New York City and Crime, A Story of Data Analysis

Brent Daniel, Richard Lavery, Anita Pinto, Rashmi Rajaguru, Jingbo Wu

Introduction

Crime has long been a story in New York City. Many of us have witnessed significant changes in both the frequency of crime and the most prevalent types of crime that have dominated New York City in the last 30 years.

Crime, in general, is a major topic for any major city. Whatever a city may have to offer, crime seems to be a major detractor for those visiting and considering living there.

It is only natural, then, we should want to give crime in New York a thorough, analytical look.

- **What** crimes are happening? (What types of crime are most frequent?)
- **Where** do crimes happen? (Which Boroughs? Maps of crime?)
- **When** do crimes occur? (Time of day? Of Month? How has crime changed over time?)
- **Why?** (Or, what other phenomena might explain crime frequency?)

(It is unsurprising that we cannot examine **Who** and **How**, as public datasets protect perpetrator and victim privacy, and don't include methods of crime).

Our analysis will focus on exploration of these themes, the **What**, **Where**, **When** and **Why** of New York City Crime.

Additionally, as we began to explore this data, we started to wonder about specific topics within the types of crime:

- **Violent Crime**: Quality of life in a city is often dictated by the lack of fear of violent crime
- **Dangerous Drugs**: Addiction problems are more prominent now and legalization of marijuana remains a hot topic

Team Members and main roles:

- Brent Daniel: Borough Crime Analysis, Crime vs. Temperature, Report Organization, Writing/Editing, Project Leadership
- Rich Lavery: Merging Meteorological and Lunar Cycle Data Looking for Relationships with Crime. Interactive Web Application Development including interactive maps and summary plots.
- Anita Pinto: Offense Categories Broad and Internal Desc with Borough Distribution, Premises/Location Analysis for Different Crime Categories, Crimes in Parks and Categories, Time Trend of Crime Categories and Attempted Crime Analysisi, Final RMD Collaboration
- Rashmi Rajaguru: Interactive Web Application development and D3 plot ,Project Data and deployment summary and challenges , Spatial Data analysis , Word Cloud based crime description (Interactive)
- Jingbo Wu: Data description; Data Quality Analysis; Partial Main Analysis including time series, frequency analysis on different time scale (annual cycle, monthly cycle, weekly cycle, daily cycle); heatmaps; frequency by Precincts

Description of Data

Main Data Set

The main source of data is from the New York City Open Data site, and can be accessed as follows:
1. Visit https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i (https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i)
2. Click Export
3. Click CSV button
4. Put it in your working directory

We downloaded the data for our own analysis on March 5, 2018. The data is updated annually in August, though it would be possible to replicate this analysis after August by limiting the data by date to the same time frame.

- This dataset includes all valid felony, misdemeanor, and violation crimes reported to the New York City Police Department (NYPD) from 2006 to the end of the year prior to the dataset creation (2016). The dataset was created on November 2, 2017.
- The dataset is provided by The New York Police Department and owned by NYC OpenData. It has 5.58 Million rows and 24 variables. Three types of variables included are number, text, and location. The 24 variables are listed below.
- Only valid complaints are included in this release. Information is accurate as of the date it was queried from the system of the record, but should be considered a close approximation of the current records, due to complaint revisions and updates. (NYPDIncidentLevelDataFootnotes.pdf)

We used the following commands to import and arrange the data for analysis:

```
library(zoo)
library(vcd)
library(dplyr)
library(tidyr)
library(tibble)
library(lattice)
library(ggplot2)
library(forcats)
library(extracat)
library(vcdExtra)
library(gridExtra)
library(tidyverse)
library(data.table)

library(lubridate)
library(RColorBrewer)
library(knitr)
library(ggthemes)

library(kableExtra)

var_names <- c("Id", "DateStart", "TimeStart", "DateEnd", "TimeEnd", "DateReport", "ClassCode", "OffenseD
esc",
               "IntClassCode", "IntOffenseDesc", "AtptCptdStatus", "Level", "Jurisdiction", "Boro", "Pct"
, "LocOfOccr", "PremDesc", "ParkName", "HousingDevName", "XCoord", "YCoord", "Lat", "Long", "Lat_Long")


crime_df <- fread("NYPD_Complaint_Data_Historic.csv",na.strings="", col.names = var_names, stringsAsFacto
rs = TRUE)

crime_df$DateStart <- as.Date(crime_df$DateStart, format='%m/%d/%Y')
crime_df$DateEnd <- as.Date(crime_df$DateEnd, format='%m/%d/%Y')
crime_df$DateReport <- as.Date(crime_df$DateReport, format='%m/%d/%Y')
```

```
crime_w_df <- crime_df
```

Supplementary Data Sets

To examine relationships with crime that went beyond the main data set, we accessed the following, other sites:

- Borough Population Data
  http://www1.nyc.gov/site/planning/data-maps/nyc-population/current-future-populations.page (http://www1.nyc.gov/site/planning/data-maps/nyc-population/current-future-populations.page)
- Phases of the Moon Data
  https://www.somacon.com/p570.php (https://www.somacon.com/p570.php)
- Police Precinct Map Data
  https://data.cityofnewyork.us/Public-Safety/Police-Precincts/78dh-3ptz/data (https://data.cityofnewyork.us/Public-Safety/Police-Precincts/78dh-3ptz/data)
- Weather Data
  https://www.ncdc.noaa.gov/ (https://www.ncdc.noaa.gov/)
- NY Unemployment Data
  https://labor.ny.gov/stats/LSLAUS.shtm (https://labor.ny.gov/stats/LSLAUS.shtm)

Analysis of Data Quality

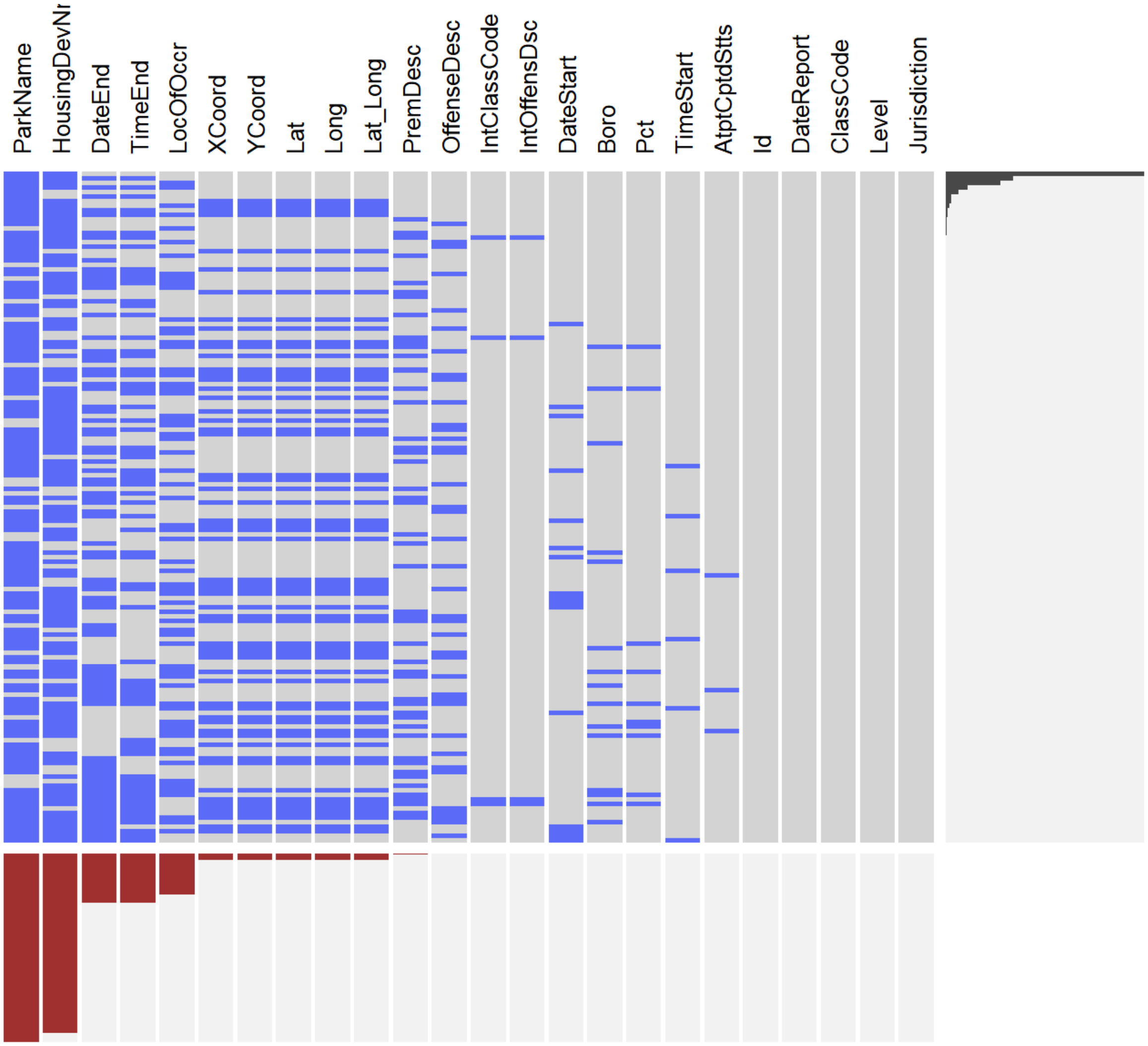
We found no significant issues in data quality that would inhibit analysis, with the possible exception of geo-location pertaining to certain types of crimes.

Each row of the data set lists a single complaint or event. The Report Date (the case reporting time) ranges from 2006-01-01 to 2016-12-31. In this section, we will investigate the patterns of missing data as well as potential errors in the data in order to determine how well we can trust the data we are using.

Missing variables Pattern using Visna

A summary graphic of missing patterns is shown below:

```
visna(crime_df,sort="b")
```



From the above Visna Plot, our observations about the missing patterns are:

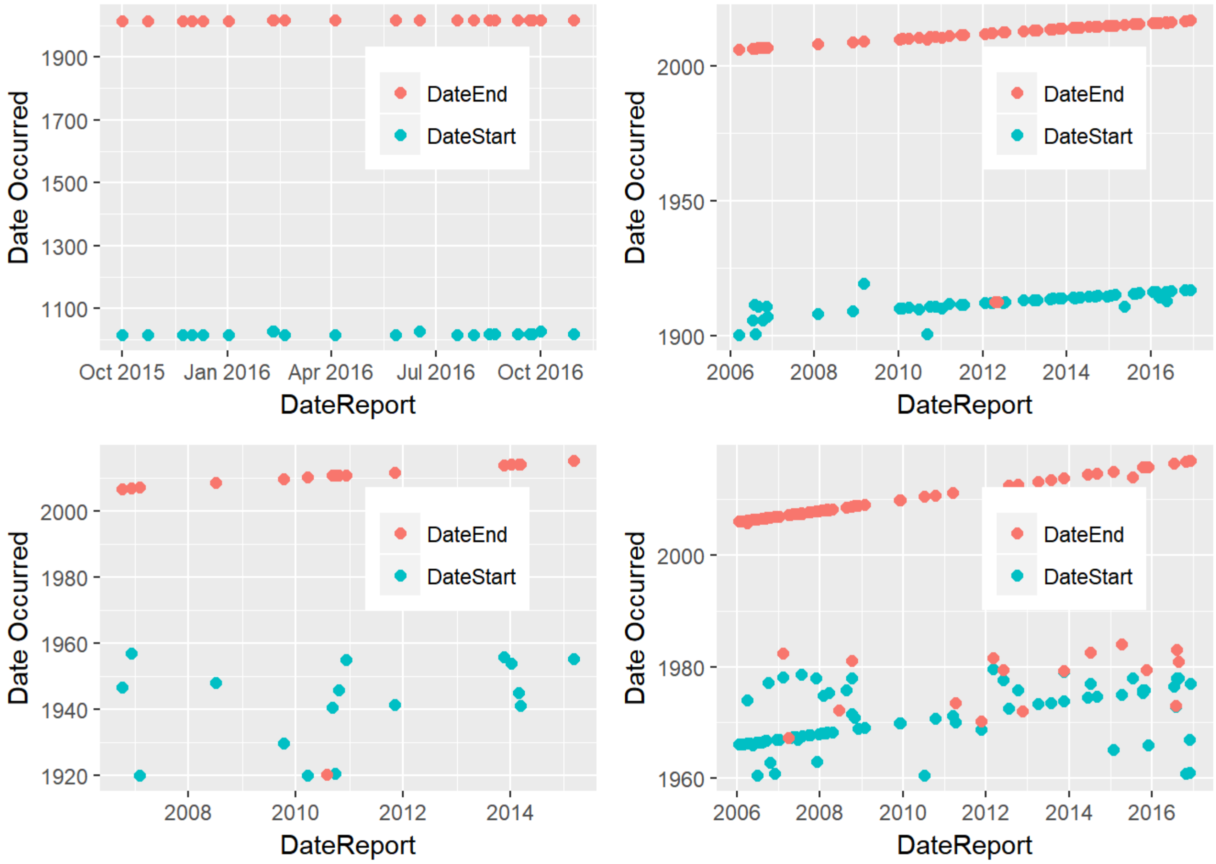
- Five of the 24 variables have no data quality concerns:
    1. Complaint Number
    2. Report Date
    3. Offense Classification Code
    4. Level of Offense
    5. Jurisdiction responsible
- *AtptCptdStatus* is an indicator of whether crime attempted or completed. There are only 7 missing cases; 5,483,869 coded as completed, and 96,159 cases indicated as attempted.
- *ParkName* is recorded if the event occurred in a park. Most of the cases doesn’t have this variable simply because most crimes did not occur in parks. We cannot estimate the percent of real missing park data.

Missing Start-Date of Crime Reported

There are total of 655 complaints missing the crime Start Date. The missing data is distributed the same as non-missing data, so we are not concerned about this random-appearing missing data.

Errors in Start-Date of Crime Reported

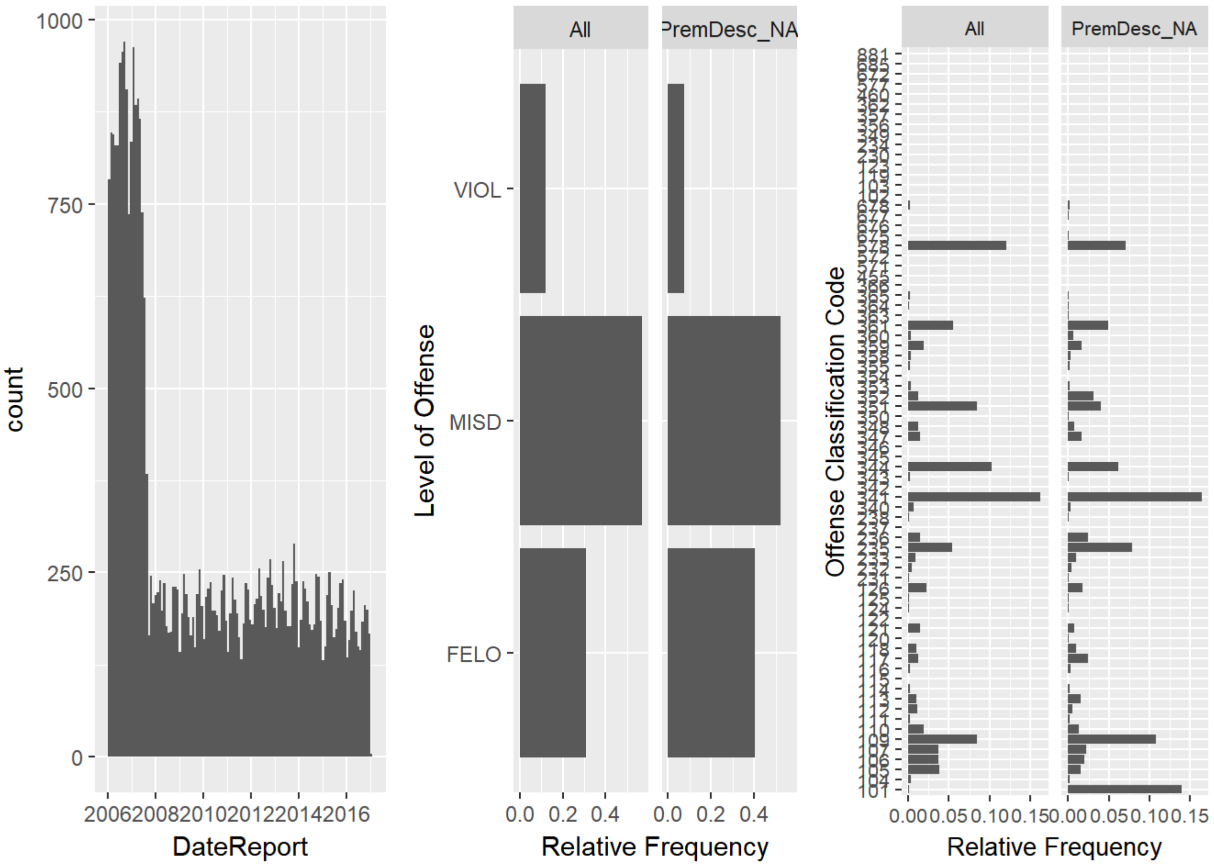
```
crime_df%>%select(DateStart,DateEnd,DateReport)->df_3DT
df_3DT%>%filter(DateStart<=as.Date("1900-01-01"))->df_3DT_Year1900
df_3DT%>%filter(DateStart>=as.Date("1900-01-01") & DateStart<=as.Date("1920-01-01"))->df_3DT_Year1900to19
20
df_3DT%>%filter(DateStart>=as.Date("1920-01-01") & DateStart<=as.Date("1960-01-01"))->df_3DT_Year1920to19
60
df_3DT%>%filter(DateStart>=as.Date("1960-01-01") & DateStart<=as.Date("1980-01-01"))->df_3DT_Year1960to19
80
#association between report date and complaint date indicating possible typo in recording the data
mytheme=theme(legend.title=element_blank(),legend.position=c(0.7,0.7))
df_3DT_Year1900%>%drop_na()%>%gather(key=DateOccur,value,-DateReport)->tmp
ggplot(tmp)+geom_point(aes(DateReport,value,color=DateOccur),size=2)+mytheme+ylab("Date Occurred")->p1
df_3DT_Year1900to1920%>%drop_na()%>%gather(key=DateOccur,value,-DateReport)->tmp
ggplot(tmp)+geom_point(aes(DateReport,value,color=DateOccur),size=2)+mytheme+ylab("Date Occurred")->p2
df_3DT_Year1920to1960%>%drop_na()%>%gather(key=DateOccur,value,-DateReport)->tmp
ggplot(tmp)+geom_point(aes(DateReport,value,color=DateOccur),size=2)+mytheme+ylab("Date Occurred")->p3
df_3DT_Year1960to1980%>%drop_na()%>%gather(key=DateOccur,value,-DateReport)->tmp
ggplot(tmp)+geom_point(aes(DateReport,value,color=DateOccur),size=2)+mytheme+ylab("Date Occurred")->p4

grid.arrange(p1,p2,p3,p4,nrow=2)
```



- There seems to be errors in *DateStart*. Some cases are shown with a year of 1015 (needless to say, a time frame not covered by this dataset). By comparing those Start Dates to the Dates of the Report, The two dates usually have very close month/date. The *DateEnd* variable also suggests the actual year to be 2015, and hence, a typo.
- The scatterplot of *DateStart* vs *DateReport* did show some strict linear correlation for many cases during some periods.
- As shown in the figure, the amount of such cases is very small. In our main analysis, we will focus on cases with *DateStart* after Jan. 1, 2000 up until Dec. 31, 2016 in total over 5.57M. Cases with *DateStart* earlier than Jan. 1, 2000 are totaled 1549 which will be ignored.

Missing Premises Description *(PremDesc)*

```
#get the reporting dates of cases with PremDesc missing and histogram over DateReport
#crime_df%>%filter(is.na(PremDesc))%>%dplyr::summarise(count=n()) #35198 missing cases
crime_df%>%select(PremDesc,DateReport)%>%filter(is.na(PremDesc))%>%select(DateReport)->tmp1
ggplot(tmp1,aes(DateReport))+ geom_histogram(bins=120)->p1

#compare pattern of crime Level between all cases vs. cases with missing PremDesc
crime_df%>%select(PremDesc,Level)%>%filter(is.na(PremDesc))%>%select(Level)%>%group_by(Level)%>%dplyr::su
mmarise(count=n())%>%mutate(RelFreq = count/sum(count))->tmp3; nr1=nrow(tmp3)
tmp3%>%mutate(type=replicate(nr1,"PremDesc_NA"))->tmp3
crime_df%>%select(Level)%>%group_by(Level)%>%dplyr::summarise(count=n())%>%mutate(RelFreq = count/sum(cou
nt))->tmp5; nr2=nrow(tmp5)
tmp5%>%mutate(type=replicate(nr2,"All"))->tmp5
rbind(tmp3,tmp5)->tmp3tmp5
tmp3tmp5%>%ggplot(aes(Level,RelFreq))+geom_bar(stat="identity")+scale_x_discrete(label=abbreviate)+
  coord_flip()+xlab("Level of Offense")+ylab("Relative Frequency")+facet_wrap(~type)->p2

#compare pattern of ClassCode between all cases vs. cases with missing PremDesc
crime_df%>%select(PremDesc,ClassCode)%>%filter(is.na(PremDesc))%>%select(ClassCode)%>%mutate(ClassCode=a
s.factor(ClassCode))%>%group_by(ClassCode)%>%dplyr::summarise(count=n())%>%mutate(RelFreq = count/sum(cou
nt))->tmp2; nrr1=nrow(tmp2)
tmp2%>%mutate(type=replicate(nrr1,"PremDesc_NA"))->tmp2
crime_df%>%select(ClassCode)%>%mutate(ClassCode=as.factor(ClassCode))%>%group_by(ClassCode)%>%dplyr::summ
arise(count=n())%>%mutate(RelFreq = count/sum(count))->tmp4; nrr2=nrow(tmp4)
tmp4%>%mutate(type=replicate(nrr2,"All"))->tmp4
rbind(tmp2,tmp4)->tmp2tmp4
tmp2tmp4%>%ggplot(aes(ClassCode,RelFreq),na.rm=FALSE)+geom_bar(stat="identity")+theme(text = element_text
(size=10))+
  coord_flip()+xlab("Offense Classification Code")+ylab("Relative Frequency")+facet_wrap(~type)->p3

grid.arrange(p1,p2,p3,nrow=1)
```



We do not rely on the description of premises much in our analyses, but readers should note that approximately 0.6% of the crimes do not have a description of the premises, with a disproportionately high number of those cases pertaining to murder/manslaughter.

- There are 35,198 cases missing the description of the premises.
- Comparison of the pattern of ClassCode between all cases vs. cases with missing PremDesc indicates that the category with ClassCode=101, which represents felony crime of MURDER & NON-NEGL. MANSLAUGHTER, have much higher frequency in the missing data.
- In 2006, there are many more cases with missing PremDesc than other years. Year 2006 had more cases overall compared to other years.

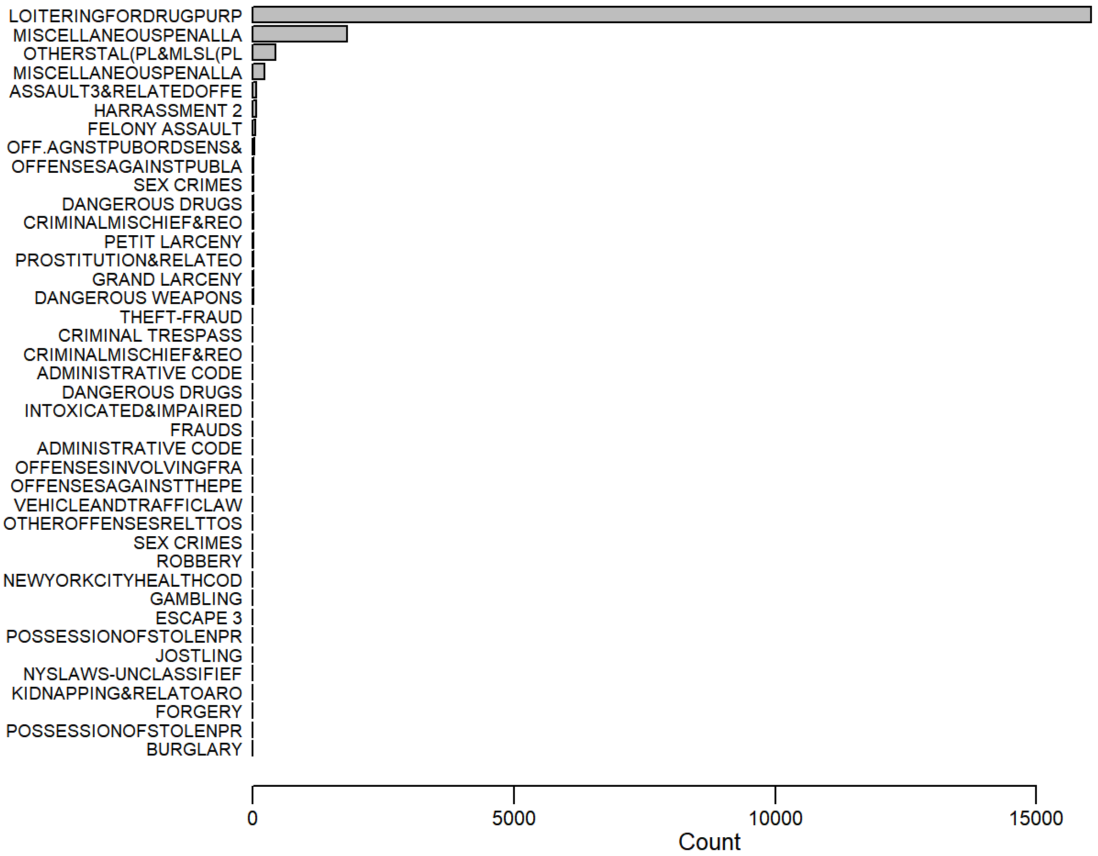
Mismatch between Precinct *(Pct)* and Borough *(Boro)*

There are a very small number of cases where Precinct or Borough are missing (0.007% and 0.008% of all cases, respectively). There is an even smaller number of cases where the Precinct shows an incorrect Borough. These mismatches will not be relevant to our analyses.

Missing Description of Offense *(OffenseDesc)*

```
#match_pct_boro can be used to fix the problem of double/triple borough names of specific Pct
crime_df %>% select(Pct,Boro)%>%group_by(Pct,Boro)%>%drop_na()%>%dplyr::summarise(count=n())%>%
  group_by(Pct)%>%dplyr::summarise(Boro=Boro[count==max(count)])->match_pct_boro
```

```
#match_code_desc can be used to retrieve OffenseDesc(when missing) from ClassCode
crime_df%>%select(ClassCode,OffenseDesc)%>%group_by(ClassCode)%>%
  dplyr::summarise(OffenseDesc=paste(unique(OffenseDesc),collapse=","))%>%
  mutate(OffenseDesc=str_replace(OffenseDesc,",NA",""))%>%mutate(OffenseDesc=str_replace(OffenseDesc,"N
A,",""))->match_code_desc

#For cases with missing OffenseDesc, how they distribute over the code ClassCode
crime_df%>%
  select(ClassCode,OffenseDesc)%>%
  filter(is.na(OffenseDesc))%>%
  group_by(ClassCode)%>%dplyr::summarise(count=n())->tmp1

#showing the supposed OffenseDesc that is missing with its ClassCode
merge(tmp1,match_code_desc,by.x="ClassCode",by.y="ClassCode")%>%arrange(dplyr::desc(count))->match_bycoun
t

par(mgp=c(1,0.3,0),mai=c(0.4,1.8,0.01,0.01))
data2<-match_bycount[order(match_bycount[,"count"]),]
barplot(data2[,"count"],names.arg=abbreviate(data2[,"OffenseDesc"],minlength=20),cex.names = 0.6,cex.axis
=0.7,cex.lab=0.8,horiz=TRUE,xlim=c(0,17500),las=1,xlab="Count")
```



Since the description of the offense can be inferred from the classification code, we should not be concerned about the missing values in this variable.

- *OffenseDesc* (with missing values) is the description of offense corresponding with key code *ClassCode* which is complete in the dataset. Code and description map each other and valid *OffenseDesc* can be infered from a map established from the dataset.
- The plot above shows missing counts of the *ClassCode* categories with *OffenseDesc* originally missing but now retrieved from the map between *ClassCode* and *OffenseDesc*.

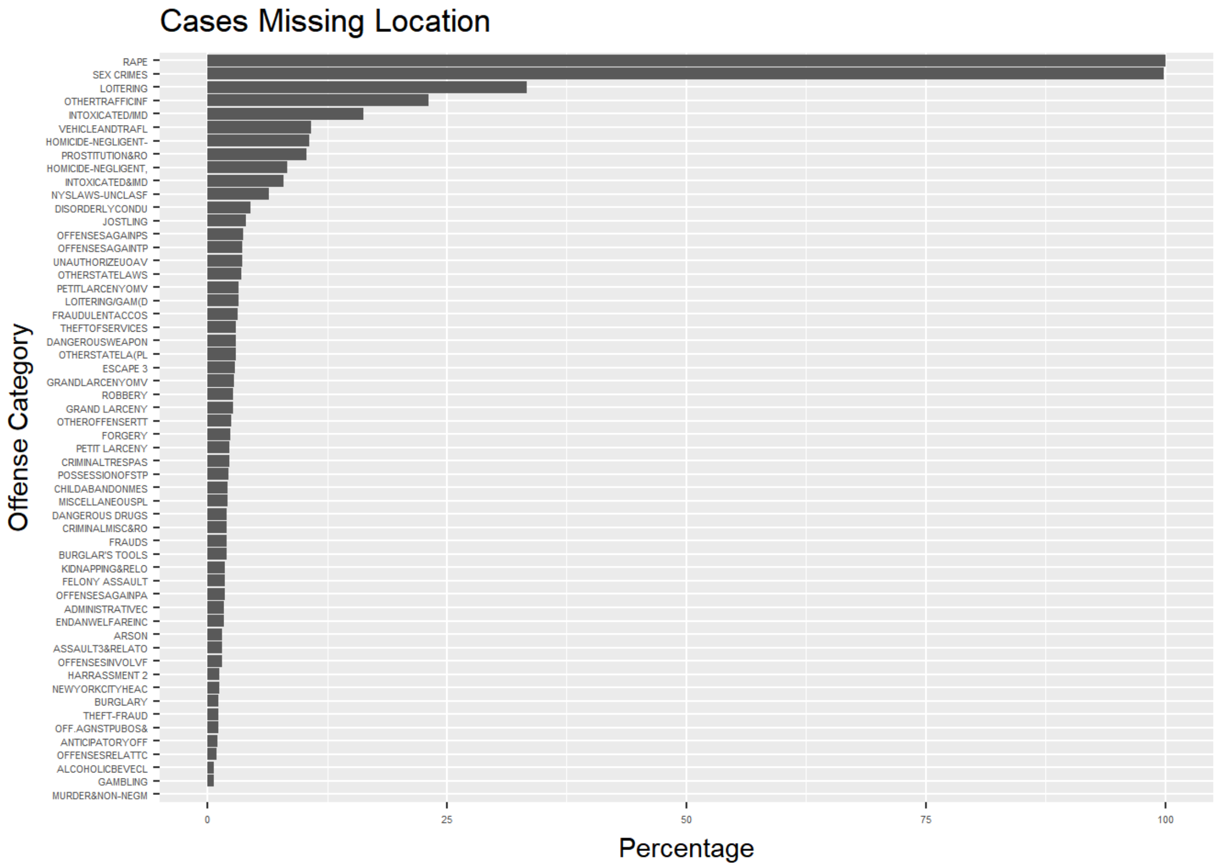
Missing Geolocation

Missing Geo-location data will have an impact on our spatial analyses, but most of the impact is that it will not bear intelligence about the location of certain types of crime.

There are five geo-location variables, and they all have the same missing pattern:

- When one is missing, all five are. We only need look at one of the variables to understand what is missing.
- The data document tells us that “to protect victim identities, rape and sex crime offenses are not geocoded”.

```
#For cases with missing geolocation (using Lat here), how they distribute over the code ClassCode

crime_df%>%select(OffenseDesc,Lat)%>%filter(is.na(Lat))%>%filter(!is.na(OffenseDesc))%>%group_by(OffenseD
esc)%>%dplyr::summarise(count=n())->tmp1

crime_df%>%select(OffenseDesc)%>%filter(!is.na(OffenseDesc))%>%group_by(OffenseDesc)%>%dplyr::summarise(t
otalcnt=n())->tmp2

merge(tmp1,tmp2,by.x="OffenseDesc",by.y="OffenseDesc",all.x=TRUE)->match_byOD
match_byOD%>%mutate(percentage=100*count/totalcnt)->match_byOD

ggplot(match_byOD, aes(reorder(OffenseDesc,percentage), percentage)) +
geom_col() + theme(axis.text = element_text(size=4))+
coord_flip() +scale_x_discrete(label=function(x) abbreviate(x, minlength=15))+
xlab("Offense Category") +
ylab("Percentage") +
ggtitle("Cases Missing Location")
```



As shown above, identity protection covers all Sex Crimes and Rape. The next few categories may have reasons for missing the location by policy (certain cases of negligent homicide), or just by difficulty in identifying the location. Since the type of crime seems to be related to missing location data, we can conclude these are “missing not at random” and should bear that in mind that particular crimes will not appear on any map or geographic analysis.

Main Analysis

Crime in New York is a rich and varied topic, a topic which allowed us to explore many angles and use many techniques. Below, you will find graphs, and the insights derived from them, organized around the concepts of:

- **What**,
- **Where**,

- **When** and
- **Why**.

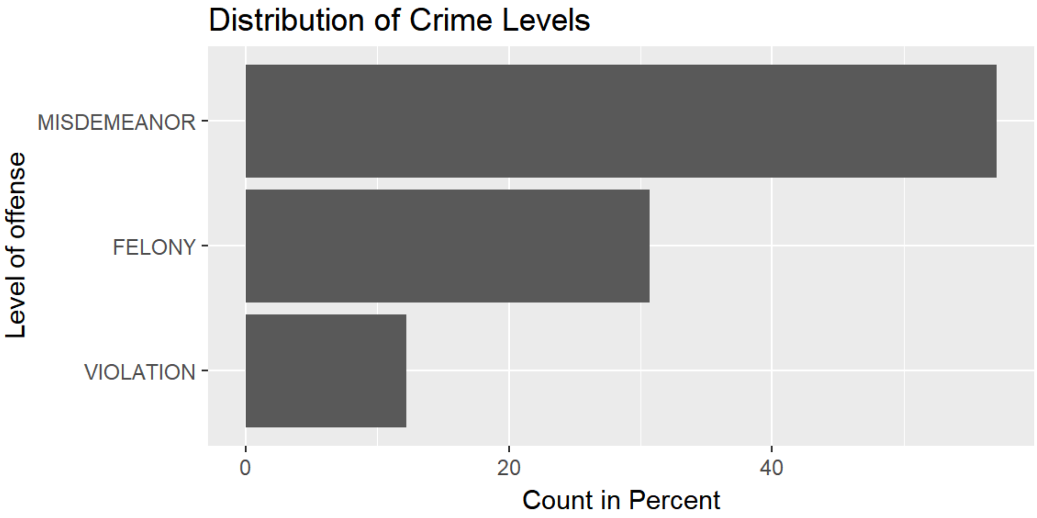
What: What kinds of crimes have been happening?

Breakdown of crime by Level (Misdemeanors, Felonies, then Violations)

First, we will examine the broadest categorization of crime: the legal level of the offense. In the raw dataset, this field is called "LAW_CAT_CD", and we have shortened that to **Level**.

```
# Remove Invalid Dates
crime_df <- crime_df %>% filter(year(DateStart)>2005)
```

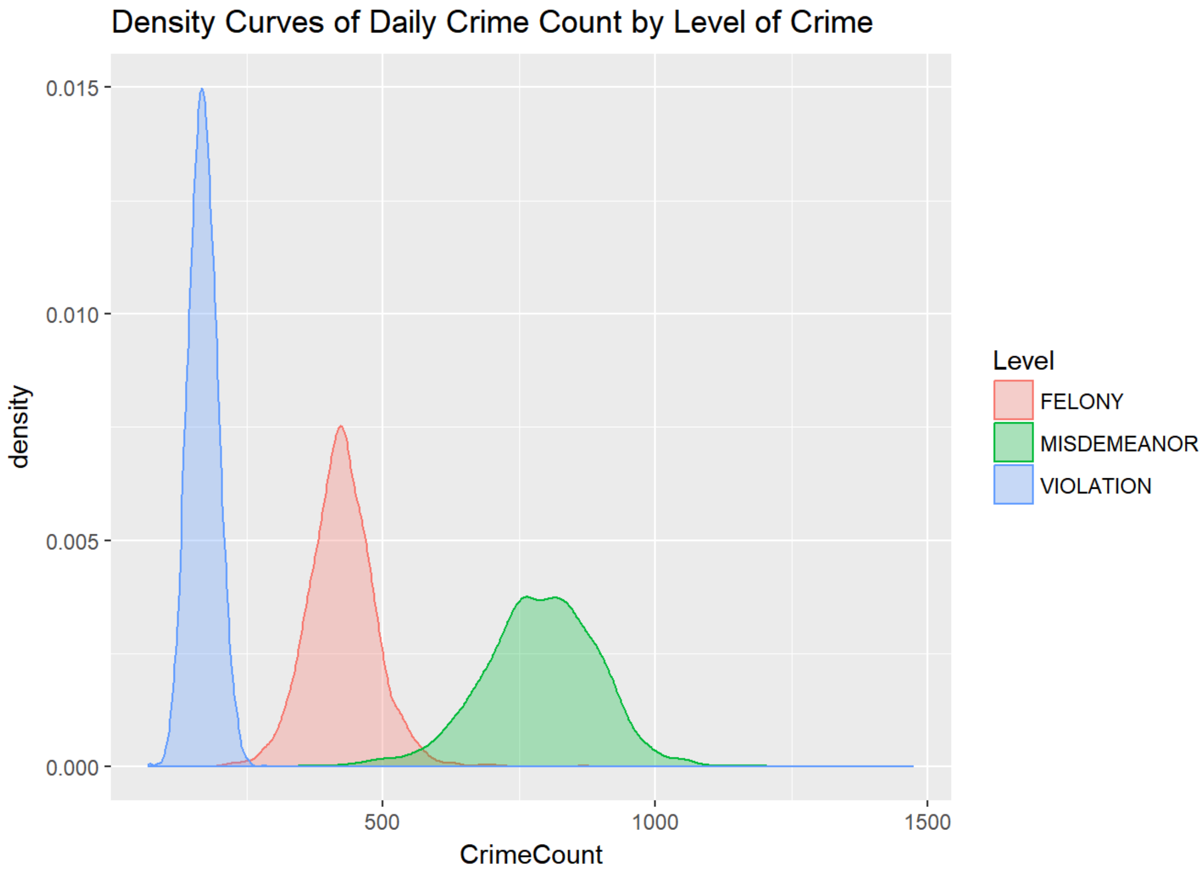
```
crime_level <- crime_df %>%
                  group_by(Level) %>%
                  summarize(count=n()) %>%
                  mutate(freq= count/sum(count)*100)

ggplot(crime_level, aes(reorder(Level, freq), freq)) +
   geom_bar(stat="identity") +
   coord_flip() +
   xlab("Level of offense") +
   ylab("Count in Percent") +
   ggtitle("Distribution of Crime Levels")
```



What we see in this graph is how much of crime is accounted for by Misdemeanors (around 58%), followed by Felonies (around 30%), and then Violations (12%).

Normal distribution of daily crime level of those levels

```
# see shape of the daily counts... normal?
daily_df <-crime_df %>%
              group_by(DateStart,Level) %>% summarize(CrimeCount=n())

ggplot(daily_df, aes(x=CrimeCount)) +
  geom_density(aes(group=Level, color=Level, fill=Level), alpha=0.3) +
  ggtitle("Density Curves of Daily Crime Count by Level of Crime")
```

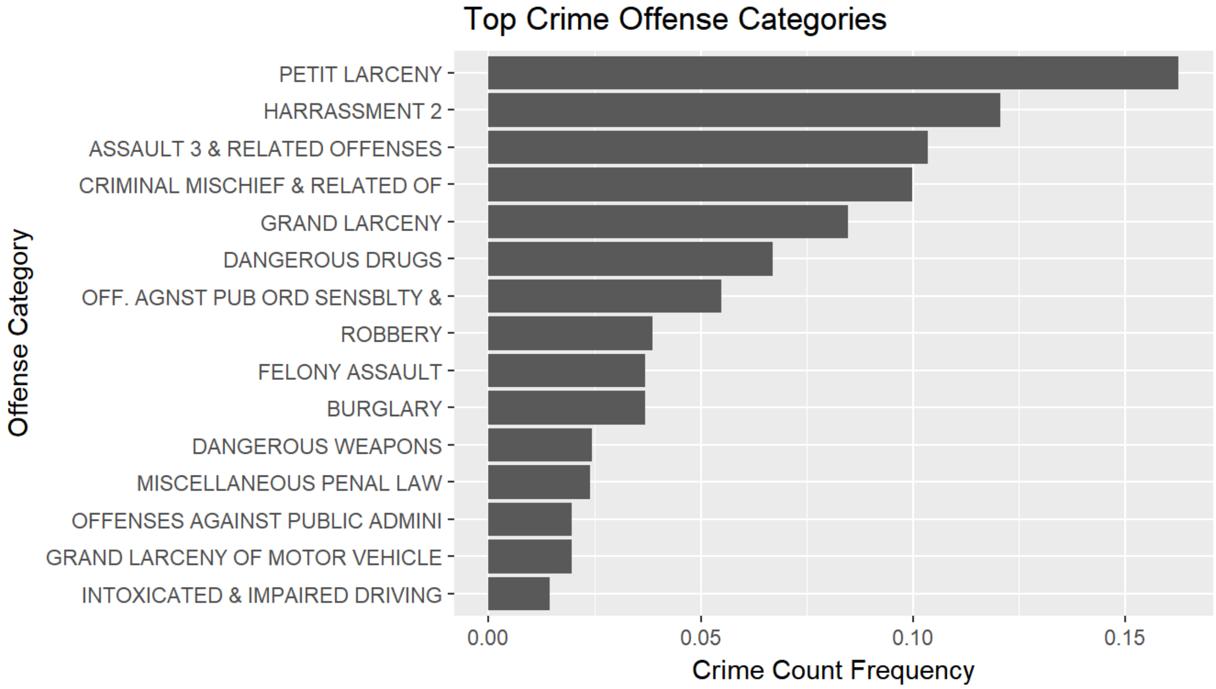


Another view of the Level of crimes is on a daily basis. We can see how the more numerous the Level of crime is, the more variation there is on a daily basis. All three Levels seem normally distributed, although the Misdemeanor Level has a bit of a plateau at the top, and there is a very, very long tail to the Violation data, suggesting a small number of days with record violations (January 1, 2010, 1,473 violations were recorded. We will discuss the January 1 phenomenon below).

Leading types of crimes

Now let’s look a layer deeper. The dataset includes codes and descriptions that give us another level of granularity in the type of offense reported. The OffenseDesc (OFNS_DESC) tells us more.

```
crime_top <- crime_df %>%
                filter(OffenseDesc!="") %>%
                group_by(OffenseDesc) %>%
                summarize(count=n()) %>%
                mutate(rel_freq = count/sum(count)) %>%
                top_n(n=15, wt=count)

ggplot(crime_top, aes(reorder(OffenseDesc,rel_freq), rel_freq)) +
  geom_col() +
  coord_flip() +
  xlab("Offense Category") +
  ylab("Crime Count Frequency") +
  ggtitle(" Top Crime Offense Categories")
```

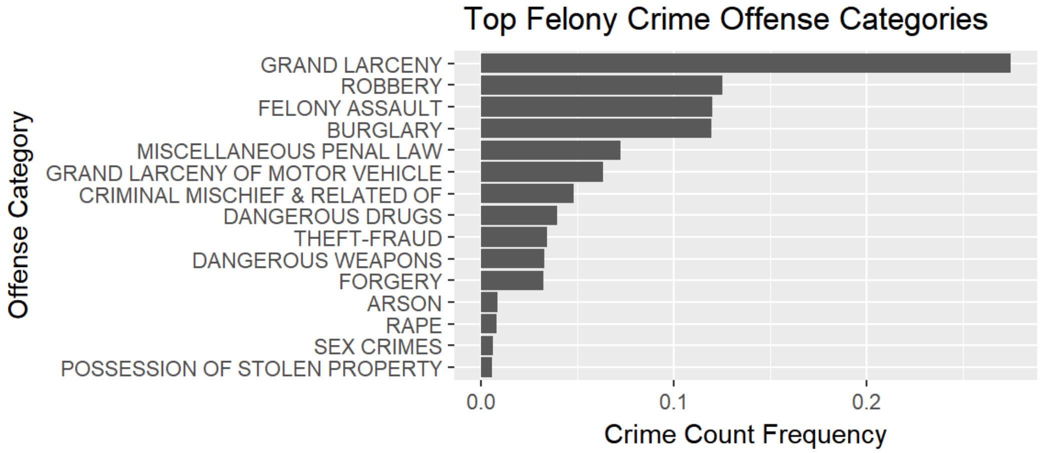


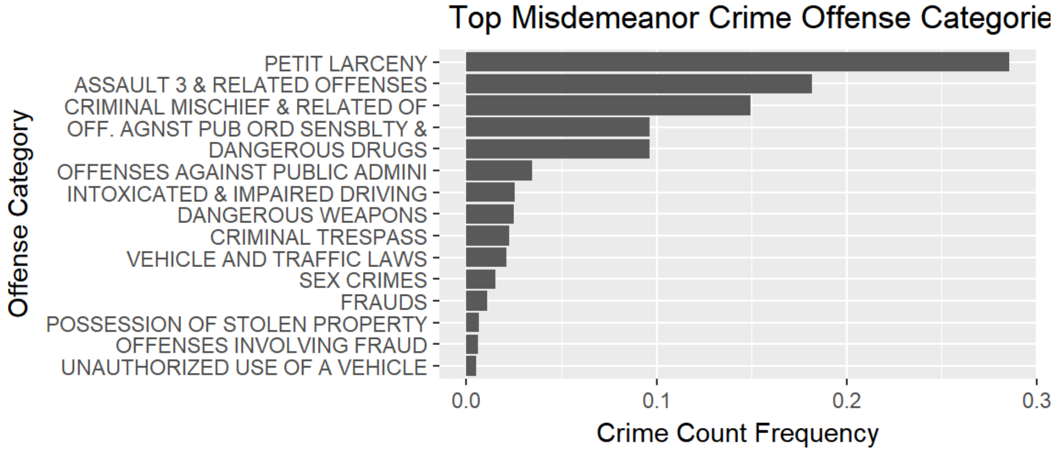
We can see from this graph that Petit Larceny accounts for the largest number of crimes in New York City, followed, but not very closely, by Harrassment 2, and then Assault 3 & Related Offenses.

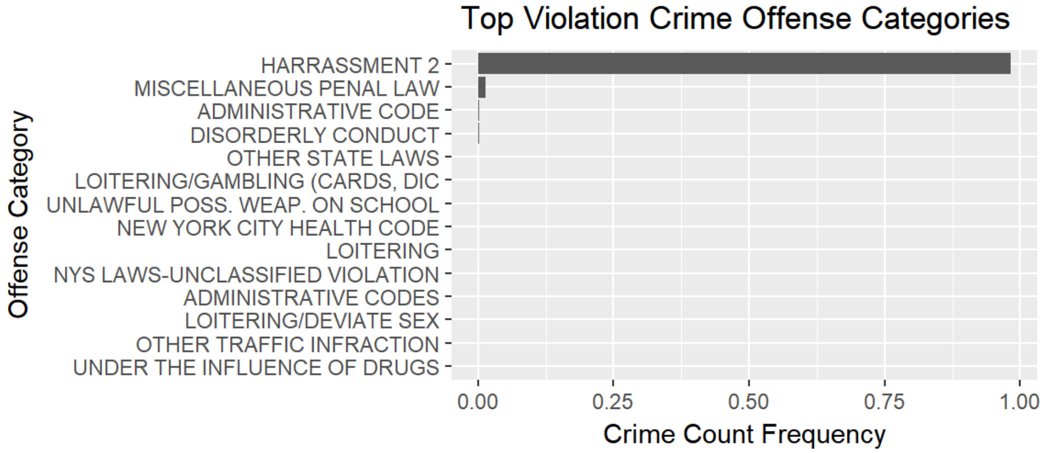
In examining this list, we had two things jump out at us: Assault, third on the list, and over 10% of all crime, seems pretty serious. We decided to take a closer look at all violent crimes because of this (and the fact that when you are concerned about crime, violent crime is the most frightening kind). Dangerous Drugs, sixth on this list, is another category of note, particularly with the way trends in drug abuse reach the news with alacrity, and a number of states have legalized use of Marijuana in recent years.

As such, we intend to examine those categories of crime in addition to trends by Level.

```
crime_top_felony <- crime_df %>%
  filter(OffenseDesc!="" & Level=="FELONY") %>%
  group_by(OffenseDesc) %>%
  summarize(count=n()) %>%
  mutate(rel_freq = count/sum(count)) %>%
  top_n(n=15, wt=count)

f <- ggplot(crime_top_felony, aes(reorder(OffenseDesc,rel_freq), rel_freq)) +
  geom_col() +
  coord_flip() +
  xlab("Offense Category") +
  ylab("Crime Count Frequency") +
  ggtitle(" Top Felony Crime Offense Categories")

crime_top_misd <- crime_df %>%
  filter(OffenseDesc!="" & Level=="MISDEMEANOR") %>%
  group_by(OffenseDesc) %>%
  summarize(count=n()) %>%
  mutate(rel_freq = count/sum(count)) %>%
  top_n(n=15, wt=count)

m <- ggplot(crime_top_misd, aes(reorder(OffenseDesc,rel_freq), rel_freq)) +
  geom_col() +
  coord_flip() +
  xlab("Offense Category") +
  ylab("Crime Count Frequency") +
  ggtitle(" Top Misdemeanor Crime Offense Categories")

crime_top_violation <- crime_df %>%
  filter(OffenseDesc!="" & Level=="VIOLATION") %>%
  group_by(OffenseDesc) %>%
  summarize(count=n()) %>%
  mutate(rel_freq = count/sum(count)) %>%
  top_n(n=15, wt=count)

v <- ggplot(crime_top_violation, aes(reorder(OffenseDesc,rel_freq), rel_freq)) +
  geom_col() +
  coord_flip() +
  xlab("Offense Category") +
  ylab("Crime Count Frequency") +
  ggtitle(" Top Violation Crime Offense Categories")

  grid.arrange(f,m,v )
```







Felonies start with Grand Larceny, Robbery, Felony Assault and Burglary, with Grand Larceny being far more frequent than any other crime. It is interesting to note how dispersed the types of Felonies get after the top 11 categories. Note that the violent crimes of Robbery, Felony Assault, and Rape appear on this list.

For Misdemeanors, we see Petit Larceny atop the list, with the top five categories accounting for 10% or more (of Misdemeanors) each. Note that Dangerous Drugs is split between Misdemeanors and Felonies, and there are Misdemeanor Assault crimes. The Violent Crimes categories on this chart include Assault 3.

This list of Violations is surprising. Following the Pareto principle, if you could stop all Harrassment (2), you'd solve for nearly all Violations! The small share of other violations is divided by a lot of categories, showing how infrequently we see them in this data.

Where: Where does crime take place?

The question of where crime happens has multiple perspectives of import:

- Where would I choose to live or work to avoid crime?

- Should we adjust policing strategies to try to reduce crime in high crime areas?

Borough Analysis

Like Level is to Crime as a whole, Borough is to The City of New York. The first question about location is the most macro: how does crime differ in the five Boroughs?

- **Total vs. Per Capita by Level**

Please visit our D3 Mosaic Plots (https://bl.ocks.org/CrimeDataNyc/raw/e3362fde2a5f94aa2fa94a524742a566/441a18c0a089ca75e03009376006eff10dd38eb8/) and review the three choices there. You should move your mouse over each block to reveal both Level and Borough.

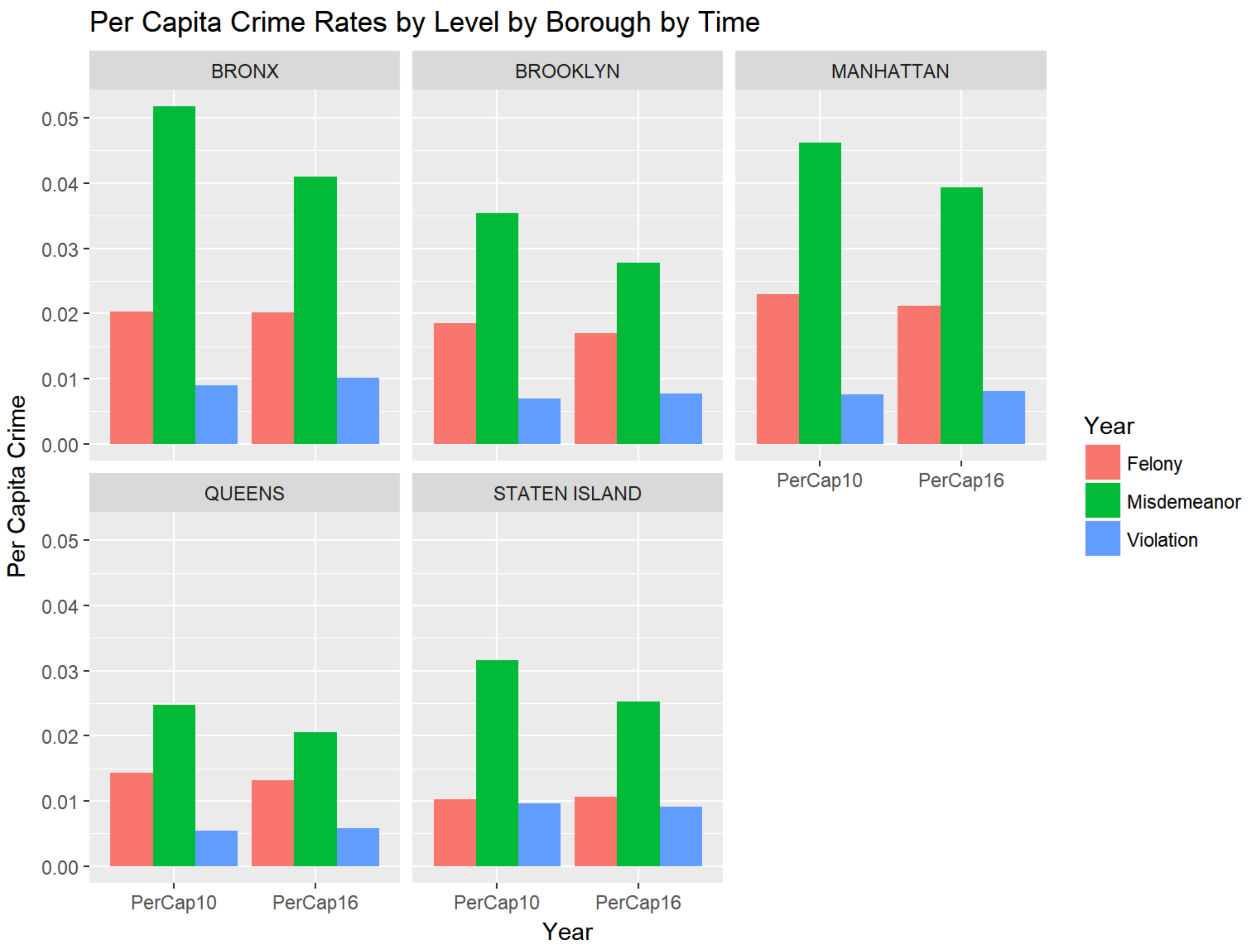
The first thing we can see is that there is more crime in Brooklyn than any other Borough, and the amount of crime in Staten Island is very small. However, crime is committed by people, and the number of people in each Borough is different. Hence, when you click for either the 2010 or 2016 Per Capita Mosaics, you notice thatthe Per Capita view shows a rather different story of crime. Staten Island, due to its small population, actually has a higher crime **rate** than some other boroughs. In fact, we see less crime per capita in Queens and Brooklyn than the overall crime totals would have us understand.

Crime by Level, Borough and Time

We can look at both Level and Time, comparing 2010 Per Capita to 2016 Per Capita.

```
# bring in Borough Population and massage it
bdf <- fread("../Data_Files/BoroughPop.csv")
bdf <- bdf[1:6,]
bdf$Boro <- c("TOTAL","BRONX","BROOKLYN","MANHATTAN","QUEENS","STATEN ISLAND")

# summarize for mosaic, per capita plots
df_bsum <-crime_df %>%
  filter(!is.na(Boro)) %>%
  group_by(Boro,Level) %>%
  summarize(Freq = n())

# merge in the borough population
df_bsum <- merge(df_bsum, bdf, by="Boro")

# per capita calculation
df_bsum$PerCap <-df_bsum$Freq/df_bsum$`2016 Estimate`


#need to rename the bdf Boro in order to make the merge work
colnames(bdf)[colnames(bdf)=="Boro"] <- "Boro"

# limit to specific years of the population data and test
# start with 2010
# summarize for mosaic, per capita plots
df_bsum2010 <-crime_df %>%
  filter(!is.na(Boro)) %>%
  filter(DateStart > "2009-12-31" & DateStart < "2011-01-01") %>%
  group_by(Boro,Level) %>%
  summarize(Freq = n())

# merge in the borough population
df_bsum2010 <- merge(df_bsum2010, bdf, by="Boro")

# per capita calculation
df_bsum2010$PerCap <-df_bsum2010$Freq/df_bsum2010$`2010 Population`

#2010 mosaic
colnames(df_bsum2010)[colnames(df_bsum2010)=="Freq"] <- "Count"
colnames(df_bsum2010)[colnames(df_bsum2010)=="PerCap"] <- "Freq"

# now 2016 Estimate
# summarize for mosaic, per capita plots
df_bsum2016 <-crime_df %>%
  filter(!is.na(Boro)) %>%
  filter(DateStart > "2015-12-31" & DateStart < "2017-01-01") %>%
  group_by(Boro,Level) %>%
  summarize(Freq = n())

# merge in the borough population
df_bsum2016 <- merge(df_bsum2016, bdf, by="Boro")

# per capita calculation
df_bsum2016$PerCap <-df_bsum2016$Freq/df_bsum2016$`2016 Estimate`

# By Per Capita -- you have to have "Freq" be the column for the thing the Mosaic will use for frequency,
so
# for Per Capita, you need to swap the Freq column names
```

```
#2016
colnames(df_bsum2016)[colnames(df_bsum2016)=="Freq"] <- "Count"
colnames(df_bsum2016)[colnames(df_bsum2016)=="PerCap"] <- "Freq"
#mosaic(Level~Boro,df_bsum2016, direction=c("v","h"), main="2016 Crime per Capita by Borough by Level")

#Plot 2010 year over 2016 year by Borough
colnames(df_bsum2010)[colnames(df_bsum2010)=="Freq"] <- "PerCap10"
colnames(df_bsum2016)[colnames(df_bsum2016)=="Freq"] <- "PerCap16"
df_bsum.pcap <- merge(df_bsum2010,df_bsum2016, by=c("Boro","Level"))
df_bsum.pcap$Count.y <- NULL
df_bsum.pcap$Borough.y <- NULL
df_bsum.pcap$"2010 Population.y" <- NULL
df_bsum.pcap$"2016 Estimate.y" <- NULL

tidy_bsum <- tidyr::gather(df_bsum.pcap, key="Year", value="PerCap", -"Boro", -"Level", -"Count.x", -"201
0 Population.x", -"2016 Estimate.x", -"Borough.x")

library(ggplot2)
ggplot(tidy_bsum, aes(x=Year, y=PerCap, fill=Level))+
  geom_bar(stat="identity",position="dodge") +
  scale_fill_discrete(name="Year",
  #breaks=c(1, 2),
  labels=c("Felony", "Misdemeanor","Violation")) +
  xlab("Year")+ylab("Per Capita Crime") +
  facet_wrap(~Boro) +
  ggtitle("Per Capita Crime Rates by Level by Borough by Time")
```



From these graphs, we see how there is an apparent drop in the rate of crime between 2010 and 2016, mostly driven by Misdemeanors (in every Borough, but most predominantly in the Bronx). We can see that the Felony rate has been mostly unchanged, except in Manhattan. Violations have gone up in every Borough except Staten Island. (We will explore more about crime-over-time in the When section.)

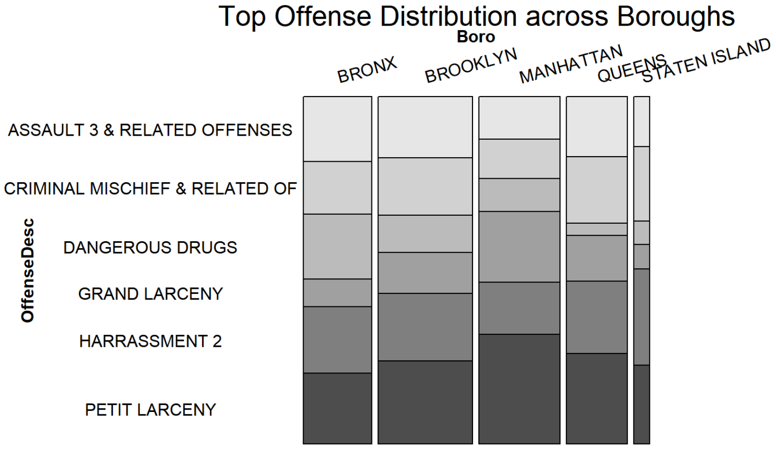
Crime over two time periods

We also found that there are some notable differences between Boroughs in the Per Capita Crime Rates. Manhattan leads the way on Felonies, followed by the Bronx. The top two are in opposite order for Misdemeanors. But it is Staten Island and the Bronx that lead in Violations. For lowest rates, Staten Island is lowest for Felonies, followed by Queens, with Queens lowest for Misdemeanors and Violations. Hence, the crime profile is quite different in each Borough.

Main Categories of Crime by Borough

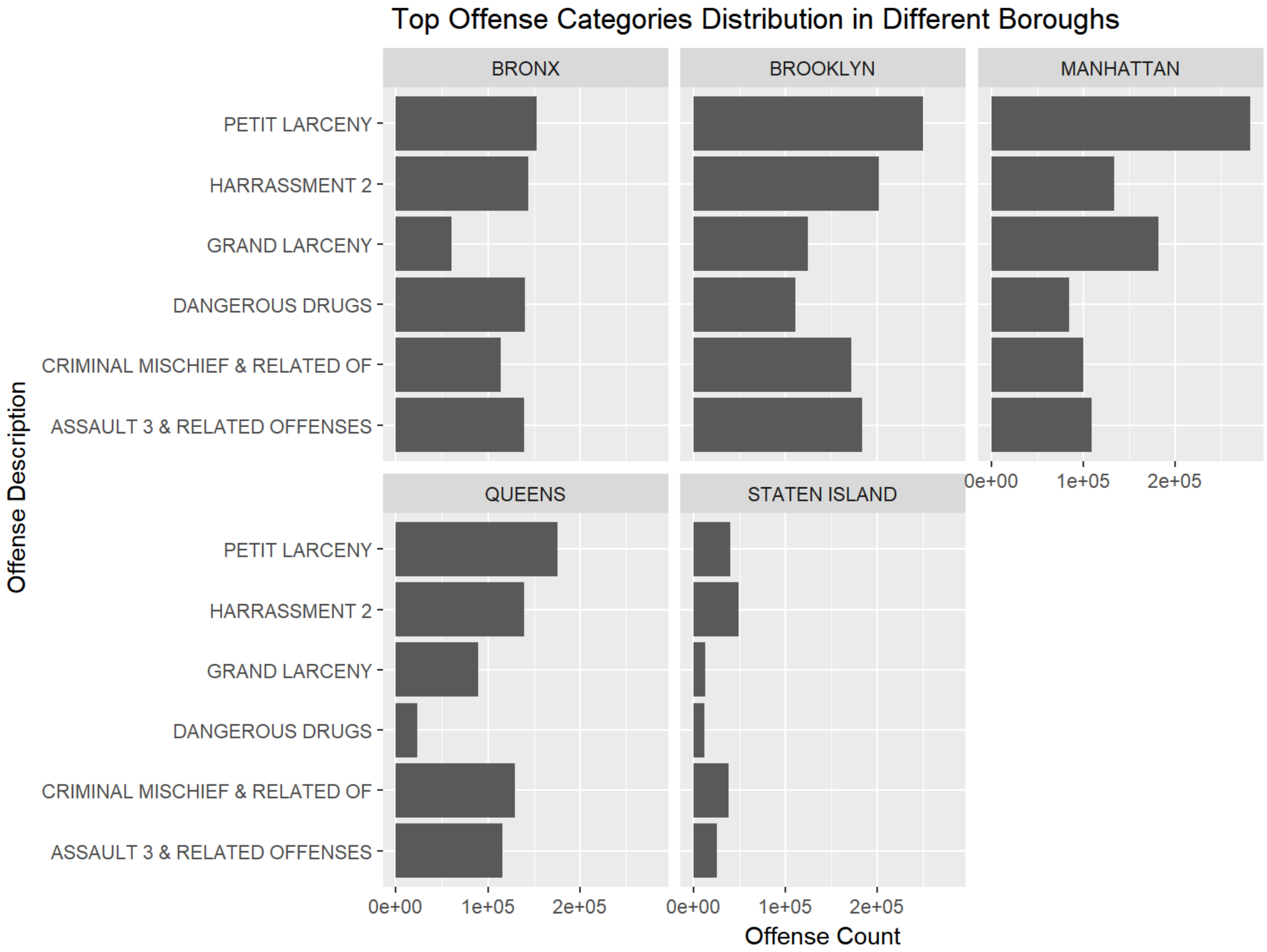
We can also get more specific about the top 6 types of crime, in terms of crime in the Boroughs.

```
  top_ofns   <- c("PETIT LARCENY", "HARRASSMENT 2", "CRIMINAL MISCHIEF & RELATED OF", "ASSAULT 3 & RELATED OFFENSES", "GRAND LARCENY", "DANGEROUS DRUGS")
  label_list   <- c("PETIT LARCENY", "HARRASSMENT 2", "CRIMINAL MISCHIEFF", "ASSAULT 3" , "GRAND LARCENY", "DANGEROUS DRUGS")

  crime_sort <- crime_df %>%
                filter(Boro != "") %>%
                filter(OffenseDesc %in% top_ofns) %>%
                group_by(Boro,OffenseDesc) %>%
                summarize(Freq=n()) %>%
                mutate(rel_freq = Freq/sum(Freq))

  crime_sort$OffenseDesc <- factor(crime_sort$OffenseDesc)

  mosaic(OffenseDesc~Boro, main ="Top Offense Distribution across Boroughs", direction=c("v"), labeling=labeling_border(rot_labels=c(15,0,0, 0), offset_labels = c(0,0,0,6.5), offset_varnames = c(1,0,0,11.6), just_labels=c("left", "left", "left", "center")), crime_sort)
```



Several observations can be made from this view:
1. Dangerous Drugs take up a disproportionately high share of crime in the Bronx, while in Queens, Dangerous Drugs account for far fewer of the crimes.
2. Larceny, both Petit and Grand, are a large share of crime in Manhattan when compared to other Boroughs.
3. Harassment 2 is the most prevalent of these top six in Staten Island

```
#doubledecker(TOP_OFFENSE~Boro, data=crime_sort)
ggplot(crime_sort, aes(OffenseDesc,Freq)) +
  geom_col() +
  ylab("Offense Count") +
  xlab("Offense Description") +
  facet_wrap(~ Boro) +
  coord_flip() +
  ggtitle(" Top Offense Categories Distribution in Different Boroughs")
```



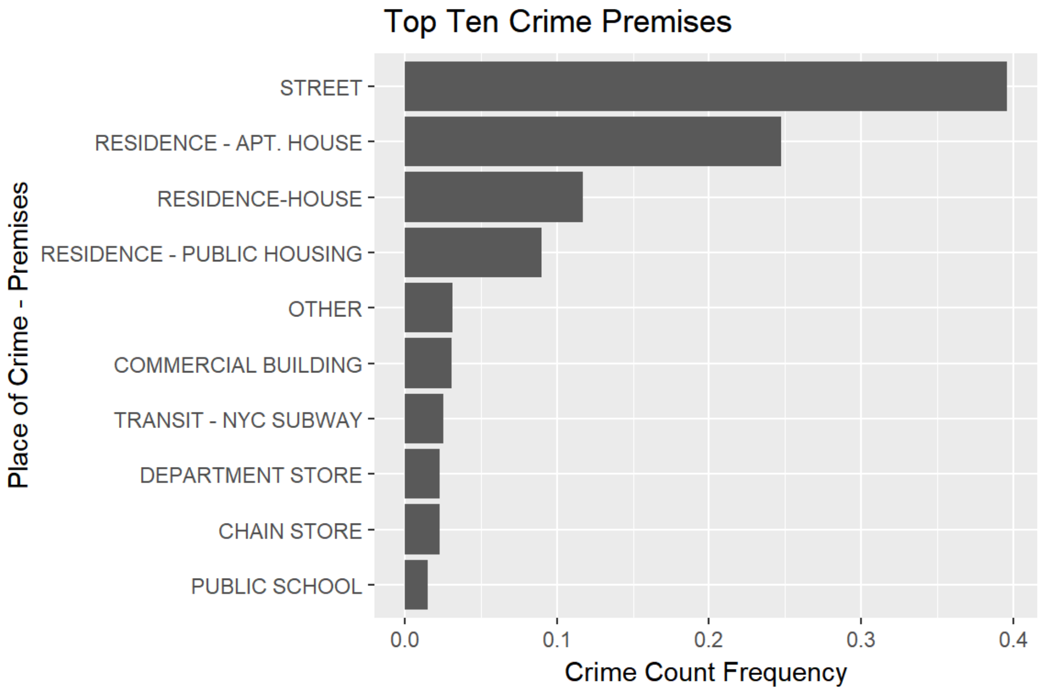
This view makes the data easier to compare these types of crime within Boroughs. Brooklyn, for instance, has the most even distribution of crime across these categories, while Manhattan has a much higher proportion of Petit Larceny than any of the other categories here.

Drugs and Violent Crime: This graph also reinforces how significantly Dangerous Drugs factor into crime in the Bronx, given how that category is nearly as prevalent as any other. It's also worth noting where Assault factors in for each Borough: In Manhattan, it's vastly outnumbered by each type of Larceny, relatively high in the Bronx and Queens, and sort of in the middle in Brooklyn.

Location Analysis

The dataset includes a field called "PREM_TYP_DESC" (which we have shortened to **PremDesc**) to indicate where the crime took place, such as on the street, in a house, etc. It turns out that about 85% of crime occurs on the Streets or within a Residence of one type or another.

```
crime_place <- crime_df %>%
               filter(!is.na(PremDesc),Level !="") %>%
               group_by(PremDesc) %>%
               summarize(count=n()) %>%
               top_n(n=10, wt=count) %>%
               mutate(rel_freq = count/sum(count))


ggplot(crime_place, aes(fct_reorder(PremDesc, rel_freq), rel_freq)) +
  geom_bar(stat="identity") +
  coord_flip() +
  ylab("Crime Count Frequency") +
  xlab("Place of Crime - Premises") +
  ggtitle(" Top Ten Crime Premises")
```



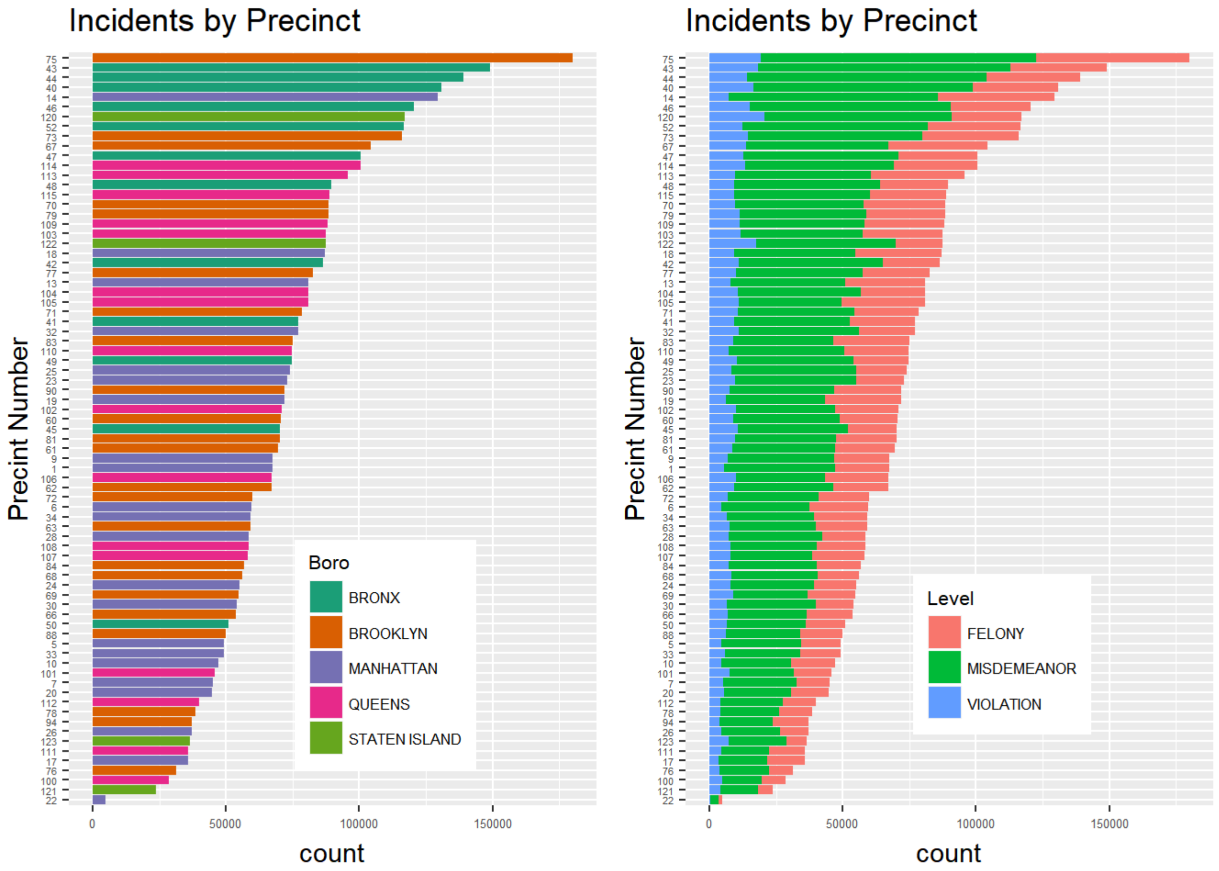
Here we see how crime on the Street is the largest, single category (nearly 40%), but if you add together the various Residence categories, Residences total 45%. This vastly outnumbers the Subway, Commercial buildings, etc. While Felonies and Misdemeanors follow the same pattern as crime, overall, more Violations happen in Apartments than on the street.

Violent Crimes and **Dangerous Drugs** Crimes also follow the same pattern as Crime overall.

Precinct differences

In terms of *where*, we can also look at the police precincts.

```
#matching Pct with Boro
crime_df %>% select(Level,Pct)%>%group_by(Level,Pct)%>%
  drop_na()%>%dplyr::summarize(count = n())%>%ungroup()->df_pct
merge(df_pct,match_pct_boro,by.x="Pct",by.y="Pct")->df_pbl
df_pbl%>%mutate(Pct=as.factor(Pct))->df_pbl

df_pbl%>%ggplot(aes(reorder(Pct, count), count,fill=Boro)) + geom_bar(stat = "identity") + xlab("Precint
 Number") + ggtitle("Incidents by Precinct") + coord_flip()+
  #scale_fill_manual(values = c("red","orange","yellow","green","blue","violet"))+
  scale_fill_brewer(palette="Dark2") +
  theme(axis.text.x = element_text(size=5, hjust = 0.5),axis.text.y=element_text(size=4),legend.position=
c(0.6,0.2),
        legend.text=element_text(size=6,hjust=0.5),legend.title=element_text(size=8),legend.key = element
_rect(size = 0.5),legend.key.size = unit(1, 'lines'))->p5
df_pbl%>%ggplot(aes(reorder(Pct, count), count,fill=Level)) + geom_bar(stat = "identity") + xlab("Precint
Number") + ggtitle("Incidents by Precinct") + coord_flip()+theme(axis.text.x = element_text(size=5, hjust
= 0.5),axis.text.y=element_text(size=4),legend.position=c(0.6,0.2),
        legend.text=element_text(size=6,hjust=0.5),legend.title=element_text(size=8),legend.key = element
_rect(size = 0.5),legend.key.size = unit(1, 'lines'))->p6
grid.arrange(p5,p6,nrow=1)
```



The graphs above shows that crime is not evenly distributed across police precincts. You can see how the Bronx has crime consolidated into a smaller number of precincts, while Manhattan has crime distributed across more numerous precincts. All Boroughs seem to have a precinct or two that handle far more crime, perhaps twice as much crime, as other precincts in the same Borough.

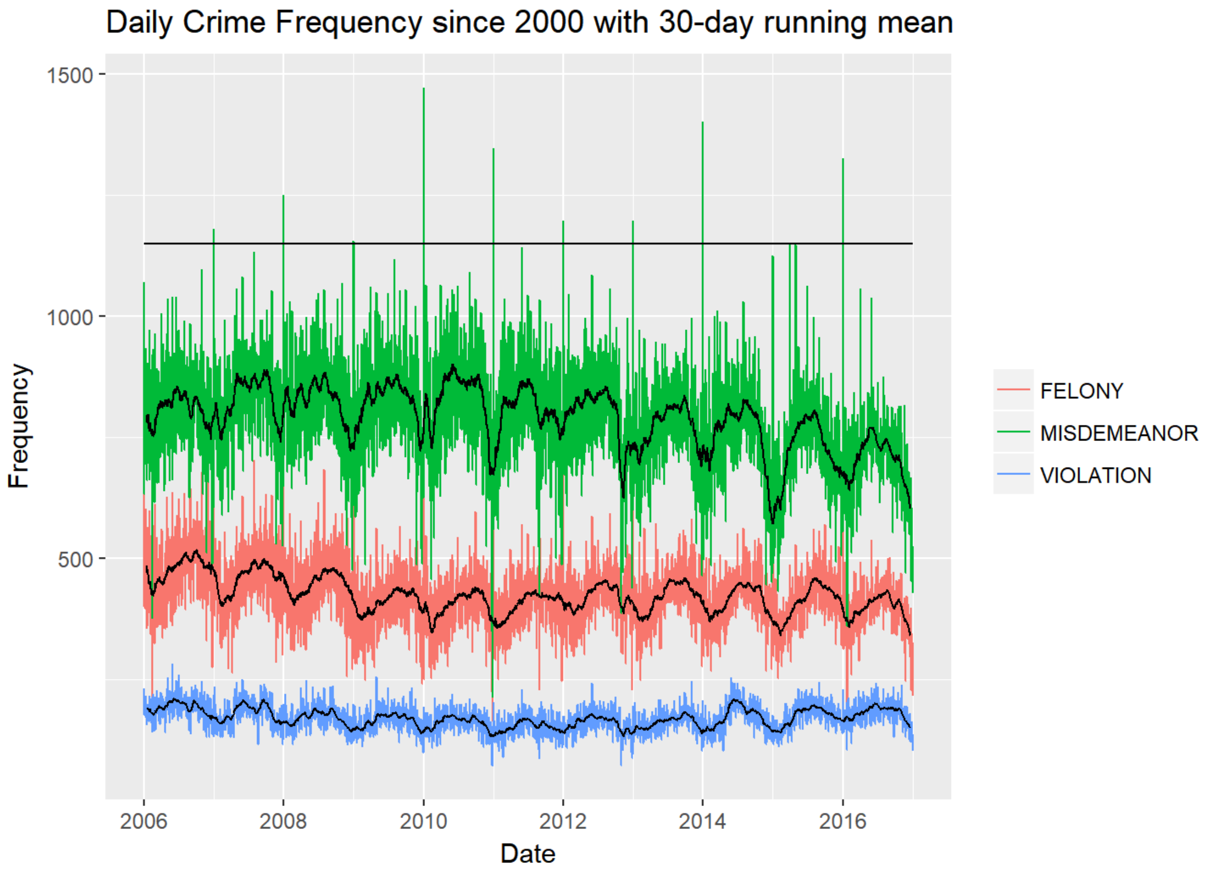
When: has crime changed over time? When, during the year, week or day, does crime take place?

General Decreasing Trend over 2006 to 2016

The total amount of crime has decreased over the past ten years.

```
#picking non-missing DateStart and filter only those after "2006-01-01", 5560408 obs.
crime_df%>%select(DateStart,Level)%>%filter(!is.na(DateStart))%>%filter(DateStart>=as.Date("2006-01-01"))
->df_Date
```

```
#time series of daily frequency of 3 crime categories 2006-2016
df_Date%>%group_by(DateStart,Level)%>%dplyr::summarise(count=n())%>%ungroup()%>%group_by(Level)%>%mutate
(mon_mean=rollmean(count,30,fill=NA))%>%ungroup()->byDateLawMean

#daily rate
byDateLawMean%>%ggplot()+
  geom_line(aes(DateStart,count,color=Level))+
  geom_line(aes(DateStart,mon_mean,group=Level))+
  ggtitle("Daily Crime Frequency since 2000 with 30-day running mean")+
  labs(x="Date",y="Frequency")+theme(legend.title=element_blank())+geom_line(aes(DateStart,count*0+1150))
```



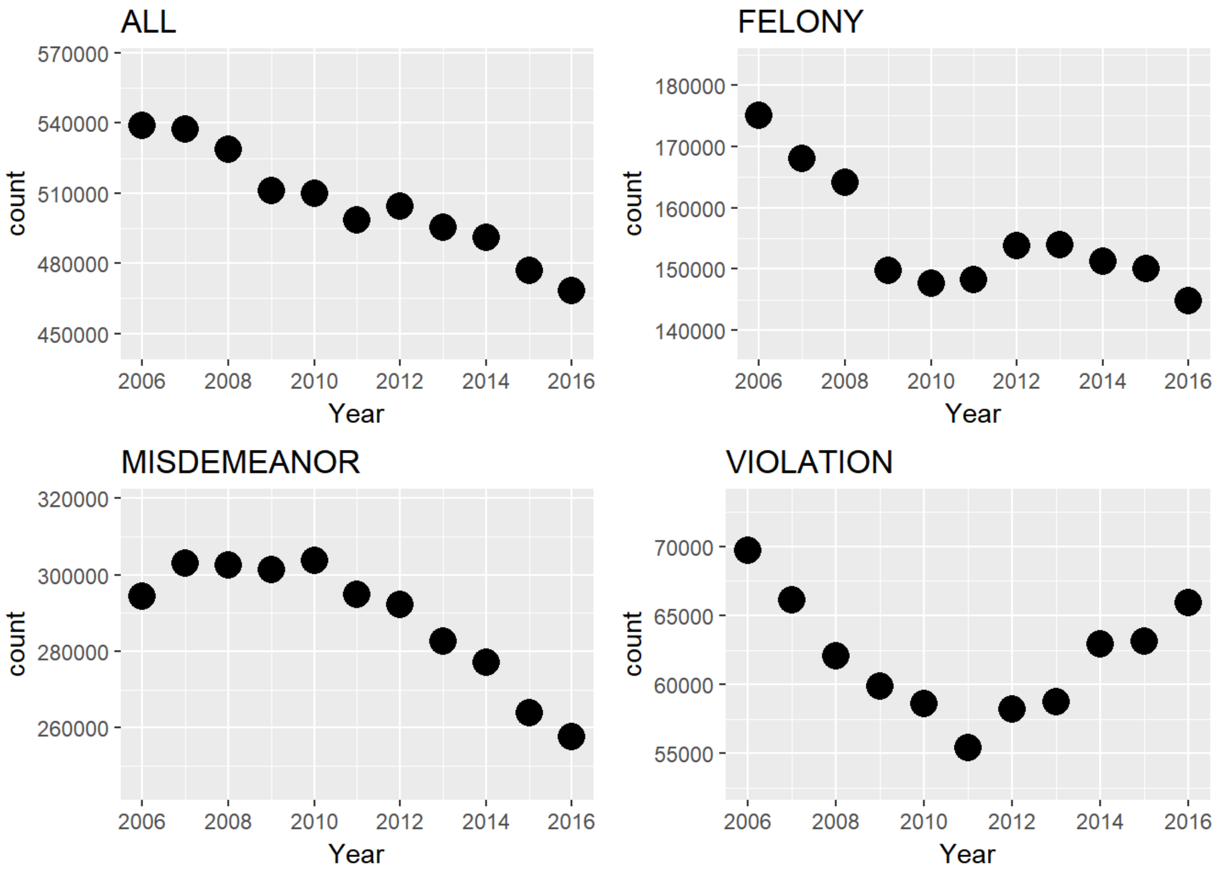
One view is this daily chart. It shows a cyclical, annual pattern, as well as an overall downward trend. We also see strange spikes around the start of every year. These spikes are, in fact, on January 1, and we see that the highest number of crimes across the entire dataset are on January 1 through the years.

This could be a data quality issue: where the date of the crime is unknown, it ends up being labelled as the first of the year. This could also be a true phenomenon associated with New Year’s Day, with more crime happening on that day. It could be some of each. We have left it as it stands in the dataset, but it shows up in the following graphs because of the large outliers that they are.

If we group the data into annual volumes, the trends become a little clearer, still.

```
crime_df%>%select(DateStart)%>%mutate(Year=year(DateStart))%>%filter(Year>=2006)%>%group_by(Year)%>%dply
r::summarise(count=n())->totalCntBySD
ggplot(totalCntBySD)+geom_point(aes(count,Year),size=5)+xlim(min(totalCntBySD$count)*0.95,max(totalCntByS
D$count)*1.05)+coord_flip()+ggtitle("ALL")->p1

crime_df%>%filter(Level=="FELONY")%>%select(DateStart)%>%mutate(Year=year(DateStart))%>%filter(Year>=2006
)%>%group_by(Year)%>%dplyr::summarise(count=n())->totalCntBySD
ggplot(totalCntBySD)+geom_point(aes(count,Year),size=5)+xlim(min(totalCntBySD$count)*0.95,max(totalCntByS
D$count)*1.05)+coord_flip()+ggtitle("FELONY")->p2

crime_df%>%filter(Level=="MISDEMEANOR")%>%select(DateStart)%>%mutate(Year=year(DateStart))%>%filter(Year>
=2006)%>%group_by(Year)%>%dplyr::summarise(count=n())->totalCntBySD
ggplot(totalCntBySD)+geom_point(aes(count,Year),size=5)+xlim(min(totalCntBySD$count)*0.95,max(totalCntByS
D$count)*1.05)+coord_flip()+ggtitle("MISDEMEANOR")->p3

crime_df%>%filter(Level=="VIOLATION")%>%select(DateStart)%>%mutate(Year=year(DateStart))%>%filter(Year>=2
006)%>%group_by(Year)%>%dplyr::summarise(count=n())->totalCntBySD
ggplot(totalCntBySD)+geom_point(aes(count,Year),size=5)+xlim(min(totalCntBySD$count)*0.95,max(totalCntByS
D$count)*1.05)+coord_flip()+ggtitle("VIOLATION")->p4
```

```
grid.arrange(p1,p2,p3,p4,nrow=2)
```



(We have zoomed in on this graph to more easily detect the differences, noting that we're looking at about 12% of the range of that axis.)

We can see that the total crime level has decreased every year except 2012, but 2012 didn't rise above 2010. We do see an overall decreasing trend.

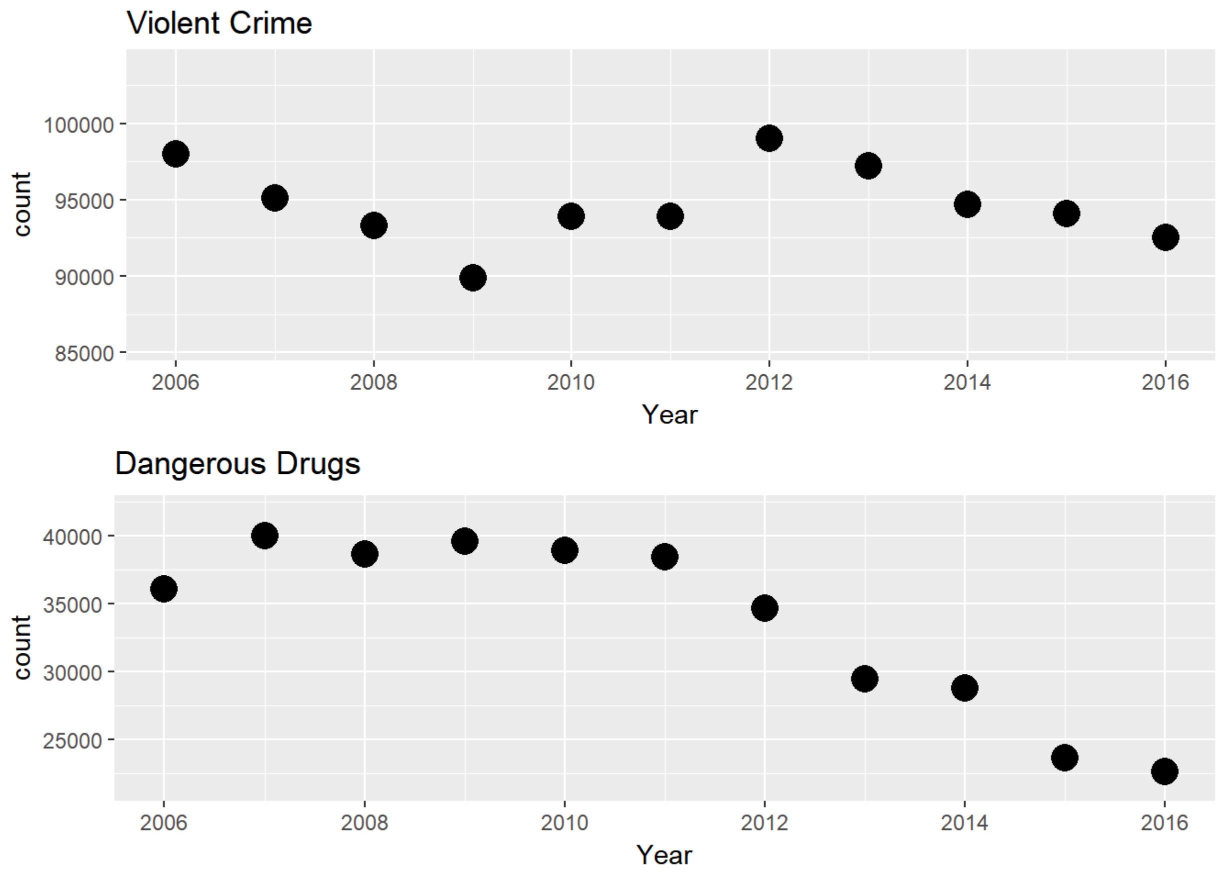
Felonies dropped considerably from 2006 to 2009, but then seemed to stabilize at a new level, approximately 12% lower than 2006.

Misdemeanors experienced a similar drop, but later. Misdemeanors held steady through 2010, then dropped in five straight years to reach, again, about a 12% drop by 2015. It is almost as if the police were focusing on Felonies, gained some, then turned their attention to Misdemeanors, but since the volume of Misdemeanors is nearly twice that of Felonies, the police would have to be twice as efficient in stopping Misdemeanors.

Violations dropped but then went back up. It seems that violations may be a matter of enforcement more than trends in attempts, so this could be related to the enforcement efforts on reducing Felonies and Misdemeanors.

```
crime_df%>%filter(OffenseDesc == "DANGEROUS DRUGS")%>%select(DateStart)%>%mutate(Year=year(DateStart))%>%
filter(Year>=2006)%>%group_by(Year)%>%dplyr::summarise(count=n())->totalCntBySD
ggplot(totalCntBySD)+geom_point(aes(count,Year),size=5)+xlim(min(totalCntBySD$count)*0.95,max(totalCntByS
D$count)*1.05)+coord_flip()+ggtitle("Dangerous Drugs")->p1

crime_df%>%filter(OffenseDesc == "ASSAULT 3 & RELATED OFFENSES" |
                                OffenseDesc == "FELONY ASSAULT" |
                                OffenseDesc == "RAPE" |
                                OffenseDesc == "ROBBERY" |
                                IntOffenseDesc == "AGGRAVATED SEXUAL ASBUSE" |
                                IntOffenseDesc == "ASSAULT 2,1,UNCLASSIFIED" |
                                IntOffenseDesc == "ASSAULT 3" |
                                IntOffenseDesc == "RAPE 1" |
                                IntOffenseDesc == "ROBBERY,OPEN AREA UNCLASSIFIED" |
                                IntOffenseDesc == "SEXUAL ABUSE" |
                                IntOffenseDesc == "SEXUAL ABUSE 3,2")%>%% select(DateStart)%>%mutate(Year=
year(DateStart))%>%filter(Year>=2006)%>%group_by(Year)%>%dplyr::summarise(count=n())->totalCntBySD
ggplot(totalCntBySD)+geom_point(aes(count,Year),size=5)+xlim(min(totalCntBySD$count)*0.95,max(totalCntByS
D$count)*1.05)+coord_flip()+ggtitle("Violent Crime")->p2
```

```
grid.arrange(p2,p1,nrow=2)
```



When we focus on **Violent Crime**, we see some features from the Felony graph (the drop from 2006 to 2009, the spike at 2012), but most years have a level around 93,000 to 94,000 such crimes which make the other years look almost like outliers, rather than part of a trend.

Dangerous Drugs, however, follows the pattern we saw in the Misdemeanor case: after 2012, the yearly rate is lower. Progress seems to have been made.

Time of year and Time of day:

Again, we'll work from the Macro (years) to the Micro. These graphs will tell us more specifics about when crime happens more (or less).

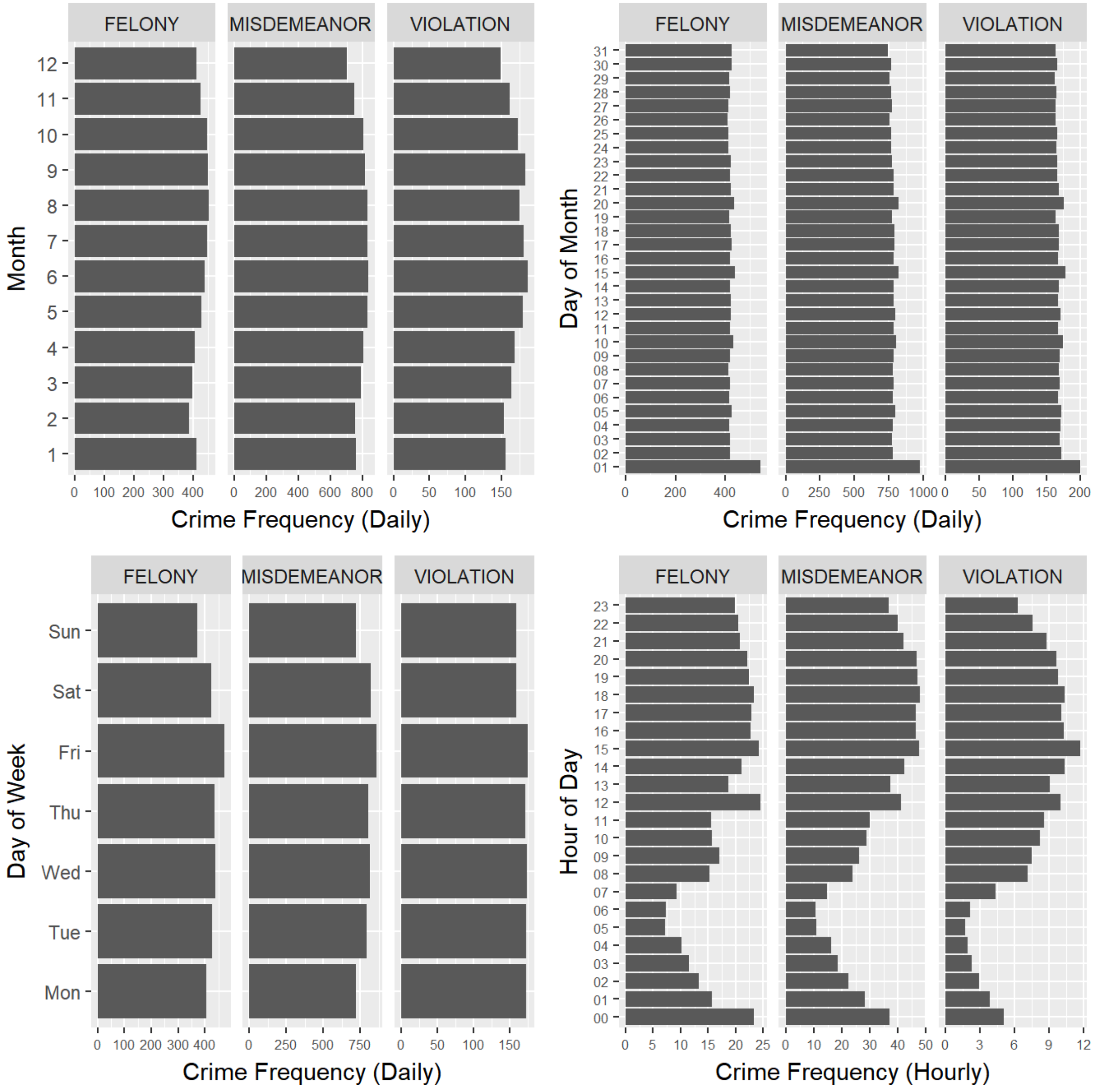
```
#frequency by month
crime_df%>%select(DateStart,Level)%>%filter(!is.na(DateStart))%>%filter(DateStart>=as.Date("2006-01-01"))->df_Date

df_Date%>%mutate(Month=as.character(month(DateStart)))%>%group_by(Month,Level)%>%dplyr::summarise(CntByMon=n())->byDateLaw_mon

byDateLaw_mon%>%mutate(Days=rep(31,3))%>%mutate(Days=ifelse(Month=="2",28,Days))%>%mutate(Days=ifelse(Month %in% c("4","6","9","11"),30,Days))->byDateLaw_mon
byDateLaw_mon%>%ggplot(aes(fct_relevel(Month,"10","11","12",after=9),CntByMon/Days/11))+geom_bar(stat="identity")+theme(axis.text.x = element_text(size=6))+coord_flip()+ylab("Crime Frequency (Daily)")+facet_wrap(~Level,scales="free_x")+xlab("Month")->p1
```

```
#frequency by day
df_Date%>%mutate(Day=as.factor(format(DateStart,"%d")))%>%group_by(Day,Level)%>%dplyr::summarise(CntByDay=n())->byDateLaw_day

#Day1-28 has the same total cnts=11yrs*12cnts/yr
#Day 29 cnts=11yrs*11cnts/yr+3cnts (leap yrs)
#Day 30 cnts=11*11; Day 31 cnts=7*11
byDateLaw_day%>%mutate(cnts=rep(12*11,3))%>%mutate(cnts=ifelse(Day=="29",11*11+3,cnts))%>%mutate(cnts=ifelse(Day=="30",11*11,cnts))%>%mutate(cnts=ifelse(Day=="31",7*11,cnts))->byDateLaw_day

byDateLaw_day%>%ggplot(aes(Day,CntByDay/cnts))+geom_bar(stat="identity")+theme(axis.text = element_text(size=6))+coord_flip()+ylab("Crime Frequency (Daily)")+facet_wrap(~Level,scales="free_x")+xlab("Day of Month")->p2
```

```
#frequency by weekday
df_Date%>%mutate(Wkday=as.factor(weekdays(DateStart,abbreviate=TRUE)))%>%group_by(Wkday,Level)%>%dplyr::summarise(CntByWkday=n())->byDateLaw_wkday

#whole 574 weeks between 2006-01-01 and 2016-12-31
nwks=574
byDateLaw_wkday%>%ggplot(aes(fct_relevel(Wkday,"Mon","Tue","Wed","Thu","Fri","Sat","Sun"),CntByWkday/nwks))+geom_bar(stat="identity")+theme(axis.text.x = element_text(size=6))+coord_flip()+ylab("Crime Frequency (Daily)")+facet_wrap(~Level,scales="free_x")+xlab("Day of Week")->p3
```

```
#picking non-missing TimeStart
crime_df%>%filter(!is.na(TimeStart))%>%filter(DateStart>=as.Date("2006-01-01"))->df_FRTM

#Frequency by hour of day, combining hour 00 and hour 24 into hour 00; 4018 days in 11yrs.
nds=4018
df_FRTM%>%mutate(Hour=as.factor(substr(TimeStart,1,2)))%>%group_by(Hour,Level)%>%dplyr::summarise(CntByHo
ur=n())->byDateLaw_hour
byDateLaw_hour$Hour[byDateLaw_hour$Hour=="24"]<-"00"
byDateLaw_hour$Hour<-factor(byDateLaw_hour$Hour)

byDateLaw_hour%>%ggplot(aes(Hour,CntByHour/nds))+geom_bar(stat="identity")+theme(axis.text = element_text
(size=6))+coord_flip()+ylab("Crime Frequency (Hourly)")+facet_wrap(~Level,scales="free_x")+xlab("Hour of
 Day")->p4
```

```
grid.arrange(p1,p2,p3,p4,nrow=2)
```



Several observations:

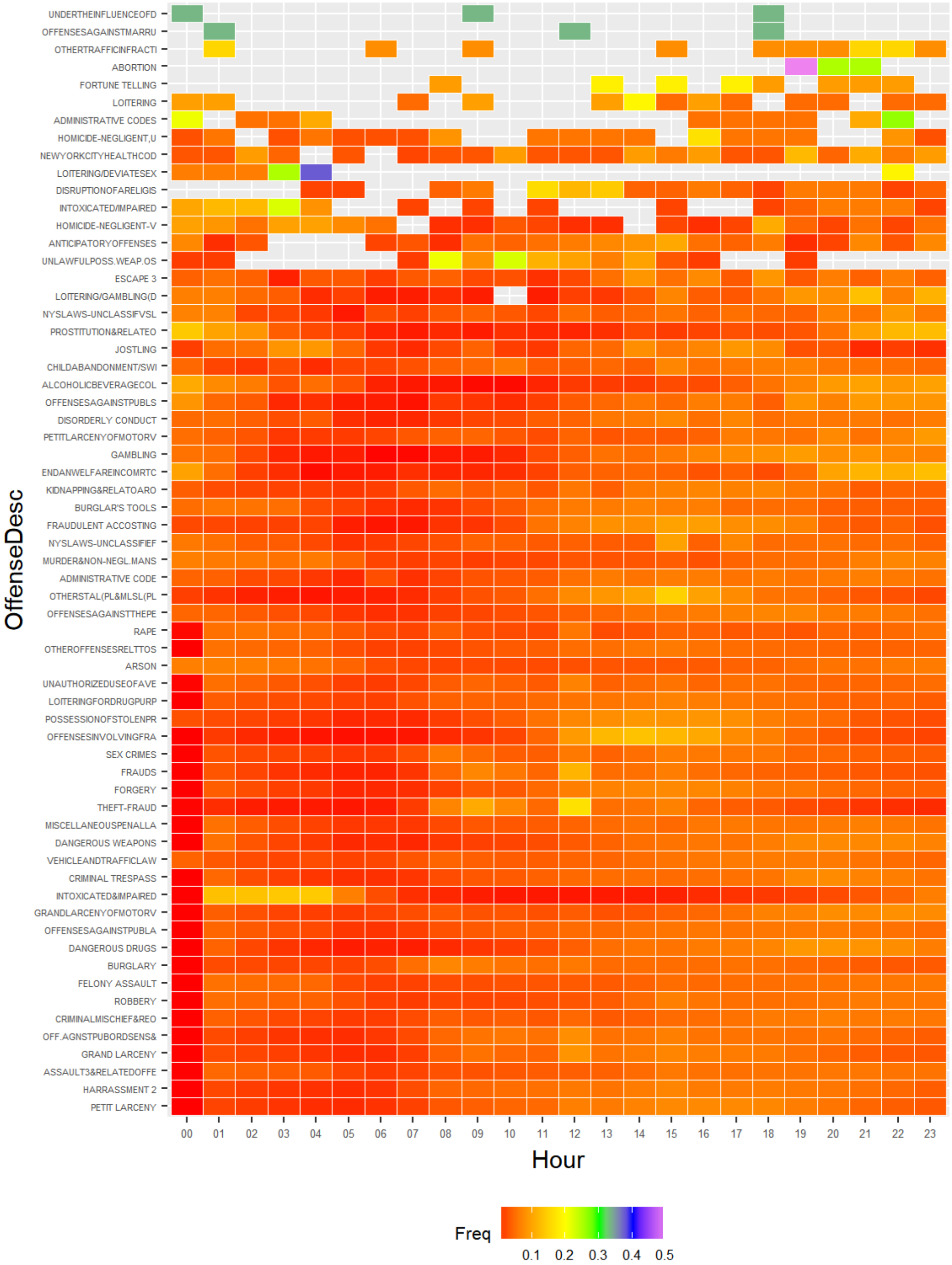
- **Month**: We can easily see that crime peaks in the summer, or when weather is warmer, as the bar graphs for all three Levels tend to bulge from June through September. The exception is in January, which may be due to the irregular

January 1 data (see above)

- **Day of the Month**: There seems to be no difference, except for the first of the month. That could indicate a similar phenomenon as the January 1 phenomenon.
- **Day of the Week**: Most crimes on Friday, least on Sunday. People going out on Friday night?
- **Hour of the Day**: a peak around lunch and again after work, a lull after the bars have been closed for an hour. In other words, times when the streets are filled with people have more crime than when most people are asleep.

We can look more closely at the Hour of the Day to see if types of crime vary by hour of the day.

```
#how the different crime types associated with time using heatmap
crime_df%>%
  select(ClassCode,TimeStart)%>%
  filter(!is.na(TimeStart))%>%
  mutate(ClassCode=as.factor(ClassCode))%>%
  mutate(Hour=as.factor(substr(TimeStart,1,2)))%>%
  group_by(ClassCode,Hour)%>%dplyr::summarise(count=n())%>%mutate(Freq=count/sum(count))->byKYbyFRTM
#combining hour 00 and hour 24 into hour 00
byKYbyFRTM$Hour[byKYbyFRTM$Hour=="24"]<-"00"
byKYbyFRTM$Hour<-factor(byKYbyFRTM$Hour)

#merging to get OffenseDesc vs TimeStart correspondence
merge(byKYbyFRTM, match_code_desc, by.x='ClassCode', by.y='ClassCode')->byKYbyFRTM_match

byKYbyFRTM_match%>%group_by(OffenseDesc)%>%dplyr::summarise(mean=mean(count),na.rm=TRUE)->desc2_desc_cnt
byKYbyFRTM_match%>%group_by(Hour)%>%dplyr::summarise(mean=mean(count),na.rm=TRUE)->Hour_desc_cnt

byKYbyFRTM_match%>%ggplot(aes(
  fct_relevel(as.factor(OffenseDesc),as.character(desc2_desc_cnt$OffenseDesc[sort(desc2_desc_cnt$mean,ind
ex.return=TRUE,decreasing=TRUE)$ix])),
  Hour,fill=Freq))+scale_fill_gradientn(colors=c("red","orange","yellow","green","blue","violet"),na.valu
e="black")+
  scale_x_discrete(label=function(x) abbreviate(x, minlength=20))+coord_flip()+geom_tile(color="white",si
ze=0.25)+
  theme(axis.text.x = element_text(size=5, hjust = 0.5),axis.text.y=element_text(size=4),legend.position=
"bottom",
        legend.text=element_text(size=6,hjust=0.5),legend.title=element_text(size=8),legend.key = element
_rect(size = 0.5),legend.key.size = unit(1, 'lines'))+ylab("Hour")+xlab("OffenseDesc")
```



A few observations from this heat map:

- That dark, blue/purple box represents “Loitering/DeviateSex” at 4am, with a green box at 3am. Just a coincidence that as bars are getting set to close, and after they close, we see a lot of people milling around, “loitering”?
- We can also see a peak for “Intoxicated/Impaired” in the 3am hour, just before bars close
- There is an odd pattern: “Under the Influence of Drugs” bunches up in three specific hours: Midnight, 9am and 6pm.
- Generally, most crime categories show a relatively even distribution across the clock, with higher proportions in the afternoon, and lower proportions in the early morning, in the hours between 5 and 7am.

Why: what factors may contribute to more (or less) crime?

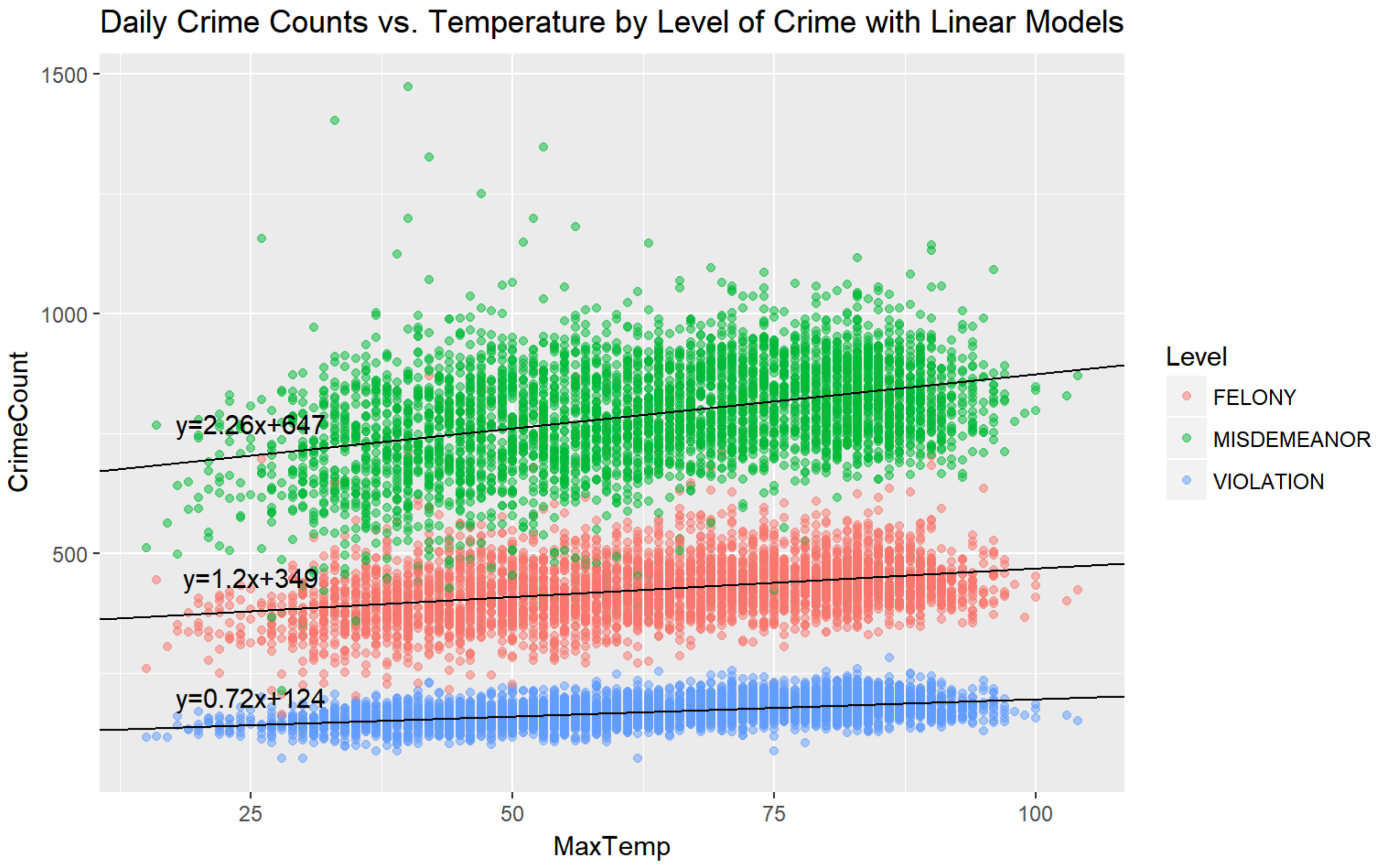
We came up with a series of ideas to test: factors that may contribute to the volume of crime. Each required we find a daily dataset with variables we could add to our main dataset.

We began with temperature data from *NOAA*.

Temperature: Hotter vs. colder

We found that in nearly all precincts, the colder days have fewer crimes committed than on days with hot weather. We can also look to see if the effect is different for the level of the crime.

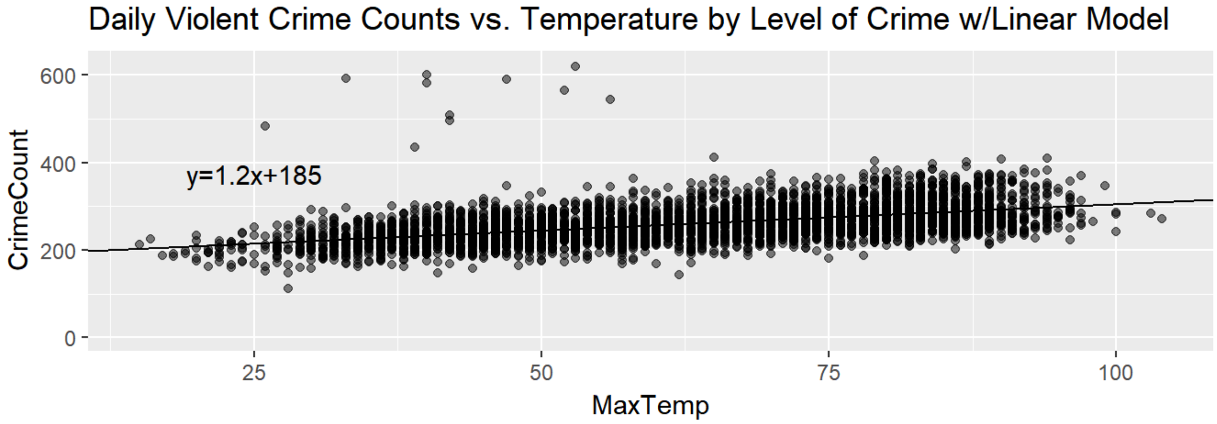
```
###########################################################################################
#Read in weather data from file
weather_select = c("DATE", "AWND", "PRCP", "SNOW", "TMAX")
weather_data <- fread("../Data_Files/nyc_weather_data.csv", na.strings="", select = weather_select, strin
gsAsFactors = FALSE)
weather_data$DATE <- as.Date(weather_data$DATE)
weather_data$AWND <- as.numeric(weather_data$AWND)
weather_data$PRCP <- as.numeric(weather_data$PRCP)
weather_data$SNOW <- as.numeric(weather_data$SNOW)
weather_data$TMAX <- as.numeric(weather_data$TMAX)

#Merge the data together
crime_w_df <- crime_w_df[weather_data, on=.(DateStart = DATE)]
#crime_w_df <-  merge(crime_df,weather_data,by="DateStart")

### Relationship between max temp and crime volume
# set up the data by day and Level
daily_df <-crime_w_df %>% group_by(DateStart,Level) %>% summarize(CrimeCount=n(),MaxTemp=mean(TMAX))
# plot it -- well, plot it later after the linear models are run so we can see the linear slopes
library(ggplot2)
# daily_df %>% ggplot(aes(x=MaxTemp, y=CrimeCount, color=Level)) + geom_point()
# linear model: Felonies
f_df <- daily_df %>% filter(Level=="FELONY")
flm <- lm(CrimeCount~MaxTemp, f_df)
# linear model: Misdemeanors
m_df <- daily_df %>% filter(Level=="MISDEMEANOR")
mlm <- lm(CrimeCount~MaxTemp, m_df)
# linear model: Violation
v_df <- daily_df %>% filter(Level=="VIOLATION")
vlm <- lm(CrimeCount~MaxTemp, v_df)
ggplot(daily_df, aes(x=MaxTemp, y=CrimeCount, color=Level)) +
  geom_point(alpha=0.5) +
  geom_abline(slope=flm[["coefficients"]][["MaxTemp"]],intercept=flm[["coefficients"]][["(Intercept)"]])
 +
  annotate("text", x= 25, y=450, label=paste0("y=",round(flm[["coefficients"]][["MaxTemp"]],2),"x+",round
(flm[["coefficients"]][["(Intercept)"]],0))) +
  geom_abline(slope=mlm[["coefficients"]][["MaxTemp"]],intercept=mlm[["coefficients"]][["(Intercept)"]])
 +
  annotate("text", x= 25, y=770, label=paste0("y=",round(mlm[["coefficients"]][["MaxTemp"]],2),"x+",round
(mlm[["coefficients"]][["(Intercept)"]],0))) +
  geom_abline(slope=vlm[["coefficients"]][["MaxTemp"]],intercept=vlm[["coefficients"]][["(Intercept)"]])
 +
  annotate("text", x= 25, y=200, label=paste0("y=",round(vlm[["coefficients"]][["MaxTemp"]],2),"x+",round
(vlm[["coefficients"]][["(Intercept)"]],0))) +
  ggtitle("Daily Crime Counts vs. Temperature by Level of Crime with Linear Models")
```

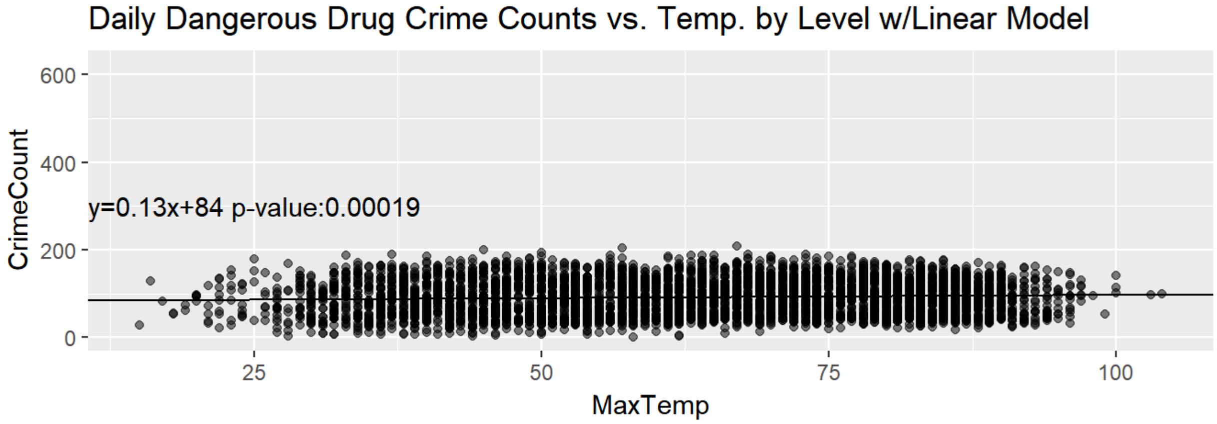


This shows us how temperature is related to crime in all three levels: the warmer it is, the more crime. The steepest slope is for Misdemeanors, meaning the greatest impact of temperature is on that level of crime, followed by Felonies.

- **Does temperature affect levels of Drug and Violent crimes?**

```
# include OffenseDesc and IntOffenseDesc into daily summary
dailies_df <-crime_w_df %>% group_by(DateStart,Level,OffenseDesc,IntOffenseDesc) %>% summarize(CrimeCount
=n(),MaxTemp=mean(TMAX))

# create a table for Violent Crimes, Daily Count
daily_violent_df <- dailies_df  %>% filter(OffenseDesc == "ASSAULT 3 & RELATED OFFENSES" |
                                           OffenseDesc == "FELONY ASSAULT" |
                                           OffenseDesc == "MURDER & NON-NEGL. MANSLAUGHTER" |
                                           OffenseDesc == "RAPE" |
                                           OffenseDesc == "ROBBERY" |
                                           IntOffenseDesc == "AGGRAVATED SEXUAL ASBUSE" |
                                           IntOffenseDesc == "ASSAULT 2,1,UNCLASSIFIED" |
                                           IntOffenseDesc == "ASSAULT 3" |
                                           IntOffenseDesc == "RAPE 1" |
                                           IntOffenseDesc == "ROBBERY,OPEN AREA UNCLASSIFIED" |
                                           IntOffenseDesc == "SEXUAL ABUSE" |
                                           IntOffenseDesc == "SEXUAL ABUSE 3,2") %>%
                                  group_by(DateStart) %>%
                                  summarize(CrimeCount=sum(CrimeCount),MaxTemp=mean(MaxTemp))


# create a table for Dangerous Drug Crimes, Daily Count
daily_drug_df <- dailies_df  %>% filter(OffenseDesc == "DANGEROUS DRUGS") %>%
                                  group_by(DateStart) %>%
                                  summarize(CrimeCount=sum(CrimeCount),MaxTemp=mean(MaxTemp))

# derive linear model
dvlm<- lm(CrimeCount~MaxTemp, daily_violent_df)

# plot
d1 <- ggplot(daily_violent_df, aes(x=MaxTemp, y=CrimeCount)) +
  geom_point(alpha=0.5) +
  geom_abline(slope=dvlm[["coefficients"]][["MaxTemp"]],intercept=dvlm[["coefficients"]][["(Intercept)"
]]) +
  annotate("text", x= 25, y=375, label=paste0("y=",round(dvlm[["coefficients"]][["MaxTemp"]],2),"x+",roun
d(dvlm[["coefficients"]][["(Intercept)"]],0)))+
  ggtitle("Daily Violent Crime Counts vs. Temperature by Level of Crime w/Linear Model")+
  scale_y_continuous(limits = c(0, 625))

# derive linear model
ddlm<- lm(CrimeCount~MaxTemp, daily_drug_df)
# grab p-value
ddlmp<-round(summary(ddlm)$coefficients[2,4],5)

# plot
d2 <- ggplot(daily_drug_df, aes(x=MaxTemp, y=CrimeCount)) +
  geom_point(alpha=0.5) +
  geom_abline(slope=ddlm[["coefficients"]][["MaxTemp"]],intercept=ddlm[["coefficients"]][["(Intercept)"
]]) +
  annotate("text", x= 25, y=300, label=paste0("y=",round(ddlm[["coefficients"]][["MaxTemp"]],2),"x+",roun
d(ddlm[["coefficients"]][["(Intercept)"]],0)," p-value:",ddlmp))+
  ggtitle("Daily Dangerous Drug Crime Counts vs. Temp. by Level w/Linear Model") +
  scale_y_continuous(limits = c(0, 625))

grid.arrange(d1,d2)
```





For both **Violent Crime** and **Dangerous Drugs**, we see an effect of temperature: the warmer it is, the more crime. However, the effect is much stronger for Violent Crimes. While there is an effect for Dangerous Drugs, it is hard to perceive from the graph, but the slope is positive and the p-value is less than .001 for the linear model.

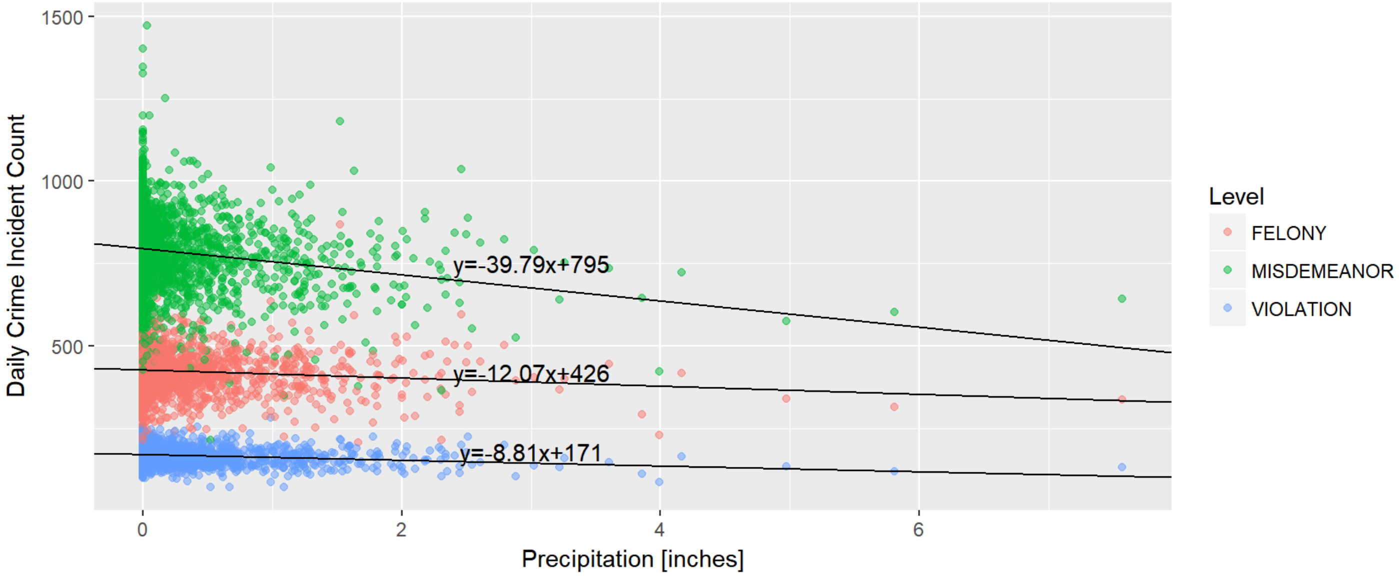
Precipitation

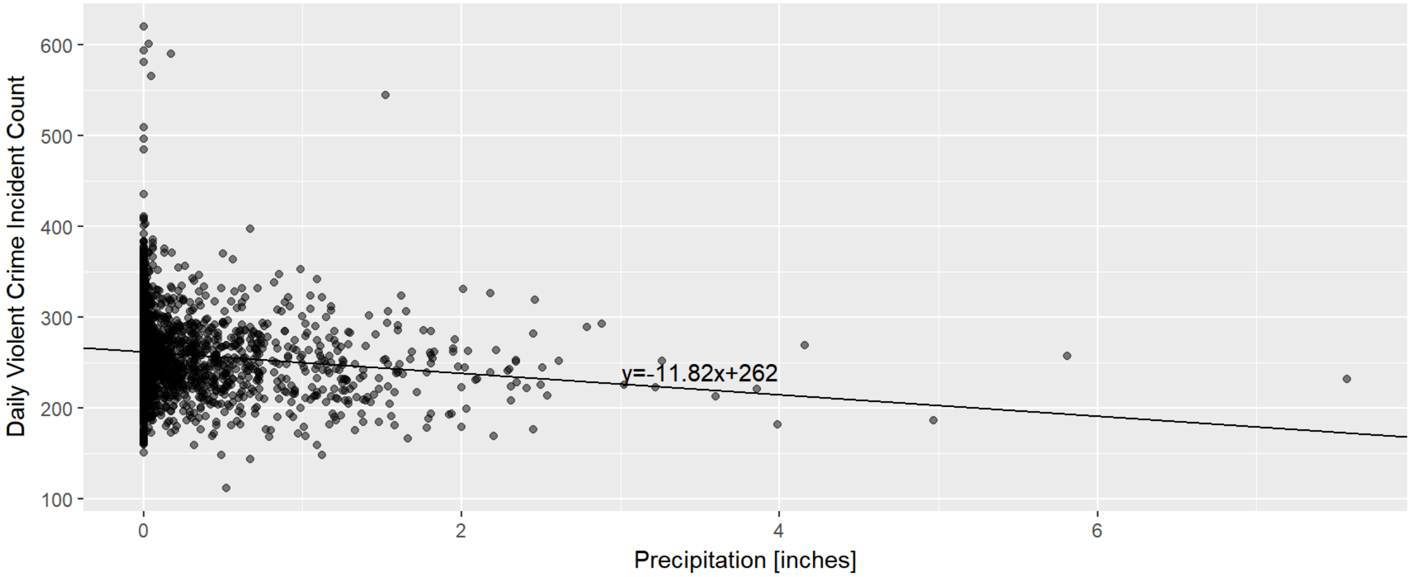
We continue with our weather data to see if rain makes a difference in crime.

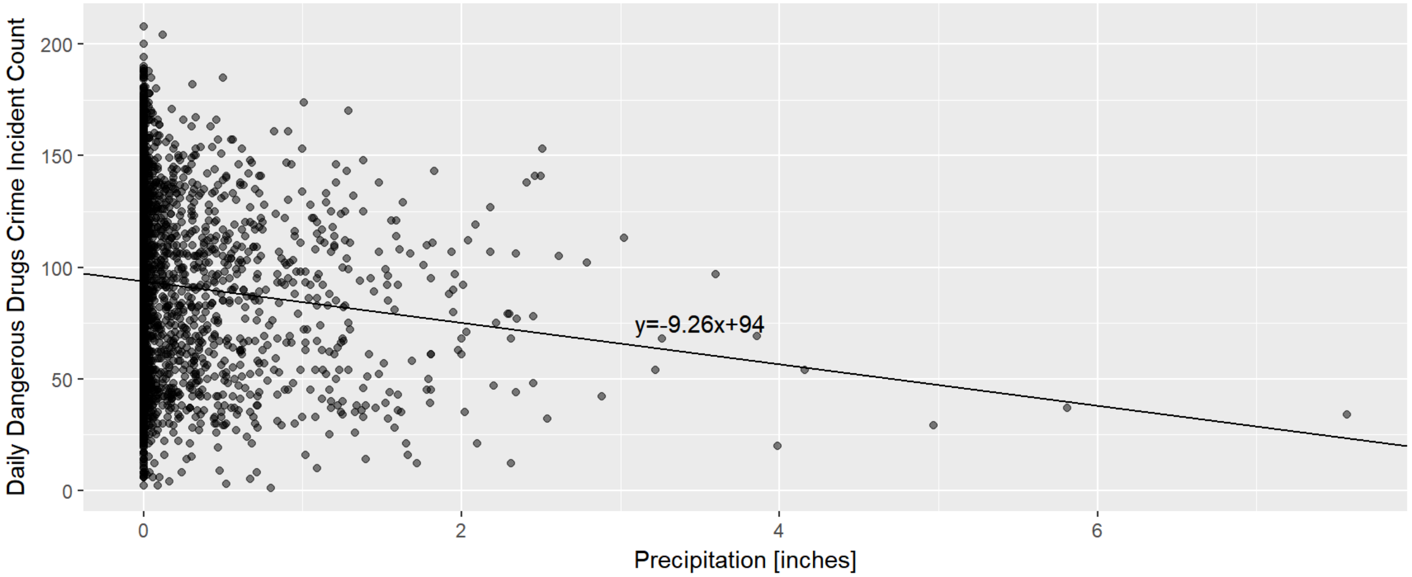
```
###############################################################################
#Read in moon phase data
moon_data <- fread("../Data_Files/nyc_moon_data.csv", na.strings="", select = c("date", "phase"), strings
AsFactors = FALSE)
moon_data$date <- as.Date(moon_data$date, format='%m/%d/%Y')
moon_data$phase <- as.factor(moon_data$phase)
full_moon_data <- moon_data %>% filter(phase == "Full Moon")

#Merge the moon phase data into the main data frame
crime_w_df <- crime_w_df %>% left_join(moon_data, by = c("DateStart" = "date"))

#Generate violent crime dataframe
#filter for violent crime
violent_crime_df <- crime_w_df %>% filter(OffenseDesc == "ASSAULT 3 & RELATED OFFENSES" |
                                          OffenseDesc == "FELONY ASSAULT" |
                                          OffenseDesc == "MURDER & NON-NEGL. MANSLAUGHTER" |
                                          OffenseDesc == "RAPE" |
                                          OffenseDesc == "ROBBERY" |
                                          IntOffenseDesc == "AGGRAVATED SEXUAL ASBUSE" |
                                          IntOffenseDesc == "ASSAULT 2,1,UNCLASSIFIED" |
                                          IntOffenseDesc == "ASSAULT 3" |
                                          IntOffenseDesc == "RAPE 1" |
                                          IntOffenseDesc == "ROBBERY,OPEN AREA UNCLASSIFIED" |
                                          IntOffenseDesc == "SEXUAL ABUSE" |
                                          IntOffenseDesc == "SEXUAL ABUSE 3,2")

drugs_crime_df <- crime_w_df %>% filter(OffenseDesc == "DANGEROUS DRUGS")
```

```
#Generate scatterplot of crime vs precipitiation
rain_summary_per_day <- crime_w_df %>% group_by(DateStart, Level) %>% summarize(Count = n()) %>% drop_na
()
#append weather data
rain_summary_per_day <- rain_summary_per_day %>% left_join(weather_data, by = c("DateStart" = "DATE")) %
>% select(DateStart, Level, Count, PRCP)
#Scatter plot of daily crimes vs. precipitation level

#Filter on Level of crime and generate linear model for each
# linear model: Felonies
f_df_rain <- rain_summary_per_day %>% filter(Level=="FELONY")
flm_rain <- lm(Count~PRCP, f_df_rain)
# linear model: Misdemeanors
m_df_rain <- rain_summary_per_day %>% filter(Level=="MISDEMEANOR")
mlm_rain <- lm(Count~PRCP, m_df_rain)
# linear model: Violation
v_df_rain <- rain_summary_per_day %>% filter(Level=="VIOLATION")
vlm_rain <- lm(Count~PRCP, v_df_rain)

#Plot data for all three crime levels vs precipitation with linear model results
s1 <- ggplot(rain_summary_per_day, aes(x=PRCP, y=Count, color=Level)) +
  geom_point(alpha=0.5) +
  geom_abline(slope=flm_rain[["coefficients"]][["PRCP"]],intercept=flm_rain[["coefficients"]][["(Intercep
t)"]]) +
  annotate("text", x= 3, y=420, label=paste0("y=",round(flm_rain[["coefficients"]][["PRCP"]],2),"x+",roun
d(flm_rain[["coefficients"]][["(Intercept)"]],0))) +
  geom_abline(slope=mlm_rain[["coefficients"]][["PRCP"]],intercept=mlm_rain[["coefficients"]][["(Intercep
t)"]]) +
  annotate("text", x= 3, y=750, label=paste0("y=",round(mlm_rain[["coefficients"]][["PRCP"]],2),"x+",roun
d(mlm_rain[["coefficients"]][["(Intercept)"]],0))) +
  geom_abline(slope=vlm_rain[["coefficients"]][["PRCP"]],intercept=vlm_rain[["coefficients"]][["(Intercep
t)"]]) +
  annotate("text", x= 3, y=180, label=paste0("y=",round(vlm_rain[["coefficients"]][["PRCP"]],2),"x+",roun
d(vlm_rain[["coefficients"]][["(Intercept)"]],0))) +
  labs(x = "Precipitation [inches]", y = "Daily Crime Incident Count", title = "Daily Crime Counts vs. Pr
ecipitation by Level of Crime with Linear Models")

#Plot data only for violent crimes vs precipitation with linear model result
#Generate scatterplot of crime vs precipitiation
vc_df_rain <- violent_crime_df %>% group_by(DateStart) %>% summarize(Count = n()) %>% drop_na()
#append weather data
vc_df_rain <- vc_df_rain %>% left_join(weather_data, by = c("DateStart" = "DATE")) %>% select(DateStart,
 Count, PRCP)
#Scatter plot of daily crimes vs. precipitation level

#Generate linear model for Violent Crime vs. Precipitation
vclm_rain <- lm(Count~PRCP, vc_df_rain)

s2 <- ggplot(vc_df_rain, aes(x=PRCP, y=Count)) +
  geom_point(alpha=0.5) +
  geom_abline(slope=vclm_rain[["coefficients"]][["PRCP"]],intercept=vclm_rain[["coefficients"]][["(Interc
ept)"]]) +
  annotate("text", x= 3.5, y=240, label=paste0("y=",round(vclm_rain[["coefficients"]][["PRCP"]],2),"x+",r
ound(vclm_rain[["coefficients"]][["(Intercept)"]],0))) +
  labs(x = "Precipitation [inches]", y = "Daily Violent Crime Incident Count", title = "Daily Violent Cri
me Counts vs. Precipitation with Linear Models")

#Plot data only for Dangerous Drugs crimes vs precipitation with linear model result
```

```
#Generate scatterplot of crime vs precipitiation
dg_df_rain <- drugs_crime_df %>% group_by(DateStart) %>% summarize(Count = n()) %>% drop_na()
#append weather data
dg_df_rain <- dg_df_rain %>% left_join(weather_data, by = c("DateStart" = "DATE")) %>% select(DateStart,
 Count, PRCP)
#Scatter plot of daily crimes vs. precipitation level

#Generate linear model for Violent Crime vs. Precipitation
dglm_rain <- lm(Count~PRCP, dg_df_rain)

s3 <- ggplot(dg_df_rain, aes(x=PRCP, y=Count)) +
  geom_point(alpha=0.5) +
  geom_abline(slope=dglm_rain[["coefficients"]][["PRCP"]],intercept=dglm_rain[["coefficients"]][["(Interc
ept)"]]) +
  annotate("text", x= 3.5, y=75, label=paste0("y=",round(dglm_rain[["coefficients"]][["PRCP"]],2),"x+",ro
und(dglm_rain[["coefficients"]][["(Intercept)"]],0))) +
  labs(x = "Precipitation [inches]", y = "Daily Dangerous Drugs Crime Incident Count", title = "Daily Dan
gerous Drugs Crime Counts vs. Precipitation with Linear Models")

grid.arrange(s1,s2,s3)
```


Daily Crime Counts vs. Precipitation by Level of Crime with Linear Models


Daily Violent Crime Counts vs. Precipitation with Linear Models


Daily Dangerous Drugs Crime Counts vs. Precipitation with Linear Models

For all three Levels of Crime, we see negative slopes, indicating that more rain suggest less crime. This is not very surprising as we would expect fewer people to be out on days with bad weather. As with Temperature, the effect is stronger for the more numerous of crimes.

A lot of the data is on days with no rain, and you can read right off the graph that there is a fairly wide range of daily crime rate. However, as the amount of rain starts to register, the density of the dots seem to shift downward.

- **Does precipitation affect levels of Drug and Violent crimes?**

When looking at the same analysis after filtering only on **Violent Crimes**, we see a relationship very similar to that of felonies, which is what we would expect considering most violent crimes are felonies.

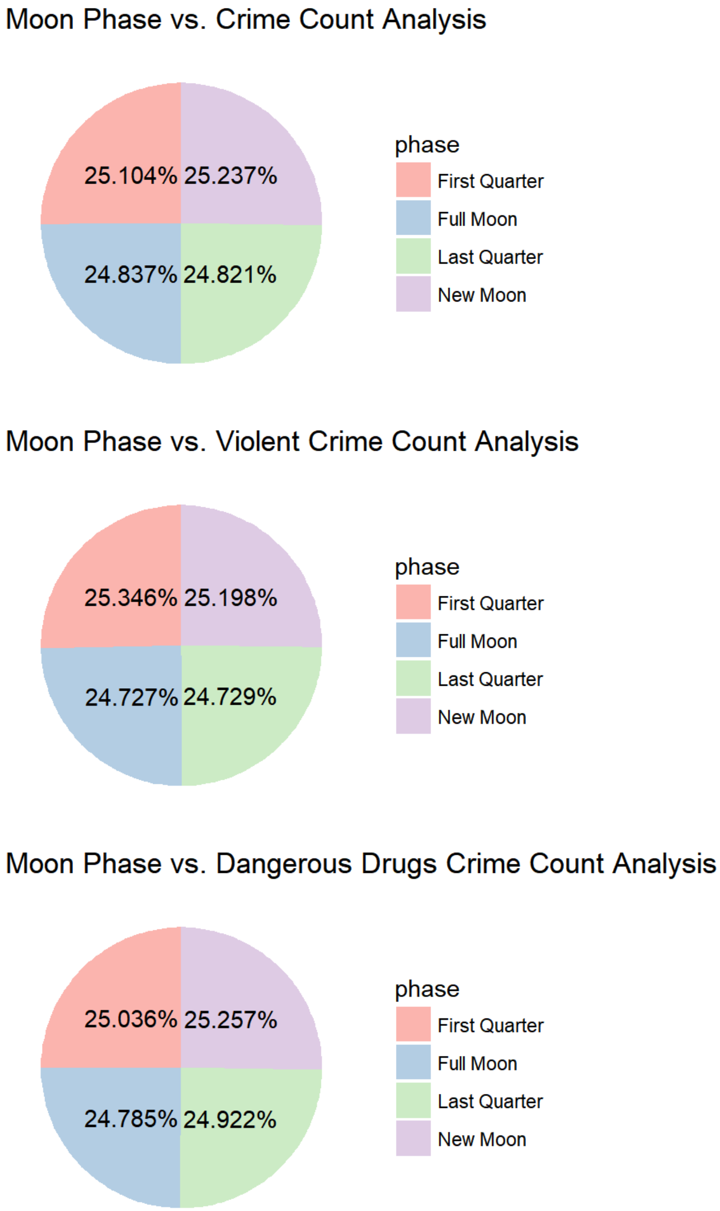
The pattern for **Dangerous Drugs** is similar: less violent crime on rainy days.

The Full Moon

We brought in data on the phases of the moon, in order to test the idea that with a full moon, we might see more crime. Articles such as in *Decoded Science* (https://www.decodedscience.org/full-moons-crime-aka-lunar-effect-real-deal-pseudoscience/41881 (https://www.decodedscience.org/full-moons-crime-aka-lunar-effect-real-deal-pseudoscience/41881)) examine the issue, but we thought we could examine it in our dataset.

```
moon_summary <- crime_w_df %>%
  filter(phase == "Full Moon" | phase == "New Moon" | phase == "First Quarter" | phase == "Last Quarter")
%>%
  group_by(DateStart, phase) %>% summarize(Count = n()) %>% drop_na()
moon_avg_crime <- moon_summary %>% group_by(phase) %>% summarize(Avg_Count = weighted.mean(Count))
moon_total_crime <- moon_summary %>% group_by(phase) %>% summarize(Total_Count = sum(Count))
moon_phase_total_count <- sum(moon_total_crime$Total_Count)
moon_total_crime <- moon_total_crime %>% mutate(Pct = Total_Count/moon_phase_total_count)

#Create a pie chart
blank_theme <- theme_minimal()+
  theme(
    axis.title.x = element_blank(),
    axis.title.y = element_blank(),
    panel.border = element_blank(),
    panel.grid=element_blank(),
    axis.ticks = element_blank(),
    #plot.title=element_text(size=14, face="bold")
  )

pie1 <- ggplot(data = moon_total_crime, aes(x="", y = Total_Count, fill = phase)) +
  geom_bar(width = 1, stat = "identity") +
  scale_fill_brewer(palette="Pastel1") +
  #scale_fill_manual(values=c("red","yellow","blue","green")) +
  coord_polar(theta = "y", start=0) +
  blank_theme +
  theme(axis.text.x=element_blank()) +
  geom_text(aes(label = scales::percent(Pct)), position = position_stack(vjust = 0.5)) +
  ggtitle("Moon Phase vs. Crime Count Analysis")

#Generate the same analysis based on Violent Crimes
vc_moon_summary <- violent_crime_df %>%
  filter(phase == "Full Moon" | phase == "New Moon" | phase == "First Quarter" | phase == "Last Quarter")
%>%
  group_by(DateStart, phase) %>% summarize(Count = n()) %>% drop_na()
vc_moon_avg_crime <- vc_moon_summary %>% group_by(phase) %>% summarize(Avg_Count = weighted.mean(Count))
vc_moon_total_crime <- vc_moon_summary %>% group_by(phase) %>% summarize(Total_Count = sum(Count))
vc_moon_phase_total_count <- sum(vc_moon_total_crime$Total_Count)
vc_moon_total_crime <- vc_moon_total_crime %>% mutate(Pct = Total_Count/vc_moon_phase_total_count)

pie2 <- ggplot(data = vc_moon_total_crime, aes(x="", y = Total_Count, fill = phase)) +
  geom_bar(width = 1, stat = "identity") +
  scale_fill_brewer(palette="Pastel1") +
  coord_polar(theta = "y", start=0) +
  blank_theme +
  theme(axis.text.x=element_blank()) +
  geom_text(aes(label = scales::percent(Pct)), position = position_stack(vjust = 0.5)) +
  ggtitle("Moon Phase vs. Violent Crime Count Analysis")

#Generate the same analysis based on Dangerous Drugs Crimes
dg_moon_summary <- drugs_crime_df %>%
  filter(phase == "Full Moon" | phase == "New Moon" | phase == "First Quarter" | phase == "Last Quarter")
%>%
  group_by(DateStart, phase) %>% summarize(Count = n()) %>% drop_na()
dg_moon_avg_crime <- dg_moon_summary %>% group_by(phase) %>% summarize(Avg_Count = weighted.mean(Count))
dg_moon_total_crime <- dg_moon_summary %>% group_by(phase) %>% summarize(Total_Count = sum(Count))
dg_moon_phase_total_count <- sum(dg_moon_total_crime$Total_Count)
dg_moon_total_crime <- dg_moon_total_crime %>% mutate(Pct = Total_Count/dg_moon_phase_total_count)
```

```
pie3 <- ggplot(data = dg_moon_total_crime, aes(x="", y = Total_Count, fill = phase)) +
  geom_bar(width = 1, stat = "identity") +
  scale_fill_brewer(palette="Pastel1") +
  coord_polar(theta = "y", start=0) +
  blank_theme +
  theme(axis.text.x=element_blank()) +
  geom_text(aes(label = scales::percent(Pct)), position = position_stack(vjust = 0.5)) +
  ggtitle("Moon Phase vs. Dangerous Drugs Crime Count Analysis")

grid.arrange(pie1,pie2,pie3)
```



Moon Phase vs. Crime Count Analysis

Moon Phase vs. Violent Crime Count Analysis

Moon Phase vs. Dangerous Drugs Crime Count Analysis

Graph note: We debated (hotly) about using Pie Charts for this data. On the one hand, a bar chart would be at least equally capable of demonstrating the lack of difference, the phase of the moon has on crime. On the other hand, we felt like we had about a thousand bar charts, and we were inspired by the shape of the moon!

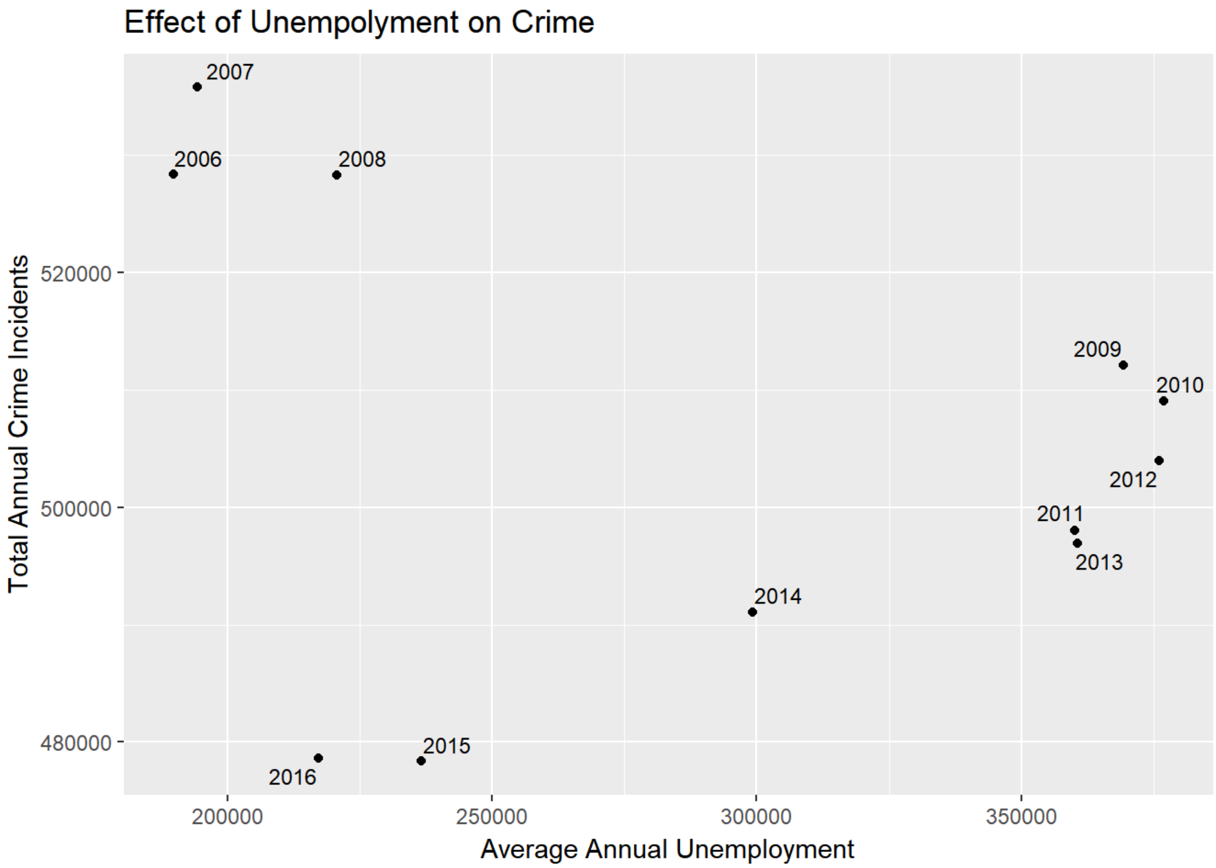
We can see from the data that crime does not appear to be affected significantly on days with a full moon. We can conclude that there is no statistical evidence to support the “old wives tale” about lunatics.

One other finding of interest: since the lunar cycle is independent of our calendar, any cyclical oddities that we found before (January 1, summer months, etc.) are assured to be just artifacts of the calendar by this data above. Why? Because the lunar slices we get here are essentially random from the perspective of our calendar, and the even distribution of crime across these slices tells us that without a calendar, we wouldn't notice any odd patterns like we saw above.

Unemployment : Does lack of employment impact crime rates?

We decided to bring in unemployment data for the city.

```
########################################################################################
#Read in UnEmployment data from file

UnEmployment_data <- fread("../Data_Files/unempCSV.csv", na.strings="", stringsAsFactors = FALSE)

UnEmployment_data$year <-as.character(UnEmployment_data$Year)
names(UnEmployment_data)[names(UnEmployment_data) == 'Ann Avg'] <- 'AnnAvg'
#UnEmployment_data
UnEmployment_data<-UnEmployment_data %>% select("year","AnnAvg")%>% drop_na()


crime_df%>%mutate(year=as.character(year(DateReport)))%>%group_by(year)%>%dplyr::summarise(CntByYear=n
())%>% drop_na()->crime_e_df


crime_e_df <- merge(x=crime_e_df,y=UnEmployment_data,by.x="year",by.y="year")

### Relationship between crime volume and unemployment
## set up the data by day and Level
crime_e_df <-crime_e_df %>% group_by(year) %>% summarize(CrimeCount=sum(CntByYear),AnnAvg,unEmpCrimeRatio
=((AnnAvg)/sum(CntByYear)))
ggplot(crime_e_df, aes(x=AnnAvg, y=CrimeCount)) + geom_point() +
  labs(x = "Average Annual Unemployment", y = "Total Annual Crime Incidents", title = "Effect of Unempoly
ment on Crime") +
  ggrepel::geom_text_repel(aes(label=year), size=3)
```



It is really tempting to ignore 2006-2008, the three dots in the upper left corner, and conclude a somewhat linear relationship where you see less crime with lower unemployment. However, with only 11 data points, discarding three as outliers may be unwise. We had a small debate on the team and settled it with a linear model that showed a p-value of 0.45. So no, we would need many more data points consistent with the 2010 to 2016 pattern to conclude there is a relationship.

Executive Summary

Introduction:

Crime has long been a story in New York City. Many of us have witnessed significant changes in both the frequency of crime and the most prevalent types of crime that have dominated New York City in the last 30 years. And crime, in general, is a major topic for any major city.

Additionally, we took a closer look at Violent Crime and Dangerous Drug crimes in our examination of data from 2006 to 2016 for the City of New York

Our main findings:

- Crime has reduced in the past ten years
- Queens has less crime per capita than Manhattan and the Bronx
- By pure count, Misdemeanors account for the majority of crime
- The weather and unemployment have an influence on crime
- Decreases in Felonies followed by Misdemeanors may reflect police enforcement policy efforts

Crime by Borough, Level of Crime proportions

Type	Overall	Bronx	Brooklyn	Manhattan	Queens	Staten Island

Type	Overall	Bronx	Brooklyn	Manhattan	Queens	Staten Island
Felonies	31%	27%	32%	32%	33%	22%
Misdemeanors	57%	61%	55%	58%	54%	60%
Violations	12%	12%	12%	10%	13%	19%
Violent Crimes and Drugs						
Violent Crime	19%	21%	21%	15%	19%	14%
Dangerous Drugs	7%	12%	7%	6%	2%	5%

Note:

[1] Above Table is auto-generated from the Crime Dataset directly

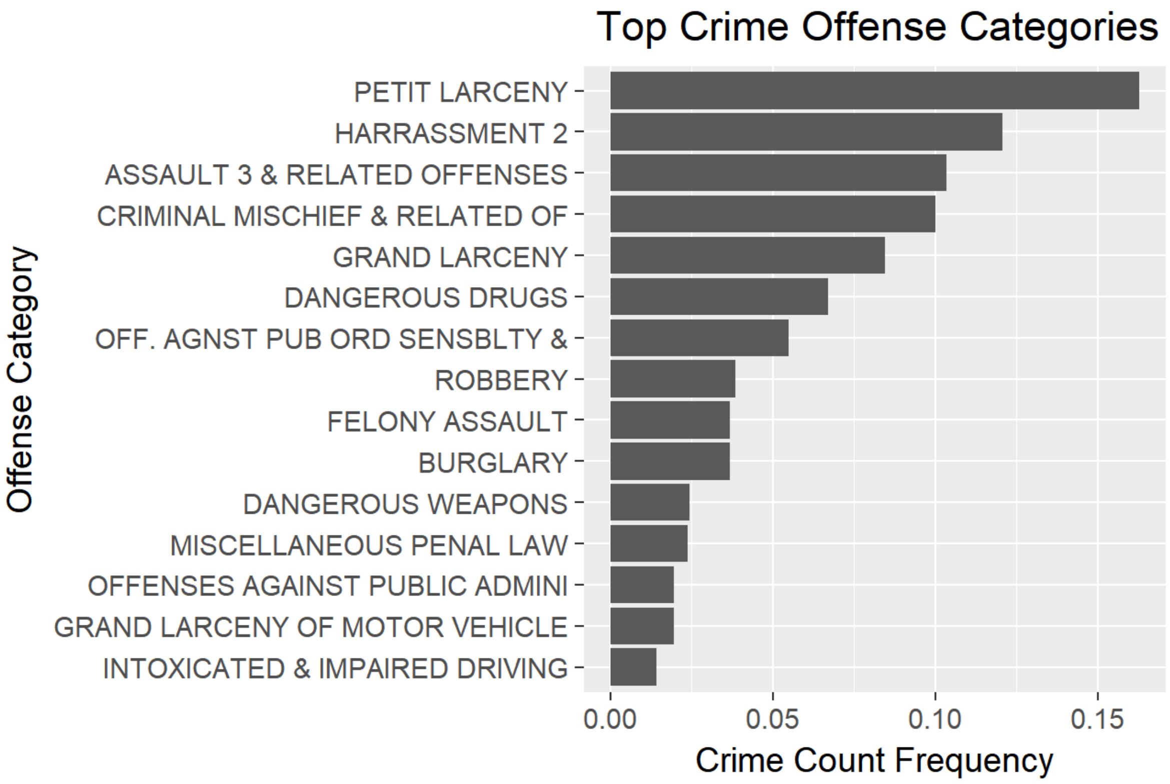
[2] Violent Crimes and Dangerous Drugs are subcategories of the Felony, Misdemeanor and Violation categories shown

[3] Felony, Misdemeanor and Violation figures of each Borough add up to 100%

What kinds of crime?

Misdemeanors make up nearly 60% of all crimes, with Felonies numbering 30%. Violations make up the remaining 10%.

Two categories of Larceny wind up in the top five; Petit (16%) and Grand (8%) make up nearly a quarter of all crimes. Assault (10%) and Dangerous Drugs (6%) fit within our focus on Violent and Drug Crimes.



These subcategories give clearer description of what kinds of crime are most frequent

Where does crime take place?

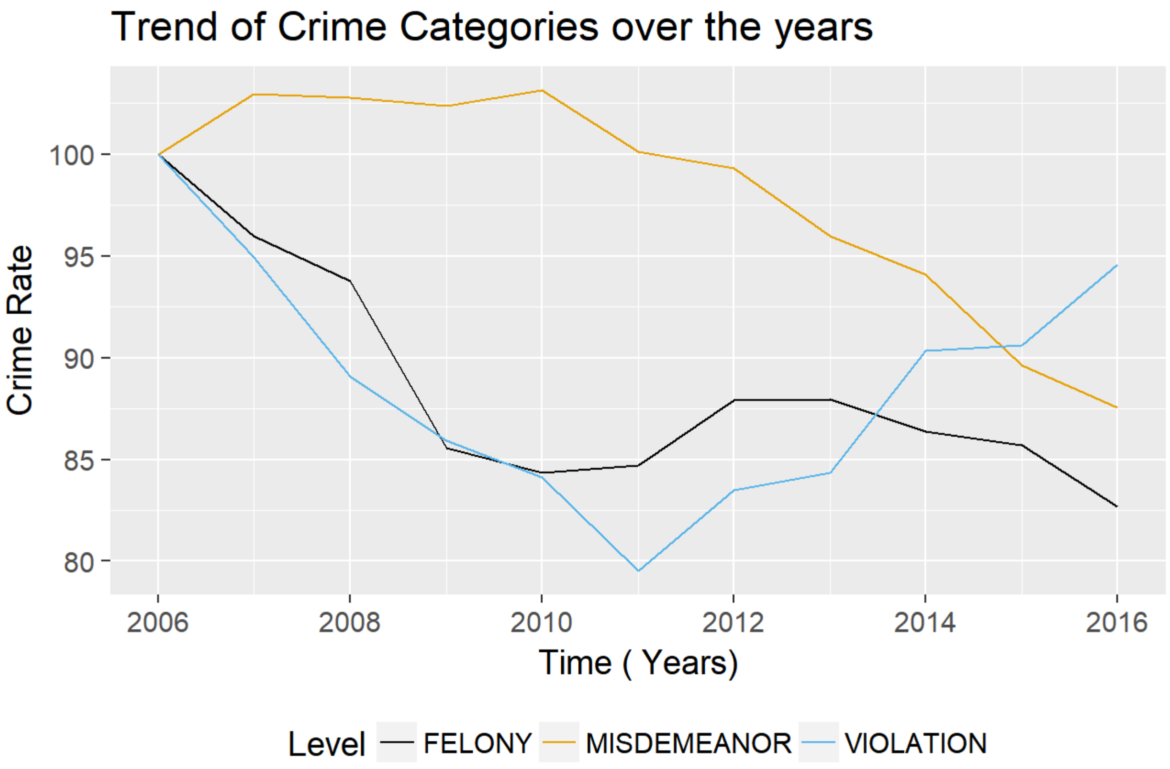
Queens is the most crime free borough when you consider crimes committed per person, though felonies are more prevalent. Manhattan and the Bronx tend to have more crime.

Streets and Residences (of one type or another) account for 85% of the locations of crime.

When does crime take place? What have been the trends?

Crime has been decreasing, in general, over the past ten years. We found that Felonies dropped first (about 12%), then a couple of years later, Misdemeanors dropped about the same amount. While Violent Crimes have been pretty steady (though lower that 2006 levels), Dangerous Drug Crimes have dropped more significantly.

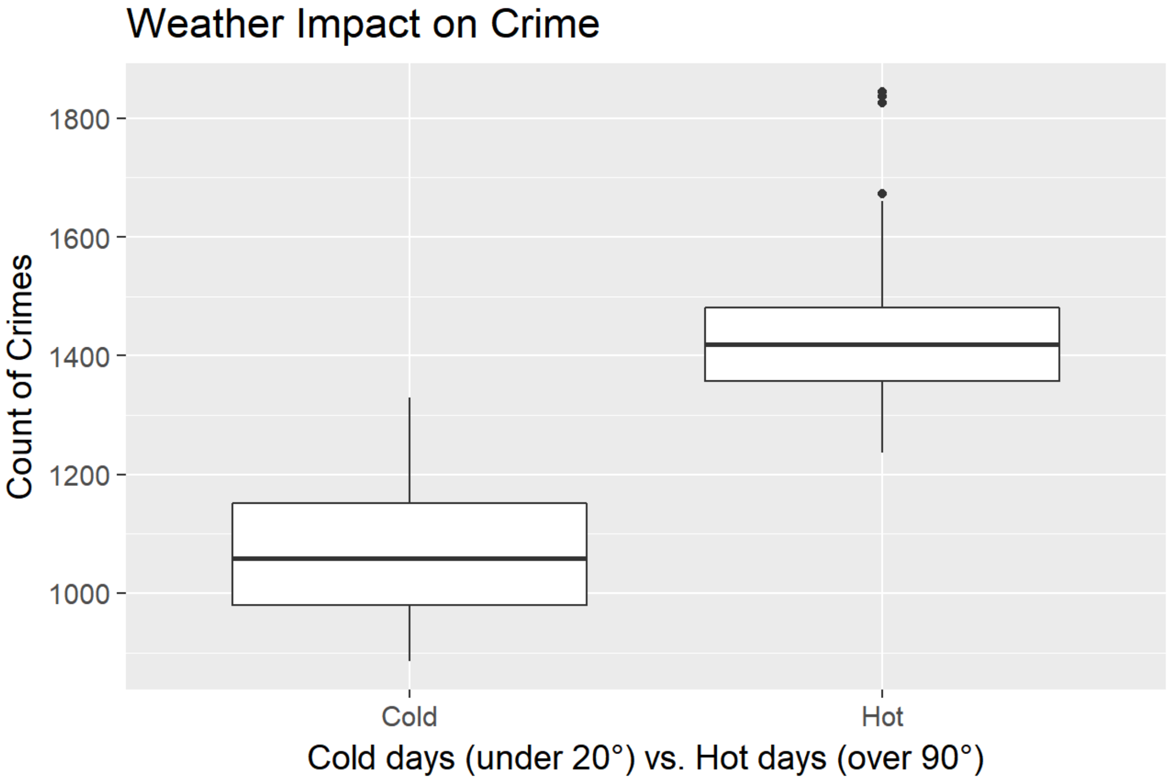
There's more crime in summer months, and less in winter. Fewer crimes between 5am and 7am, more during lunchtime, drive time. Certain types of crime (loitering, intoxication) happen around bar-closing hours.



Misdemeanors and Felonies both end up more than 12% lower than 2006, with Felonies dropping first

Why does crime happen? What contributes?

The weather impacts crime! There's more crime when it is hotter, especially misdemeanors. There's less crime when it rains. Violent Crimes happen more often when it is hot, less often when rainy. Dangerous Drug crimes are less affected by the heat, but do happen less often with rain.



The number of crimes on Hot days is clearly higher than on Cold days

The phase of the moon has no impact on crime. *None.*

We did not find evidence that Unemployment has a relationship with crime.

Interactive Component

Shiny Based Interactive App

- Overview:
  Our goal with the interactive application was to give the user the ability to explore both spatial relationships as well as summary statistics relating to our dataset on NYC crime. The application was developed using the Shiny application development environment provided in R-studio. Application URL: https://nycrimeanalysis.shinyapps.io/CrimeDataNYCApp/ (https://nycrimeanalysis.shinyapps.io/CrimeDataNYCApp/)
  There are two tabs for the user to select. The user input selections on the left side panel change depending on which tab the user selects.
- Maps Tab:
  The first tab, labeled “Maps” provides the user with the option to select a specific date and borough of interest. If the user would like to consider all five boroughs, that is also an option. The application will filter the dataset based on the user’s selections and render an interactive map, the type of which is determined by the user’s selection under the “Select Map Type” pull down menu.
  In addition, below the user input selections on the left side panel, a word cloud is displayed indicating the relative frequency of different offense descriptions.
  If “Point” is selected, the user can view the locations of every crime incident that occurred on that day. Clusters of crimes are grouped together with a label indicating the number of incidents in that cluster. As the user zooms in or clicks on clusters the individual crime incidents become visible. As the user mouses over the point a tooltip popup will display the level of the crime, the precinct it occurred in, and a brief description of the offense.
  If “Density” is selected, the user will see a two-dimensional density plot of the crimes that occurred on that day. This allows the user to identify areas of the city with the highest density and lowest density of crime.

If “Precincts” is selected, the user will see a static display of the geographic outlines for all 77 police precincts. As the user mouses over each polygon, a tooltip pop up will display the precinct number.
All maps are interactive, such that the user can pan and zoom as needed.

- Summary Tab:
  Under the “Summary” Tab, the user will have the option to generate summary statistics for a range of dates they are interested in. The horizontal bar chart on the right shows the total crime incidents reported over the selected date range and allows the user to break up the categories with a primary and secondary variable selection. As the user mouses over each bar, a pop-up will display the number of crime incidents in that section. Below the user input section on the left side panel is a summary table showing the data plotted in the right main panel.

D3 Based Interactive App

- Overview
  This app provides a basic interactive feature to compare Borough Crime density across different categories like Felony, Misdemeanor and Violations. plot helps summarize the following features:

- There are three user choice options available:
  - Total Crime by Borough
  - 2010 Per Capita (Census)
  - 2016 Per Capita (Estimate)
- Users can hover over each block, seeing a summary of Borough, Level of crime, and the number/percentage

Tools used:

- Ordinal scale scaleBand schemeCategory20 used for filling the color.
- Data feed archived via d3 Json API
- On order to run the page locally please run the python -m SimpleHTTPServer to activate the local server
- Best browser compatibility safari chrome FF latest versions.

Application URL:
https://bl.ocks.org/CrimeDataNyc/raw/e3362fde2a5f94aa2fa94a524742a566/441a18c0a089ca75e03009376006eff10dd38eb8/ (https://bl.ocks.org/CrimeDataNyc/raw/e3362fde2a5f94aa2fa94a524742a566/441a18c0a089ca75e03009376006eff10dd38eb8/)

Conclusion

All our current work and analysis done in this project can be found in our Project repository on Github -> https://github.com/Columbia-CVN-STAT5702/CrimeDataNYC (https://github.com/Columbia-CVN-STAT5702/CrimeDataNYC)

Limitations:

- **Time**: There are seemingly endless avenues we could explore, and directions we could go. We aspired to change a set of two graphs into an interactive D3 visualization of 4 to 6 graphs. We also aspired to bring in more and more datasets to test more hypotheses, such as, “did the stop-and-frisk” policy influence crime? We simply ran out of time.
- **Brevity**: We could have had twice as many graphs, as this data was pretty rich and nuanced, and our explorations took multiple directions and multiple techniques for visualizing the data
- **Technical issues**:
  - Free instance of Shiny cloud allow 1GB per account to host the application. Working with large/extended dataset drag us in to many memory sensitive and deployment challenges. Therefore, we re-scoped our application using limiting the number of plots.
  - Different browsers resulted in different response from the ShinyApp

Future Directions:

- Testing more hypotheses with other data sets
  - Stop and Frisk
  - More economic indicators
- Deeper dive on sociological measures and testing against Violent Crime and Drugs
- Converting other sets of graphs into interactive graphs, possibly through D3

- Create our own Project Website from Github

Lessons Learned

- Didn’t take full advantage of github to coordinate works from group members, a lot of merging were done manually. Need to get more familiar with how github can help to merging contribution from team members.
- Teaming with 5 part-time people from different time zones and with only online meeting required more coordination
- A five person team created a task of prioritization and editing down from lots and lots and lots more graphs we used to explore this data

Debates:

- We debated whether we should use a pie chart or not
- We debated if we should change the column names of the dataset
- We debated to remove most of our detailed Data Quality Analysis section

And with that, we submit this project. - Anita, Brent, Jingbo, Rashmi and Rich

:)