NYPD Project

2025-05-13

NYPD Shooting Project

In this project, we will analyze a dataset from the New York City Police Department containing information pertaining shootings. Our goal will be to train a linear model that will predict whether a shooting led to a murder.

Loading the data

First, we will load the data

```
nypd_shooting <- read.csv('https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD')
```

Initial Exploratory Data Analysis

Now that we have loaded our data, let's proceed with some Exploratory Data Analysis. First, we need to see what data we have available.

```
str(nypd_shooting)
```

```
## 'data.frame':    29744 obs. of  21 variables:
##  $ INCIDENT_KEY          : int  231974218 177934247 255028563 25384540 72616285 85875439 797
80323 85744504 142324890 152868707 ...
##  $ OCCUR_DATE            : chr  "08/09/2021" "04/07/2018" "12/02/2022" "11/19/2006" ...
##  $ OCCUR_TIME            : chr  "01:06:00" "19:48:00" "22:57:00" "01:50:00" ...
##  $ BORO                  : chr  "BRONX" "BROOKLYN" "BRONX" "BROOKLYN" ...
##  $ LOC_OF_OCCUR_DESC     : chr  "" "" "OUTSIDE" "" ...
##  $ PRECINCT              : int  40 79 47 66 46 42 71 69 75 69 ...
##  $ JURISDICTION_CODE     : int  0 0 0 0 0 2 0 2 0 0 ...
##  $ LOC_CLASSFCTN_DESC    : chr  "" "" "STREET" "" ...
##  $ LOCATION_DESC         : chr  "" "" "GROCERY/BODEGA" "PVT HOUSE" ...
##  $ STATISTICAL_MURDER_FLAG: chr  "false" "true" "false" "true" ...
##  $ PERP_AGE_GROUP        : chr  "" "25-44" "(null)" "UNKNOWN" ...
##  $ PERP_SEX              : chr  "" "M" "(null)" "U" ...
##  $ PERP_RACE             : chr  "" "WHITE HISPANIC" "(null)" "UNKNOWN" ...
##  $ VIC_AGE_GROUP         : chr  "18-24" "25-44" "25-44" "18-24" ...
##  $ VIC_SEX               : chr  "M" "M" "M" "M" ...
##  $ VIC_RACE              : chr  "BLACK" "BLACK" "BLACK" "BLACK" ...
##  $ X_COORD_CD            : chr  "1006343" "1000082.937500000000000" "1020691" "985107.312500
000000000" ...
##  $ Y_COORD_CD            : chr  "234270" "189064.671875000000000" "257125" "173349.796875000
000000" ...
##  $ Latitude              : num  40.8 40.7 40.9 40.6 40.8 ...
##  $ Longitude             : num  -73.9 -73.9 -73.9 -74 -73.9 ...
##  $ Lon_Lat               : chr  "POINT (-73.92019278899994 40.80967347200004)" "POINT (-73.9
4291302299996 40.685609672000055)" "POINT (-73.868233 40.872349)" "POINT (-73.99691224999998 40.
642489932000046)" ...
```

We can see that we have a combination of numerical and categorical variables. Moreover, the variable “STATISTICAL_MURDER_FLAG” will serve as our target. That is, this is the column we will attempt to predict based on other columns.

We see that some entries are null. So let’s remove those Now, because we think that OCCUR_TIME is important for our model (and we will validate this after we have trained our initial model) we will extract the hour from it.

```
nypd_shooting$datetime_str <- paste(nypd_shooting$OCCUR_DATE, nypd_shooting$OCCUR_TIME)
nypd_shooting$datetime <- mdy_hms(nypd_shooting$datetime_str)
nypd_shooting$HOUR <- hour(nypd_shooting$datetime)
```

Now, notice that the column LOC_OF_OCCUR_DESC has a null value ““. So we will remove it.

```
nypd_shooting <- nypd_shooting %>% filter(LOC_OF_OCCUR_DESC != "")
```

Similarly, we will do the same for the columns LOC_CLASSFCTN_DESC, LOCATION_DESC, PERP_AGE_GROUP, PERP_SEX, and PERP_RAC where the value “(null)” will be deleted

```
nypd_shooting <- nypd_shooting %>% filter(LOC_CLASSFCTN_DESC != "(null)")
nypd_shooting <- nypd_shooting %>% filter(LOCATION_DESC != "(null)")
nypd_shooting <- nypd_shooting %>% filter(PERP_AGE_GROUP != "(null)")
nypd_shooting <- nypd_shooting %>% filter(PERP_SEX != "(null)")
nypd_shooting <- nypd_shooting %>% filter(PERP_RACE != "(null)")
```

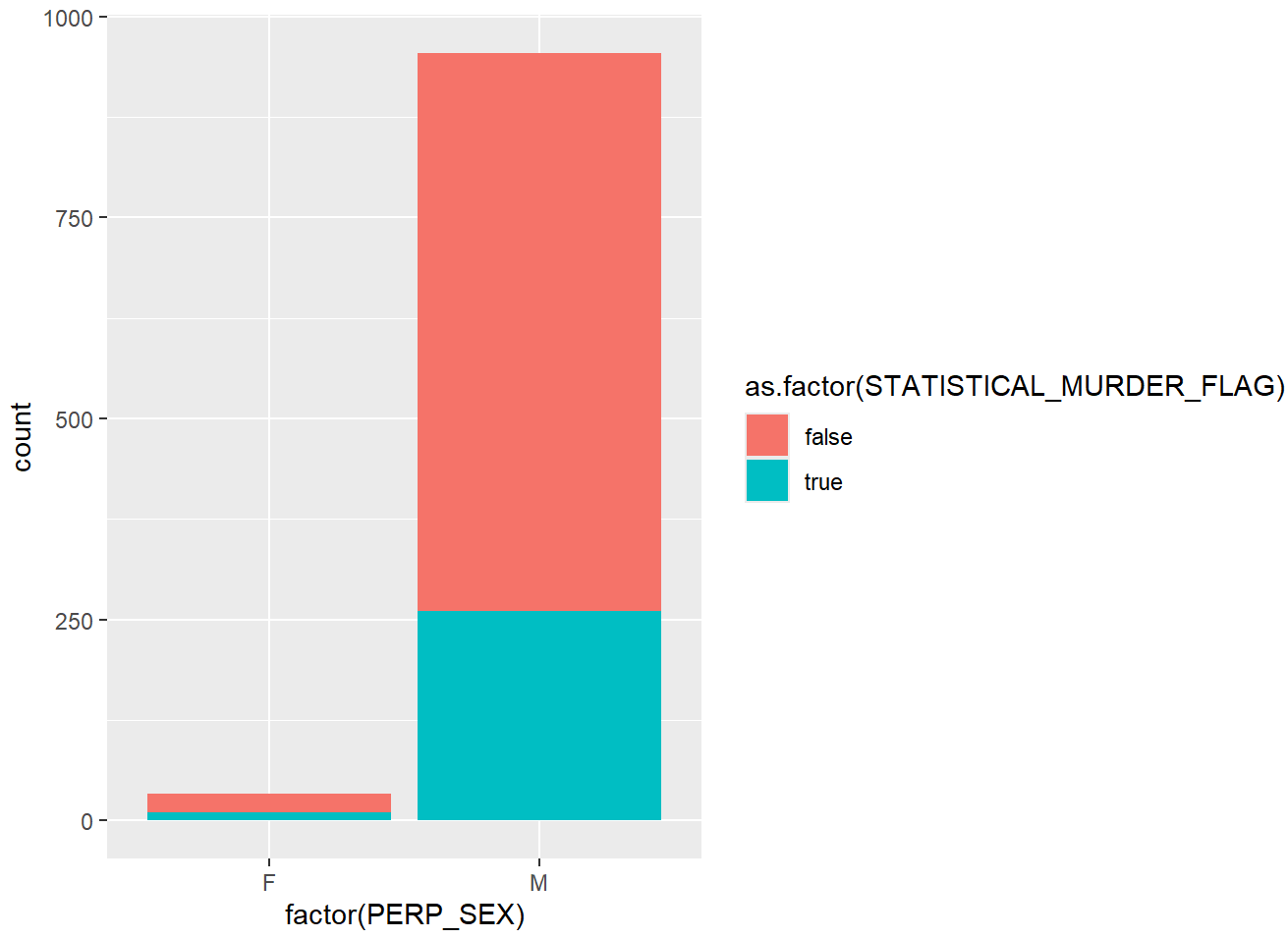
Now, with that out of the way, let's only keep the variables that make sense for our modelling while removing those that are redundant. For instance, "Lon_Lat" will be removed since it has the same information as "Latitude" and "Longitude".

```
nypd_shooting<-nypd_shooting[, c("HOUR", "BORO", "LOC_OF_OCCUR_DESC", "LOC_CLASSFCTN_DESC", "LOC
ATION_DESC","STATISTICAL_MURDER_FLAG","PERP_AGE_GROUP", "PERP_SEX", "PERP_RACE", "VIC_AGE_GROU
P", "VIC_SEX", "VIC_RACE", "Latitude", "Longitude")]
```

Visualizations

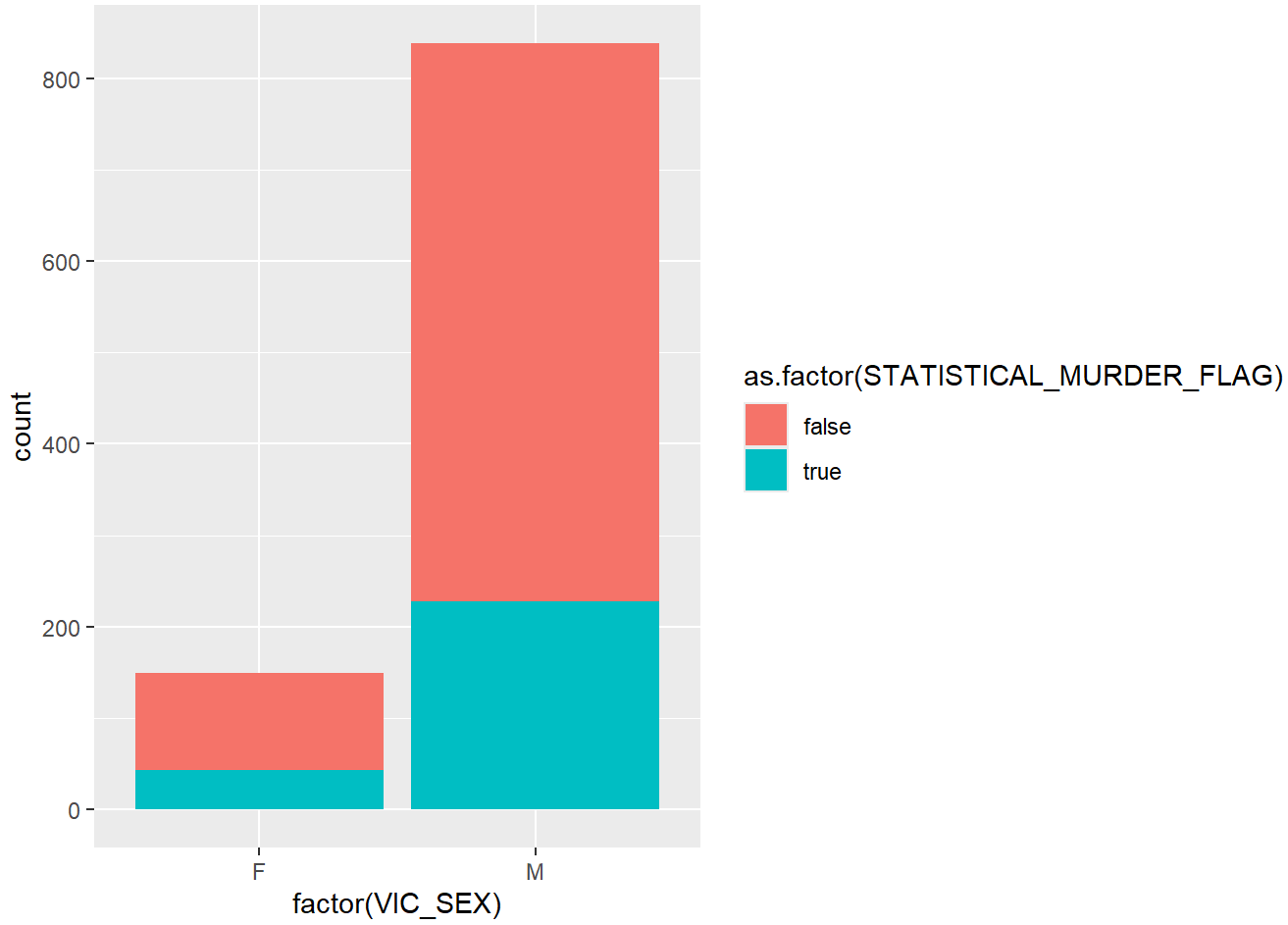
The first visualization we are going to conduct is to see the difference between the number of male perpetrator vs the number of female ones

```
ggplot(nypd_shooting,  aes(x = factor(PERP_SEX), fill = as.factor(STATISTICAL_MURDER_FLAG)))+geo
m_bar(position = "stack")
```



Now, we are going to plot a visualization to find the distribution in gender among the victims

```
ggplot(nypd_shooting,  aes(x = factor(VIC_SEX), fill = as.factor(STATISTICAL_MURDER_FLAG)))+geom
_bar(position = "stack")
```



From the visuals, we can see that the majority of perpetrators are male and the majority of victims are also male.

Modelling

Now that we have done some EDA and visualizations, let's get started with the modelling. First, we will convert our target variable from true and false to 1 and 0

```
nypd_shooting$STATISTICAL_MURDER_FLAG = recode(nypd_shooting$STATISTICAL_MURDER_FLAG, "true"=1,
"false"=0)
```

Now we can build our first model

```
model = glm(STATISTICAL_MURDER_FLAG~.,family=binomial, data = nypd_shooting)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

Let's produce a summary of the model to see which features are statistically significant by looking at the p-values. We will set alpha to be 0.05. That is, only the features with a p-value lower than 0.05 will be significant for us.

```
summary(model)
```

```
##
## Call:
## glm(formula = STATISTICAL_MURDER_FLAG ~ ., family = binomial,
##     data = nypd_shooting)
##
## Coefficients:
##                                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)                          1.481e+02  3.962e+03   0.037  0.97018
## HOUR                                 5.367e-03  1.134e-02   0.473  0.63608
## BOROBROOKLYN                         3.687e-01  5.063e-01   0.728  0.46644
## BOROMANHATTAN                        2.751e-01  3.100e-01   0.887  0.37492
## BOROQUEENS                          -3.329e-01  4.939e-01  -0.674  0.50022
## BOROSTATEN ISLAND                    9.839e-01  9.170e-01   1.073  0.28331
## LOC_OF_OCCUR_DESCOUTSIDE            -6.166e-01  2.187e-01  -2.820  0.00481
## LOC_CLASSFCTN_DESCDWELLING           1.258e+00  4.170e-01   3.017  0.00256
## LOC_CLASSFCTN_DESCHOUSING            1.285e+00  1.250e+00   1.028  0.30393
## LOC_CLASSFCTN_DESCOTHER             -1.360e+01  1.318e+03  -0.010  0.99177
## LOC_CLASSFCTN_DESCPARKING LOT       -5.183e-01  1.306e+00  -0.397  0.69154
## LOC_CLASSFCTN_DESCSTREET             8.280e-01  3.396e-01   2.438  0.01475
## LOCATION_DESCBEAUTY/NAIL SALON       1.046e+00  7.979e-01   1.311  0.18969
## LOCATION_DESCCANDY STORE            -1.660e+01  2.224e+03  -0.007  0.99405
## LOCATION_DESCCHAIN STORE            -1.621e+01  2.797e+03  -0.006  0.99538
## LOCATION_DESCCOMMERCIAL BLDG         3.847e-01  6.525e-01   0.590  0.55545
## LOCATION_DESCDRUG STORE              1.827e+01  2.282e+03   0.008  0.99361
## LOCATION_DESCFACTORY/WAREHOUSE       2.032e+01  2.797e+03   0.007  0.99420
## LOCATION_DESCFAST FOOD               5.793e-01  6.417e-01   0.903  0.36670
## LOCATION_DESCGAS STATION            -1.574e+00  1.132e+00  -1.391  0.16434
## LOCATION_DESCGROCERY/BODEGA          2.916e-01  4.386e-01   0.665  0.50617
## LOCATION_DESCGYM/FITNESS FACILITY   -2.644e+00  4.170e+03  -0.001  0.99949
## LOCATION_DESCHOSPITAL               -1.553e+01  1.391e+03  -0.011  0.99110
## LOCATION_DESCHOTEL/MOTEL            -1.687e+01  1.616e+03  -0.010  0.99167
## LOCATION_DESCJEWELRY STORE          -1.673e+01  3.956e+03  -0.004  0.99663
## LOCATION_DESCLIQUOR STORE            1.836e+00  9.940e-01   1.847  0.06475
## LOCATION_DESCMULTI DWELL - APT BUILD -6.448e-01  4.902e-01  -1.315  0.18837
## LOCATION_DESCMULTI DWELL - PUBLIC HOUS -1.238e+00  1.278e+00  -0.969  0.33277
## LOCATION_DESCPVT HOUSE              -4.423e-01  5.316e-01  -0.832  0.40542
## LOCATION_DESCRESTAURANT/DINER        3.001e-01  6.877e-01   0.436  0.66254
## LOCATION_DESCSHOE STORE             -1.689e+01  3.956e+03  -0.004  0.99659
## LOCATION_DESCSMALL MERCHANT          8.284e-01  7.726e-01   1.072  0.28365
## LOCATION_DESCSOCIAL CLUB/POLICY LOCATI -1.630e+01  2.794e+03  -0.006  0.99535
## LOCATION_DESCSTORE UNCLASSIFIED      1.165e+00  1.516e+00   0.769  0.44216
## LOCATION_DESCSUPERMARKET            -1.660e+01  2.791e+03  -0.006  0.99526
## LOCATION_DESCTELECOMM. STORE         2.227e+00  9.796e-01   2.274  0.02298
## LOCATION_DESCVIDEO STORE             1.904e+01  1.607e+03   0.012  0.99055
## PERP_AGE_GROUP18-24                 -2.763e-01  3.063e-01  -0.902  0.36703
## PERP_AGE_GROUP25-44                 -2.285e-01  3.047e-01  -0.750  0.45323
## PERP_AGE_GROUP45-64                  3.824e-01  3.708e-01   1.031  0.30244
## PERP_AGE_GROUP65+                    3.453e-02  1.341e+00   0.026  0.97947
## PERP_SEXM                            1.467e-01  4.221e-01   0.348  0.72817
## PERP_RACEBLACK                       9.614e-01  7.153e-01   1.344  0.17893
## PERP_RACEBLACK HISPANIC              1.126e+00  7.585e-01   1.484  0.13770
## PERP_RACEWHITE                       7.710e-01  1.039e+00   0.742  0.45811
```

```
## PERP_RACEWHITE HISPANIC                1.219e+00  7.341e-01   1.661  0.09679
## VIC_AGE_GROUP18-24                    -2.850e-02  3.340e-01  -0.085  0.93202
## VIC_AGE_GROUP25-44                     3.162e-01  3.163e-01   1.000  0.31748
## VIC_AGE_GROUP45-64                     5.119e-01  3.651e-01   1.402  0.16087
## VIC_AGE_GROUP65+                       5.307e-01  6.512e-01   0.815  0.41506
## VIC_AGE_GROUPUNKNOWN                  -1.668e+01  2.773e+03  -0.006  0.99520
## VIC_SEXM                              -4.999e-02  2.227e-01  -0.224  0.82240
## VIC_RACEASIAN / PACIFIC ISLANDER       1.802e+01  3.956e+03   0.005  0.99637
## VIC_RACEBLACK                          1.720e+01  3.956e+03   0.004  0.99653
## VIC_RACEBLACK HISPANIC                 1.682e+01  3.956e+03   0.004  0.99661
## VIC_RACEWHITE                          1.736e+01  3.956e+03   0.004  0.99650
## VIC_RACEWHITE HISPANIC                 1.710e+01  3.956e+03   0.004  0.99655
## Latitude                               2.613e+00  2.500e+00   1.045  0.29597
## Longitude                              3.711e+00  2.585e+00   1.436  0.15114
##
## (Intercept)
## HOUR
## BOROBROOKLYN
## BOROMANHATTAN
## BOROQUEENS
## BOROSTATEN ISLAND
## LOC_OF_OCCUR_DESCOUTSIDE              **
## LOC_CLASSFCTN_DESCDWELLING            **
## LOC_CLASSFCTN_DESCHOUSING
## LOC_CLASSFCTN_DESCOTHER
## LOC_CLASSFCTN_DESCPARKING LOT
## LOC_CLASSFCTN_DESCSTREET              *
## LOCATION_DESCBEAUTY/NAIL SALON
## LOCATION_DESCCANDY STORE
## LOCATION_DESCCHAIN STORE
## LOCATION_DESCCOMMERCIAL BLDG
## LOCATION_DESCDRUG STORE
## LOCATION_DESCFACTORY/WAREHOUSE
## LOCATION_DESCFAST FOOD
## LOCATION_DESCGAS STATION
## LOCATION_DESCGROCERY/BODEGA
## LOCATION_DESCGYM/FITNESS FACILITY
## LOCATION_DESCHOSPITAL
## LOCATION_DESCHOTEL/MOTEL
## LOCATION_DESCJEWELRY STORE
## LOCATION_DESCLIQUOR STORE             .
## LOCATION_DESCMULTI DWELL - APT BUILD
## LOCATION_DESCMULTI DWELL - PUBLIC HOUS
## LOCATION_DESCPVT HOUSE
## LOCATION_DESCRESTAURANT/DINER
## LOCATION_DESCSHOE STORE
## LOCATION_DESCSMALL MERCHANT
## LOCATION_DESCSOCIAL CLUB/POLICY LOCATI
## LOCATION_DESCSTORE UNCLASSIFIED
## LOCATION_DESCSUPERMARKET
## LOCATION_DESCTELECOMM. STORE          *
## LOCATION_DESCVIDEO STORE
```

```
## PERP_AGE_GROUP18-24
## PERP_AGE_GROUP25-44
## PERP_AGE_GROUP45-64
## PERP_AGE_GROUP65+
## PERP_SEXM
## PERP_RACEBLACK
## PERP_RACEBLACK HISPANIC
## PERP_RACEWHITE
## PERP_RACEWHITE HISPANIC                .
## VIC_AGE_GROUP18-24
## VIC_AGE_GROUP25-44
## VIC_AGE_GROUP45-64
## VIC_AGE_GROUP65+
## VIC_AGE_GROUPUNKNOWN
## VIC_SEXM
## VIC_RACEASIAN / PACIFIC ISLANDER
## VIC_RACEBLACK
## VIC_RACEBLACK HISPANIC
## VIC_RACEWHITE
## VIC_RACEWHITE HISPANIC
## Latitude
## Longitude
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1144.11  on 970  degrees of freedom
## Residual deviance:  997.14  on 912  degrees of freedom
##   (17 observations deleted due to missingness)
## AIC: 1115.1
##
## Number of Fisher Scoring iterations: 16
```

We can see that only LOC_OF_OCCUR_DESC, LOC_CLASSFCTN_DESC, and LOCATION_DESC are statistically significant. So let's build a new model with these variables

```
model_red = glm(STATISTICAL_MURDER_FLAG~LOC_OF_OCCUR_DESC+LOC_CLASSFCTN_DESC+LOCATION_DESC,famil
y=binomial, data = nypd_shooting)
```

```
summary(model_red)
```

```
##
## Call:
## glm(formula = STATISTICAL_MURDER_FLAG ~ LOC_OF_OCCUR_DESC + LOC_CLASSFCTN_DESC +
##     LOCATION_DESC, family = binomial, data = nypd_shooting)
##
## Coefficients:
##                                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)                         -1.27341    0.37137  -3.429 0.000606
## LOC_OF_OCCUR_DESCOUTSIDE            -0.71739    0.20486  -3.502 0.000462
## LOC_CLASSFCTN_DESCDWELLING           1.08264    0.39155   2.765 0.005692
## LOC_CLASSFCTN_DESCHOUSING            1.11239    1.22472   0.908 0.363733
## LOC_CLASSFCTN_DESCOTHER            -12.74639  784.88814  -0.016 0.987043
## LOC_CLASSFCTN_DESCPARKING LOT       -0.06062    1.23753  -0.049 0.960930
## LOC_CLASSFCTN_DESCSTREET             0.83881    0.32187   2.606 0.009159
## LOCATION_DESCBEAUTY/NAIL SALON       1.31511    0.73620   1.786 0.074041
## LOCATION_DESCCANDY STORE           -14.84484 1364.54014  -0.011 0.991320
## LOCATION_DESCCHAIN STORE           -14.57527 1696.73439  -0.009 0.993146
## LOCATION_DESCCOMMERCIAL BLDG         0.44198    0.61623   0.717 0.473229
## LOCATION_DESCDRUG STORE             17.71805 1385.37784   0.013 0.989796
## LOCATION_DESCFACTORY/WAREHOUSE      18.55686 1696.73439   0.011 0.991274
## LOCATION_DESCFAST FOOD               0.66363    0.53604   1.238 0.215705
## LOCATION_DESCGAS STATION            -1.19508    1.11077  -1.076 0.281973
## LOCATION_DESCGROCERY/BODEGA          0.61106    0.39672   1.540 0.123494
## LOCATION_DESCGYM/FITNESS FACILITY   -2.54627 2524.65133  -0.001 0.999195
## LOCATION_DESCHOSPITAL              -13.81355  784.95594  -0.018 0.985960
## LOCATION_DESCHOTEL/MOTEL            -0.35867    1.20879  -0.297 0.766685
## LOCATION_DESCJEWELRY STORE         -15.29266 2399.54475  -0.006 0.994915
## LOCATION_DESCLIQUOR STORE            1.99765    0.94842   2.106 0.035179
## LOCATION_DESCMULTI DWELL - APT BUILD -0.22020    0.44843  -0.491 0.623386
## LOCATION_DESCMULTI DWELL - PUBLIC HOUS -0.83148  1.24919  -0.666 0.505658
## LOCATION_DESCPVT HOUSE              -0.10223    0.48382  -0.211 0.832656
## LOCATION_DESCRESTAURANT/DINER        0.66236    0.64906   1.020 0.307502
## LOCATION_DESCSHOE STORE            -14.57527 2399.54475  -0.006 0.995154
## LOCATION_DESCSMALL MERCHANT          0.83477    0.71484   1.168 0.242897
## LOCATION_DESCSOCIAL CLUB/POLICY LOCATI -15.29266 1696.73438 -0.009 0.992809
## LOCATION_DESCSTORE UNCLASSIFIED      1.21269    1.45613   0.833 0.404945
## LOCATION_DESCSUPERMARKET           -15.41409 1696.73439  -0.009 0.992752
## LOCATION_DESCTELECOMM. STORE         2.56278    0.95328   2.688 0.007180
## LOCATION_DESCVIDEO STORE            17.83947  979.61010   0.018 0.985471
##
## (Intercept)                        ***
## LOC_OF_OCCUR_DESCOUTSIDE           ***
## LOC_CLASSFCTN_DESCDWELLING         **
## LOC_CLASSFCTN_DESCHOUSING
## LOC_CLASSFCTN_DESCOTHER
## LOC_CLASSFCTN_DESCPARKING LOT
## LOC_CLASSFCTN_DESCSTREET           **
## LOCATION_DESCBEAUTY/NAIL SALON     .
## LOCATION_DESCCANDY STORE
## LOCATION_DESCCHAIN STORE
## LOCATION_DESCCOMMERCIAL BLDG
## LOCATION_DESCDRUG STORE
```

```
## LOCATION_DESCFACTORY/WAREHOUSE
## LOCATION_DESCFAST FOOD
## LOCATION_DESCGAS STATION
## LOCATION_DESCGROCERY/BODEGA
## LOCATION_DESCGYM/FITNESS FACILITY
## LOCATION_DESCHOSPITAL
## LOCATION_DESCHOTEL/MOTEL
## LOCATION_DESCJEWELRY STORE
## LOCATION_DESCLIQUOR STORE              *
## LOCATION_DESCMULTI DWELL - APT BUILD
## LOCATION_DESCMULTI DWELL - PUBLIC HOUS
## LOCATION_DESCPVT HOUSE
## LOCATION_DESCRESTAURANT/DINER
## LOCATION_DESCSHOE STORE
## LOCATION_DESCSMALL MERCHANT
## LOCATION_DESCSOCIAL CLUB/POLICY LOCATI
## LOCATION_DESCSTORE UNCLASSIFIED
## LOCATION_DESCSUPERMARKET
## LOCATION_DESCTELECOMM. STORE           **
## LOCATION_DESCVIDEO STORE
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1160.9  on 987  degrees of freedom
## Residual deviance: 1056.3  on 956  degrees of freedom
## AIC: 1120.3
##
## Number of Fisher Scoring iterations: 15
```

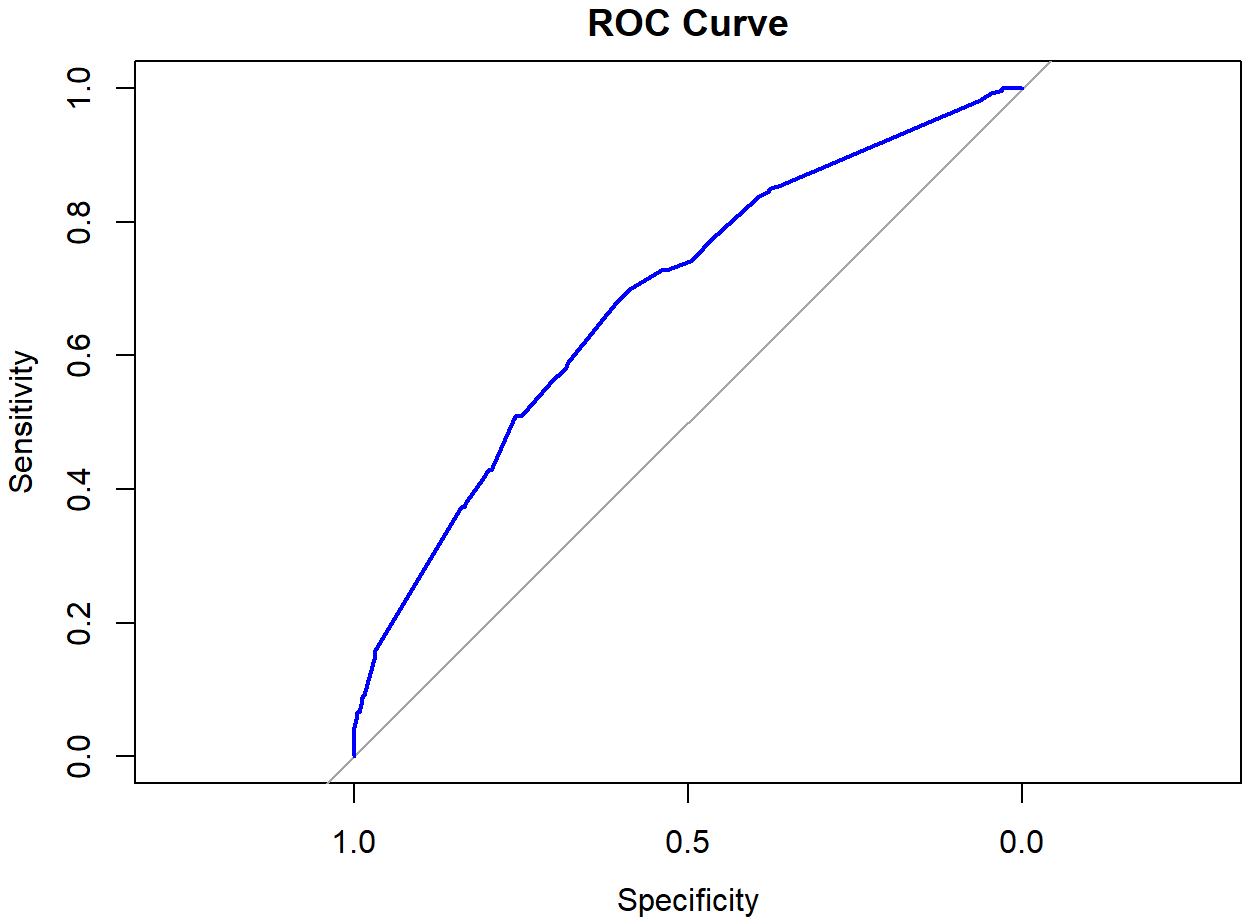
Finally, let's plot a ROC curve of our model to see how much better does it perform compared to random guess.

```
pred_probs <- predict(model_red, type="response")
roc_obj <- roc(nypd_shooting$STATISTICAL_MURDER_FLAG, pred_probs)
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

```
plot(roc_obj, col = "blue", lwd = 2, main = "ROC Curve")
```



Conclusion

We see that our model predicts better than a random guess. Thus, we have trained a sound model to predict whether a shooting will lead to a murder.