Homework #4 Q3

Nan You

Question 3

Begin the analysis of one variable in the dataset you are using the final project. As this is an individual homework assignment, each group member should choose a different variable. Choose three visualizations as appropriate to show the distribution of the variable, conditioned on another variable if desired (for example, the distribution of income by region). Write a few sentences describing what you found and what new questions your visualizations have generated. (Faceted graphs count as one graph; graphs put together with grid.arrange() or similar count as multiple graphs.)

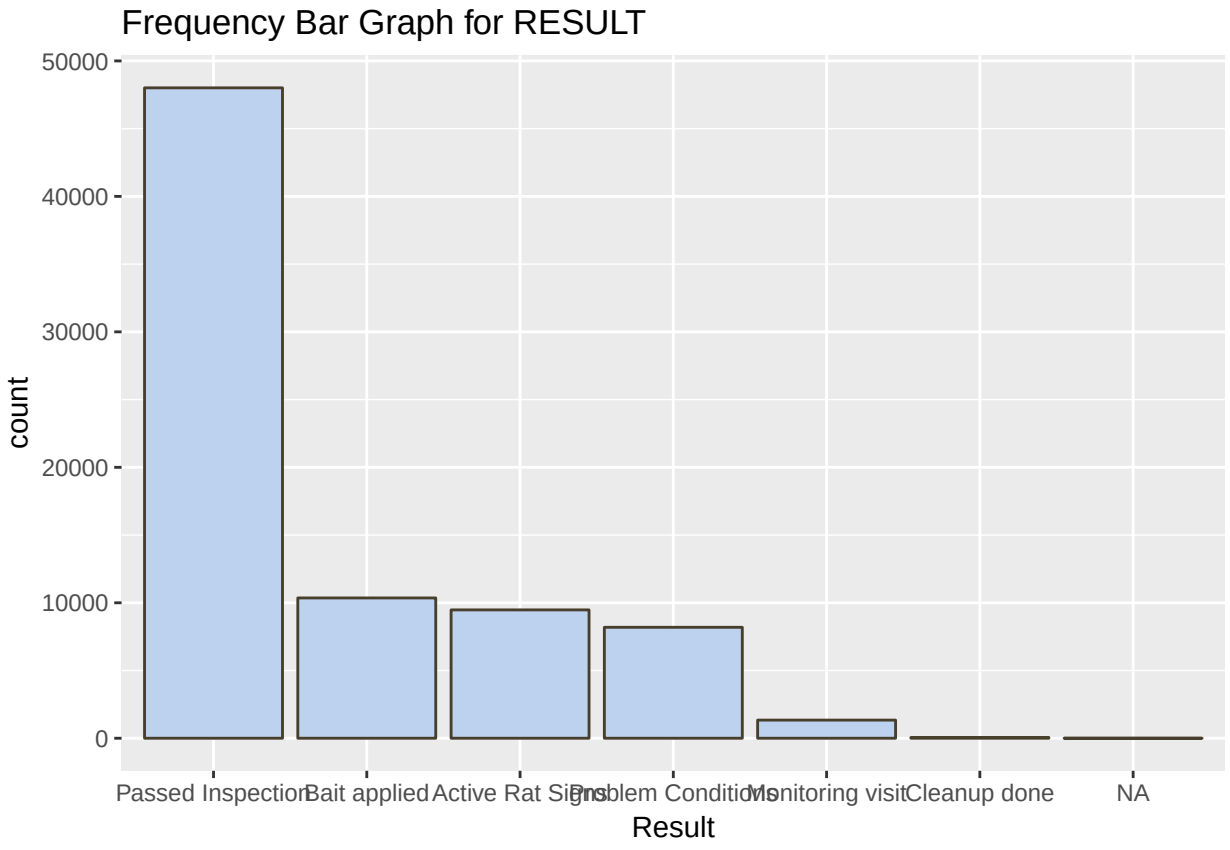
```
library(ggplot2)
library(forcats)

set.seed(1)
data <- read.csv("Rodent_Inspection_77K_Sample_Clean.csv")
summary(data)
```

```
##        X1            INSPECTION_TYPE  JOB_TICKET_OR_WORK_ORDER_ID
##  Min.   :      6   BAIT      :11694   Min.   :     39
##  1st Qu.: 386658   CLEAN_UPS :   60   1st Qu.: 309780
##  Median : 773912   COMPLIANCE:11154   Median : 717475
##  Mean   : 773730   INITIAL   :54523   Mean   : 803479
##  3rd Qu.:1160606                      3rd Qu.:1298474
##  Max.   :1548621                      Max.   :1855306
##
##        JOB_ID       JOB_PROGRESS          BBL              BORO_CODE
##  PO983685 :   11   Min.   :  1.000   Min.   :1.000e+09   Min.   :1.000
##  PO1216458:   10   1st Qu.:  1.000   1st Qu.:1.019e+09   1st Qu.:1.000
##  PO1082226:    9   Median :  1.000   Median :2.031e+09   Median :2.000
##  PO1053654:    7   Mean   :  1.954   Mean   :2.228e+09   Mean   :2.202
##  PO1081990:    7   3rd Qu.:  2.000   3rd Qu.:3.021e+09   3rd Qu.:3.000
##  PO1053651:    6   Max.   :170.000   Max.   :5.080e+09   Max.   :5.000
##  (Other)  :77381
##      BLOCK            LOT          HOUSE_NUMBER             STREET_NAME
##  Min.   :    0   Min.   :   0.0   null   :  483   BROADWAY        :  993
##  1st Qu.: 1386   1st Qu.:  17.0   0      :  214   3 AVENUE        :  616
##  Median : 2243   Median :  36.0   15     :  186   GRAND CONCOURSE :  532
##  Mean   : 2592   Mean   : 193.9   25     :  176   AMSTERDAM AVENUE:  481
##  3rd Qu.: 3221   3rd Qu.:  60.0   55     :  160   2 AVENUE        :  467
##  Max.   :16296   Max.   :8901.0   (Other):74844   (Other)         :74307
##                                   NA's   : 1368   NA's            :   35
##     ZIP_CODE        X_COORD             Y_COORD             LATITUDE
##  Min.   :    0   Min.   :        0   Min.   :        0   Min.   :-85.44
##  1st Qu.:10032   1st Qu.:   994848   1st Qu.:   194278   1st Qu.: 40.70
##  Median :10458   Median :  1003631   Median :   219722   Median : 40.77
##  Mean   :10562   Mean   :  1031461   Mean   :   243369   Mean   : 40.74
##  3rd Qu.:11212   3rd Qu.:  1012652   3rd Qu.:   243098   3rd Qu.: 40.83
##  Max.   :11694   Max.   :246169952   Max.   :246170928   Max.   : 40.91
##  NA's   :3       NA's   :510         NA's   :510         NA's   :209
##    LONGITUDE               BOROUGH        INSPECTION_DATE
```

```
##  Min.   :-155.78   Bronx        :24064   2017/12/1 :  107
##  1st Qu.: -73.96   Brooklyn     :18263   2018/2/9  :   91
##  Median : -73.93   Manhattan    :25043   2017/11/30:   90
##  Mean   : -73.93   Queens       : 7743   2017/11/21:   89
##  3rd Qu.: -73.90   Staten Island: 2318   2018/3/16 :   89
##  Max.   :  56.31                         2018/1/16 :   84
##  NA's   :209                             (Other)   :76881
##                 RESULT        APPROVED_DATE
##  Active Rat Signs  : 9475   2018/3/19:  165
##  Bait applied      :10356   2017/12/6:  147
##  Cleanup done      :   60   2018/2/12:  141
##  Monitoring visit  : 1338   2018/2/21:  137
##  Passed Inspection :48010   2017/4/12:  135
##  Problem Conditions: 8190   2018/6/20:  130
##  NA's              :    2   (Other)  :76576
##                                    LOCATION
##  (40.7180216447923, -73.9934340556085):   17
##  (0.0, 0.0)                           :   13
##  (40.721502812148, -73.9779563028835) :   12
##  (40.7578602598906, -73.8302509522777):   12
##  (40.7138194794758, -74.0055768832927):   11
##  (Other)                              :77153
##  NA's                                 :  213
```

```
ggplot(data,aes(x = fct_infreq(RESULT))) +
  geom_bar(fill = "lightsteelblue2", color = "#473e2c") +
  ggtitle("Frequency Bar Graph for RESULT") + xlab("Result")
```



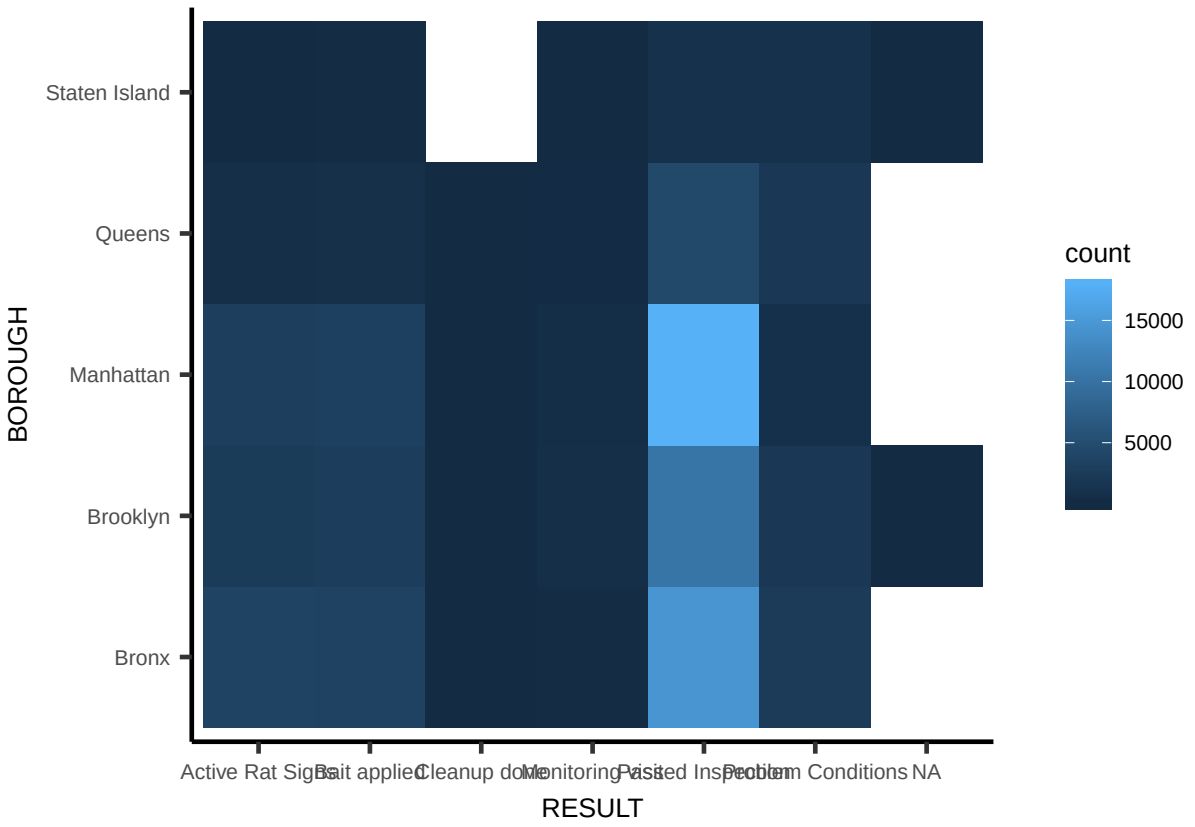
The bar graph shows the distribution of `RESULT`. `Passed Inspection` appears to be significantly high frequency. Bait applied, Active Rat Sign, and Problem Conditions follow, and these three results have closed number of counts.

```
library(vcd)
library(grid)
library(tidyverse)

counts <- data %>%
  group_by(RESULT, BOROUGH) %>%
  summarise(Freq = n()) %>%
  ungroup() %>%
  arrange(desc(Freq)) %>%
  complete(RESULT, BOROUGH)
counts$RESULT <- factor(counts$RESULT,
                                levels = c('Passed Inspection','Bait applied','Active Rat Sign','Proble
counts$BOROUGH <- factor(counts$BOROUGH,
                         levels=c('Manhattan','Bronx','Brooklyn','Queens','Staten Island'))
mosaic(factor(RESULT)~BOROUGH, counts,
       direction = c('v','h'),gp_labels=(gpar(fontsize=10)))
```

The masaic plot appears that there is some association between `RESULT` and `BOROUGH`. Manhattan appears to have the largest proportion of Passed Inspection, Bronx has the second largest, Brooklyn follows, while Staten Island has the smallest.

```
library(hexbin)
library(dplyr)
ggplot(data, aes(RESULT, BOROUGH)) +
  theme_classic(18) +
  geom_bin2d(bins = 15)+
  theme(text=element_text(size=10))
```



The heatmap shows the approximate count of each result in each borough. The count of Passed Inspection appears relatively very high in Manhattan, Bronx, and Brooklyn, which is consistent with the findings in the mosaic plot.

New question: how does the distribution of `RESULT` differ by smaller areas?