DAT (DATA ANALYTICS TEAM) PRESENTS:

Predicting Crime in Queens.

William D. Lombardi

Adam Miller

Max Gerstman

James Lu

WHO ARE WE?

OUR OBJECTIVE:

Using a variety of predictive modeling techniques, to what degree of accuracy can we predict the level of crime that will be committed in the bars and nightclubs of Queens, NY?

OUR DATASET:

NEW YORK CITY CRIMES
2014-2015.

PROBLEM FACED:

OVER 1 MILLION CRIME RECORDS

Dimensionality reduction

DATA CLEANING/PRE-PREPROCESSING:

```r
nycrime_data <- read.csv(filepath)
nycrime_data <- nycrime_data[nycrime_data$BORO_NM == 'QUEENS',]
nycrime_data <- nycrime_data[nycrime_data$PREM_TYP_DESC == 'BAR/NIGHT CLUB',]
nycrime_data <- na.omit(nycrime_data)

write.csv(nycrime_data, 'Clean_NYPD_Data.csv')
nycrime_data <- read.csv('Clean_NYPD_Data.csv')

num <- c(3,4,9,12,13,16,17)

nycrime_data.clean <- subset(nycrime_data, select = c(num))
colnames(nycrime_data.clean) <- c('Date','Time','Desc','Success','LVO','Precinct','Occurence')
nycrime_data.clean$Precinct <- as.factor(nycrime_data.clean$Precinct)
nycrime_data.clean$Date <- as.Date(nycrime_data.clean$Date, "%m/%d/%Y")

nycrime_data.clean <- nycrime_data.clean %>%
mutate(month = format(Date, "%m"), year = format(Date, "%Y")) %>%
group_by(month,year)
nycrime_data.clean <- nycrime_data.clean[,-1]

z <- lapply(str_split(nycrime_data.clean$Time, ":"), '[',1)
nycrime_data.clean$Hour <- as.character(z)
nycrime_data.clean <- nycrime_data.clean[,-1]
cols <- c('month','year','Hour')
nycrime_data.clean[cols] <- lapply(nycrime_data.clean[cols], factor)
```
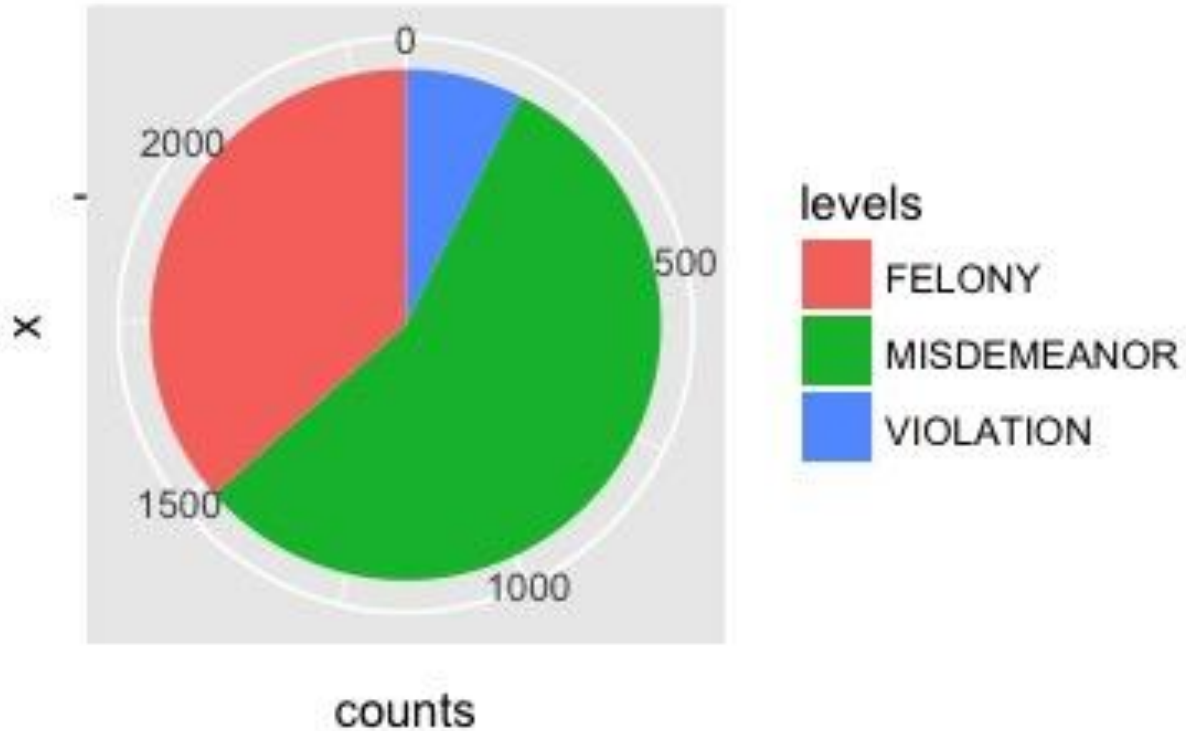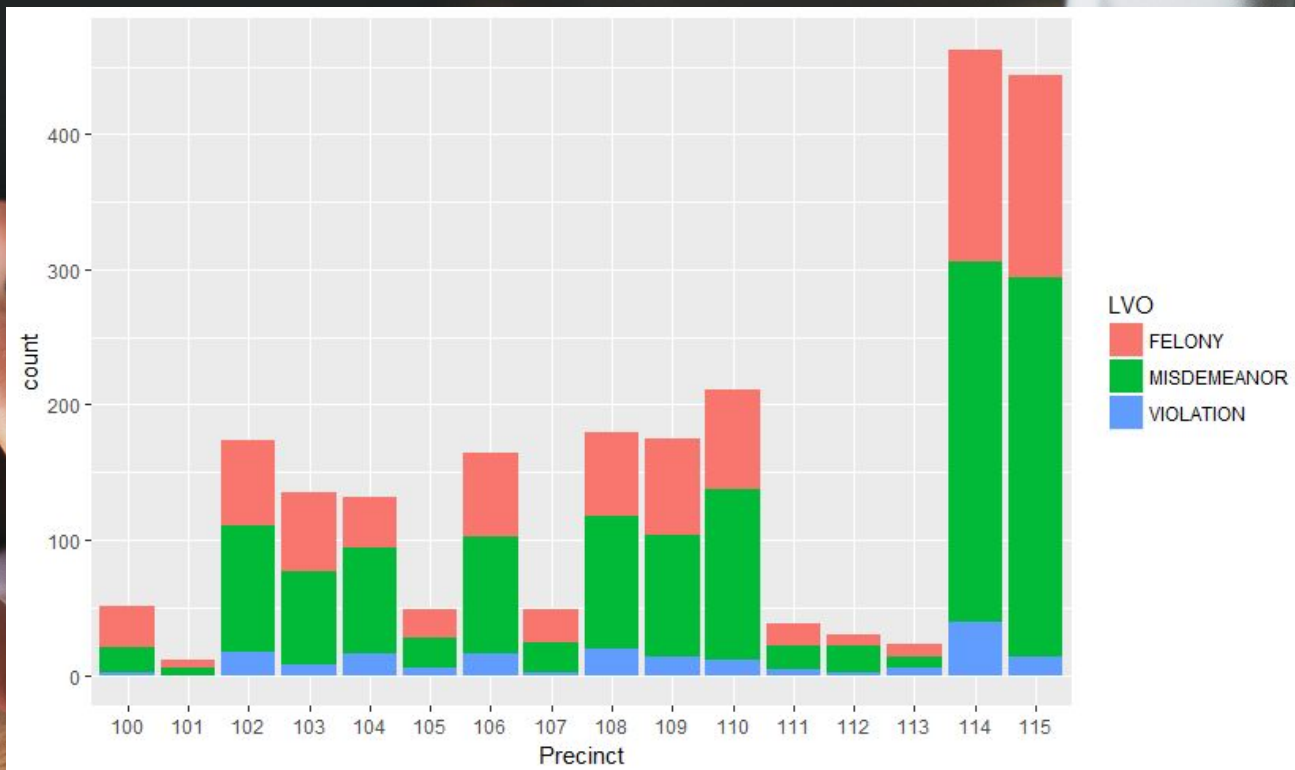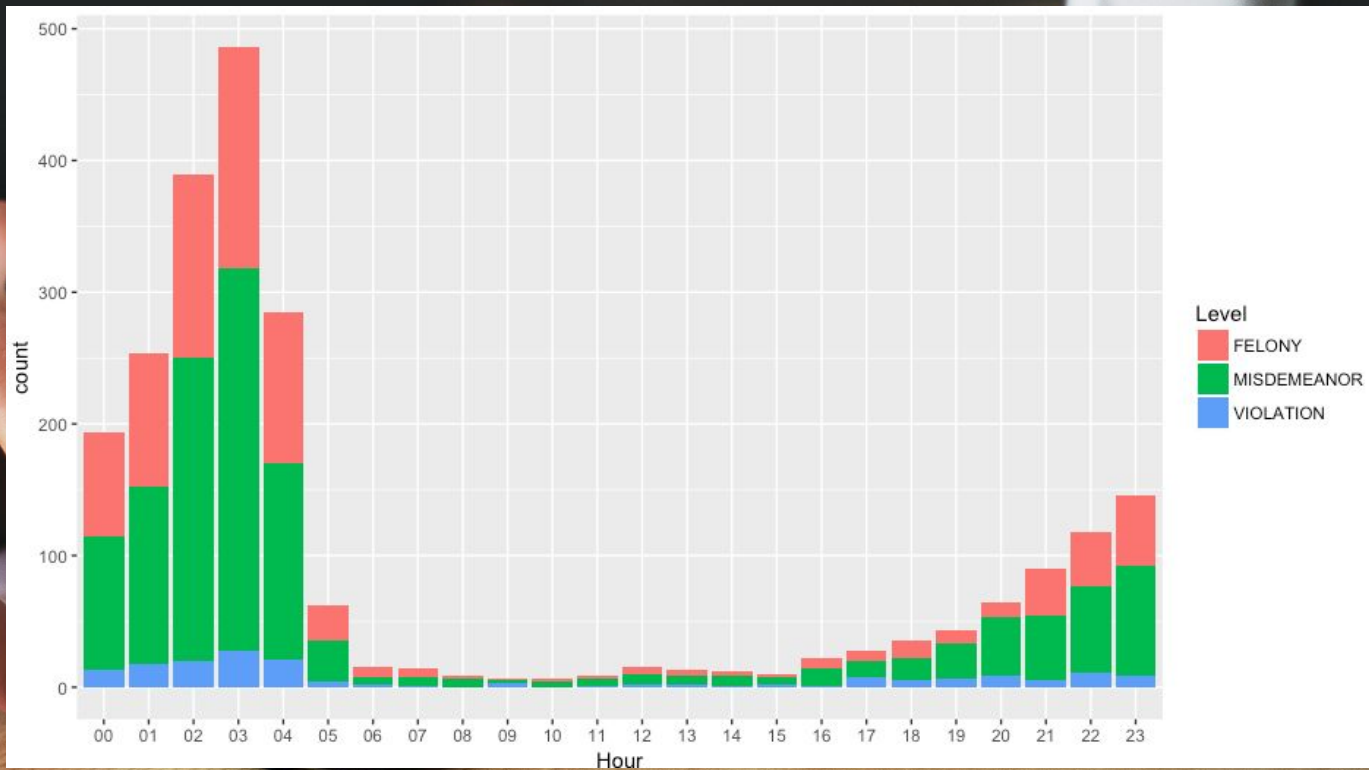
DESCRIPTIVE STATISTICS

DESCRIPTIVE STATISTICS:

DESCRIPTIVE STATISTICS:

CMPLNT_FR_DT	Exact date of occurrence for the reported event (or starting date of occurrence, if CMPLNT_TO_DT exists)
CMPLNT_FR_TM	Exact time of occurrence for the reported event (or starting time of occurrence, if CMPLNT_TO_TM exists)
OFNS_DESC	Description of offense corresponding with key code
CRM_ATPT_CPTD_CD	Indicator of whether crime was successfully completed or attempted, but failed or was interrupted prematurely
LAW_CAT_CD	Level of offense: felony, misdemeanor, violation
ADDR_PCT_CD	The precinct in which the incident occurred
LOC_OF_OCCUR_DESC	Specific location of occurrence in or around the premises; inside, opposite of, front of, rear of

PREDICTIVE MODELING
- Decision Tree (RPart)
- Naive Bayes
- Support Vector Machine

DECISION TREE:

```
nycrime.dtmodel <- rpart(LVO ~ ., data = nycrime_data.clean,
                          method='class',
                          minsplit = 1,
                          minbucket =5)
```

OUR RESULTS:

Classification Tree for Queens Crime LVO

94.8%
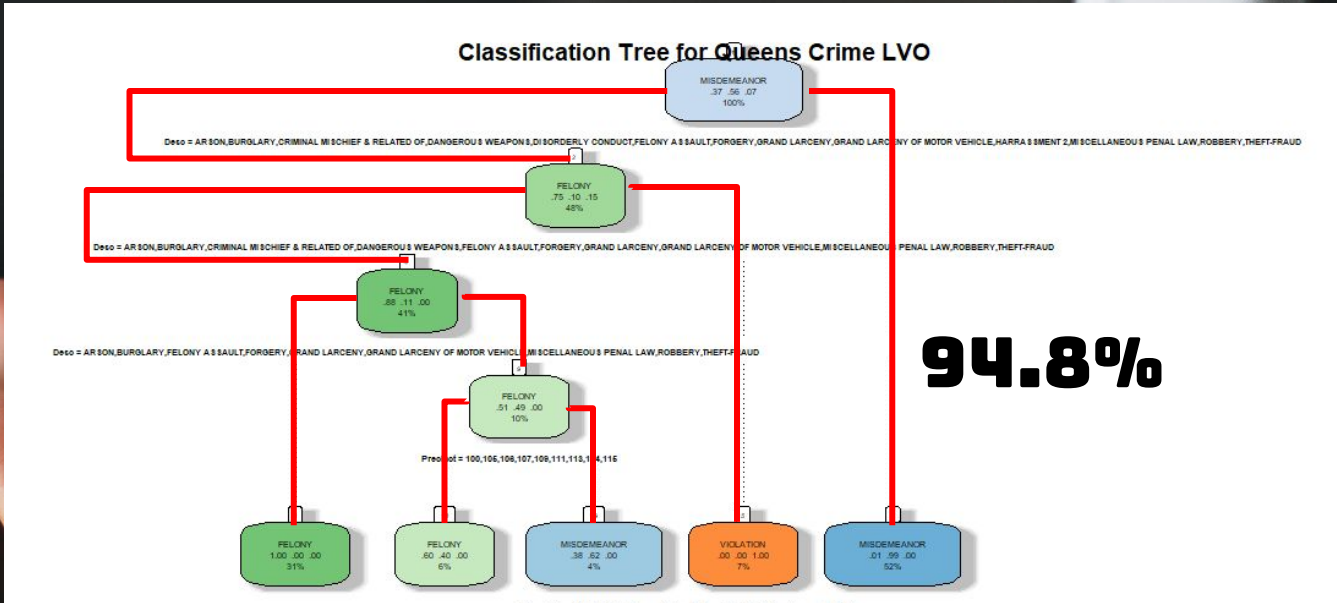
Prediction	Reference		
	FELONY	MISDEMEANOR	VIOLATION
FELONY	164	12	0
MISDEMEANOR	12	235	0
VIOLATION	0	0	43

MAIN SPLITTING ATTRIBUTES
- Crime Description
- Precinct

NAIVE BAYES:

Initial

```
nycrime.nb <- train(x = train_x, y = train_y, method = 'nb', trControl=trainControl(method='cv', number=10))
nycrime.nb

nycrime.nb.predict <- predict(nycrime.nb$finalModel, test_x)

confusiontable.nb <- table(test_y,nycrime.nb.predict$class,dnn=c('Actual','Prediction'))
accuracy_initial <- sum(diag(confusiontable.nb)) / sum(confusiontable.nb)
accuracy_initial
```

DT Predictors

```
x <- nycrime_data.clean[,c(1,4)]
train_x <- x[index,]
test_x <- x[-index,]

nycrime.nb.updated <- train(x = train_x, y = train_y, method = 'nb', trControl=trainControl(method='cv', number=10))
nycrime.nb.updated

nycrime.nb.updated.predict <- predict(nycrime.nb.updated$finalModel, test_x)

confusiontable.nb.updated <- table(test_y,nycrime.nb.updated.predict$class,dnn=c('Actual','Prediction'))
accuracy_updated <- sum(diag(confusiontable.nb.updated)) / sum(confusiontable.nb.updated)
accuracy_updated
```

OUR RESULTS:

Initial

```
                    Prediction
Actual          FELONY MISDEMEANOR VIOLATION
    FELONY         161          18         0
    MISDEMEANOR      7         250         0
    VIOLATION        0           0        30
```

DT Predictors

```
                    Prediction
Actual          FELONY MISDEMEANOR VIOLATION
    FELONY         161          18         0
    MISDEMEANOR      8         249         0
    VIOLATION        0           0        30
```

SUPPORT VECTOR MACHINES:

Initial

```
train_data <- nycrime_data.clean[index,]
test_data <- nycrime_data.clean[-index,]

nycrime.svm <- svm(LVO~., data=train_data)
nycrime.svm.predict <- predict(nycrime.svm, test_data)
mean(nycrime.svm.predict == test_data$LVO)
```

DT Predictors

```
nycrime.svm.dt <- svm(LVO~Desc+Precinct, data=train_data)
nycrime.svm.dt.predict <- predict(nycrime.svm.dt, test_data)
mean(nycrime.svm.dt.predict == test_data$LVO)
```

OUR RESULTS:

Initial

```
                Reference
Prediction     FELONY MISDEMEANOR VIOLATION
FELONY          172        21        0
MISDEMEANOR       7       236        0
VIOLATION         0         0       30
```

DT Predictors

```
                Reference
Prediction     FELONY MISDEMEANOR VIOLATION
FELONY          149         0        0
MISDEMEANOR      30       257        0
VIOLATION         0         0       30
```

CONCLUSIONS:

- **ALL MODELS RELATIVELY ROBUST**
  - Achieved >93% for each algorithm used
  - With exception of decision tree, all built in very little time
- **EITHER LITTLE OR NO DIFFERENCE BETWEEN MODEL ITERATIONS**
- **OBSERVED OR INFERRED BIASES**
  - Crime is much more prevalent in certain precincts over others
  - May be at the behest of the opinion of whoever was the first responder
  - Majority of crime committed between the hours of 1-3 AM

QUESTIONS?