NYC_Rats_Project

Nahel Rifai

5/6/2017

This project revolves around the issue of rats in NYC. We analyze if rats have been increasing over time in all of the NYC boroughs. We will use the following datasets: 1. NYC Rodent Inspection Dataset (each data point represents a report submitted by a person who has seen a rat in their apartment) and 2. Restaurant Inspection Results (each data point represent a restaurant inspection and its results - sometimes, rat nests are found in restaurants and this where it would be reported).

The main task of this project is to find whether there is a correlation between rats being found in apartments and rat nests being found in restaurants. One would initially think that if there are Active Rat Signs in an apartment, then there should also be a higher probability of finding rats in restaurants nearby. But this project finds that there is actually a negative correlation. When rats are reported, and NOT found by the inspectors, there is a higher probability that a restaurant nearby has a rat nest.

Setting the Working Directory and Loading all Libraries

```
setwd("~/Desktop/Prasanna/Rats/files")
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following object is masked from 'package:base':
##
##     date
```

```
library(magrittr)
library(ggplot2)
library(nnet)

rat_data = NULL
rat_data = read.csv("Nahel Rifai", header = TRUE) ##loading main rat data file into Rstudio
```

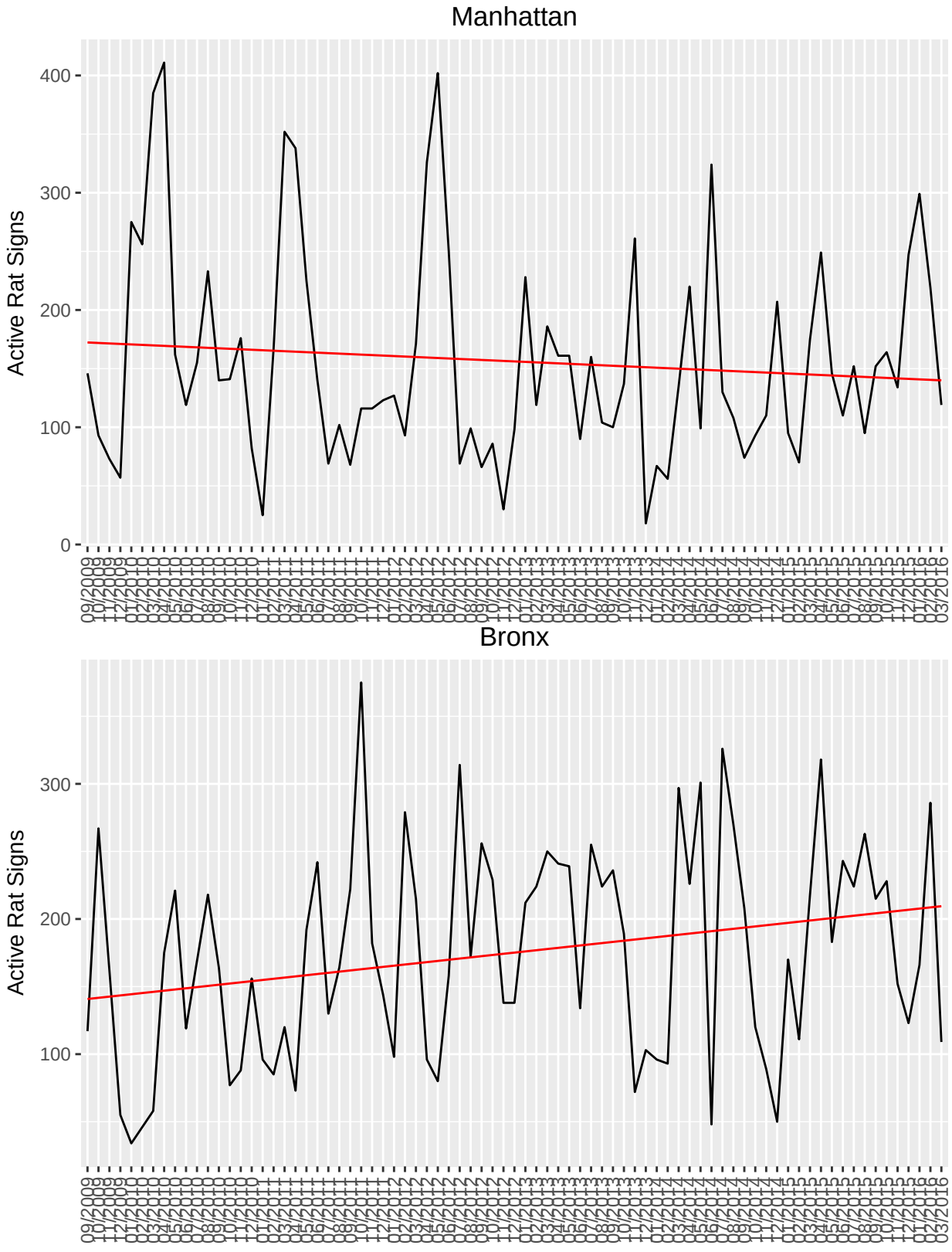
Part 1. Descriptive Statistics and Figures

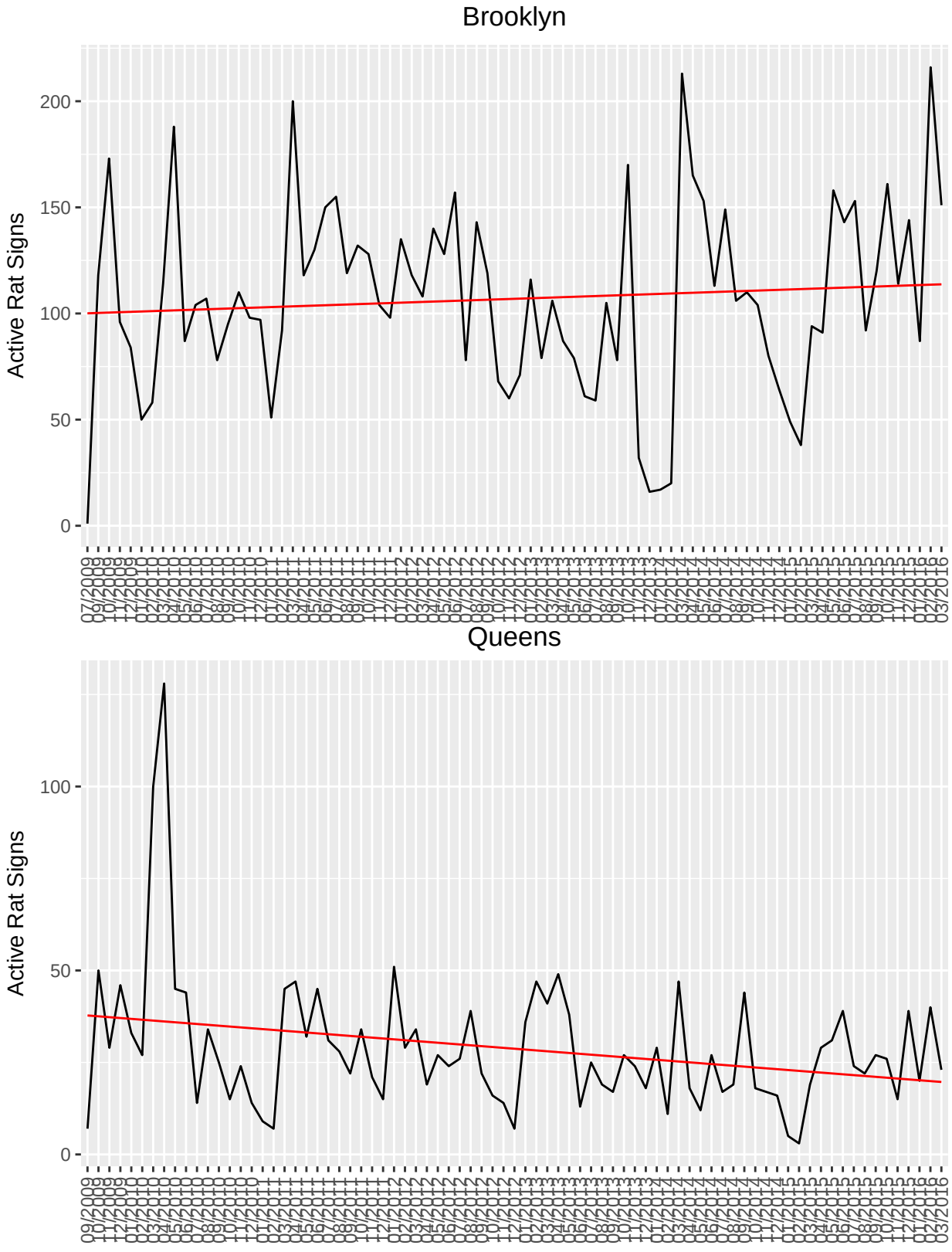
The "Active Rat Signs" variable indicates signs of rats were reported in an apartment and, when the inspector was sent, he/she actually found a rat nest.

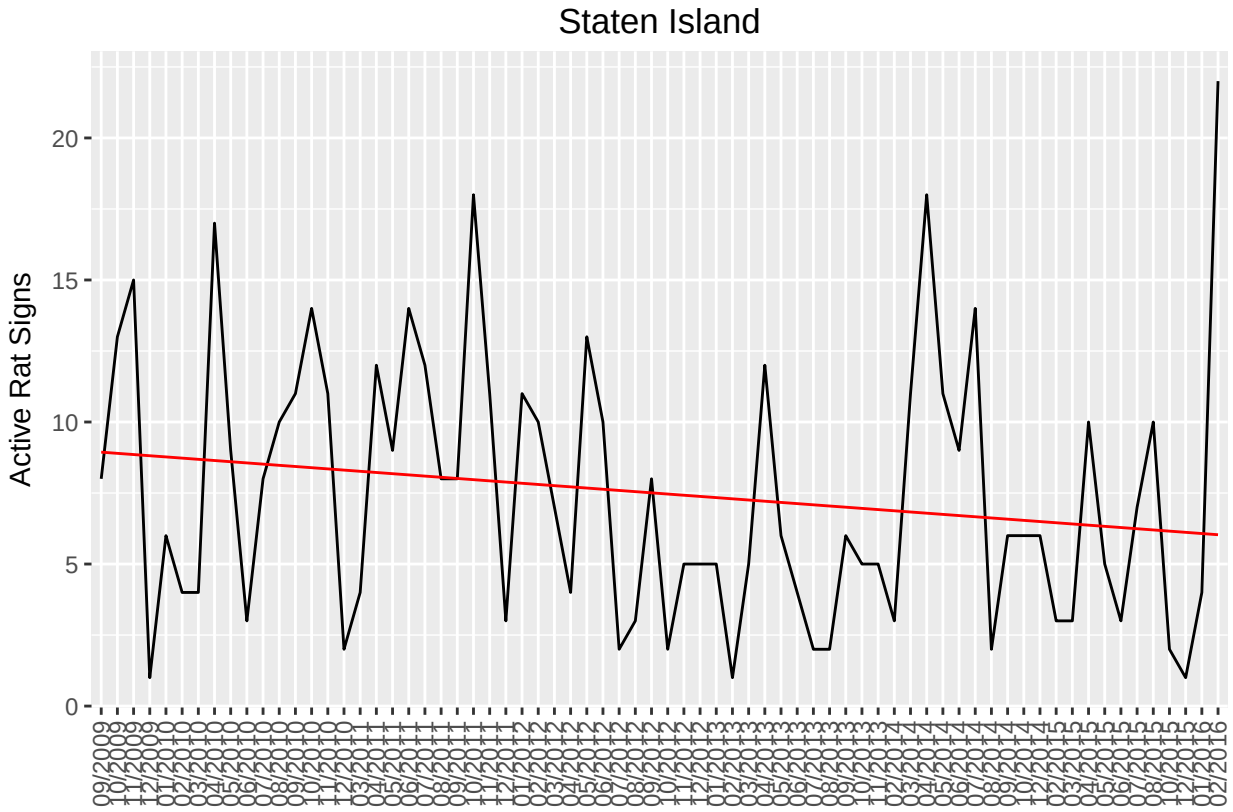
a) Active Rat Signs by Borough Over Time

```
desiredMonth = substr(as.character(rat_data$INSPECTION_DATE), 1,2)
##Stripping out the month from inspection date column
```

```r
desiredYear = substr(as.character(rat_data$INSPECTION_DATE), 7,10)
##Stripping out the year from inspection date column
month_year = paste(desiredMonth,desiredYear,sep = "/")
##create a variable with month/ear
rat_data$monthYear = c(month_year)
#create a column with a month/year for each row

initial_filter = NULL
initial_filter = filter(rat_data, RESULT == "Active Rat Signs" & INSPECTION_TYPE == "INITIAL")
## data frame with instances when rats were found

for (i in 1:5) {

  ##this loop creates a graph of active rat signs per month/year for every borough
  bor_filter = filter(initial_filter, BORO_CODE == i)

  ##filter out the borough
  temp = bor_filter %>% group_by(monthYear) %>% tally()

  ##amount of instances per month per year
  data_ordered = temp[order(as.Date(paste0(temp$monthYear,'/1'),"%m/%Y/%d")),]

  ##in order to sort by time and not alphabetically
  ##we need to create a m/d/y (we added a dummy day 1 for every month)

  colnames(data_ordered)[2] = "ActiveRatSightings"

  ## change column name from "n" to "Active Rat Sightings"
  data_ordered$monthYear = factor(data_ordered$monthYear, levels = data_ordered$monthYear)

  ##telling R this is the order we want, do not re-arrange when plotting graph
  timegraph =
    ggplot(data_ordered, aes(x= monthYear, y=ActiveRatSightings, group=1)) +
    geom_line() +
    theme(axis.text.x=element_text(angle=90,hjust=1,vjust=0.5))+ xlab("") +
    ylab("Active Rat Signs") +
    stat_smooth(method = "lm", col = "red", size = 0.5, se = FALSE) +
    ggtitle(bor_filter$BOROUGH[i]) +
    theme(plot.title = element_text(hjust = 0.5))

  print(timegraph)
}
```


Manhattan
Active Rat Signs
400
300
200
100
0
Bronx
Active Rat Signs
300
200
100


Brooklyn
Active Rat Signs
200
150
100
50
0
Queens
Active Rat Signs
100
50
0

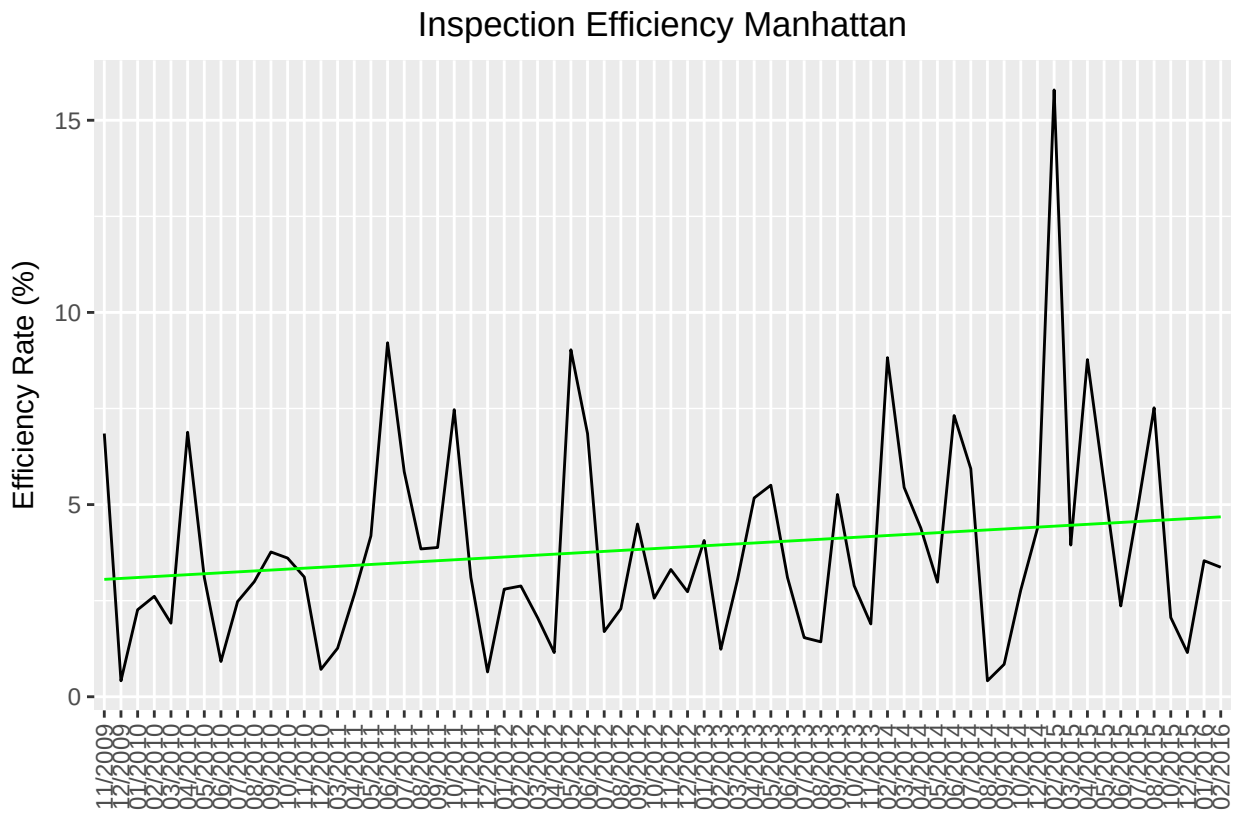


b) Number of inspections yielding "Active Rat Signs"

-Active Rat Signs/All Rat Reports; we'll call this "Efficiency of Inspection"

```
inspection_filter = filter(rat_data, BORO_CODE ==5)
inspection_filter$sign_logic = (inspection_filter$RESULT == "Active Rat Signs"
                                & inspection_filter$INSPECTION_TYPE == "INITIAL")

##for every effective inspection, add a TRUE on a new column
inspection_filter$sign_number = as.integer(as.logical(inspection_filter$sign))

## Convert TRUE to 1, and FALSE to 0
inspection_filter$sign_logic = NULL ##delete TRUE/FALSE column, keep binary column

temp_1 = aggregate(inspection_filter$sign_number,
                   by=list(monthYear= inspection_filter$monthYear), FUN=sum)

##using the binary column, add all 1s and 0s to create
##an aggregate of all effective inspections per month
temp_2 = inspection_filter %>% group_by(monthYear) %>% tally()

##then simply generate the number of all the
##inspections per month (what's going to be the denominator)
merged_table = merge(temp_1, temp_2, by = "monthYear")

##we create a dataframe that has columns month/Year, effective inspections,
##and total num. of inspections
merged_table$efficiency = merged_table$x/merged_table$n
## we create the "efficiency rate by doing effective/total inspections
```

```r
colnames(merged_table) = c("monthYear", "Rats Count", "Inspections Count", "Efficiency")
## change column names

merged_table$`Rats Count`= NULL
##delete effective inspections, only interested on efficiency rate now

merged_table$`Inspections Count` = NULL
##delete total inspections, only interested on efficiency rate now

merged_table = merged_table[order(as.Date(paste0(merged_table$monthYear,'/1'),"%m/%Y/%d")),]
##to order by date instead of alphabetial

merged_table<-merged_table[!(merged_table$Efficiency==0 | merged_table$Efficiency==1),]
##clean up for erroneous results (probability of 0% or 100%)

merged_table <- merged_table[3:nrow(merged_table), ]
##cleaning up first two rows anomalous results

merged_table$monthYear = factor(merged_table$monthYear, levels = merged_table$monthYear)
##tell R this is the order we want, don't change order when plotting

ggplot(merged_table, aes(x= monthYear, y=Efficiency*100, group=1)) + geom_line() +
  theme(axis.text.x=element_text(angle=90,hjust=1,vjust=0.5)) + xlab("") +
  ylab("Efficiency Rate (%)") +
  stat_smooth(method = "lm", col = "green", size = 0.5, se = FALSE) +
  ggtitle("Inspection Efficiency Manhattan") +
  theme(plot.title = element_text(hjust = 0.5))
```



Part 2. We'll now see "Hot Spots" of Rats in Manhattan by Zipcode

These are the worst zipcodes in terms of "Active Rat Signs"

```
borough_filter = filter(rat_data, BORO_CODE ==5 &  RESULT == "Active Rat Signs"
                        & INSPECTION_TYPE == "INITIAL")
zipcode_count = borough_filter %>% group_by(ZIP_CODE) %>% tally()
##tally up active signs per borough
top_ten = zipcode_count[order(zipcode_count$n, decreasing= T),]
##orden by amount, decreasing
top_ten = top_ten[1:10,] ##filter top ten
top_ten
```

```
## # A tibble: 10 × 2
##    ZIP_CODE     n
##       <dbl> <int>
## 1     10301    82
## 2     10312    71
## 3     10314    62
## 4     10304    61
## 5     10303    50
## 6     10305    49
## 7     10302    36
## 8     10306    36
## 9     10309    21
## 10    10310    21
```

Part 3. The Logistic Model

Using the Restaurant Inspection Database, we create a new column called "Rat Violation" that takes 1 if the inspector found a rat nest in the Restaurant or 0 if they didn't. We will now see if there is a correlation between "Active Rat Signs" in apartments and "Rat Violations" in restaurants in the same zip code.

These are the two possible hypotheses:

1. One argument is that lower rat inspection yields in homes mean that there are indeed fewer rats in the area, which can, in turn, mean that local restaurants are less likely to have violations based on rodents.
2. An alternative argument is that a lower rat inspection hit rate in homes can indicate that many people are seeing rats, but inspectors are not finding them in homes, which suggests that they may be nesting at local establishments (especially restaurants because they are attracted to food).

a) Cleaning and preparing data for regression model

```
rat_data$Year = c(desiredYear)
##Identifying Restaurants with rat problems

restaurants_data = NULL
restaurants_data = read.csv("DOHMH_New_York_City_Restaurant_Inspection_Results.csv", header = TRUE)
##load restaurants inspections database

restaurants_data$rat_logic = restaurants_data$VIOLATION.CODE == "04L"  |
  restaurants_data$VIOLATION.CODE == "04K"  | restaurants_data$VIOLATION.CODE == "08A)"
##for every restaurant with rat problems, add TRUE on a separate column, else FALSE

restaurants_data$rat_binary = as.integer(as.logical(restaurants_data$rat_logic))
##convert TRUE/FALSE column to 0s and 1s
restaurants_data$rat_logic = NULL
```

```
##general rat data, likelihood of rat by zip code by year

rat_data$Month = c(desiredMonth)
##first, create a new variable Month on the original rat database
##(previously we had month/year, and Year, now we only want Month separate as well)

##number of rat incidents,
##tally up number of rat inspections by zipcode, month, and year

total_rat_inspections = filter(rat_data, INSPECTION_TYPE == "INITIAL")
inspections_zip_group = total_rat_inspections %>%
  group_by(ZIP_CODE, Month, Year) %>% tally()
inspections_zip_group =
  inspections_zip_group[!(inspections_zip_group$ZIP_CODE < 1000
                          | inspections_zip_group$ZIP_CODE > 99950),]
inspections_zip_group <-
  inspections_zip_group[complete.cases(inspections_zip_group), ]
##remove NAs
colnames(inspections_zip_group)[4] = "total_inspections"

##total number of rat sightings, tally up number of active rat signs,
##by zipcode, month and year

rat_sightings = filter(rat_data, RESULT == "Active Rat Signs"
                       & INSPECTION_TYPE == "INITIAL")
rat_zip_group = rat_sightings %>%
  group_by(ZIP_CODE, Month, Year) %>% tally()
rat_zip_group = rat_zip_group[!(rat_zip_group$ZIP_CODE < 1000
                                | rat_zip_group$ZIP_CODE > 99950),]
rat_zip_group <- rat_zip_group[complete.cases(rat_zip_group), ] ##remove NAs
colnames(rat_zip_group)[4] = "rat_sightings"

##merge two dataframes (first one contains total inspections by combo of zip, month,
##and year, second one is active rat signs), find likelihood

merged_likelihood = merge(inspections_zip_group, rat_zip_group,
                          by=c("ZIP_CODE","Month", "Year"))
merged_likelihood$likelihood =
  merged_likelihood$rat_sightings/merged_likelihood$total_inspections
##now that we merged by zip,month and year,
##lets find "likelihood" of finding rats per reported
##incident for every zip,month, year
merged_likelihood$rat_sightings = NULL
merged_likelihood$total_inspections = NULL
colnames(merged_likelihood)[1] = "ZIPCODE" ## we change name so that we can merge
##it later with restaurant data frame
## "merged likelihood" contains likelihood of inspectors findings rats when they
##were reported (every row has a separate combo of zipcode, month, and year)

## Now we go back to the "Restaurants database",
##for every restaurant row we'll add Month and Year of the inspection

rest_desiredMonth = substr(as.character(restaurants_data$INSPECTION.DATE), 1,2)
```

```
rest_desiredYear = substr(as.character(restaurants_data$INSPECTION.DATE), 7,10)
restaurants_data$Month = c(rest_desiredMonth)
restaurants_data$Year = c(rest_desiredYear)
merged_final = merge(merged_likelihood, restaurants_data,
                     by=c("ZIPCODE","Month", "Year"))
##by merging with Restaurants database by Zip,
##month, and year, every "restaurant" row will have the likelihood of
##inspectors finding rats in apartments in that area
clean_table =
  merged_final[,c("Month", "Year", "ZIPCODE", "likelihood", "rat_binary")]
##we clean the table so we're only left with columns that matter for
##regression, every row still represents a restaurant inspection
colnames(clean_table)[5] = "RatViolation"
colnames(clean_table)[4] = "ActiveRatSightings"
```

b) Creating and Running the logistic regression

```
set.seed(4382)
clean_table$Month <- as.factor(clean_table$Month)
##set month as factor for regression
clean_table$Year = as.factor(clean_table$Year)
##set year as factor for regression

model.final = glm(RatViolation ~ ActiveRatSightings + Month + Year,
                  data=clean_table,
                  family = binomial(link="logit")
)

summary(model.final)
```

```
##
## Call:
## glm(formula = RatViolation ~ ActiveRatSightings + Month + Year,
##     family = binomial(link = "logit"), data = clean_table)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.4531  -0.4212  -0.3983  -0.3816   2.3645
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -10.42486   48.70465  -0.214 0.830514
## ActiveRatSightings  -0.09307    0.02743  -3.393 0.000692 ***
## Month02              0.00743    0.03526   0.211 0.833124
## Month03             -0.04206    0.03432  -1.226 0.220377
## Month04             -0.10770    0.03524  -3.057 0.002239 **
## Month05             -0.21548    0.03671  -5.870 4.35e-09 ***
## Month06             -0.24082    0.03683  -6.538 6.23e-11 ***
## Month07             -0.37223    0.03928  -9.476  < 2e-16 ***
## Month08             -0.31696    0.03949  -8.026 1.00e-15 ***
## Month09             -0.29269    0.03785  -7.733 1.05e-14 ***
## Month10             -0.27320    0.03761  -7.265 3.74e-13 ***
## Month11             -0.14494    0.03906  -3.711 0.000206 ***
## Month12             -0.07143    0.03793  -1.883 0.059652 .
```

```
## Year2011             8.18077   48.70589   0.168 0.866613
## Year2012             8.18074   48.70466   0.168 0.866610
## Year2013             8.19496   48.70465   0.168 0.866380
## Year2014             8.15775   48.70465   0.167 0.866981
## Year2015             8.17704   48.70465   0.168 0.866670
## Year2016             8.02599   48.70465   0.165 0.869110
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 137715  on 250192  degrees of freedom
## Residual deviance: 137466  on 250174  degrees of freedom
## AIC: 137504
##
## Number of Fisher Scoring iterations: 9
```

Final Findings:

The regression tells us the following: The variable "ActiveRatSightings" is statistically significant as P is very close to 0. The other statistically significant variables are months 5 to 11 (May to November). The other variables appear to be insignificant.

What lessons do we learn from this? ActiveRatSightings: This variable is negatively correlated to restaurants having a "rat nest". Thus, this would mean that the lower the "ActiveRatSightings" in the zip code, the higher the chance of there being a "nest" in a restaurant nearby. This supports Prof. Tambe's argument that "a lower rat inspection hit rate in homes can indicate that many people are seeing rats, but inspectors are not finding them in homes, which suggests that they may be nesting at local establishments (especially restaurants because they are attracted to food)". Hence, the Sanitary department should expect to find more "rat nests" in eating establishments located in zip codes where the "ActiveRatSightings" are lower May to November: We can see the warmer months of the year are statistically significant in our regression (specially July). Since they are negatively correlated, this means that, all else being equal, during warmer months there is actually less probability of finding a "nest" in a restaurant. This makes sense because during the warmer months of the year people go out much more. When people go out they leave much more food/trash on the the streets (think picnics at Central Park). Therefore rats also go out to the streets on warmer months as there is more supply of food there. This is why the inspectors are finding less rat nests inside the restaurants during the warmer months of the year