Week 8: Discrete data and GLM(M) regression

The Solution

26 March 2019

```
pacman::p_load(s20x, tidyverse, lme4, car, multcomp,
               AICcmodavg, gridExtra, xtable)
```

Data

Structure

```
## 'data.frame':    241 obs. of  10 variables:
##  $ Address       : Factor w/ 60 levels " 1 LAKESIDE AV",..: 33 56 57 35 41 56 44 48 28 58 ...
##  $ Date          : Factor w/ 186 levels " 1/1/2013"," 1/22/2012",..: 39 97 41 161 134 174 6 100 49 3
##  $ Time          : Factor w/ 130 levels " 16:00","0:00",..: 27 17 42 108 32 31 82 82 39 82 ...
##  $ SpecificCharge: Factor w/ 96 levels " 13 ASSAULT",..: 61 61 61 93 61 61 25 25 4 4 ...
##  $ ChargeType    : Factor w/ 10 levels "DOMESTIC VIOLENCE",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ GameDay       : Factor w/ 2 levels "No","Yes": 1 1 1 2 1 1 2 2 1 1 ...
##  $ Venue         : Factor w/ 2 levels "FirstEnergyStadium",..: 1 1 1 1 1 1 2 2 2 2 ...
##  $ AirTemp       : int  20 26 32 66 68 88 26 32 35 60 ...
##  $ Day           : Factor w/ 18 levels " Friday"," Friday ",..: 1 10 11 18 5 11 5 17 17 3 ...
##  $ Event         : Factor w/ 4 levels "Baseball","Basketball",..: NA NA NA 4 NA NA 2 NA NA NA ...
```

Compare discrete data with simple variance structure

Crime by game days and event type

```
gd.gg <-
  clev.d %>%
    filter(Venue == "GatewayProgessiveField") %>%
      ggplot() + theme_bw(20) +
        geom_bar(aes(x=GameDay), stat="count", fill="#377eb8")
et.gg <-
    clev.d %>%
    filter(Venue == "GatewayProgessiveField") %>%
      tidyr::drop_na(Event) %>%
        ggplot() + theme_bw(20) +
          geom_bar(aes(x=Event), stat="count", fill="#377eb8")
grid.arrange(gd.gg, et.gg, nrow =1 )
```

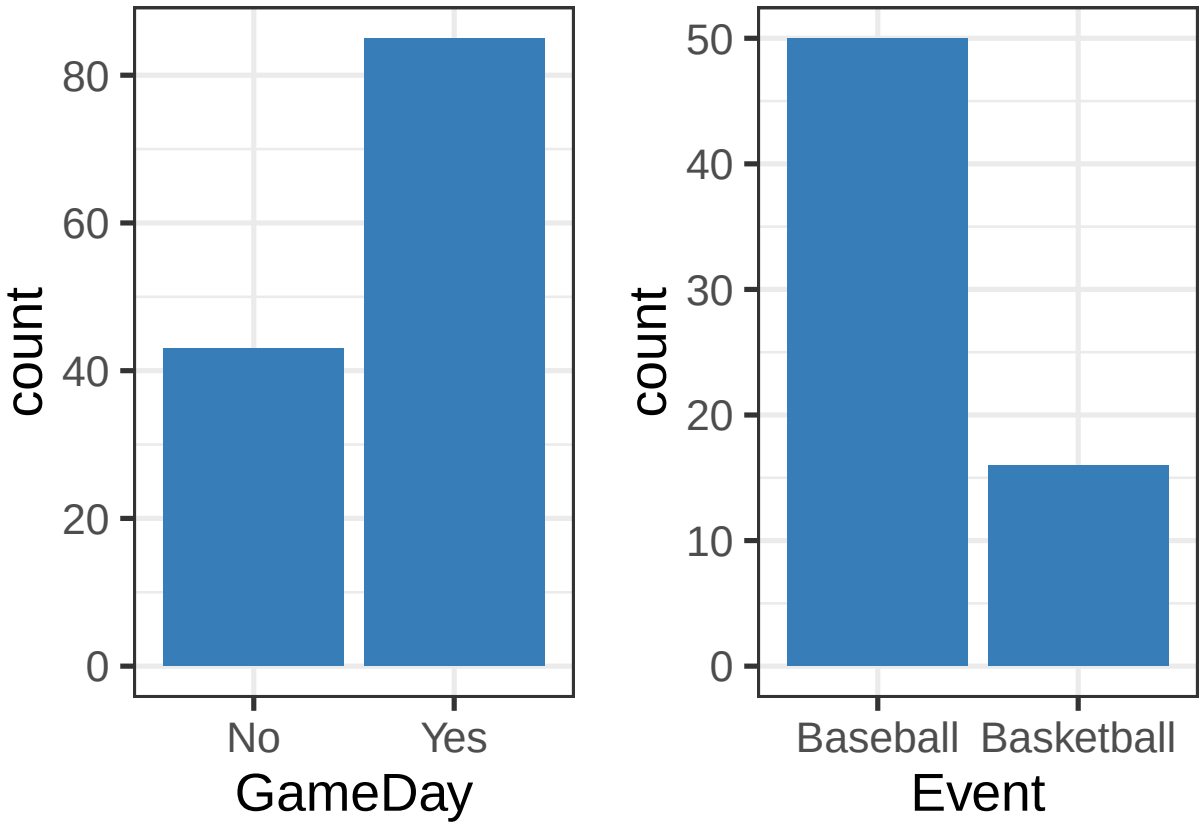

Figure 1: Crime data around the Quicken Loan Arena/Progessive Field major league sports venues in Cleveland (Gateway Park). LEFT: More charges are filed on game days than non-game days. RIGHT: More charges are filed on days with baseball games than basketball games

Crime by event type

Basketball and baseball are played at the Gateway Complex, in the Quicken Loans Arena and Progressive Field, respectively. It appears more charges are filed on days with baseball games than basketball games. There are two options to test this hypothesis:

Option 1: The basic χ^2 route:

```
clev.d %>%
   filter(Venue == "GatewayProgessiveField") %>%
     drop_na(Event) %>%
       xtabs(~ Event, data=. ) %>%
         chisq.test(.) %>%
           pander::pander(caption="
             Results of Chi-squared test showing baseball games
             incur significantly greater crime than basketball
             games around the Gateway Complex in Cleveland.")
```

Table 1: Results of Chi-squared test showing baseball games incur significantly greater crime than basketball games around the Gateway Complex in Cleveland.

Test statistic	df	P value
101	3	9.331e-22 * * *

Option 2: GLM with Poisson distribution:

```
clev.d %>%
  filter(Venue == "GatewayProgessiveField") %>%
    drop_na(Event) %>%
      group_by(Event) %>%
        summarize(charges = length(ChargeType)) %>%
glm(charges ~ Event, data=.,
            family=poisson(link = "log")) %>%
Anova(.) %>%
  pander::pander(caption="
            Results of GLM showing baseball games incur significantly
            greater crime than basketball games around the
                        Gateway Complex in Cleveland.")
```

Table 2: Results of GLM showing baseball games incur significantly greater crime than basketball games around the Gateway Complex in Cleveland.

	LR Chisq	Df	Pr(>Chisq)
Event	18.39	1	1.804e-05

In each case, results indicate there are significantly more charges filed on days with baseball games then basketball games. My colleague and I have a couple thoughts on this. One, there are just a lot more baseball games played in a season than basketball games, and some rate like charges/event would be necessary. Also, NBA games are expensive! That's a different clientele than a bunch of bros getting cheap outfield seats and putting down a bunch of stadium beers before hitting the bars after the game and starting fights and committing property damage.

Compare discrete data with non-independent variance

What special considerations must we give were we to model data from both venues together?

One must use a model sensitive to the fact that data from each venue are not independent. Thus venue should be included as a random effect to ensure variance is not pooled across venues in the model.

Data presentation

```
clev.d %>%
  group_by(Venue, GameDay) %>%
    summarize(charges = length(ChargeType)) %>%
      knitr::kable(caption="Counts of total charges at two major league sports venues
                   in Cleveland on game days and non-game days.")
```

Table 3: Counts of total charges at two major league sports venues in Cleveland on game days and non-game days.

Venue	GameDay	charges
FirstEnergyStadium	No	43
FirstEnergyStadium	Yes	70
GatewayProgessiveField	No	43
GatewayProgessiveField	Yes	85

```
clev.d %>%
     ggplot() + theme_bw(20) +
       geom_bar(aes(x=GameDay, fill=Venue), stat="count") +
 theme(legend.position = "top",
       legend.direction = "vertical")
```

Statistical testing and interpretation

- It is important that our model include Venue as a random effect and use a discrete distribution to account for the count data.
- A `glmer` with a Poisson distribution is appropriate for these data. \

```
mod.dat <-
  clev.d %>%
    group_by(Venue, GameDay) %>%
      summarize(charges = length(ChargeType))

  NullMod <- glmer(charges ~ 1 + (1|Venue), mod.dat,
                   family=poisson(link = "log"))
  GameDay <- glmer(charges ~ GameDay + (1|Venue), mod.dat,
                family=poisson(link = "log"))
  anova(NullMod, GameDay) %>%
    pander::pander(caption="Across both major leage sports venues in Cleveland, more
                   charges are filed on game days than non-game days.")
```

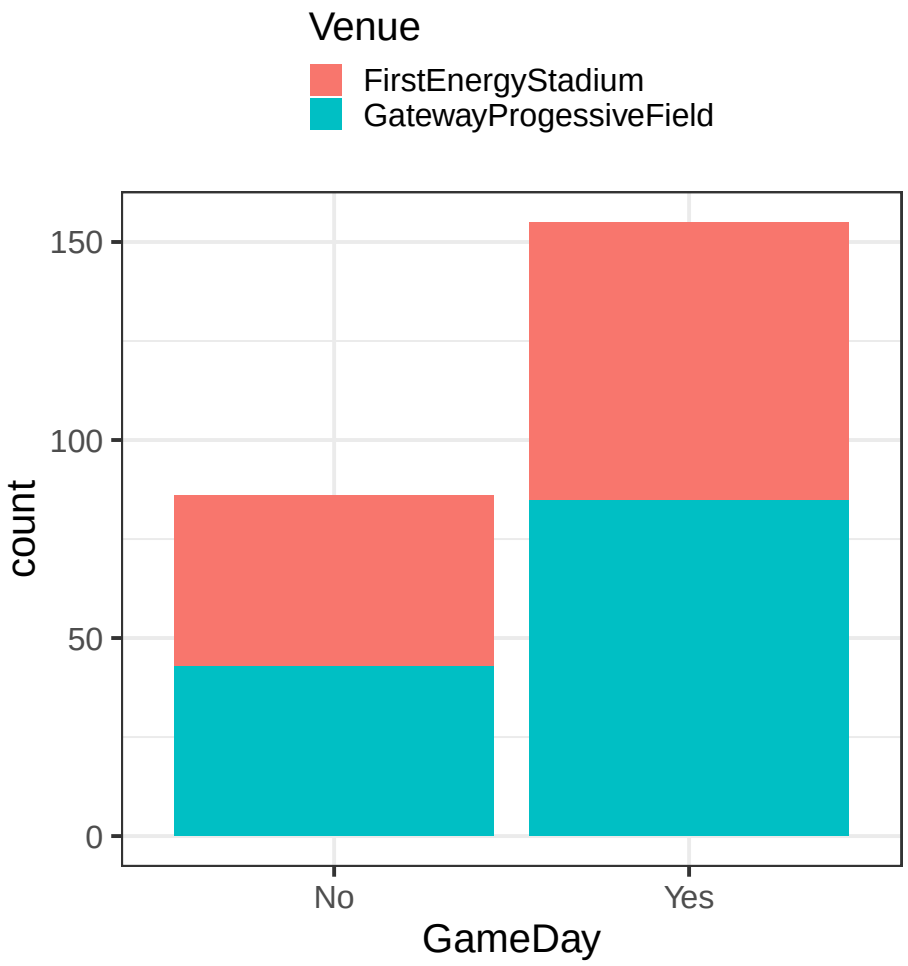

Figure 2: Total charges filed on game days and non-game days at both major league sports venues in Cleveland.

Table 4: Across both major leage sports venues in Cleveland, more charges are filed on game days than non-game days.

	Df	AIC	BIC	logLik	deviance	Chisq	Chi Df	Pr(>Chisq)
NullMod	2	49.07	47.84	-22.53	45.07	NA	NA	NA
GameDay	3	31.03	29.19	-12.52	25.03	20.03	1	7.606e-06

Going further

- The number of charges does vary with charge type ($\chi^2 = 167$, $P < 0.001$).
- One might need to use another distribution but the Poisson continues to work well for these data.

```
  clev.d %>%
    group_by(ChargeType, GameDay) %>%
    summarize(ChargeCount = length(ChargeType) )  %>%
    ggplot() + theme_bw(16) +
    geom_bar(aes(x=reorder(ChargeType, -ChargeCount, sum),
                 y=ChargeCount, fill=GameDay), stat="identity") +
    labs(x="Charge type",
         y=" Number of charges filed") +
    theme(axis.text.x = element_text(angle=45, hjust=1),
          panel.grid.major.x = element_blank())
```

```
type.d <-
  clev.d %>%
    group_by(Venue, ChargeType, GameDay) %>%
    summarize(ChargeCount = length(ChargeType) )

type.glm <- glmer(ChargeCount ~ ChargeType + (1|Venue), type.d,
              family=poisson(link = "log"))
Anova(type.glm) %>%
  xtable(caption="Statistical evidence that the number of charges varies
         with charge type.") %>%
          print(comment=FALSE, include.rownames=FALSE)
```

Chisq	Df	Pr(>Chisq)
166.48	9	0.0000

Table 5: Statistical evidence that the number of charges varies with charge type.

When focusing on just the three most-reported charge types:

- There are significant differences between the number of charges per type. In post-hoc pairwise comparisons, Violent crimes were more frequent than both property damage and resisting arrest ($P < 0.001$ for each), but a trend towards fewer charges for resisting arrest than property damage was not statistically significant ($P = 0.19$).
- AIC_c-based model selection (Table 8) indicated that game day had an increasing effect on the number of charges for the three most-frequent charge types (95% CI: 0.54–1.23).
- Charges for violent crimes were the most frequent, while rates for resisting arrest and property damage were substantially aobve zero but not different from each other (Fig. 4).

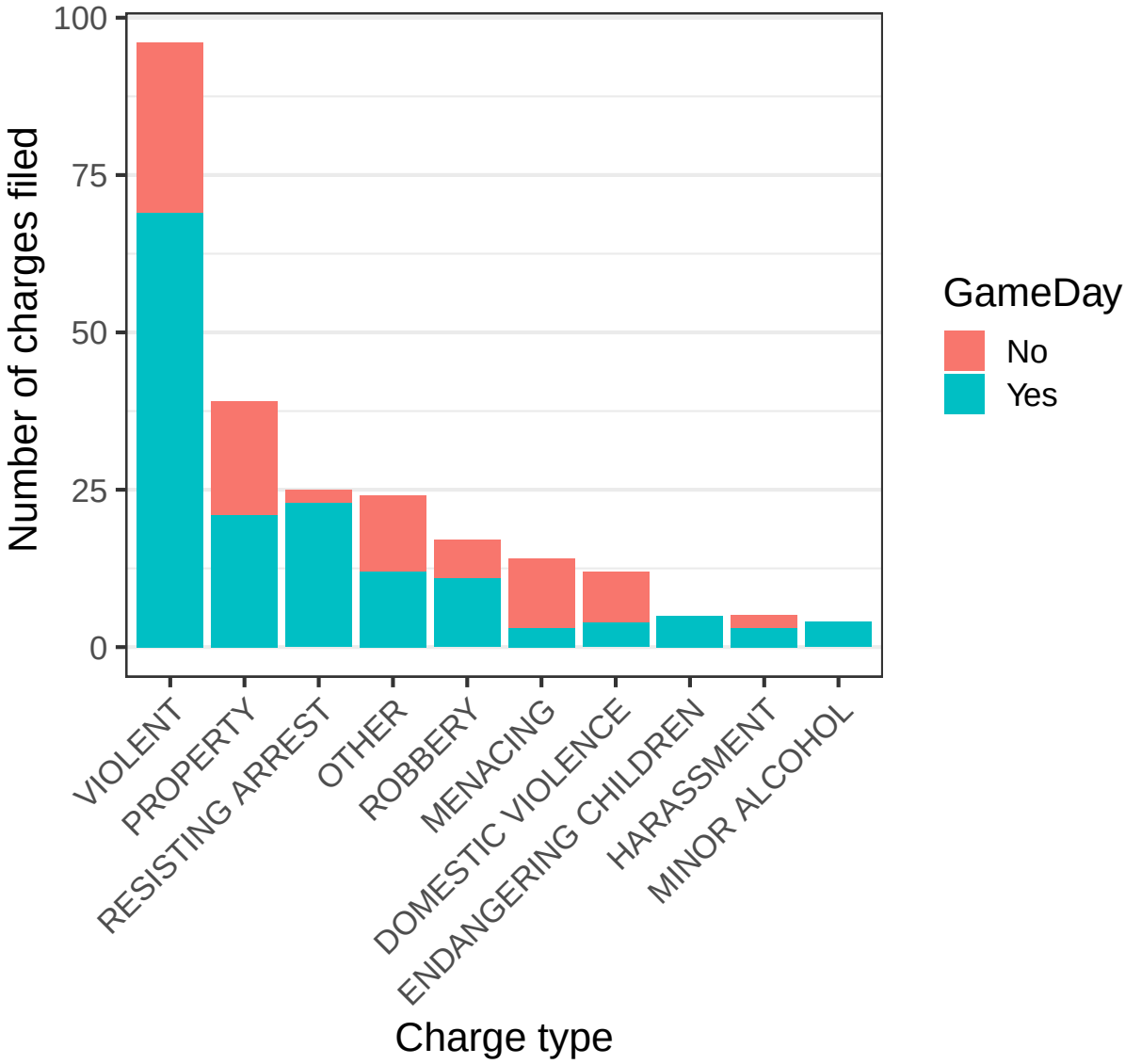

Figure 3: Frequency of charges by charge type and whether they occurred on a game day or not.

```
Top3.d <-
  clev.d %>%
   filter(ChargeType %in% c("VIOLENT",
                            "PROPERTY",
                            "RESISTING ARREST")) %>%
    group_by(Venue, ChargeType, GameDay) %>%
    summarize(ChargeCount = length(ChargeType) )

type2.glm <- glmer(ChargeCount ~ ChargeType + (1|Venue), Top3.d,
              family=poisson(link = "log"))

glht.res <- cbind(round(summary(glht(type2.glm,
                        linfct=mcp(ChargeType ="Tukey")))$test$coefficients,2),
                  round(summary(glht(type2.glm,
                        linfct=mcp(ChargeType ="Tukey")))$test$pvalues,2))
colnames(glht.res) <- c( "Coefficient", "P-value")
xtable(glht.res, caption="Results of post-hoc pairwise comparison of the
       top three most-reported charge types.") %>%
          print(comment=FALSE) #, include.rownames=FALSE)
```

	Coefficient	P-value
RESISTING ARREST - PROPERTY	-0.44	0.19
VIOLENT - PROPERTY	0.90	0.00
VIOLENT - RESISTING ARREST	1.35	0.00

Table 6: Results of post-hoc pairwise comparison of the top three most-reported charge types.

```
  type2.null <- glmer(ChargeCount ~ 1 + (1|Venue), Top3.d,
                      family=poisson(link = "log"))
  type2.gd<- glmer(ChargeCount ~ GameDay + (1|Venue), Top3.d,
                     family=poisson(link = "log"))
  type2.ct<- glmer(ChargeCount ~ ChargeType + (1|Venue), Top3.d,
                     family=poisson(link = "log"))
  type2.add<- glmer(ChargeCount ~ ChargeType + GameDay + (1|Venue), Top3.d,
                      family=poisson(link = "log"))
  type2.int<- glmer(ChargeCount ~ GameDay * ChargeType + (1|Venue), Top3.d,
                     family=poisson(link = "log"))

  cand.mod.names <- c("type2.null", "type2.gd", "type2.ct",
                      "type2.add", "type2.int")
  cand.mods <- list( )
  for(i in 1:length(cand.mod.names)) {
    cand.mods[[i]] <- get(cand.mod.names[i]) }

        aictab(cand.set = cand.mods,
               modnames = cand.mod.names) %>%
     rownames_to_column( var = "Model") %>%
      xtable(caption="AICc table showing that the additive model, with ChargeType and
             GameDay terms, is the only competitive model in the model set.") %>%
          print(comment=FALSE, include.rownames=FALSE)
```

```
    confint(type2.add) %>%
  as.data.frame %>%
           round(., 2) %>%
```

Model	K	K.1	AICc	ModelLik weight	log-Likelihood
4	type2.add	5.00	91.92	1.00	0.99
5	type2.int	7.00	102.01	0.01	0.01
3	type2.ct	4.00	113.68	0.00	0.00
2	type2.gd	3.00	131.47	0.00	0.00
1	type2.null	2.00	155.86	0.00	0.00

Table 7: AICc table showing that the additive model, with ChargeType and GameDay terms, is the only competitive model in the model set.

```
    rownames_to_column( var = "term") %>%
      slice(-1) %>%
        mutate(estimate = round(fixef(type2.add),2 ) ,
               term = c("Charge type: Property",
                        "Charge type: Resisting Arrest",
                        "Charge type: Violent",
                        "Game day") ) %>%
      xtable(caption="In the context of the assignment, the take-home of this analysis
             is that game day has a positive, non-zero effect on the number of charges filed.", label="
          print(comment=FALSE, include.rownames=FALSE)
```

term	2.5 %	97.5 %	estimate
Charge type: Property	1.33	2.12	1.75
Charge type: Resisting Arrest	-0.96	0.05	-0.44
Charge type: Violent	0.54	1.28	0.90
Game day	0.54	1.23	0.88

Table 8: In the context of the assignment, the take-home of this analysis is that game day has a positive, non-zero effect on the number of charges filed.

```
  t3.d <- Top3.d %>%  filter(GameDay =="Yes")
 type3.glm <-  glmer(ChargeCount ~ 0 + ChargeType + (1|Venue), data=t3.d,
                     family=poisson(link = "log"))
confint(type3.glm) %>%
  as.data.frame %>%
    rownames_to_column( var = "term") %>%
      slice(-1) %>% # Cuts the (Intercept) row
        mutate(estimate = fixef(type3.glm),
               term = gsub("[[:lower:]]","", term),
               term = gsub(substr(term, 1,2), "", term) ) %>%
        plyr::rename(c("2.5 %" = "lower",
                       "97.5 %" = "upper")) %>%
    ggplot() +
      coord_flip() + theme_bw(20) +
      geom_hline(yintercept = 0) +
      geom_errorbar(aes(x=term,
                        ymin=lower,
                        ymax=upper),
                        width=0.25, size=1.25, color="#377eb8") +
      geom_point(aes(x=term,
                     y=estimate),
                 size=5, pch=21, stroke=2,
                 bg="#377eb8", color="white")
```

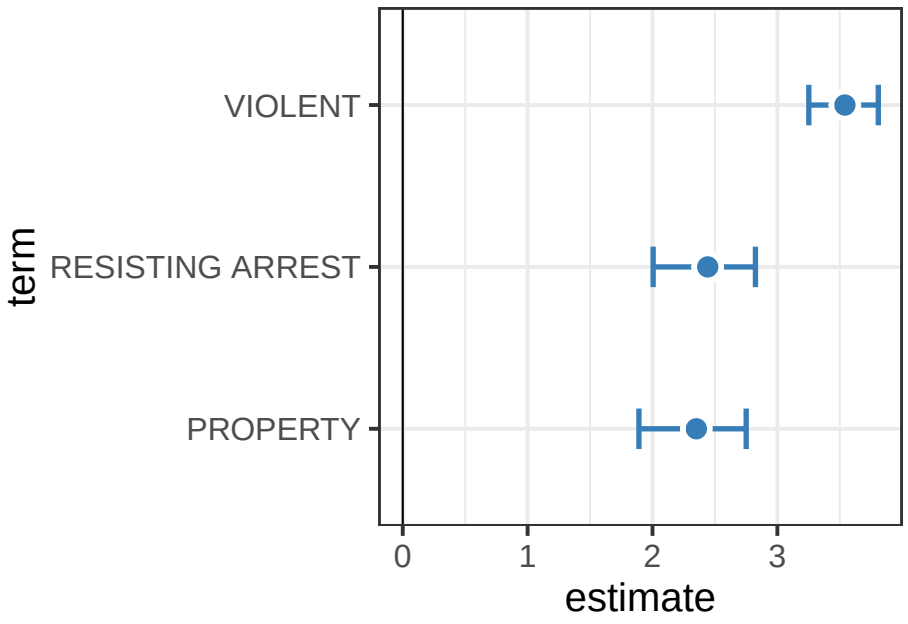

Figure 4: Charges for violent crimes are the most likely to be reported on game days around major league sports venues in Cleveland. The difference between these and the table above is that the data have been limited to charges filed on game days.

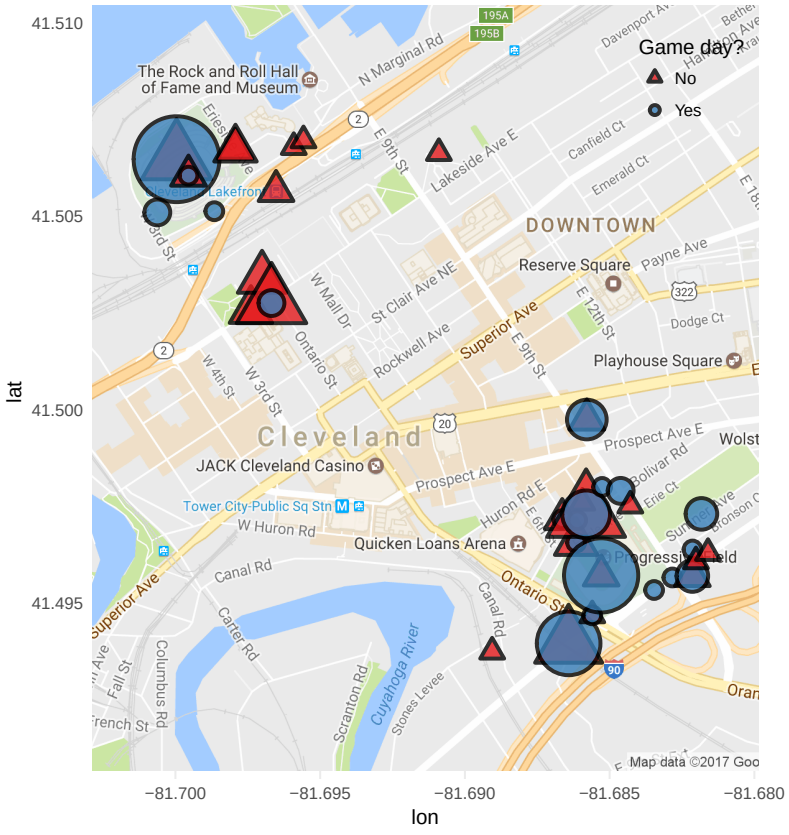

Figure 5: A map of downtown Cleveland showing crime clusters in the neighborhood of each major league sports venue. Blue circles denote charges filed on game days while red triangles denote charges filed on non-game days. Symbol size scales with number of charges reported at each location.