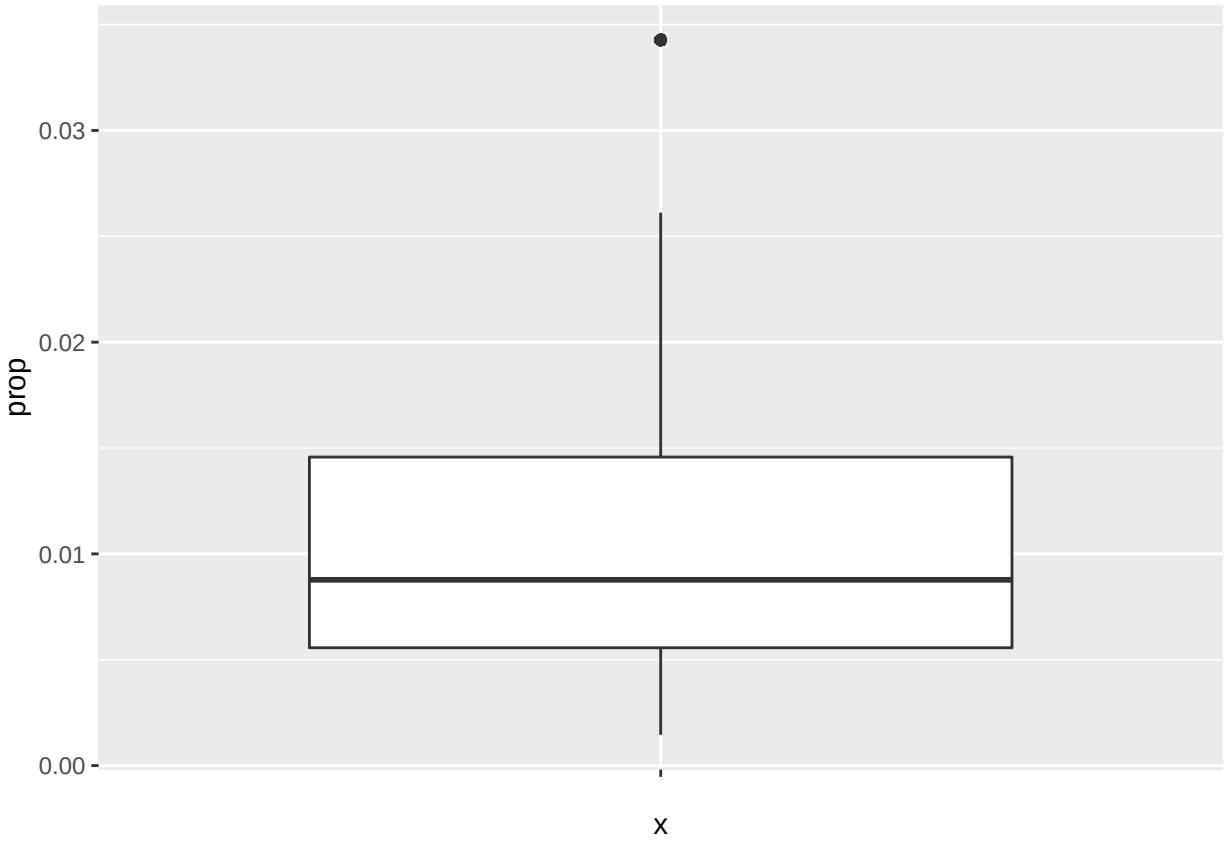
What, if any, is the most common type of Break and Enters (B&E) that occur in high-rate B&E neighbouroods, and does this type occur more often than in low-rate B&E neighbourhoods?

#create a new variable that represents the proportion of all reported B&E's that occurred in this neighbourhood.

```
grouped_hood <- break_and_enters %>%
  filter(!is.na(Neighbourhood)) %>%
  group_by(Neighbourhood) %>%
  mutate(number = n()) %>%
  mutate(prop = number/43302)
break_and_enters_prop <- merge(break_and_enters,grouped_hood,by=c("X1", "Index", "event_unique_id", "oc
```

```
ggplot(break_and_enters_prop,
aes(x = "", y=prop)) +
geom_boxplot()
```



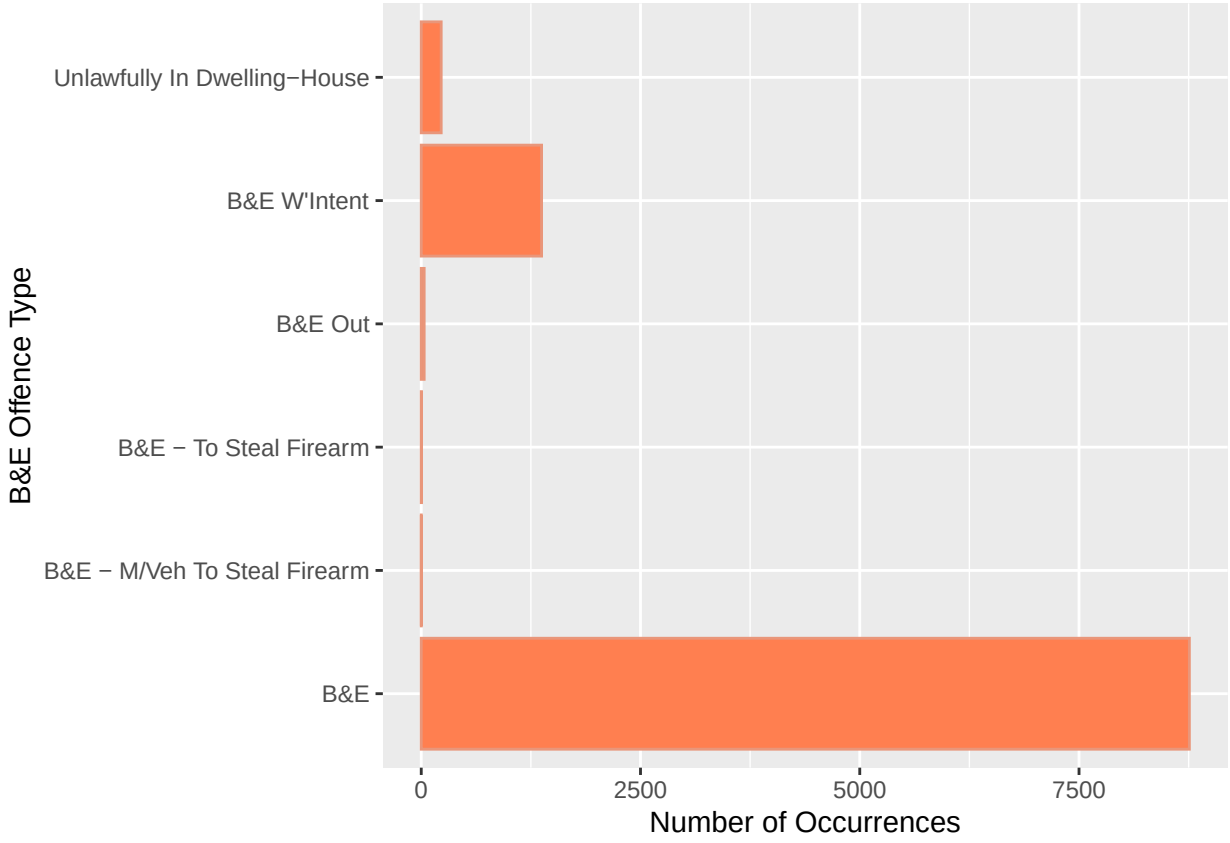

```
break_and_enters_prop %>%
summarise(n=n(),
Q1=quantile(prop, 0.25),
med=median(prop),
mean=mean(prop),
Q3=quantile(prop, 0.75),
max=max(prop))
```

```
##       n          Q1         med       mean         Q3        max
## 1 43302 0.005565563 0.008775576 0.01074941 0.01457208 0.03427093
```

```
b_and_e_top <- break_and_enters_prop %>%
  filter(prop >= 0.01457208)
b_and_e_lower <- break_and_enters_prop %>%
  filter(prop <= 0.005565563)
b_and_e_extremes <- rbind(b_and_e_top, b_and_e_lower)
b_and_e_extremes <- b_and_e_extremes %>% mutate(level = ifelse(prop >= 0.01457208, "high", "low"))
```

First, we try to identify which type of B&E occurs the most often in high-rate neighborhoods. To do this, we use a simple bar graph.

```
ggplot(b_and_e_top, aes(x = offence)) +
  geom_bar(color="darksalmon", fill="coral") +
  labs(x = "B&E Offence Type", y = "Number of Occurrences") +
  coord_flip()
```



As we can see from this graph, B&E is the most common offence type in neighborhoods with high rates.

Next, we want to see if this trend is different from that of the low-rate neighborhoods. This can be done through a hypothesis test. Our hypothesis test will seek to H_0: $P_{\text{diff}} = 0$ H_a: P_{diff} != 0 where H_0 represents the null hypothesis, H_a the alternative hypothesis, and P_{diff} the difference between the proportion of a certain offence type in high B&E rate neighborhoods and that of low B&E rate neighborhoods.

```
prop_bande <- b_and_e_extremes %>%
  group_by(level) %>%
  summarise(n = n(), num_bne = sum(offence == "B&E"), prop_bne = num_bne/n)
test_stat <- prop_bande %>%
  summarise(test_stat = diff(prop_bne))
test_stat
```

```
## # A tibble: 1 x 1
##   test_stat
##       <dbl>
## 1   -0.0122
```

```
set.seed(123)
repetitions <- 1000
simulated_differences <- rep(NA, times=repetitions)
for(i in 1:repetitions){
  simdata <- b_and_e_extremes %>%
    mutate(level=sample(level)) %>%
    group_by(level) %>%
    summarise(n = n(), num_bne = sum(offence == "B&E"), prop_bne = num_bne/n)
  sim_prop_diff <- simdata %>% summarise(value = diff(prop_bne))
  simulated_differences[i] <- as.numeric(sim_prop_diff)
}
sim <- tibble(prop_diff = simulated_differences)
```

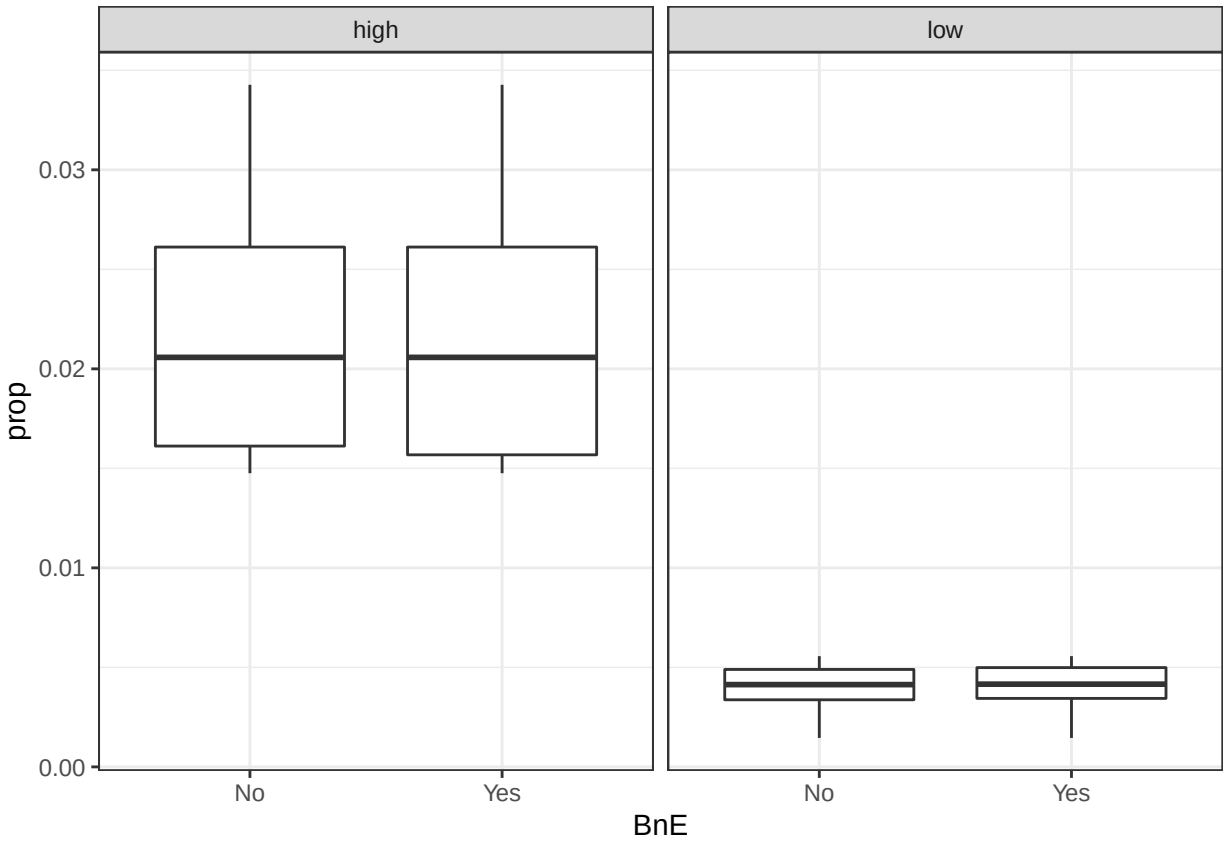
```
#ggplot(sim, aes(x=prop_diff)) +
#geom_histogram(binwidth = 0.005, fill="gray", color="black") +
#labs(x = "Difference in proportion of B&E (offence) for neighborhoods
   #  with high B&E rates versus those with low B&E rates") +
 # geom_vline(xintercept=c(test_stat, -test_stat), color="red")
sim %>%
  filter(prop_diff >= abs(test_stat) | prop_diff <= -abs(test_stat))%>%
  summarise(pvalue = n() / repetitions)
```

```
## # A tibble: 1 x 1
##   pvalue
##    <dbl>
## 1      0
```

Since the p-value is less than 0.0001, we have very strong evidence against the null hypothesis that there is no difference in proportion of B&E offences between neighborhoods with high B&E rates and those with low B&E rates. Thus we can conclude that there is a difference in proportion of the B&E type offences between high and low rate B&E rate neighbourhoods.

Next we will show how the proportion of Break and Enters differs between high and low rate neighbourhoods on a plot:

```
b_and_e_extremes <- b_and_e_extremes %>%
  mutate(BnE = ifelse(offence == "B&E", "Yes", "No"))
ggplot(b_and_e_extremes, aes(x=BnE, y=prop)) +
  geom_boxplot() + theme_bw() + facet_wrap(~level)
```



From the high rate neighbourhoods it is evident ...

Further, we will use another method - linear regresssion model to predict the proportion of BnE using the BnE as a predictor. The regression equation will be: $\text{BnEprop}_i = \beta B_0 + \beta B_1 \text{I}$ Where our dependent variable is BnEprop for the ith observation, I is the independent variable, βB_0 is the intercept parameter and βB_1 is the slope parameter. We are interested to test the Null hypothesis: $\beta B_1 = 0$ Alternate Hypothesis: βB_1 != 0. We will assess whether there is a linear association between the two variables.

```
summary(lm(prop ~ BnE, data=b_and_e_extremes))$coefficients
```

```
##                Estimate   Std. Error   t value   Pr(>|t|)
## (Intercept) 0.0120897403 0.0001625682 74.367193 0.00000000
## BnEYes      0.0003489016 0.0001777800  1.962547 0.04971151
```

From the model, we can see that the p-value is

<<<<<<< HEAD

======= >>>>>>> 09d98f4bdb7dcd8b96e2980566b3dc32811776bb