The Correlation Between Weather and Crime

Introduction

I have used the National Oceanic and Atmospheric Association(NOAA) weather dataset and Chicago Crime Dataset from the City of Chicago. I chose these because they both have daily entries, and I was able to merge the datasets by day. These will allow to me find a day by day relationship between weather and crime, as I felt that monthly was too far apart to find meaningful relationships. I also liked the crime dataset because it labels the type of each crime. This way I am able to investigate the relationship between weather and different types of crimes.

Getting a Weather Dataset

I used the Weather Dataset from the National Oceanic and Atmospheric Association(NOAA). I was able to find a dataset for daily weather reports in Chicago from 1949 to 2009, which I filtered down to only the 2008 reports. I only keps values such as precipitation, snow, and temperatur max and mins, along with the key values year, month, and day.

```
weather <- read.dly("ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/daily/all/USW00094846.dly")

# keep only 2008 entries
weather_2008 <- weather %>%
  filter(YEAR == 2008)

# Select variables of interest
weather_2008 <- weather_2008 %>%
  select(YEAR, MONTH, DAY, PRCP.VALUE, SNOW.VALUE,SNWD.VALUE, TMAX.VALUE, TMIN.VALUE)

# check out the dataset
save(weather_2008, file='weather.RData')
rm(weather)
```

Gathering the Chicago Crime Dataset

I used the official dataset from the City of Chicago, recording all reported crimes in 2008. I then split the date into year, month, and day to prepare for merging with the weather datset. Finally, I only kept the variables year, month, day, as well as the type of crime commited.

```
# Download one year of crime data from the open data portal of city of Chicago
# NOTE: This may take a while depening on the strength of your internet connection
# First I ran read_csv() to find the default col_types() then I updated them to this:
type=cols( `CASE#` = col_character(),
           `DATE  OF OCCURRENCE` = col_datetime(format="%m/%d/%Y %I:%M:%S %p"),
           BLOCK = col_factor(),
           IUCR = col_factor(),
           `PRIMARY DESCRIPTION` = col_factor(),
           `SECONDARY DESCRIPTION` = col_factor(),
           `LOCATION DESCRIPTION` = col_factor(),
           ARREST = col_factor(),
           DOMESTIC = col_factor(),
           BEAT = col_factor(),
           WARD = col_factor(),
           `FBI CD` = col_factor(),
           `X COORDINATE` = col_double(),
           `Y COORDINATE` = col_double(),
           LATITUDE = col_double(),
           LONGITUDE = col_double(),
           LOCATION = col_character()
)

# Read in data
crime_raw <- read_csv('Crimes_-_2008.csv', na='',col_types = type)

# Fix column names
names(crime_raw)<-str_to_lower(names(crime_raw)) %>%
  str_replace_all(" ","_") %>%
  str_replace_all("__","_") %>%
  str_replace_all("#","_num")

crime_2008 <- crime_raw %>%
  separate(date, c('MONTH', 'DAY', 'YEAR'), sep = c('/'))

crime_2008 <- crime_2008 %>%
  separate(YEAR, c('YEAR', 'TIME'), sep = c(' '))

crime_2008$YEAR <- as.numeric(crime_2008$YEAR)
crime_2008$MONTH <- as.numeric(crime_2008$MONTH)
crime_2008$DAY <- as.numeric(crime_2008$DAY)

crime_2008 <- crime_2008 %>% select(YEAR, MONTH, DAY, primary_type)

crime_2008 <- crime_2008 %>% arrange(YEAR)

save(crime_2008, file='crime.RData')

rm(crime_raw)
```

Merging the Crime and Weather Datasets

Next, I merged the two datasets by key ID's Year, Month, Day. First, I inspected the two datasets. The crime dataset had 427142 observations, while weather_2008 had 366 (one for each day as 2008 was a leap year). Performing an antijoin on the two resulted in a table with 0 observations, meaning that there were no mismatches between the datasets!. I performed a left join on the crime dataset with the weather dataset, resulting in a table with the same number of observations as the crime datasets since no rows had to be dropped. Finally, I reverted to a singular date columm to make plotting a bit easier.

```
## inspect weather dataset

load('weather.RData')
load('crime.RData')
head(weather_2008)%>%
  kbl() %>%
  kable_styling(bootstrap_options = c("striped", "hover", "condensed"))
```

	YEAR	MONTH	DAY	PRCP.VALUE	SNOW.VALUE	SNWD.VALUE	TMAX.VALUE	TMIN.VALUE
745	2008	1	1	1.3	20	0	-2.8	-15.6
745.1	2008	1	2	0.0	0	127	-6.7	-16.7
745.2	2008	1	3	0.0	0	102	-5.0	-17.8
745.3	2008	1	4	0.0	0	102	2.2	-5.6
745.4	2008	1	5	0.0	0	51	6.1	1.7
745.5	2008	1	6	0.3	0	0	15.6	5.0

```
#inspect crime dataset
head(crime_2008)%>%
  kbl() %>%
  kable_styling(bootstrap_options = c("striped", "hover", "condensed"))
```

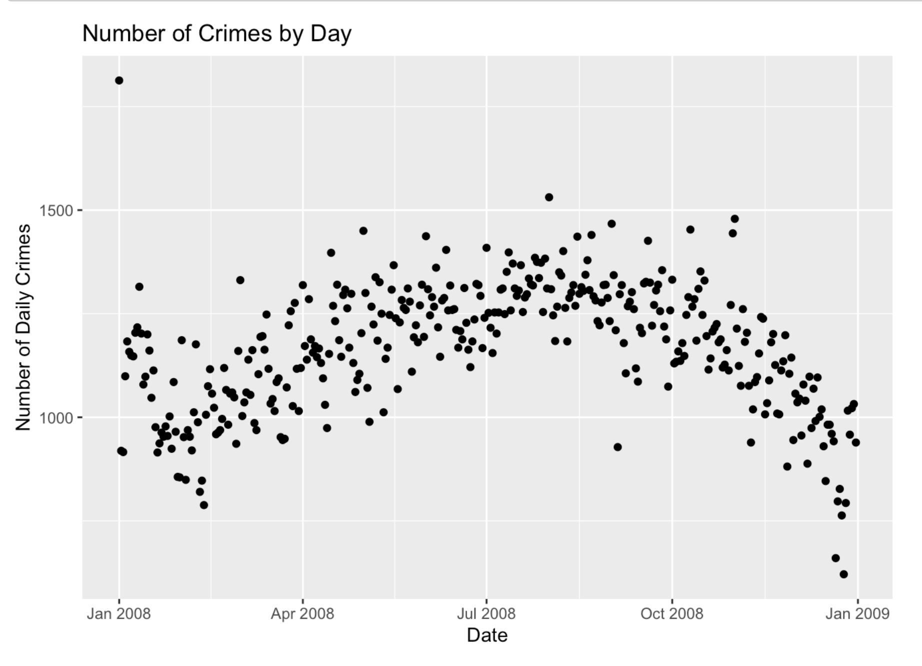
YEAR	MONTH	DAY	primary_type
2008	1	1	CRIM SEXUAL ASSAULT
2008	10	24	DECEPTIVE PRACTICE
2008	7	24	SEX OFFENSE
2008	1	1	OFFENSE INVOLVING CHILDREN
2008	2	26	MOTOR VEHICLE THEFT
2008	6	28	THEFT

```
merged_data <- left_join(crime_2008, weather_2008, by = c('YEAR', 'MONTH', 'DAY'))

merged_data$date <- as.Date(with(merged_data,paste(YEAR,MONTH,DAY,sep="-")),"%Y-%m-%d")
merged_data <- merged_data %>%
  select(-c('YEAR', 'MONTH' ,'DAY'))
```

Check if the number of crimes per day varies over time

Next, I checked to make sure that the number of crimes had some sort of vartiation over time.

```
merged_data %>%
  group_by(date)%>%
  count()%>%
  ggplot(aes(x = date, y = n)) +
  geom_point() +
  labs(
    x = 'Date',
    y = 'Number of Daily Crimes'
  )+
  ggtitle('Number of Crimes by Day')
```

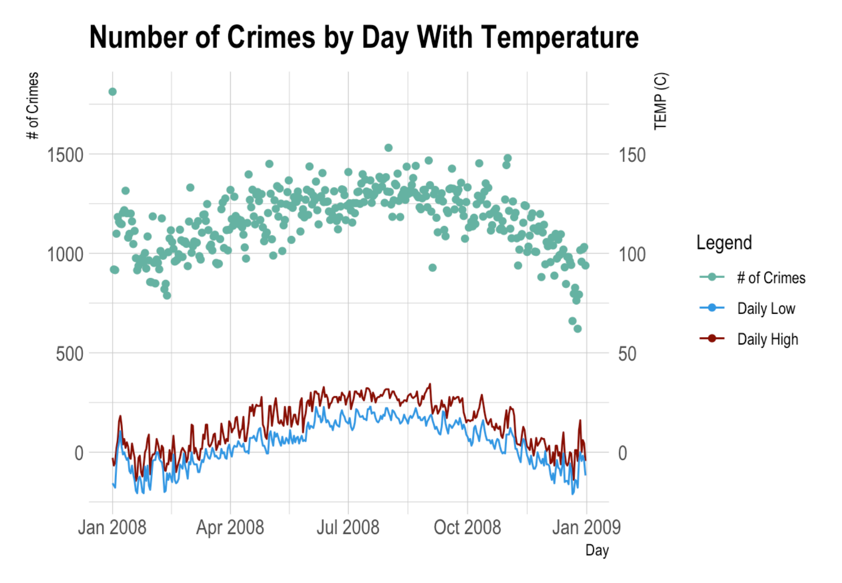


We can see that the amount of crime varies thoughout the year, with a peak in the middle of the year. Let's inspect this further

Inspect Relationship Between Temperature and Crime

Next, I wanted to plot the high and low temperature of each day on the same graph, to see if there was a correlation. I created a plot with 2 y axes: one for crime count and another for temperature.

```
crimeColor <- "#69b3a2"
lowTempColor <- rgb(0.2, 0.6, 0.9, 1)
highTempColor<- '#8a0303'

colors = c('# of Crimes'=crimeColor, 'Daily Low'=lowTempColor, 'Daily High'=highTempColor)
coeff = (10)
coeff2 = (4)
c2f  = 9/5

merged_data %>%
  group_by(date, TMAX.VALUE, TMIN.VALUE)%>%
  count()%>%
  ggplot(aes(x = date)) +

  geom_point(aes(y = n ,color = '# of Crimes')) +
  geom_line(aes(y = TMAX.VALUE * coeff , color = 'Daily High') ) +
  geom_line(aes(y = TMIN.VALUE * coeff , color = 'Daily Low')) +
  scale_y_continuous(

    # Features of the first axis
    name = "# of Crimes",

    # Add a second axis and specify its features
    sec.axis = sec_axis(~ . / coeff, name="TEMP (C)")
  ) +
  labs(x = "Day",
        color = "Legend")+
  scale_color_manual(values = colors)+
  ggtitle('Number of Crimes by Day With Temperature')+
  theme_ipsum()
```

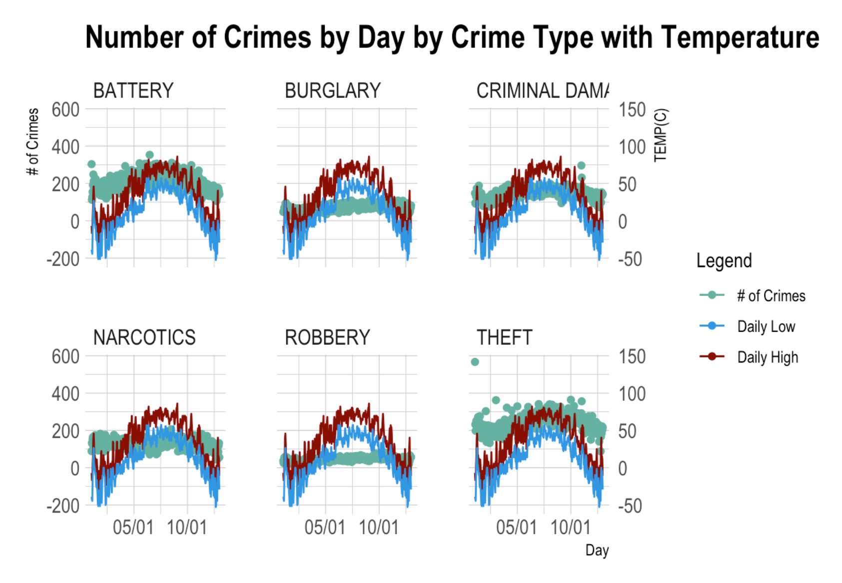


There seems to be a pretty strong correlation between the number of crimes and temperature. As temperature goes up, the number of crimes goes up with it. Let's see if that holds true for all types of criminal activity

Breaking up Crimes per Day by type of Crime

Seeing that there was a fairly strong correlation with crime and temparature, I wanted to see if this held true for all types of crimes. The crime dataset already labels each crime with a type, so I picked the six most interesting crimes to plot along with the same high and low temperatures

```
unique_days = unique(merged_data$date)
merged_data %>%
  filter(primary_type %in%
           c('BURGLARY', 'BATTERY', 'CRIMINAL DAMAGE', 'THEFT', 'NARCOTICS', 'ROBBERY'))%>%
  group_by(date, primary_type, TMAX.VALUE, TMIN.VALUE)%>%
  count()%>%
  ggplot(aes(x = date)) +
  geom_point(aes(y = n ,color = '# of Crimes')) +
  geom_line(aes(y = TMAX.VALUE * coeff , color = 'Daily High') ) +
  geom_line(aes(y = TMIN.VALUE * coeff , color = 'Daily Low')) +
  scale_y_continuous(

    # Features of the first axis
    name = "# of Crimes",

    # Add a second axis and specify its features
    sec.axis = sec_axis(~ . / coeff2, name="TEMP(C)")
  ) +
   labs(x = "Day",
        color = "Legend")+
  scale_color_manual(values = colors)+
  ggtitle('Number of Crimes by Day by Crime Type with Temperature')+
  theme_ipsum() +
  scale_x_date(breaks = '5 months', date_labels = '%m/%d')+
  facet_wrap(~primary_type)
```

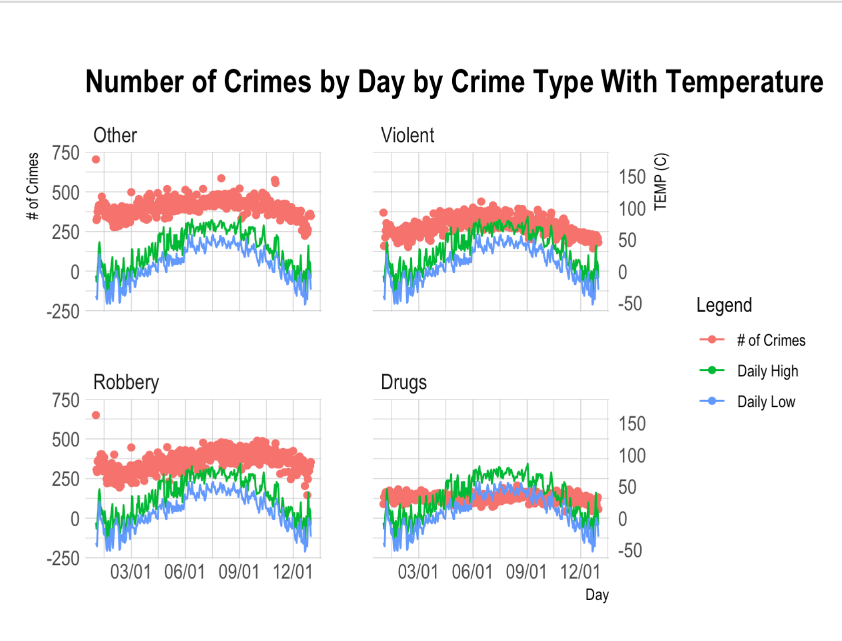


It seems that Battery ,Theft ,and Criminal Damage seem to follow the temparature trends, while the rest seem to stay fairly constant thorughout the year.

Grouping some Crime types together

Some of the Crimes listed have very similar types, so let's see if we can group more crimes together. We will create 4 larger groups of crimes: Violent, Robbery, and Drug Crimes, and a broad section for crimes not related to these

```
key <- c("Other", "Violent", "Robbery", "Drugs")
names(key) <- c("0", "1", "2", "3")

merged_data %>%
  group_by(date, primary_type, TMAX.VALUE, TMIN.VALUE)%>%
  mutate(
    type = primary_type %in% c('ARSON', 'ASSAULT', 'BATTERY', 'KIDNAPPING', 'INTIMIDATION')+ #Cr
imes we consider violent
    primary_type %in% c('BURGLARY', 'MOTOR_VEHICLE_THEFT', 'ROBBERY', 'THEFT')*2 + #Crimes we co
nsider robbery
    primary_type %in% c('LIQUOR_LAW_VIOLATION', 'NARCOTICS', 'OTHER_NARCOTIC_VIOLATION')*3 #Crim
es we consider drug use
  ) %>%

  group_by(date, type, TMAX.VALUE, TMIN.VALUE) %>%
  count() %>%
  ggplot(aes(x = date, y = value)) +
  geom_point(aes(y = n ,color = '# of Crimes')) +
  geom_line(aes(y = TMAX.VALUE * coeff , color = 'Daily High') ) +
  geom_line(aes(y = TMIN.VALUE * coeff , color = 'Daily Low')) +
  ggtitle('Number of Crimes by Day by Crime Type With Temperature')+
  scale_y_continuous(

    # Features of the first axis
    name = "# of Crimes",

    # Add a second axis and specify its features
    sec.axis = sec_axis(~ . / coeff2, name="TEMP (C)")
  ) +
  #scale_x_discrete(guide = guide_axis(n.dodge=3))+
  labs(x = "Day",
        color = "Legend")+
  scale_x_date(breaks = '3 months', date_labels = '%m/%d', )+
  theme_ipsum() +
  facet_wrap(~type, labeller = labeller(type = key))
```



All crimes seems to follow the same temperature trend except drug use, Which stays constant throughout the year. This shows how reql drug addiction is, as drug crimes stays constant

Crimes per Day v.s Precipitation

Finally, I investigated the relationship between daily crime and daily precipitation

```
range = range(merged_data$PRCP.VALUE)

precip_table <- merged_data %>%
  drop_na(PRCP.VALUE)%>%
  group_by(date, PRCP.VALUE)%>%
  count()%>%
  summarize(
    n_crimes = n,
  )%>%
  group_by(preciptiation_ranges=cut(PRCP.VALUE, breaks= seq(range[1], range[2], by = 6, inlude.l
owest = TRUE)))%>%
  summarize(mean_crimes = mean(n_crimes)) %>%
  drop_na(preciptiation_ranges)
```

```
## `summarise()` has grouped output by 'date'. You can override using the
## `.groups` argument.
```

```
  precip_table %>%
    kbl() %>%
    kable_styling(bootstrap_options = c("striped", "hover", "condensed"))
```

preciptiation_ranges	mean_crimes
(0,6]	1135.284
(6,12]	1114.167
(12,18]	1091.417
(18,24]	1232.429
(24,30]	1188.667
(30,36]	1184.000
(36,42]	1104.667
(42,48]	1016.000
(60,66]	1184.000
(72,78]	928.000

```
  precip_table %>%
    ggplot(aes(x = preciptiation_ranges, y = mean_crimes))+
    geom_point() +
    ggtitle("Crimes per Day with Precipitation") +
    labs(
      y = "# of Crimes",
      x = "Precipitation (mm)"
    )
```


```
  summary(merged_data$PRCP.VALUE)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   0.000   3.405   1.500 168.700
```

After investigating the plot, the relationship between precipitation and crimes per day doesn't seem to be as drastic as expected. There is some dropoff on days with heavier rain, but there are also very few days with that much precipitation, so this could just be a result of a small sample size.

Citations

City of chicago: Data Portal: City of chicago: Data Portal. Chicago. (n.d.). Retrieved March 30, 2022, from https://data.cityofchicago.org/

Index of /pub/data. (n.d.). Retrieved March 30, 2022, from https://www1.ncdc.noaa.gov/pub/data/