NYPD Shooting Data

The question of Interests

Can we predict the likelihood of a statistical murder flag being true or false based on the borough and location of occurrence in the NYPD Shooting Incident dataset?

Import Libraries:

In [5]:
```
library(tidyverse)
```

Load the Dataset:

In [6]:
```
url <- "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"
NYPD = read.csv(url)
```

In [8]:
```
head(NYPD)
```

	INCIDENT_KEY	OCCUR_DATE	OCCUR_TIME	BORO	LOC_OF_OCCUR_DESC	PRECINCT	JURISDICTION_CODE	LOC_CLASSFCTN_DESC	LOC/
	<int>	<chr>	<chr>	<chr>	<chr>	<int>	<int>	<chr>	
1	228798151	05/27/2021	21:30:00	QUEENS		105	0		
2	137471050	06/27/2014	17:40:00	BRONX		40	0		
3	147998800	11/21/2015	03:56:00	QUEENS		108	0		
4	146837977	10/09/2015	18:30:00	BRONX		44	0		
5	58921844	02/19/2009	22:58:00	BRONX		47	0		
6	219559682	10/21/2020	21:36:00	BROOKLYN		81	0		

Data Tidying:

In [9]:
```
NYPD <- na.omit(NYPD)
NYPD <- unique(NYPD)
```

Data Exploration:

In [12]:
```
summary(NYPD)
```

```
 INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME            BORO
Min.   :  9953245   Length:27300       Length:27300       Length:27300
1st Qu.: 63859933   Class :character   Class :character   Class :character
Median : 90340495   Mode  :character   Mode  :character   Mode  :character
Mean   :120812778
3rd Qu.:188587325
Max.   :261190187
LOC_OF_OCCUR_DESC     PRECINCT      JURISDICTION_CODE LOC_CLASSFCTN_DESC
Length:27300       Min.   :  1.00   Min.   :0.000     Length:27300
Class :character   1st Qu.: 44.00   1st Qu.:0.000     Class :character
Mode  :character   Median : 68.00   Median :0.000     Mode  :character
                   Mean   : 65.64   Mean   :0.327
                   3rd Qu.: 81.00   3rd Qu.:0.000
                   Max.   :123.00   Max.   :2.000
LOCATION_DESC      STATISTICAL_MURDER_FLAG PERP_AGE_GROUP
Length:27300       Length:27300            Length:27300
Class :character   Class :character        Class :character
Mode  :character   Mode  :character        Mode  :character



  PERP_SEX          PERP_RACE         VIC_AGE_GROUP        VIC_SEX
Length:27300       Length:27300       Length:27300       Length:27300
Class :character   Class :character   Class :character   Class :character
Mode  :character   Mode  :character   Mode  :character   Mode  :character

  VIC_RACE           X_COORD_CD       Y_COORD_CD        Latitude
Length:27300       Min.   : 914928   Min.   :125757   Min.   :40.51
Class :character   1st Qu.:1000033   1st Qu.:182832   1st Qu.:40.67
Mode  :character   Median :1007742   Median :194478   Median :40.70
                   Mean   :1009451   Mean   :208128   Mean   :40.74
                   3rd Qu.:1016838   3rd Qu.:239518   3rd Qu.:40.82
                   Max.   :1066815   Max.   :271128   Max.   :40.91
  Longitude        Lon_Lat
Min.   :-74.25   Length:27300
1st Qu.:-73.94   Class :character
Median :-73.92   Mode  :character
Mean   :-73.91
3rd Qu.:-73.88
Max.   :-73.70
```

In [13]:
```
table(NYPD$BORO)
```

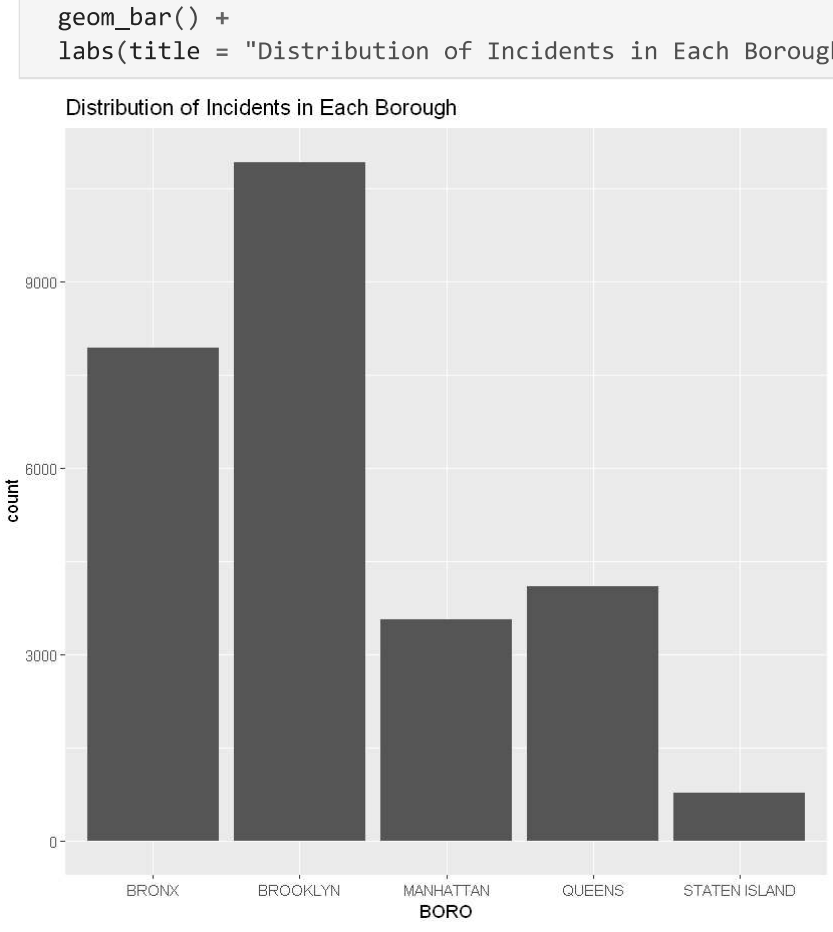
```
       BRONX      BROOKLYN     MANHATTAN        QUEENS STATEN ISLAND
        7937         10929          3567          4091           776
```

In [14]:
```
table(NYPD$LOC_OF_OCCUR_DESC)
```

```
       INSIDE OUTSIDE
 25594     242    1464
```

Visualization:

In [11]:
```
# Plotting the distribution of incidents in each borough
ggplot(NYPD, aes(x = BORO)) +
  geom_bar() +
  labs(title = "Distribution of Incidents in Each Borough")
```



Analysis:

In [15]:
```
# Analyzing incidents by borough and location description
analysis_result <- NYPD %>%
  group_by(BORO, LOC_OF_OCCUR_DESC) %>%
  summarise(incident_count = n()) %>%
  arrange(desc(incident_count))

head(analysis_result)
```

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

A grouped_df: 6 × 3

BORO	LOC_OF_OCCUR_DESC	incident_count
<chr>	<chr>	<int>
BROOKLYN		10365
BRONX		7402
QUEENS		3827
MANHATTAN		3264
STATEN ISLAND		736
BRONX	OUTSIDE	471

Modeling:

In [18]:
```
table(NYPD$STATISTICAL_MURDER_FLAG)
```

```
false  true
22034  5266
```

In [20]:
```
# Assuming 'TRUE' indicates a positive outcome (1), and 'FALSE' indicates a negative outcome (0)
NYPD$STATISTICAL_MURDER_FLAG <- as.factor(NYPD$STATISTICAL_MURDER_FLAG)
NYPD$STATISTICAL_MURDER_FLAG <- as.numeric(NYPD$STATISTICAL_MURDER_FLAG) - 1
table(NYPD$STATISTICAL_MURDER_FLAG)
```

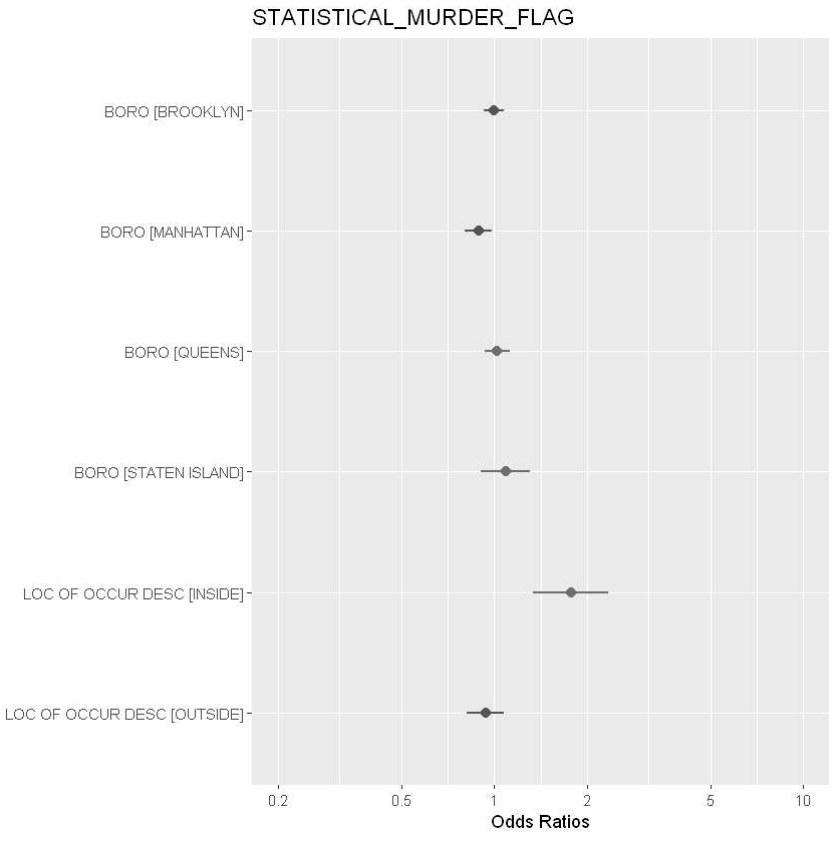
```
   0    1
22034  5266
```

In [21]:
```
# Logistic Regression predicting a binary outcome
model <- glm(STATISTICAL_MURDER_FLAG ~ BORO + LOC_OF_OCCUR_DESC, data = NYPD, family = "binomial")
summary(model)
```

```
Call:
glm(formula = STATISTICAL_MURDER_FLAG ~ BORO + LOC_OF_OCCUR_DESC,
    family = "binomial", data = NYPD)

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)              -1.424118   0.028702 -49.618  < 2e-16 ***
BOROBROOKLYN             -0.002753   0.037309  -0.074    0.941
BOROMANHATTAN            -0.116417   0.052304  -2.226    0.026 *
BOROQUEENS                0.023038   0.048433   0.476    0.634
BOROSTATEN ISLAND         0.086510   0.092821   0.932    0.351
LOC_OF_OCCUR_DESCINSIDE   0.574111   0.141535   4.056 4.99e-05 ***
LOC_OF_OCCUR_DESCOUTSIDE -0.064517   0.069707  -0.926    0.355
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 26775  on 27299  degrees of freedom
Residual deviance: 26751  on 27293  degrees of freedom
AIC: 26765

Number of Fisher Scoring iterations: 4
```

In [28]:
```
# Visualize the coefficients
plot_model(model, type = "std", ci_method = "wald")
```

```
Profiled confidence intervals may take longer time to compute.
  Use `ci_method="wald"` for faster computation of CIs.
```



Conclusion

- The model suggests that the borough of Manhattan (BOROMANHATTAN) and incidents occurring inside are significant predictors of a shooting incident being flagged as a statistical murder.
- Other boroughs and incidents occurring outside do not show significant associations.