables in Section 3.3.

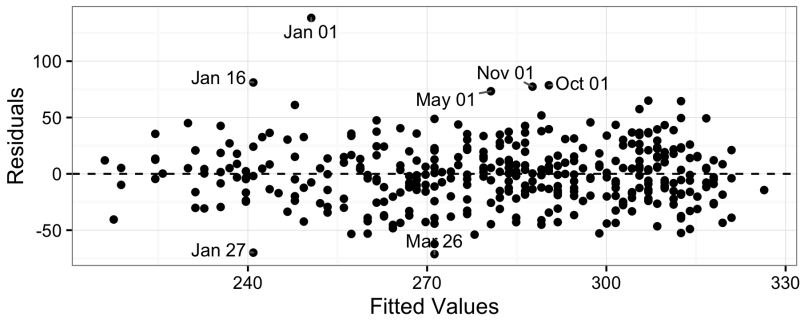


Figure 3: Residual plot for the regression of `felonies` on `temp_min_degF`.

There are some potentially problematic observations with high leverage or large studentized residuals, but there was no change in interpretation after removing these observations from the dataset. Also note that we tested higher order terms of `temp_min_degF` in a multiple regression, but only the first-order term was significant.

## 3.3 Statistical Analysis

We next incorporate additional covariates, performing a multiple linear regression of felonies on the variables shown in the regression summary below:

```
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              210.71208    5.66649  37.186  < 2e-16 ***
temp_min_degF              1.38099    0.08768  15.750  < 2e-16 ***
any_precip1              -21.01037    8.59459  -2.445 0.014990 *
is_holiday1               -6.00627    7.68276  -0.782 0.434865
is_school_day1             5.94836    3.74662   1.588 0.113258
day_of_week2               0.06874    5.73913   0.012 0.990450
day_of_week3              -3.29691    5.70419  -0.578 0.563647
day_of_week4              -3.46950    5.71773  -0.607 0.544376
day_of_week5              -0.17653    5.68349  -0.031 0.975239
day_of_week6              19.01055    5.69764   3.337 0.000938 ***
day_of_week7              14.21319    5.02454   2.829 0.004940 **
temp_min_degF:any_precip1  0.15666    0.16512   0.949 0.343385
```

Temperature, the precipitation indicator, and day-of-week are all significant (removing `day_of_week` gives a p-value of $3 \times 10^{-6}$ in an extra sum of squares F-test). The holiday indicator, school day indicator, and the interaction between temperature and precipitation are not significant.

Possibly rephrase this paragraph as a summary of results, in terms of the first question of interest. e.g., "after accounting for temp, precip and day of week are significant, but holiday, school day, and interaction aren't (see p-values in the regression summary)."

### 3.4 Model Checking and Improvement

## 4 Modeling the Day-to-Day Change in the Number of Felonies Per Day

### 4.1 Assumptions

### 4.2 Statistical Analysis

### 4.3 Model Checking and Improvement

## 5 Conclusion

*Todo list