DETER COVID-19 DS3 Analysis

1 Introduction

In this analysis, we use a dataset of 5,163 records documenting the behavior of 6,199 individuals around 19 hospitals and urgent care clinics across 4 of New York's 5 boroughs at the height of the city's COVID-19 outbreak in the Spring of 2020 (March 22-May 19). The records were collected opportunistically by 16 student observers over 1,500 hours across all days of the week and hours of the day. Of all the variables available for discovery and investigation, the binary variable of touch object or not and whether the observer wears PPE(Personal Protective Equipment) or not attracts my attention. As some recent research stats that some effective way of avoiding COVID-19 transmission is to wear a Mask and touch less object. I would like to investigate the relationship between other predictors such as gender, facility type has a significant impact on whether the observer uses PPE(Personal Protective Equipment) or touch an object. After some exploration, we decide to use fit multiple regression to the model as well as apply the multi-level model to the dataset. We would want to potentially provide some suggestions on how to increase the percentage of People who wear Personal Protective Equipment or decrease the percentage of People who touch fewer objects.

2 Data Wrangling and Visualization

2.1 Data Preprocessing

To discover the relationship between different predictors variable and our outcome variable Touch Object and PPE(Personal Protective Equipment), all data have been preprocessed for model fitting and visualization. Touch Object and PPE(Personal Protective Equipment) Use are recoded into a binary variable with 'No' as the reference level. Also, we generated the length of each record using the MAR mechanism to get the start and end time for each record. Thus, we can generate a binary variable of work and off-work hours for the time when each observer is being recorded.

2.2 Descriptive Analysis and Visualization

In this section, we decide to perform some basic analysis to discover the possible impact of gender, observers on our outcome variable Touch_Binary. Since different data are collected by different data recorders from different medical facilities in a different borough. It is natural to consider whether different facilities and observers could lead to the difference in the distribution of the percentage of observers who touched objects.

```
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(tidyr)
Master<- read.csv('Master_spread_dta.csv')
#Observer, facility count
Master %>% group_by(Observer,Facility_Name) %>% count()
```

```
## # A tibble: 19 x 3
## # Groups:   Observer, Facility_Name [19]
##    Observer Facility_Name                                    n
##    <fct>    <fct>                                        <int>
##  1 AH       Flatbush - CityMD Urgent Care                   19
##  2 AK       Elmhurst Hospital Center Emergency Room         90
##  3 AK       Fair Medical Care                               42
##  4 BD       CityMD Pelham Parkway Urgent Care              446
##  5 BD       Montefiore Hospital                              7
##  6 CH       Parkchester CityMD Urgent Care                 244
##  7 DC       23rd St CityMD Urgent Care                     335
##  8 IS       Wyckoff Heights Medical Center                 631
##  9 JF       Montefiore- Albert Einstein Medical Campus     436
## 10 JP       ModernMD urgent care-SE williamsburg           188
## 11 KS       Mount Sinai Beth Israel                        299
## 12 MV       Flushing Hospital Medical Center               391
## 13 NT       NYC Health + Hospitals/Bellevue                239
## 14 SP       NYU Langone Hospital Brooklyn                  543
## 15 TT       CityMD West 42nd Urgent Care                    52
## 16 TT       Mount Sinai West                                45
## 17 VJ       CityMD Fresh Meadows Urgent Care               318
## 18 VN       NYU Langone Hospital Brooklyn                  680
## 19 WQ       The Brooklyn Hospital Center                   158
```

```
#Day
Master %>% group_by(Day) %>% count()
```

```
## # A tibble: 7 x 2
## # Groups:   Day [7]
##   Day       n
##   <fct> <int>
## 1 Fri     856
## 2 Mon     699
## 3 Sat     628
## 4 Sun     450
## 5 Thu     850
## 6 Tue     893
## 7 Wed     787
```

```
#Day Type
Master %>% group_by(Day_Type) %>% count()
```

```
## # A tibble: 2 x 2
```

```
## # Groups:   Day_Type [2]
##   Day_Type     n
##   <fct>    <int>
## 1 Weekday   4085
## 2 Weekend   1078
```

```
#Time missing
Master %>% group_by(Time_Missing) %>% count()
```

```
## # A tibble: 2 x 2
## # Groups:   Time_Missing [2]
##   Time_Missing     n
##          <int> <int>
## 1            0   766
## 2            1  4397
```

```
#Time length
Master %>% dplyr::select(Observer,Length) %>% na.omit() %>% group_by(Observer) %>% summarise(Mean_lengt
```

```
## # A tibble: 16 x 4
##    Observer Mean_length maximum_length minimum_length
##    <fct>          <dbl>          <int>          <int>
##  1 AH              3.95             15              1
##  2 AK             10.9              35              0
##  3 BD              4.74             23              0
##  4 CH              2.97             26              0
##  5 DC              7.19             28              0
##  6 IS              3.54             22              0
##  7 JF              6.91             19              0
##  8 JP              1.23             16              0
##  9 KS              2.88             73              0
## 10 MV              7.07             24              0
## 11 NT             16.4              59              3
## 12 SP              7                37              0
## 13 TT             29.9             135              0
## 14 VJ              2.88             69              0
## 15 VN              4.36             38              0
## 16 WQ             11.2              33              0
```

```
#Time length longer than an hour
Master %>% filter(Length>60 | length_generated>60) %>% group_by(Observer) %>% count()
```

```
## # A tibble: 3 x 2
## # Groups:   Observer [3]
##   Observer     n
##   <fct>    <int>
## 1 KS           1
## 2 TT           9
## 3 VJ           2
```

```r
#Time type
Master %>% dplyr::select(Observer,time_type) %>% count(Observer,time_type)
```

```
## # A tibble: 31 x 3
##    Observer time_type        n
##    <fct>    <fct>        <int>
##  1 AH       Work_hour       19
##  2 AK       Offwork_hour    28
##  3 AK       Work_hour      104
##  4 BD       Offwork_hour    95
##  5 BD       Work_hour      358
##  6 CH       Offwork_hour   100
##  7 CH       Work_hour      144
##  8 DC       Offwork_hour   122
##  9 DC       Work_hour      213
## 10 IS       Offwork_hour   286
## # ... with 21 more rows
```

```r
Master %>% group_by(time_type) %>% count()
```

```
## # A tibble: 2 x 2
## # Groups:   time_type [2]
##   time_type        n
##   <fct>        <int>
## 1 Offwork_hour  1801
## 2 Work_hour     3362
```

```r
#Number of people
summary(Master$Number_of_People)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.000   1.000   1.201   1.000   9.000
```

```r
#Gender
Master %>% dplyr::select(Observer,Facility_Address,Gender) %>% na.omit() %>% group_by(Observer,Facility
```

```
## # A tibble: 40 x 4
## # Groups:   Observer, Facility_Address [20]
##    Observer Facility_Address                                      Gender       n
##    <fct>    <fct>                                                 <fct>    <int>
##  1 AH       2125 Nostrand Ave, Brooklyn NY                        "Female"     8
##  2 AH       2125 Nostrand Ave, Brooklyn NY                        "Male"      11
##  3 AK       10406 Sutter Ave, Ozone Park, NY 11417                "Female"    24
##  4 AK       10406 Sutter Ave, Ozone Park, NY 11417                "Female~     1
##  5 AK       10406 Sutter Ave, Ozone Park, NY 11417                "Male"      17
##  6 AK       79-01 Broadway, Queens, NY 11373                      "Female"    39
##  7 AK       79-01 Broadway, Queens, NY 11373                      "Male"      44
##  8 BD       2178 White Plains Rd, The Bronx, NY 10462             "Female"   231
##  9 BD       2178 White Plains Rd, The Bronx, NY 10462             "Male"     208
## 10 BD       Greene Medical Arts Pavilion, 3400 Bainbridge Ave, T~ "Female"     4
## # ... with 30 more rows
```

```
Master %>% dplyr::select(Gender) %>% summary()
```

```
##      Gender
##  Female :2581
##  Female :   1
##  Male   :2488
##  NA's   :  93
```

```
#Destination
Master %>% dplyr::select(Final_Destination) %>% summary()
```

```
##         Final_Destination
##  Personal Vehicle:1214
##  Hospital        : 707
##  Parking Lot     : 415
##  Restaurant      : 399
##  Street          : 371
##  Subway          : 365
##  (Other)         :1692
```

```
#PPE Use
Master %>% dplyr::select(PPE_Use) %>% summary()
```

```
##  PPE_Use
##  No  : 463
##  Yes :2123
##  NA's:2577
```

```
#Final Destination
Master %>% dplyr::select(Observer,Facility_Address,Final_Destination) %>% group_by(Observer,Facility_Ad
```

```
## # A tibble: 238 x 4
## # Groups:   Observer, Facility_Address [20]
##    Observer Facility_Address                          Final_Destination     n
##    <fct>    <fct>                                     <fct>             <int>
##  1 AH       2125 Nostrand Ave, Brooklyn NY            Bus Stop              4
##  2 AH       2125 Nostrand Ave, Brooklyn NY            Grocery Store         2
##  3 AH       2125 Nostrand Ave, Brooklyn NY            Other                 5
##  4 AH       2125 Nostrand Ave, Brooklyn NY            Personal Vehicle      2
##  5 AH       2125 Nostrand Ave, Brooklyn NY            Restaurant            1
##  6 AH       2125 Nostrand Ave, Brooklyn NY            Store                 1
##  7 AH       2125 Nostrand Ave, Brooklyn NY            Subway                2
##  8 AH       2125 Nostrand Ave, Brooklyn NY            Urgent Care           2
##  9 AK       10406 Sutter Ave, Ozone Park, NY 11417    Bus Stop              1
## 10 AK       10406 Sutter Ave, Ozone Park, NY 11417    Hospital              4
## # ... with 228 more rows
```

```
Master %>% group_by(Final_Destination) %>% count()
```

```
## # A tibble: 26 x 2
## # Groups:   Final_Destination [26]
##    Final_Destination     n
##    <fct>             <int>
##  1 Ambulance            32
##  2 Bank                 48
##  3 Bicycle               8
##  4 Building              1
##  5 Building(Unknown)     6
##  6 Bus Stop            338
##  7 Citibike             23
##  8 Deli                 26
##  9 Food truck           30
## 10 Grocery Store       302
## # ... with 16 more rows
```

```
#Touch Object
Master %>% group_by(Touch_Binary) %>% count()
```

```
## # A tibble: 3 x 2
## # Groups:   Touch_Binary [3]
##   Touch_Binary     n
##   <fct>        <int>
## 1 "No"          2215
## 2 "No "            1
## 3 "Yes"         2947
```

```
Master %>% group_by(Observer,Facility_Name) %>% count(Touch_Binary)
```

```
## # A tibble: 39 x 4
## # Groups:   Observer, Facility_Name [19]
##    Observer Facility_Name                             Touch_Binary     n
##    <fct>    <fct>                                     <fct>        <int>
##  1 AH       Flatbush - CityMD Urgent Care             No               7
##  2 AH       Flatbush - CityMD Urgent Care             Yes             12
##  3 AK       Elmhurst Hospital Center Emergency Room   No              23
##  4 AK       Elmhurst Hospital Center Emergency Room   Yes             67
##  5 AK       Fair Medical Care                         No               6
##  6 AK       Fair Medical Care                         Yes             36
##  7 BD       CityMD Pelham Parkway Urgent Care         No             253
##  8 BD       CityMD Pelham Parkway Urgent Care         Yes            193
##  9 BD       Montefiore Hospital                       No               6
## 10 BD       Montefiore Hospital                       Yes              1
## # ... with 29 more rows
```

```
#Borough
Master %>% group_by(Borough) %>% count()
```

```
## # A tibble: 4 x 2
## # Groups:   Borough [4]
##   Borough       n
##   <fct>     <int>
```

```
## 1 Bronx      1133
## 2 Brooklyn   2219
## 3 Manhattan   970
## 4 Queens      841
```

```
#Facility Type
Master %>% group_by(Facility_Type) %>% count()
```

```
## # A tibble: 2 x 2
## # Groups:   Facility_Type [2]
##   Facility_Type     n
##   <fct>         <int>
## 1 H              3429
## 2 U              1734
```

```
#number of observers in each Borough
Master %>% group_by(Borough) %>% summarise(Number_of_Observer=length(unique(Observer)))
```

```
## # A tibble: 4 x 2
##   Borough   Number_of_Observer
##   <fct>                  <int>
## 1 Bronx                      3
## 2 Brooklyn                   6
## 3 Manhattan                  4
## 4 Queens                     3
```

```
#Final Destination count
Master %>% group_by(Final_Destination) %>% count()
```

```
## # A tibble: 26 x 2
## # Groups:   Final_Destination [26]
##    Final_Destination     n
##    <fct>             <int>
##  1 Ambulance            32
##  2 Bank                 48
##  3 Bicycle               8
##  4 Building              1
##  5 Building(Unknown)     6
##  6 Bus Stop            338
##  7 Citibike             23
##  8 Deli                 26
##  9 Food truck           30
## 10 Grocery Store       302
## # ... with 16 more rows
```

```
library(ggplot2)
library(sqldf)
```

```
## Loading required package: gsubfn
```

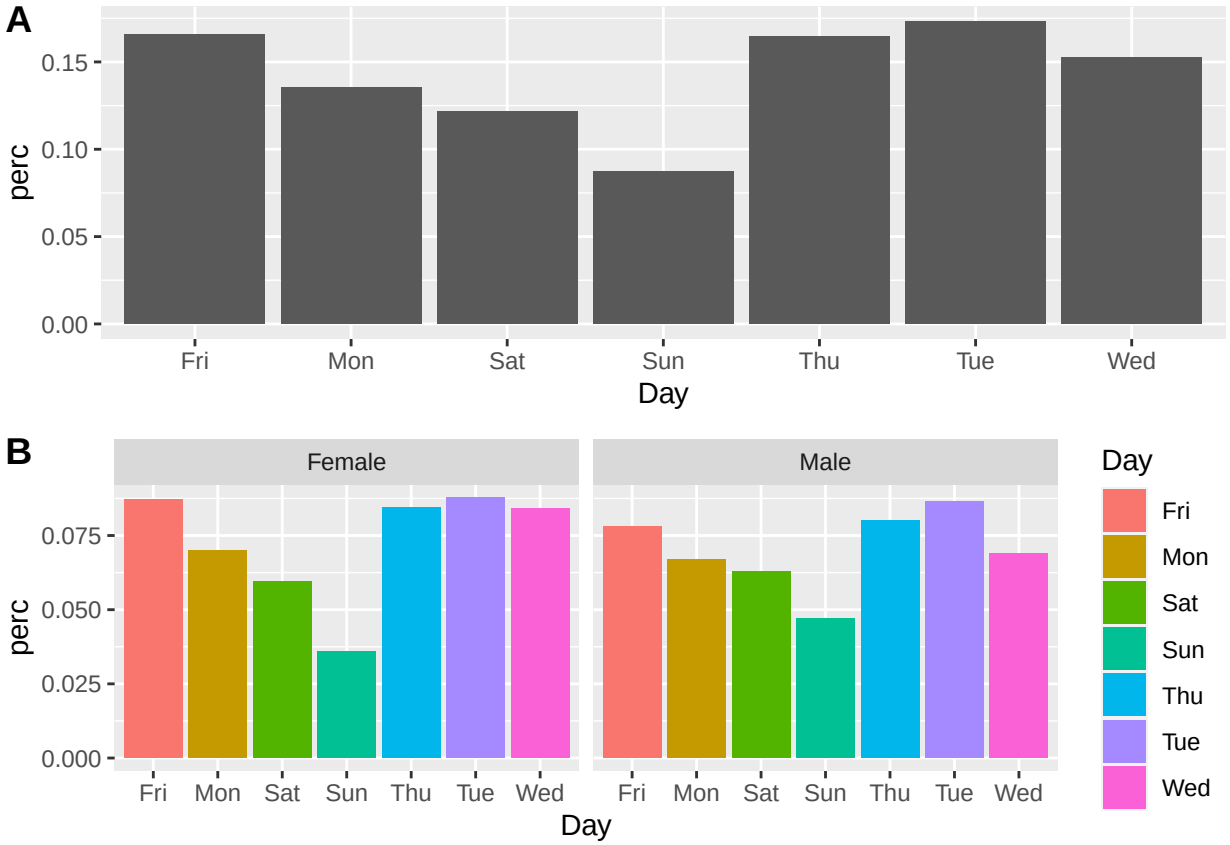
```
## Loading required package: proto
```

```
## Loading required package: RSQLite
```

```
library(ggpubr)
```

```
## Loading required package: magrittr

##
## Attaching package: 'magrittr'

## The following object is masked from 'package:tidyr':
##
##     extract
```

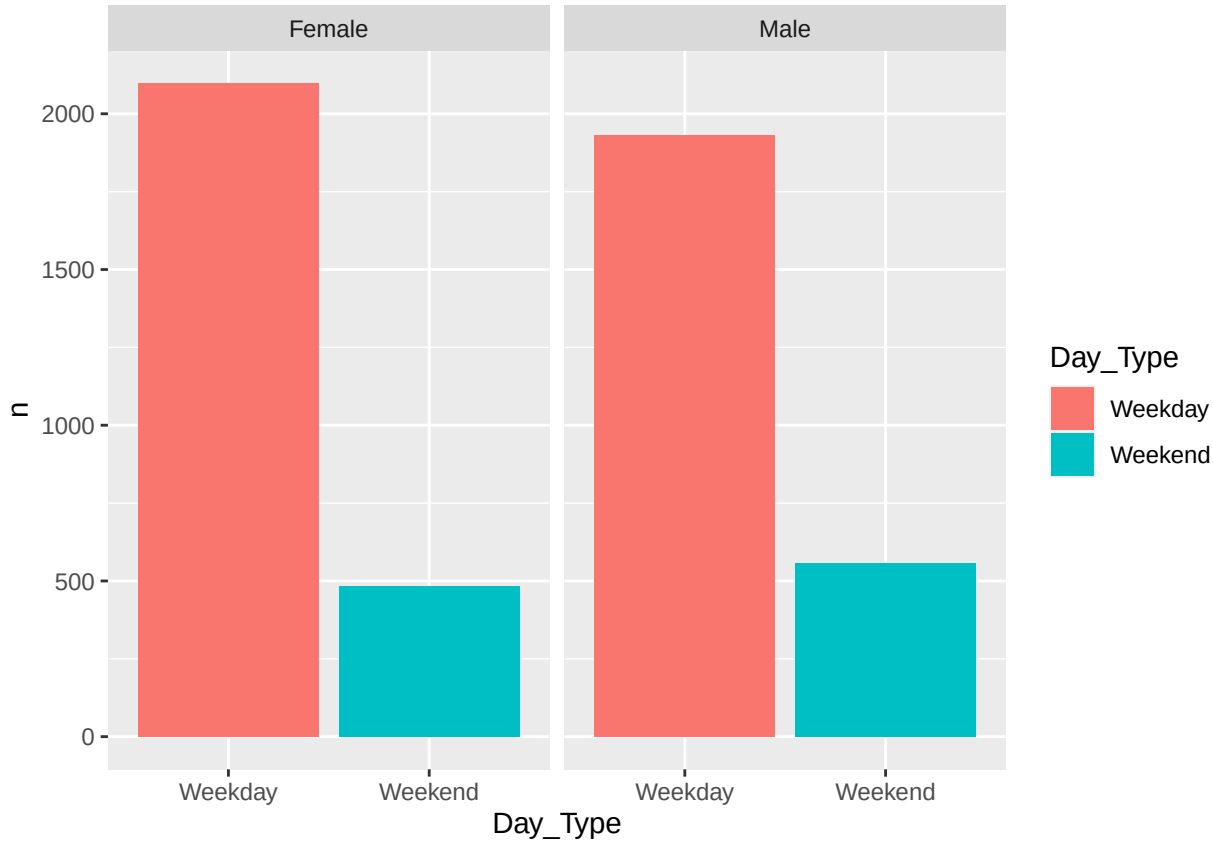
```
#day distribution and day type
Day_dta<- as.data.frame(Master %>% group_by(Day) %>% count())
Day_dta<- Day_dta %>% mutate(perc= n/sum(n))
Day_dta<- Day_dta %>% arrange(perc)
Day_dis<- ggplot(Day_dta,aes(x=Day,y=perc))+geom_bar(stat='identity')

Day_dta_2<- as.data.frame(Master %>% filter(is.na(Gender)==F) %>% group_by(Day, Gender) %>% count())
Day_dta_2<- Day_dta_2 %>% mutate(perc= n/sum(n))
Day_dta_2<- Day_dta_2[c(-10),]
Day_gen<- ggplot(Day_dta_2,aes(x=Day,y=perc))+geom_bar(stat='identity',aes(fill=Day))+facet_grid(~Gende

ggarrange(Day_dis, Day_gen,labels=c('A','B'),ncol=1,nrow=2)
```



From the graph above, we can conclude that the percentage of the records collected on a different day does

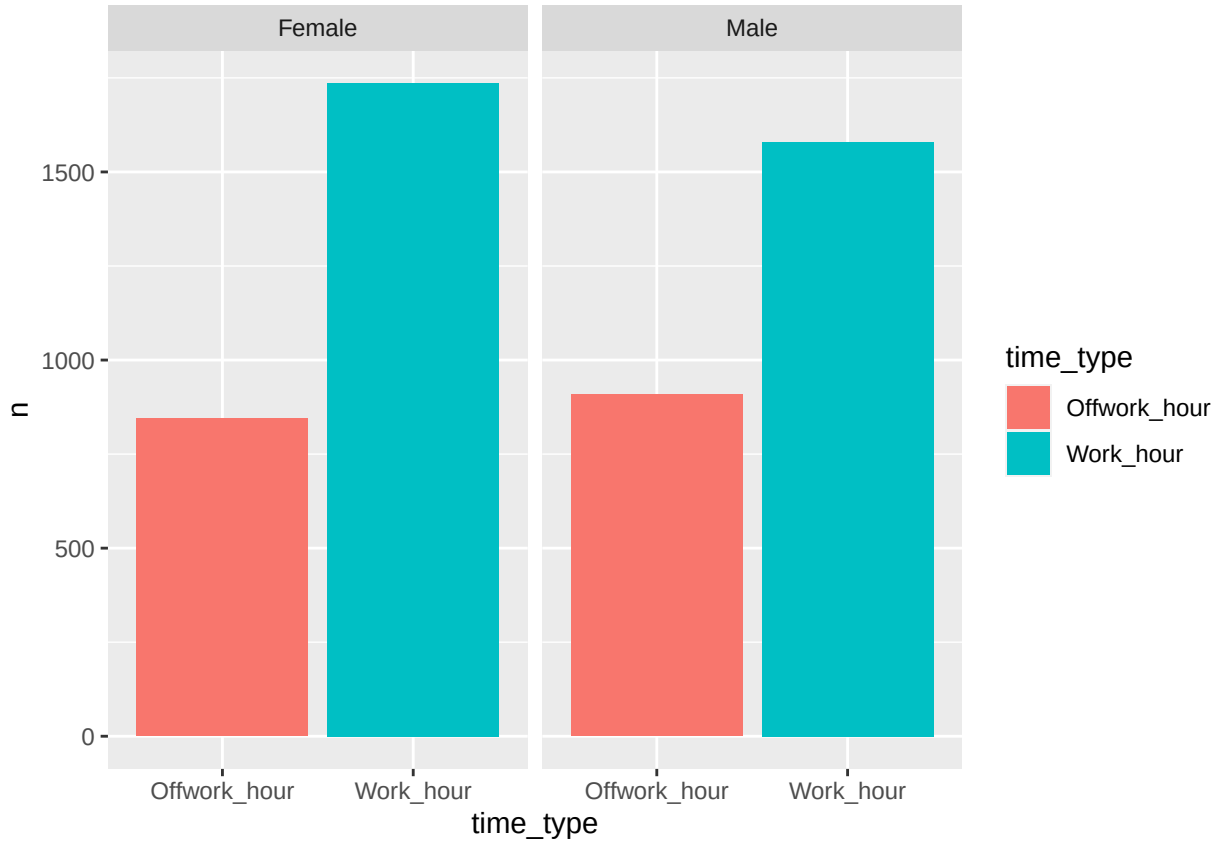
not seem to vary a lot for male and female comparing to the overall distribution. However, we are going to investigate a little bit more into the distribution of the number of records collected on a different day of the week by looking at weekday and weekend.

```
Day_type_dta<- as.data.frame(Master %>% filter(is.na(Gender)==F) %>% group_by(Day_Type,Gender) %>% coun
Day_type_dta<- Day_type_dta[c(-2),]
ggplot(Day_type_dta,aes(x=Day_Type,y=n))+geom_bar(stat='identity',aes(fill=Day_Type))+facet_grid(~Gende
```

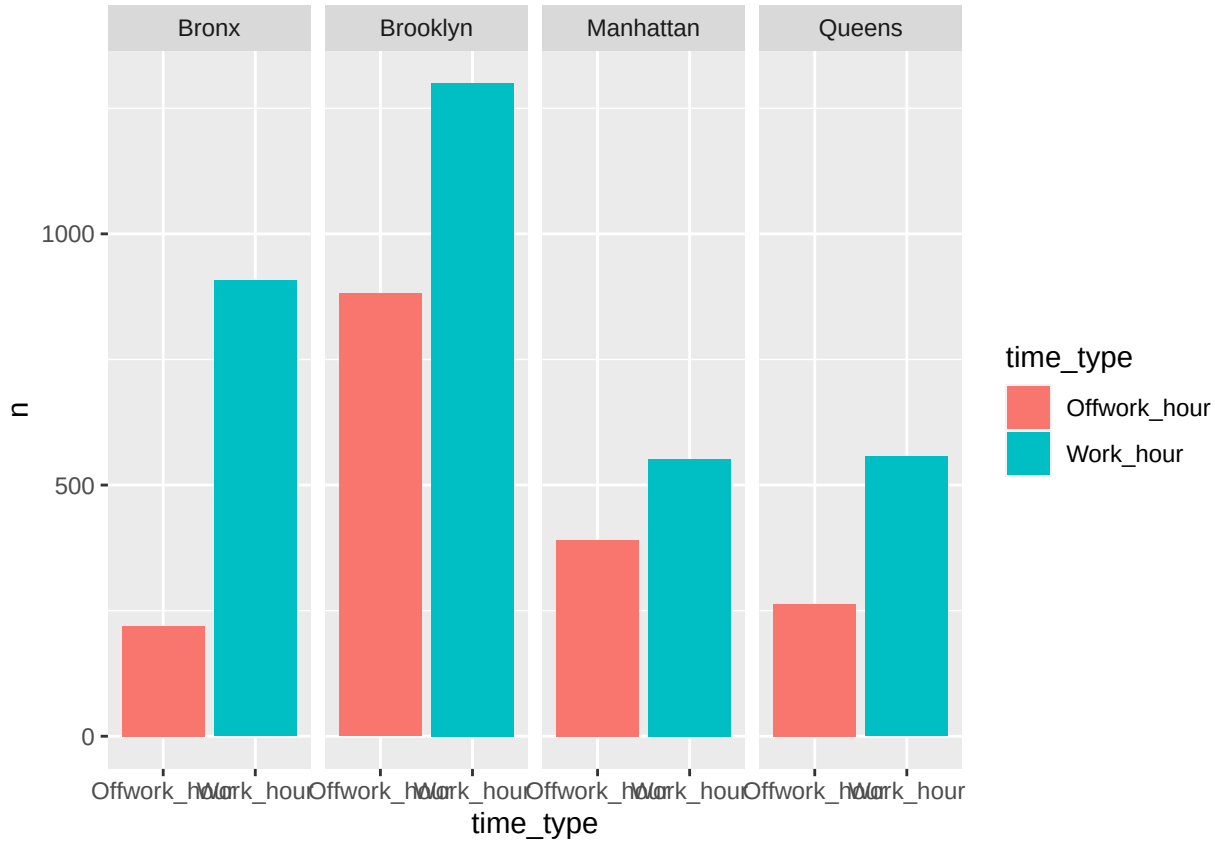


We find out that the number of records collected on weekdays is slightly higher than the male, however, more statistical testing needs to be conducted for us to conclude whether if there is a significant difference between the percentage of records collected on weekdays and weekends for male and female.

```
time_type_dta<- as.data.frame(Master %>% filter(is.na(Gender)==F) %>% group_by(time_type,Gender) %>% co
time_type_dta<- time_type_dta[c(-4),]
ggplot(time_type_dta,aes(x=time_type,y=n))+geom_bar(stat='identity',aes(fill=time_type))+facet_grid(~Ge
```




```
#per borough
Day_type_dta_2<- as.data.frame(Master %>% filter(is.na(Gender)==F) %>% group_by(time_type,Borough) %>% 
ggplot(Day_type_dta_2,aes(x=time_type,y=n))+geom_bar(stat='identity',aes(fill=time_type))+facet_grid(~Bo
```

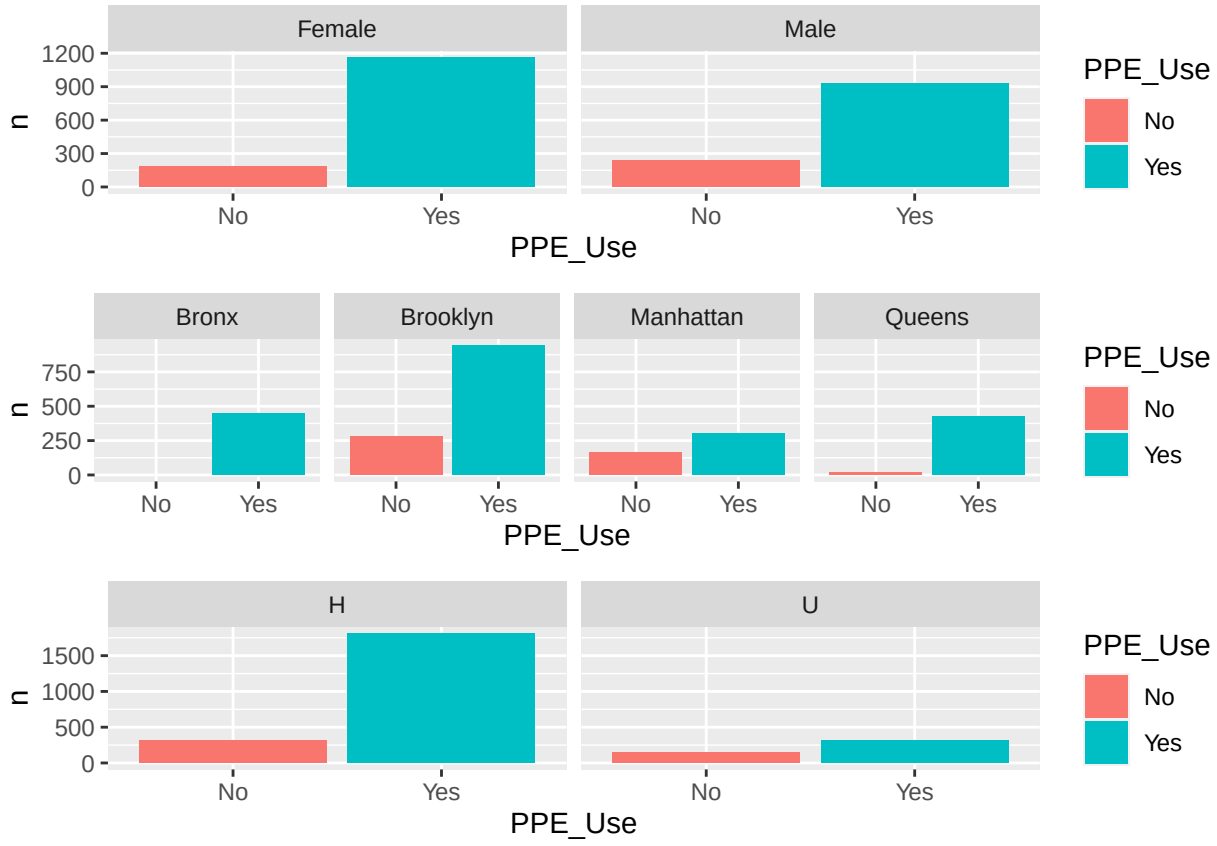


We also look at the distribution of the number of records collected during the off-work hour and work hour(which is defined as from 9:00 am to 5:00 pm). We discover that more records of the female seem to be collected during work hour than male. For Bronx and Brooklyn, the difference in the number of records collected seems to be greater compared to the other two borough Manhattan and Queens.

Furthermore, we are going to look at the distribution of our outcome variable Touch_Binary and PPE_Use by observers' gender, facility type, and borough.

```
PPE_dta<- as.data.frame(Master %>% filter(is.na(Gender)==F & is.na(PPE_Use)==F ) %>% group_by(PPE_Use,G
ppe_gen<- ggplot(PPE_dta,aes(x=PPE_Use,y=n))+geom_bar(stat='identity',aes(fill=PPE_Use))+facet_grid(~Ge

PPE_dta_2<- as.data.frame(Master %>% filter(is.na(Borough)==F & is.na(PPE_Use)==F) %>% group_by(PPE_Use
ppe_bor<- ggplot(PPE_dta_2,aes(x=PPE_Use,y=n))+geom_bar(stat='identity',aes(fill=PPE_Use))+facet_grid(~

PPE_dta_3<- as.data.frame(Master %>% filter(is.na(Facility_Type)==F & is.na(PPE_Use)==F) %>% group_by(P
ppe_fac<- ggplot(PPE_dta_3,aes(x=PPE_Use,y=n))+geom_bar(stat='identity',aes(fill=PPE_Use))+facet_grid(~

Touch_dta<- as.data.frame(Master %>% filter(is.na(Facility_Type)==F & is.na(Touch_Binary)==F) %>% group_
Touch_dta<- Touch_dta[c(-3),]
touch_fac<- ggplot(Touch_dta,aes(x=Touch_Binary,y=n))+geom_bar(stat='identity',aes(fill=Touch_Binary))+

Touch_dta_2<- as.data.frame(Master %>% filter(is.na(Gender)==F & is.na(Touch_Binary)==F) %>% group_by(T
Touch_dta_2<- Touch_dta_2[c(-3,-5),]
touch_gen<- ggplot(Touch_dta_2,aes(x=Touch_Binary,y=n))+geom_bar(stat='identity',aes(fill=Touch_Binary)

Touch_dta_3<- as.data.frame(Master %>% filter(is.na(Borough)==F & is.na(Touch_Binary)==F) %>% group_by(
Touch_dta_3<- Touch_dta_3[c(-5),]
```

```
touch_bor<- ggplot(Touch_dta_3,aes(x=Touch_Binary,y=n))+geom_bar(stat='identity',aes(fill=Touch_Binary))

ggarrange(ppe_gen,ppe_bor,ppe_fac,nrow=3,ncol=1)
```




```
ggarrange(touch_gen,touch_bor,touch_fac,nrow=3,ncol=1)
```

For PPE(Personal Protective Equipment), the use of PPE does not seem to differ much by observers' gender. However, except Manhattan, all three other boroughs seem to have a greater amount of records of Observers wearing PPE than those who don't. The use of PPE(Personal Protective Equipment) differs significantly by facility type. Records that are collected at the hospital have a significantly higher PPE usage than records that are collected at the urgent care.

For Touch_Binary(whether the observer touched the object or not), the distribution of whether they touch the object or not does not seems to have a significant difference by gender. Interestingly, for the Bronx, the number of records where observers did not touch any object is greater than those who did touch. This is different than all other three boroughs which all have a higher amount of records who touched objects than those who didn't. Also, the distribution of 'YES' and 'NO' for the Touch_Binary variable does not seem to differ at the hospital and the urgent care.

3 3D-Visualization

After some initial discovery and descriptive analysis, a 3D visualization seems to be more beneficial for presenting the distribution of our outcome variables. Therefore, we created 3D-visualizations for Touch_Binary and PPE_Use with the x-axis being the Gender percentage, y-axis being out outcome variable, z-axis being the population density, and color by Borough with extra information in the description box.

```
library(plotly)
```

```
##
## Attaching package: 'plotly'
```

```
## The following object is masked from 'package:ggplot2':
##
##     last_plot

## The following object is masked from 'package:stats':
##
##     filter

## The following object is masked from 'package:graphics':
##
##     layout
```

```
#3d bubble plot (Touch object count/Gender count/Population density  per Borough)
bubble_dta<- as.data.frame(Master %>% filter(is.na(Touch_Binary)==F) %>% group_by(Zipcode,Touch_Binary)
bubble_dta<- bubble_dta[c(-6),]
bubble_dta_2<- bubble_dta[,c(-2)]
bubble_dta<- bubble_dta_2 %>% group_by(Zipcode) %>% mutate(Touch_object_Perc=n/sum(n)*100)
bubble_dta_3<- as.data.frame(Master %>% filter(is.na(Gender)==F) %>% group_by(Zipcode,Gender) %>% count
bubble_dta_3<- bubble_dta_3[c(-8),c(-2)]
bubble_dta_3<- bubble_dta_3 %>% group_by(Zipcode) %>% mutate(Gender_Perc=n/sum(n)*100)
Master_popdens<- Master %>% group_by(Zipcode) %>% distinct(Pop_Dens)
bubble_dta<- as.data.frame(cbind(bubble_dta$Zipcode,bubble_dta$Touch_object_Perc, bubble_dta_3$Gender_P
colnames(bubble_dta)<- c('Zipcode','Touch_Object_Percentage','Male_Percentage')
bubble_dta<- bubble_dta[c(1,3,5,7,9,11,13,15,17,19,21,23,25,27,29,31,33),]
bubble_dta<- sqldf('select b.Zipcode, Touch_Object_Percentage, Male_Percentage, Pop_Dens from bubble_dt
record_count<- Master %>% group_by(Borough, Zipcode) %>% count()
colnames(record_count)<- c('Borough','Zipcode','Record_count')
bubble_dta<- sqldf('select * from bubble_dta b inner join record_count r on b.Zipcode=r.Zipcode')
bubble_dta<- bubble_dta[,c(-6)]
bubble_dta$size<- bubble_dta$Record_count
bubble_dta$Zipcode<- as.factor(bubble_dta$Zipcode)
colors <- c('#4AC6B7', '#1972A4', '#965F8A', '#FF7070', '#C61951')
fig_touch<- plot_ly(bubble_dta,x=~Touch_Object_Percentage, y=~Male_Percentage, z=~Pop_Dens, color=~Borou
        marker=list(symbol='circle',sizemode='diameter'), sizes = c(5,150),
        text=~paste('Zipcode:', Zipcode, '<br>Borough:', Borough, '<br>Touch Object Percentage:', Touch_
fig_touch<- fig_touch %>% layout(title= 'COVID-19 Touch Object Percentage v. Male_Percentage, by Zipcod
  xaxis=list(title= 'Percentage of Object that touched object(%)',
             gridcolor= 'rgb(255,255,255)',
             zerolinewidth = 1,
             ticklen = 5,
             gridwidth = 2),
  yaxis=list(title='Percentage of Male in observers(%)',
             gridcolor= 'rgb(255,255,255)',
             zerolinewidth = 1,
             ticklen= 5,
             gridwidth = 2),
  zaxis=list(title='Population Density(by Zipcode)',
             gridcolor= 'rgb(255,255,255)',
             zerolinewidth= 1,
             ticklen= 5,
             gridwidth = 2)),
  paper_bgcolor= 'rgb(243,243,243)',
  plot_bgcolor='rgb(243,243,243)')
fig_touch
```

```
## No trace type specified:
##   Based on info supplied, a 'scatter3d' trace seems appropriate.
##   Read more about this trace type -> https://plot.ly/r/reference/#scatter3d

## No scatter3d mode specifed:
##   Setting the mode to markers
##   Read more about this attribute -> https://plot.ly/r/reference/#scatter-mode

## Warning: `line.width` does not currently support multiple values.

## Warning: `line.width` does not currently support multiple values.

## Warning: `line.width` does not currently support multiple values.

## Warning: `line.width` does not currently support multiple values.
```

```
#3d bubble plot (PPE Use count/Gender Percentage/Population density per Borough)
bubble_dta_ppe<- as.data.frame(Master %>% filter(PPE_Use=='Yes') %>% group_by(Zipcode,PPE_Use) %>% coun
colnames(bubble_dta_ppe)<- c('Zipcode','PPE_Use','PPE_Yes')
ppe_count<- Master %>% filter(PPE_Re_Bi=='Recorded') %>% group_by(Zipcode) %>% count()
colnames(ppe_count)<- c('Zipcode','Total_PPE')
bubble_dta_ppe_new<- sqldf('select * from bubble_dta_ppe b inner join ppe_count p on b.Zipcode=p.Zipcod
bubble_dta_ppe_new$Percentage_PPE_Use<- bubble_dta_ppe_new$PPE_Yes/bubble_dta_ppe_new$Total_PPE
```

```r
bubble_dta_ppe_new<- bubble_dta_ppe_new[,c(-2,-4)]
bubble_dta_ppe_2<- bubble_dta[,c(1,3,4,5)]
bubble_dta_ppe_final<- sqldf('select * from bubble_dta_ppe_new b inner join bubble_dta_ppe_2 b2 on b.Zi
bubble_dta_ppe_final<- bubble_dta_ppe_final[,c(-5)]
bubble_dta_ppe_final$size<- bubble_dta_ppe_final$PPE_Yes
fig_ppe<- plot_ly(bubble_dta_ppe_final,x=~Percentage_PPE_Use, y=~Male_Percentage, z=~Pop_Dens, color=~B
        marker=list(symbol='circle',sizemode='diameter'), sizes = c(5,150),
        text=~paste('Zipcode:', Zipcode, '<br>Borough:', Borough, '<br>PPE Use Percentage:', Percentage
fig_ppe<- fig_ppe %>% layout(title= 'COVID-19 PPE Use Percentage v. Male_Percentage, by Zipcode',scene=
  xaxis=list(title= 'Percentage of Object that Use PPE(%)',
             gridcolor= 'rgb(255,255,255)',
             zerolinewidth = 1,
             ticklen = 5,
             gridwidth = 2),
  yaxis=list(title='Percentage of Male in observers(%)',
             gridcolor= 'rgb(255,255,255)',
             zerolinewidth = 1,
             ticklen= 5,
             gridwidth = 2),
  zaxis=list(title='Population Density(by Zipcode)',
             gridcolor= 'rgb(255,255,255)',
             zerolinewidth= 1,
             ticklen= 5,
             gridwidth = 2)),
  paper_bgcolor= 'rgb(243,243,243)',
  plot_bgcolor='rgb(243,243,243)')
fig_ppe
```

```
## No trace type specified:
##   Based on info supplied, a 'scatter3d' trace seems appropriate.
##   Read more about this trace type -> https://plot.ly/r/reference/#scatter3d
## No scatter3d mode specifed:
##   Setting the mode to markers
##   Read more about this attribute -> https://plot.ly/r/reference/#scatter-mode

## Warning: `line.width` does not currently support multiple values.

## Warning: `line.width` does not currently support multiple values.

## Warning: `line.width` does not currently support multiple values.

## Warning: `line.width` does not currently support multiple values.
```

4 Regression and Statistical Testing

In the beginning, we cleaned the data a little bit more for better regression fitting later in this section.

```
Master_reg<- Master %>% filter(is.na(Gender)==F & is.na(Touch_Binary)==F)
Master_reg$Gender<- relevel(Master_reg$Gender,ref='Female')
Master_reg$Temp<- as.numeric(substr(Master_reg$Temp,start=1, stop=2))
Master_reg$Humidity<- as.numeric(Master_reg$Humidity)/100
Master_reg$Pop_Dens<- as.numeric(substr(Master_reg$Pop_Dens,1,6))
Master_reg$time_type<- relevel(Master_reg$time_type,ref='Work_hour')
Master_reg$Touch_Binary<- ifelse(Master_reg$Touch_Binary=='Yes',1,0)
```

We fit our data into three different regression models to discover which variable seems to have an impact on our outcome variable Touch_Binary.

4.1 Linear Regression

```
Model_Touch_Object_LR<- lm(Touch_Binary~Day_Type+time_type+Number_of_People+Gender+Borough+Facility_Typ
summary(Model_Touch_Object_LR)
```

```
##
## Call:
## lm(formula = Touch_Binary ~ Day_Type + time_type + Number_of_People +
```

```
##     Gender + Borough + Facility_Type + Temp + Humidity + log(Pop_Dens),
##     data = Master_reg)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9479 -0.5216  0.2594  0.4014  0.9148
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -1.3319551  0.2319083  -5.743 9.82e-09 ***
## Day_TypeWeekend         0.1543220  0.0236927   6.513 8.06e-11 ***
## time_typeOffwork_hour -0.1011445  0.0199395  -5.073 4.07e-07 ***
## Number_of_People       -0.0192164  0.0135551  -1.418 0.156353
## GenderFemale            0.3087874  0.4722208   0.654 0.513203
## GenderMale              0.0473405  0.0133444   3.548 0.000392 ***
## BoroughBrooklyn         0.2224941  0.0212831  10.454  < 2e-16 ***
## BoroughManhattan        0.1345230  0.0287873   4.673 3.05e-06 ***
## BoroughQueens           0.3968235  0.0219814  18.053  < 2e-16 ***
## Facility_TypeU          0.1019245  0.0160710   6.342 2.46e-10 ***
## Temp                   -0.0024762  0.0008703  -2.845 0.004455 **
## Humidity                0.1391262  0.0306257   4.543 5.68e-06 ***
## log(Pop_Dens)           0.1642391  0.0210961   7.785 8.38e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4716 on 5057 degrees of freedom
## Multiple R-squared:  0.09426,    Adjusted R-squared:  0.09212
## F-statistic: 43.86 on 12 and 5057 DF,  p-value: < 2.2e-16
```

After fitting the linear regression model, we find out that the number of people is not significant in predicting the outcome variable touch_binary with the p-value of 0.156. Also, to accommodate the better fitting of the data, we use the log of population density instead of the original population density. Gender, Borough, Facility type, temperature, humidity, and population density all seem to be correlated to whether observers touch an object or not.

4.2 Logistic Regression(Binomial)

```
Model_Touch_Object_Bi<- glm(Touch_Binary~Observer+Day_Type+time_type+Number_of_People+Gender+Facility_Ty
summary(Model_Touch_Object_Bi)
```

```
##
## Call:
## glm(formula = Touch_Binary ~ Observer + Day_Type + time_type +
##     Number_of_People + Gender + Facility_Type + Temp + Humidity +
##     log(Pop_Dens), family = "binomial", data = Master_reg)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1802  -1.0397   0.4918   0.9222   2.1011
##
## Coefficients:
```

```
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)              1.348794   4.859283   0.278 0.781342
## ObserverAK               0.676917   0.538671   1.257 0.208883
## ObserverBD              -0.889052   0.506062  -1.757 0.078951 .
## ObserverCH               0.090920   0.517160   0.176 0.860445
## ObserverDC               0.629896   0.609587   1.033 0.301456
## ObserverIS               1.283656   0.661955   1.939 0.052478 .
## ObserverJF              -2.578045   0.680370  -3.789 0.000151 ***
## ObserverJP              -0.377931   0.534553  -0.707 0.479564
## ObserverKS              -1.119490   0.797152  -1.404 0.160211
## ObserverMV               0.623017   0.660162   0.944 0.345306
## ObserverNT              -0.109062   0.811749  -0.134 0.893123
## ObserverSP              -1.179824   0.668744  -1.764 0.077692 .
## ObserverTT              -0.101407   0.611440  -0.166 0.868276
## ObserverVJ               0.029794   0.618184   0.048 0.961560
## ObserverVN              -0.979209   0.667235  -1.468 0.142223
## ObserverWQ               0.021691   0.650227   0.033 0.973388
## Day_TypeWeekend          0.218567   0.114574   1.908 0.056436 .
## time_typeOffwork_hour   -0.200606   0.094709  -2.118 0.034164 *
## Number_of_People        -0.086408   0.065741  -1.314 0.188722
## GenderFemale            10.545140 196.968035   0.054 0.957304
## GenderMale               0.191455   0.064190   2.983 0.002858 **
## Facility_TypeU          -0.247451   0.408776  -0.605 0.544949
## Temp                    -0.012143   0.004168  -2.914 0.003570 **
## Humidity                 0.405659   0.148891   2.725 0.006439 **
## log(Pop_Dens)            0.001010   0.461969   0.002 0.998256
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6925.9  on 5069  degrees of freedom
## Residual deviance: 5862.6  on 5045  degrees of freedom
## AIC: 5912.6
##
## Number of Fisher Scoring iterations: 10
```

```
Model_Touch_Object_Area_Bi<-glm(Touch_Binary~Day_Type+time_type+Number_of_People+Gender+Borough+Facility
summary(Model_Touch_Object_Area_Bi)
```

```
##
## Call:
## glm(formula = Touch_Binary ~ Day_Type + time_type + Number_of_People +
##     Gender + Borough + Facility_Type + Temp + Humidity + log(Pop_Dens),
##     family = binomial(link = "logit"), data = Master_reg)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0891  -1.2091   0.7684   1.0109   2.0216
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)             -8.871272   1.121059  -7.913 2.51e-15 ***
## Day_TypeWeekend          0.670637   0.104889   6.394 1.62e-10 ***
```

```
## time_typeOffwork_hour  -0.436313   0.087711  -4.974 6.54e-07 ***
## Number_of_People       -0.084853   0.060708  -1.398 0.162191
## GenderFemale           10.715933 196.967709   0.054 0.956613
## GenderMale              0.213672   0.060082   3.556 0.000376 ***
## BoroughBrooklyn         0.968932   0.097274   9.961  < 2e-16 ***
## BoroughManhattan        0.530687   0.129016   4.113 3.90e-05 ***
## BoroughQueens           1.798205   0.107363  16.749  < 2e-16 ***
## Facility_TypeU          0.476155   0.075395   6.315 2.69e-10 ***
## Temp                   -0.010864   0.003893  -2.791 0.005255 **
## Humidity                0.641912   0.139445   4.603 4.16e-06 ***
## log(Pop_Dens)           0.794526   0.101531   7.825 5.06e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6925.9  on 5069  degrees of freedom
## Residual deviance: 6426.3  on 5057  degrees of freedom
## AIC: 6452.3
##
## Number of Fisher Scoring iterations: 10
```

```
summary(Model_Touch_Object_Area_Bi)
```

```
##
## Call:
## glm(formula = Touch_Binary ~ Day_Type + time_type + Number_of_People +
##     Gender + Borough + Facility_Type + Temp + Humidity + log(Pop_Dens),
##     family = binomial(link = "logit"), data = Master_reg)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0891  -1.2091   0.7684   1.0109   2.0216
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -8.871272   1.121059  -7.913 2.51e-15 ***
## Day_TypeWeekend         0.670637   0.104889   6.394 1.62e-10 ***
## time_typeOffwork_hour  -0.436313   0.087711  -4.974 6.54e-07 ***
## Number_of_People       -0.084853   0.060708  -1.398 0.162191
## GenderFemale           10.715933 196.967709   0.054 0.956613
## GenderMale              0.213672   0.060082   3.556 0.000376 ***
## BoroughBrooklyn         0.968932   0.097274   9.961  < 2e-16 ***
## BoroughManhattan        0.530687   0.129016   4.113 3.90e-05 ***
## BoroughQueens           1.798205   0.107363  16.749  < 2e-16 ***
## Facility_TypeU          0.476155   0.075395   6.315 2.69e-10 ***
## Temp                   -0.010864   0.003893  -2.791 0.005255 **
## Humidity                0.641912   0.139445   4.603 4.16e-06 ***
## log(Pop_Dens)           0.794526   0.101531   7.825 5.06e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 6925.9  on 5069  degrees of freedom
## Residual deviance: 6426.3  on 5057  degrees of freedom
## AIC: 6452.3
##
## Number of Fisher Scoring iterations: 10
```

Initially, we fit the regression on different data collectors to see if the tendency of recording different behavior of observers. We find that collectors with the ID 'JF' seem to directly influence our outcome variable touch_binary. However, in this section, we will neglect the influence of collectors on its fixed effect. We will address those impacts in our multi-level model in the later section. After fitting the logistic regression model, we find out that observers in Queens and Brooklyn seem to have a higher chance of touching an object with a coefficient of 1.798205 and 0.968932(p-value of 0). Also, the higher than temperature is, the less likely observers will touch any object(coefficient of -0.010864).

4.3 Poisson Regression

```
Model_Touch_Object_Pos<- glm(Touch_Binary~Day_Type+time_type+Number_of_People+Gender+Borough+Facility_Ty
summary(Model_Touch_Object_Pos)
```

```
##
## Call:
## glm(formula = Touch_Binary ~ Day_Type + time_type + Number_of_People +
##     Gender + Borough + Facility_Type + Temp + Humidity + log(Pop_Dens),
##     family = poisson(link = "log"), data = Master_reg)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5120  -1.0083   0.2743   0.4901   1.2741
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -4.921030   0.693159  -7.099 1.25e-12 ***
## Day_TypeWeekend        0.279124   0.067254   4.150 3.32e-05 ***
## time_typeOffwork_hour -0.170161   0.058406  -2.913 0.003575 **
## Number_of_People      -0.039312   0.039342  -0.999 0.317683
## GenderFemale           0.403991   1.001993   0.403 0.686810
## GenderMale             0.085775   0.037499   2.287 0.022173 *
## BoroughBrooklyn        0.498811   0.064676   7.712 1.23e-14 ***
## BoroughManhattan       0.283085   0.081003   3.495 0.000475 ***
## BoroughQueens          0.783829   0.065711  11.928  < 2e-16 ***
## Facility_TypeU         0.229946   0.045113   5.097 3.45e-07 ***
## Temp                  -0.004172   0.002423  -1.722 0.085106 .
## Humidity               0.255238   0.085663   2.980 0.002886 **
## log(Pop_Dens)          0.369528   0.062043   5.956 2.58e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 3244.5  on 5069  degrees of freedom
## Residual deviance: 3018.2  on 5057  degrees of freedom
```

```
## AIC: 8834.2
##
## Number of Fisher Scoring iterations: 5
```

```
Model_Touch_Object_Pos_2<- glm(Touch_Binary~Day_Type+time_type+Gender+Borough+Facility_Type+Humidity+log
summary(Model_Touch_Object_Pos_2)
```

```
##
## Call:
## glm(formula = Touch_Binary ~ Day_Type + time_type + Gender +
##     Borough + Facility_Type + Humidity + log(Pop_Dens), family = poisson(link = "log"),
##     data = Master_reg)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5468  -1.0132   0.2852   0.4873   1.2894
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -5.27977    0.66698  -7.916 2.45e-15 ***
## Day_TypeWeekend         0.25932    0.06633   3.909 9.26e-05 ***
## time_typeOffwork_hour -0.16266    0.05823  -2.793 0.005218 **
## GenderFemale            0.37220    1.00150   0.372 0.710157
## GenderMale              0.08587    0.03748   2.291 0.021969 *
## BoroughBrooklyn         0.48769    0.06447   7.565 3.88e-14 ***
## BoroughManhattan        0.27841    0.08089   3.442 0.000578 ***
## BoroughQueens           0.77253    0.06541  11.810  < 2e-16 ***
## Facility_TypeU          0.22611    0.04510   5.014 5.34e-07 ***
## Humidity                0.31471    0.07905   3.981 6.86e-05 ***
## log(Pop_Dens)           0.37632    0.06188   6.081 1.19e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 3244.5  on 5069  degrees of freedom
## Residual deviance: 3022.1  on 5059  degrees of freedom
## AIC: 8834.1
##
## Number of Fisher Scoring iterations: 5
```

```
Model_Touch_Object_Pos_3<- glm(Touch_Binary~Day_Type+time_type+Gender+Borough+Facility_Type+I(Temp*Humi
summary(Model_Touch_Object_Pos_3)
```

```
##
## Call:
## glm(formula = Touch_Binary ~ Day_Type + time_type + Gender +
##     Borough + Facility_Type + I(Temp * Humidity) + log(Pop_Dens),
##     family = poisson(link = "log"), data = Master_reg)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5273  -1.0155   0.2859   0.4860   1.2853
```

```
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -5.279320   0.666816  -7.917 2.43e-15 ***
## Day_TypeWeekend         0.249183   0.066159   3.766 0.000166 ***
## time_typeOffwork_hour -0.159071   0.058204  -2.733 0.006277 **
## GenderFemale            0.366020   1.001499   0.365 0.714759
## GenderMale              0.085372   0.037485   2.277 0.022758 *
## BoroughBrooklyn         0.486766   0.064474   7.550 4.36e-14 ***
## BoroughManhattan        0.277825   0.080861   3.436 0.000591 ***
## BoroughQueens           0.766959   0.065376  11.732  < 2e-16 ***
## Facility_TypeU          0.225437   0.045109   4.998 5.81e-07 ***
## I(Temp * Humidity)      0.005586   0.001582   3.531 0.000414 ***
## log(Pop_Dens)           0.377252   0.061853   6.099 1.07e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 3244.5  on 5069  degrees of freedom
## Residual deviance: 3025.5  on 5059  degrees of freedom
## AIC: 8837.5
##
## Number of Fisher Scoring iterations: 5
```

In the Poisson regression model, the temperature seems to be not correlated with our outcome variable (p=0.09). However, to consider the fact that we want to use humidity as a measurement for the possibility of raining. It is more accurate if we add an interaction term of temperature and humidity to our Poisson model as the possibility of raining. Also, it works better to balance the effect of temperature since its range is in a certain interval that does not start from 0.

4.4. Model Comparison

In this section, We compared three models to see which one our data fits better using the AIC score as the criteria.

For the linear model, it is more straightforward in its interpretation as the coefficient directly shows the impact of certain predictors. However, it does not account for the fact that our outcome variable is binary.

For the Poisson regression model, although allowing an interaction term between humidity and temperature allows our model to be more accurate in its prediction. However, the AIC score of the Poisson model is 8837.5, which is significantly bigger than the logistic model with the AIC score of 6452.3. Thus, I believe the logistic regression model(Binomial) is a better-fitted model for predicting whether an observer touches any object or not.

5 Multi-level Modelling

As mentioned before, the tendency or habit of each data collector might result in the difference of our outcome variable whether an observer touch object or not. Thus, we fitted a multi-level model to account for such an effect.

First, we fit an unconditional mean model using our dataset using borough as the first level, and observer as the level nested under borough.

```
library(lme4)
```

```
## Loading required package: Matrix
```

```
##
## Attaching package: 'Matrix'
```

```
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
```

```
library(lmerTest)
```

```
##
## Attaching package: 'lmerTest'
```

```
## The following object is masked from 'package:lme4':
##
##     lmer
```

```
## The following object is masked from 'package:stats':
##
##     step
```

```
library(car)
```

```
## Loading required package: carData
```

```
## Registered S3 methods overwritten by 'car':
##   method                          from
##   influence.merMod                lme4
##   cooks.distance.influence.merMod lme4
##   dfbeta.influence.merMod         lme4
##   dfbetas.influence.merMod        lme4
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
Unconditional_Mean_Model<- lmer(Touch_Binary~(1|Borough/Observer),data=Master_reg)
summary(Unconditional_Mean_Model)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: Touch_Binary ~ (1 | Borough/Observer)
##    Data: Master_reg
```

```
##
## REML criterion at convergence: 6288.2
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.9862 -0.9615  0.2509  0.7583  1.8965
##
## Random effects:
##  Groups           Name        Variance Std.Dev.
##  Observer:Borough (Intercept) 0.033083 0.18189
##  Borough          (Intercept) 0.004379 0.06617
##  Residual                     0.199831 0.44702
## Number of obs: 5070, groups:  Observer:Borough, 16; Borough, 4
##
## Fixed effects:
##             Estimate Std. Error      df t value Pr(>|t|)
## (Intercept)  0.59529    0.05746 1.82578   10.36   0.0124 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Then, we add the borough level predictors to the model.

```
#Borough level predictor
Model_borough<- lmer(Touch_Binary~log(E_UNEMP)+RPL_THEMES+(1|Borough)+(1|Observer),data=Master_reg)
```

```
## boundary (singular) fit: see ?isSingular
```

```
summary(Model_borough)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: Touch_Binary ~ log(E_UNEMP) + RPL_THEMES + (1 | Borough) + (1 |
##     Observer)
##    Data: Master_reg
##
## REML criterion at convergence: 6284.5
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.9858 -0.9571  0.2512  0.7676  1.9000
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Observer (Intercept) 0.0297   0.1723
##  Borough  (Intercept) 0.0000   0.0000
##  Residual             0.1998   0.4470
## Number of obs: 5070, groups:  Observer, 16; Borough, 4
##
## Fixed effects:
##              Estimate Std. Error      df t value Pr(>|t|)
## (Intercept)   -4.0178     3.1963 13.3976  -1.257   0.2302
## log(E_UNEMP)   0.4878     0.3080 13.3566   1.584   0.1367
## RPL_THEMES    -1.0258     0.4515 13.0769  -2.272   0.0406 *
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) l(E_UN
## lg(E_UNEMP) -0.996
## RPL_THEMES   0.649 -0.710
## convergence code: 0
## boundary (singular) fit: see ?isSingular
```

```
anova(Model_borough,Unconditional_Mean_Model,refit=F)
```

```
## Data: Master_reg
## Models:
## Unconditional_Mean_Model: Touch_Binary ~ (1 | Borough/Observer)
## Model_borough: Touch_Binary ~ log(E_UNEMP) + RPL_THEMES + (1 | Borough) + (1 |
## Model_borough:     Observer)
##                           Df    AIC    BIC  logLik deviance  Chisq Chi Df
## Unconditional_Mean_Model  4 6296.2 6322.4 -3144.1   6288.2
## Model_borough             6 6296.5 6335.7 -3142.2   6284.5 3.7833      2
##                           Pr(>Chisq)
## Unconditional_Mean_Model
## Model_borough                 0.1508
```

We have added the observer level variable into the multi-level model as well.

```
Model_observer<- lmer(Touch_Binary~Facility_Type+log(Pop_Dens)+log(E_UNEMP)+RPL_THEMES+(1|Borough)+(1|O
```

```
## boundary (singular) fit: see ?isSingular
```

```
summary(Model_observer)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: Touch_Binary ~ Facility_Type + log(Pop_Dens) + log(E_UNEMP) +
##     RPL_THEMES + (1 | Borough) + (1 | Observer)
##    Data: Master_reg
##
## REML criterion at convergence: 6291.6
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.9858 -0.9572  0.2509  0.7665  1.9002
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Observer (Intercept) 0.03146  0.1774
##  Borough  (Intercept) 0.00000  0.0000
##  Residual             0.19988  0.4471
## Number of obs: 5070, groups:  Observer, 16; Borough, 4
##
## Fixed effects:
```

```
##                Estimate Std. Error       df t value Pr(>|t|)
## (Intercept)    -3.88684    3.51274 13.94859  -1.106   0.2872
## Facility_TypeU  0.01012    0.06400 58.64177   0.158   0.8749
## log(Pop_Dens)  -0.01049    0.06754 84.48306  -0.155   0.8769
## log(E_UNEMP)    0.48694    0.32026 12.09143   1.520   0.1541
## RPL_THEMES     -1.04114    0.46968 11.78323  -2.217   0.0471 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) Fcl_TU l(P_D) l(E_UN
## Faclty_TypU -0.106
## log(Pp_Dns) -0.340  0.034
## lg(E_UNEMP) -0.970  0.097  0.115
## RPL_THEMES   0.565 -0.076  0.127 -0.687
## convergence code: 0
## boundary (singular) fit: see ?isSingular
```

```
linearHypothesis(Model_observer,c('Facility_TypeU','log(Pop_Dens)'))
```

```
## Linear hypothesis test
##
## Hypothesis:
## Facility_TypeU = 0
## log(Pop_Dens) = 0
##
## Model 1: restricted model
## Model 2: Touch_Binary ~ Facility_Type + log(Pop_Dens) + log(E_UNEMP) +
##     RPL_THEMES + (1 | Borough) + (1 | Observer)
##
##   Df  Chisq Pr(>Chisq)
## 1
## 2  2 0.0509     0.9749
```

Finally, we added the individual record level variable into this multi-level model.

```
Model_individual<-lmer(Touch_Binary~Day_Type+time_type+Number_of_People+Gender+Temp+Humidity+Facility_Ty
```

```
## boundary (singular) fit: see ?isSingular
```

```
summary(Model_individual)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: Touch_Binary ~ Day_Type + time_type + Number_of_People + Gender +
##     Temp + Humidity + Facility_Type + log(Pop_Dens) + log(E_UNEMP) +
##     RPL_THEMES + (1 | Borough) + (1 | Observer)
##    Data: Master_reg
##
## REML criterion at convergence: 6296
##
## Scaled residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -2.0650 -0.9518  0.2782  0.7825  2.0490
##
## Random effects:
##  Groups   Name        Variance  Std.Dev.
##  Observer (Intercept) 2.915e-02 1.707e-01
##  Borough  (Intercept) 2.452e-11 4.952e-06
##  Residual             1.986e-01 4.457e-01
## Number of obs: 5070, groups:  Observer, 16; Borough, 4
##
## Fixed effects:
##                          Estimate Std. Error         df t value Pr(>|t|)
## (Intercept)           -4.678e+00  3.400e+00  1.422e+01  -1.376  0.19007
## Day_TypeWeekend        4.887e-02  2.317e-02  5.051e+03   2.109  0.03497 *
## time_typeOffwork_hour -4.379e-02  1.939e-02  5.056e+03  -2.259  0.02395 *
## Number_of_People      -1.579e-02  1.289e-02  5.051e+03  -1.225  0.22062
## GenderFemale           2.855e-01  4.496e-01  5.054e+03   0.635  0.52541
## GenderMale             3.784e-02  1.271e-02  5.051e+03   2.978  0.00292 **
## Temp                  -2.451e-03  8.331e-04  5.056e+03  -2.942  0.00328 **
## Humidity               7.999e-02  2.921e-02  5.058e+03   2.738  0.00619 **
## Facility_TypeU         1.766e-02  6.306e-02  5.459e+01   0.280  0.78046
## log(Pop_Dens)          2.133e-02  6.698e-02  7.614e+01   0.318  0.75103
## log(E_UNEMP)           5.364e-01  3.089e-01  1.229e+01   1.736  0.10747
## RPL_THEMES            -1.041e+00  4.528e-01  1.196e+01  -2.298  0.04039 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) Dy_TyW tm_tO_ Nmb__P GndrFm GndrMl Temp   Humdty Fcl_TU
## Day_TypWknd -0.026
## tm_typOffw_  0.010 -0.688
## Numbr_f_Ppl -0.004 -0.001  0.002
## GenderFemal -0.029  0.005  0.001 -0.020
## GenderMale  -0.004  0.007 -0.015  0.024  0.013
## Temp         0.013 -0.168  0.087  0.031 -0.009 -0.006
## Humidity    -0.028  0.037 -0.046  0.046  0.006  0.006  0.384
## Faclty_TypU -0.112  0.046  0.013 -0.008  0.000 -0.004  0.015  0.035
## log(Pp_Dns) -0.350  0.034 -0.022 -0.005  0.095 -0.006 -0.047  0.037  0.040
## lg(E_UNEMP) -0.969  0.020 -0.008 -0.001  0.007  0.003 -0.017  0.011  0.101
## RPL_THEMES   0.559 -0.004  0.010  0.005  0.018  0.003  0.000  0.002 -0.076
##             l(P_D) l(E_UN
## Day_TypWknd
## tm_typOffw_
## Numbr_f_Ppl
## GenderFemal
## GenderMale
## Temp
## Humidity
## Faclty_TypU
## log(Pp_Dns)
## lg(E_UNEMP)  0.120
## RPL_THEMES   0.131 -0.685
## convergence code: 0
## boundary (singular) fit: see ?isSingular
```

```
anova(Model_individual,Model_borough)
```

```
## refitting model(s) with ML (instead of REML)
```

```
## Data: Master_reg
## Models:
## Model_borough: Touch_Binary ~ log(E_UNEMP) + RPL_THEMES + (1 | Borough) + (1 |
## Model_borough:     Observer)
## Model_individual: Touch_Binary ~ Day_Type + time_type + Number_of_People + Gender +
## Model_individual:     Temp + Humidity + Facility_Type + log(Pop_Dens) + log(E_UNEMP) +
## Model_individual:     RPL_THEMES + (1 | Borough) + (1 | Observer)
##                  Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## Model_borough     6 6290.8 6330.0 -3139.4   6278.8
## Model_individual 15 6269.0 6366.9 -3119.5   6239.0 39.799      9  8.261e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

After comparing those models, we can carefully conclude that the model includes individual-level record variable are a better fit using our data. To discover the fixed and random effect at the Borough level and observer level, we add Observer level predictor at the borough level to learn its random effect among different boroughs.

```
Model_Facility_Type<- lmer(Touch_Binary~Day_Type+time_type+Number_of_People+Gender+Temp+Humidity+Facili
```

```
## boundary (singular) fit: see ?isSingular
```

```
summary(Model_Facility_Type)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: Touch_Binary ~ Day_Type + time_type + Number_of_People + Gender +
##     Temp + Humidity + Facility_Type + log(Pop_Dens) + log(E_UNEMP) +
##     RPL_THEMES + (Facility_Type | Borough) + (1 | Observer)
##    Data: Master_reg
##
## REML criterion at convergence: 6291.8
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.0662 -0.9523  0.2791  0.7794  2.0570
##
## Random effects:
##  Groups   Name           Variance Std.Dev. Corr
##  Observer (Intercept)    0.02812  0.1677
##  Borough  (Intercept)    0.01041  0.1020
##           Facility_TypeU 0.04935  0.2221   -1.00
##  Residual                0.19835  0.4454
## Number of obs: 5070, groups:  Observer, 16; Borough, 4
##
## Fixed effects:
##                         Estimate Std. Error         df t value Pr(>|t|)
```

```
## (Intercept)           -4.882e+00  3.385e+00  9.965e+00  -1.442  0.17989
## Day_TypeWeekend        4.876e-02  2.316e-02  5.050e+03   2.105  0.03532 *
## time_typeOffwork_hour -4.327e-02  1.938e-02  5.055e+03  -2.233  0.02562 *
## Number_of_People      -1.571e-02  1.288e-02  5.049e+03  -1.219  0.22274
## GenderFemale           2.443e-01  4.495e-01  5.051e+03   0.543  0.58688
## GenderMale             3.783e-02  1.270e-02  5.049e+03   2.979  0.00291 **
## Temp                  -2.492e-03  8.329e-04  5.026e+03  -2.992  0.00279 **
## Humidity               8.033e-02  2.919e-02  5.055e+03   2.752  0.00594 **
## Facility_TypeU         4.076e-02  1.311e-01  2.628e+00   0.311  0.77894
## log(Pop_Dens)         -4.622e-02  7.223e-02  7.841e+01  -0.640  0.52414
## log(E_UNEMP)           6.399e-01  3.099e-01  1.142e+01   2.065  0.06244 .
## RPL_THEMES            -1.338e+00  4.650e-01  5.205e+00  -2.877  0.03314 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) Dy_TyW tm_tO_ Nmb__P GndrFm GndrMl Temp   Humdty Fcl_TU
## Day_TypWknd -0.023
## tm_typOffw_  0.006 -0.688
## Numbr_f_Ppl -0.005 -0.001  0.003
## GenderFemal -0.029  0.005  0.001 -0.020
## GenderMale  -0.004  0.007 -0.015  0.024  0.013
## Temp         0.012 -0.168  0.087  0.031 -0.009 -0.005
## Humidity    -0.029  0.037 -0.045  0.046  0.006  0.006  0.384
## Faclty_TypU -0.079  0.015  0.008 -0.004  0.001 -0.005  0.000  0.016
## log(Pp_Dns) -0.316  0.028 -0.024 -0.007  0.100 -0.008 -0.041  0.032  0.041
## lg(E_UNEMP) -0.962  0.018 -0.003  0.001  0.003  0.004 -0.017  0.012  0.052
## RPL_THEMES   0.535 -0.007  0.007  0.003  0.027  0.001  0.000  0.000 -0.012
##             l(P_D) l(E_UN
## Day_TypWknd
## tm_typOffw_
## Numbr_f_Ppl
## GenderFemal
## GenderMale
## Temp
## Humidity
## Faclty_TypU
## log(Pp_Dns)
## lg(E_UNEMP)  0.058
## RPL_THEMES   0.228 -0.683
## convergence code: 0
## boundary (singular) fit: see ?isSingular
```

```
anova(Model_Facility_Type,Model_individual,refit=F)
```

```
## Data: Master_reg
## Models:
## Model_individual: Touch_Binary ~ Day_Type + time_type + Number_of_People + Gender +
## Model_individual:     Temp + Humidity + Facility_Type + log(Pop_Dens) + log(E_UNEMP) +
## Model_individual:     RPL_THEMES + (1 | Borough) + (1 | Observer)
## Model_Facility_Type: Touch_Binary ~ Day_Type + time_type + Number_of_People + Gender +
## Model_Facility_Type:     Temp + Humidity + Facility_Type + log(Pop_Dens) + log(E_UNEMP) +
## Model_Facility_Type:     RPL_THEMES + (Facility_Type | Borough) + (1 | Observer)
##                      Df    AIC    BIC  logLik deviance Chisq Chi Df Pr(>Chisq)
```

```
## Model_individual    15 6326.0 6423.9 -3148.0   6296.0
## Model_Facility_Type 17 6325.8 6436.9 -3145.9   6291.8 4.129      2     0.1269
```

```
Model_Pop<- lmer(Touch_Binary~Day_Type+time_type+Number_of_People+Gender+Temp+Humidity+Facility_Type+log
```

```
## boundary (singular) fit: see ?isSingular
```

```
## Warning: Model failed to converge with 1 negative eigenvalue: -9.3e-01
```

```
summary(Model_Pop)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: Touch_Binary ~ Day_Type + time_type + Number_of_People + Gender +
##     Temp + Humidity + Facility_Type + log(Pop_Dens) + log(E_UNEMP) +
##     RPL_THEMES + (log(Pop_Dens) | Borough) + (1 | Observer)
##    Data: Master_reg
##
## REML criterion at convergence: 6291.8
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.0656 -0.9528  0.2795  0.7808  2.0581
##
## Random effects:
##  Groups   Name          Variance Std.Dev. Corr
##  Observer (Intercept)    0.02519 0.1587
##  Borough  (Intercept)   13.99591 3.7411
##           log(Pop_Dens)  0.12427 0.3525   -1.00
##  Residual                0.19839 0.4454
## Number of obs: 5070, groups:  Observer, 16; Borough, 4
##
## Fixed effects:
##                           Estimate Std. Error         df t value Pr(>|t|)
## (Intercept)           -3.720e+00  4.751e+00  9.070e+00  -0.783  0.45364
## Day_TypeWeekend        4.906e-02  2.315e-02  5.047e+03   2.119  0.03412 *
## time_typeOffwork_hour -4.370e-02  1.937e-02  5.057e+03  -2.256  0.02414 *
## Number_of_People      -1.566e-02  1.288e-02  5.051e+03  -1.216  0.22420
## GenderFemale           2.449e-01  4.500e-01  5.054e+03   0.544  0.58633
## GenderMale             3.789e-02  1.270e-02  5.053e+03   2.984  0.00286 **
## Temp                  -2.471e-03  8.329e-04  5.050e+03  -2.967  0.00302 **
## Humidity               8.011e-02  2.919e-02  5.057e+03   2.744  0.00608 **
## Facility_TypeU        -6.902e-02  7.355e-02  4.797e+01  -0.938  0.35272
## log(Pop_Dens)          4.114e-02  1.991e-01  1.821e+02   0.207  0.83653
## log(E_UNEMP)           4.315e-01  3.810e-01  3.959e+00   1.133  0.32129
## RPL_THEMES            -9.437e-01  4.538e-01  1.082e+01  -2.079  0.06217 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) Dy_TyW tm_tO_ Nmb__P GndrFm GndrMl Temp   Humdty Fcl_TU
## Day_TypWknd -0.024
```

```
## tm_typOffw_  0.004 -0.688
## Numbr_f_Ppl -0.009 -0.001  0.002
## GenderFemal -0.008  0.005  0.001 -0.020
## GenderMale  -0.003  0.007 -0.015  0.024  0.013
## Temp        -0.010 -0.168  0.087  0.031 -0.010 -0.006
## Humidity    -0.026  0.037 -0.046  0.046  0.006  0.006  0.384
## Faclty_TypU -0.317  0.040  0.012 -0.005  0.006 -0.005  0.025  0.033
## log(Pp_Dns) -0.471  0.008 -0.006 -0.002  0.013  0.000 -0.006  0.009  0.028
## lg(E_UNEMP) -0.885  0.024 -0.004  0.006  0.002  0.002  0.001  0.017  0.325
## RPL_THEMES   0.155 -0.003  0.012  0.008  0.000  0.004  0.013  0.003  0.017
##             l(P_D) l(E_UN
## Day_TypWknd
## tm_typOffw_
## Numbr_f_Ppl
## GenderFemal
## GenderMale
## Temp
## Humidity
## Faclty_TypU
## log(Pp_Dns)
## lg(E_UNEMP)  0.015
## RPL_THEMES   0.123 -0.325
## convergence code: 0
## boundary (singular) fit: see ?isSingular
```

```
anova(Model_Pop,Model_individual,refit=F)
```

```
## Data: Master_reg
## Models:
## Model_individual: Touch_Binary ~ Day_Type + time_type + Number_of_People + Gender +
## Model_individual:     Temp + Humidity + Facility_Type + log(Pop_Dens) + log(E_UNEMP) +
## Model_individual:     RPL_THEMES + (1 | Borough) + (1 | Observer)
## Model_Pop: Touch_Binary ~ Day_Type + time_type + Number_of_People + Gender +
## Model_Pop:     Temp + Humidity + Facility_Type + log(Pop_Dens) + log(E_UNEMP) +
## Model_Pop:     RPL_THEMES + (log(Pop_Dens) | Borough) + (1 | Observer)
##                   Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## Model_individual 15 6326.0 6423.9 -3148.0   6296.0
## Model_Pop        17 6325.8 6436.9 -3145.9   6291.8 4.1286      2     0.1269
```

We have also added the individual level predictor at the observer level to investigate its random effect by the observer.

```
#add individual record level predictor at Observer level
Model_Day_type<- lmer(Touch_Binary~Day_Type+time_type+Number_of_People+Gender+Temp+Humidity+Facility_Ty
```

```
## boundary (singular) fit: see ?isSingular
```

```
summary(Model_Day_type)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: Touch_Binary ~ Day_Type + time_type + Number_of_People + Gender +
```

```
##     Temp + Humidity + Facility_Type + log(Pop_Dens) + log(E_UNEMP) +
##     RPL_THEMES + (1 | Borough) + (Day_Type | Observer)
##    Data: Master_reg
##
## REML criterion at convergence: 6293.6
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.0627 -0.9489  0.2810  0.7920  2.0521
##
## Random effects:
##  Groups   Name            Variance Std.Dev. Corr
##  Observer (Intercept)     0.03186  0.1785
##           Day_TypeWeekend 0.00184  0.0429   -0.55
##  Borough  (Intercept)     0.00000  0.0000
##  Residual                 0.19838  0.4454
## Number of obs: 5070, groups:  Observer, 16; Borough, 4
##
## Fixed effects:
##                          Estimate Std. Error         df t value Pr(>|t|)
## (Intercept)           -5.430e+00  3.370e+00  1.534e+01  -1.611  0.12753
## Day_TypeWeekend        4.523e-02  2.670e-02  2.512e+01   1.694  0.10261
## time_typeOffwork_hour -3.990e-02  1.946e-02  4.852e+03  -2.051  0.04036 *
## Number_of_People      -1.564e-02  1.289e-02  5.050e+03  -1.214  0.22480
## GenderFemale           2.856e-01  4.493e-01  5.048e+03   0.636  0.52502
## GenderMale             3.821e-02  1.270e-02  5.047e+03   3.009  0.00264 **
## Temp                  -2.432e-03  8.372e-04  4.012e+03  -2.905  0.00370 **
## Humidity               8.165e-02  2.924e-02  4.999e+03   2.792  0.00525 **
## Facility_TypeU        -6.650e-03  6.370e-02  5.514e+01  -0.104  0.91724
## log(Pop_Dens)          2.599e-02  6.645e-02  7.519e+01   0.391  0.69682
## log(E_UNEMP)           6.038e-01  3.058e-01  1.312e+01   1.974  0.06977 .
## RPL_THEMES            -1.094e+00  4.479e-01  1.250e+01  -2.443  0.03027 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) Dy_TyW tm_tO_ Nmb__P GndrFm GndrMl Temp   Humdty Fcl_TU
## Day_TypWknd -0.011
## tm_typOffw_  0.003 -0.601
## Numbr_f_Ppl -0.004 -0.004  0.002
## GenderFemal -0.031  0.004  0.001 -0.020
## GenderMale  -0.006  0.004 -0.014  0.024  0.013
## Temp         0.017 -0.146  0.088  0.031 -0.009 -0.006
## Humidity    -0.029  0.027 -0.043  0.046  0.006  0.006  0.386
## Faclty_TypU -0.126  0.063  0.003 -0.008  0.003 -0.007  0.019  0.039
## log(Pp_Dns) -0.359  0.008 -0.022 -0.004  0.093 -0.005 -0.041  0.039  0.061
## lg(E_UNEMP) -0.969  0.008  0.000 -0.001  0.009  0.004 -0.024  0.011  0.111
## RPL_THEMES   0.569 -0.003  0.002  0.008  0.014  0.004  0.007  0.006 -0.082
##             l(P_D) l(E_UN
## Day_TypWknd
## tm_typOffw_
## Numbr_f_Ppl
## GenderFemal
## GenderMale
```

```
## Temp
## Humidity
## Faclty_TypU
## log(Pp_Dns)
## lg(E_UNEMP)  0.131
## RPL_THEMES   0.114 -0.691
## convergence code: 0
## boundary (singular) fit: see ?isSingular
```

```
anova(Model_individual,Model_Day_type,refit=F)
```

```
## Data: Master_reg
## Models:
## Model_individual: Touch_Binary ~ Day_Type + time_type + Number_of_People + Gender +
## Model_individual:     Temp + Humidity + Facility_Type + log(Pop_Dens) + log(E_UNEMP) +
## Model_individual:     RPL_THEMES + (1 | Borough) + (1 | Observer)
## Model_Day_type: Touch_Binary ~ Day_Type + time_type + Number_of_People + Gender +
## Model_Day_type:     Temp + Humidity + Facility_Type + log(Pop_Dens) + log(E_UNEMP) +
## Model_Day_type:     RPL_THEMES + (1 | Borough) + (Day_Type | Observer)
##                   Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## Model_individual 15 6326.0 6423.9 -3148.0   6296.0
## Model_Day_type   17 6327.6 6438.6 -3146.8   6293.6 2.3753      2     0.3049
```

```
Model_time_type<-lmer(Touch_Binary~Day_Type+time_type+Number_of_People+Gender+Temp+Humidity+Facility_Ty
```

```
## boundary (singular) fit: see ?isSingular
```

```
summary(Model_time_type)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: Touch_Binary ~ Day_Type + time_type + Number_of_People + Gender +
##     Temp + Humidity + Facility_Type + log(Pop_Dens) + log(E_UNEMP) +
##     RPL_THEMES + (1 | Borough) + (time_type | Observer)
##    Data: Master_reg
##
## REML criterion at convergence: 6293.3
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.0531 -0.9434  0.2777  0.7886  2.0509
##
## Random effects:
##  Groups   Name                 Variance Std.Dev. Corr
##  Observer (Intercept)          0.026353 0.16234
##           time_typeOffwork_hour 0.001207 0.03475  0.41
##  Borough  (Intercept)          0.000000 0.00000
##  Residual                      0.198409 0.44543
## Number of obs: 5070, groups:  Observer, 16; Borough, 4
##
## Fixed effects:
##                         Estimate Std. Error         df t value Pr(>|t|)
```

```
## (Intercept)           -4.631e+00  3.257e+00  1.375e+01  -1.422  0.17737
## Day_TypeWeekend        4.364e-02  2.379e-02  6.038e+02   1.835  0.06707 .
## time_typeOffwork_hour -4.065e-02  2.198e-02  1.926e+01  -1.849  0.07981 .
## Number_of_People      -1.556e-02  1.289e-02  5.046e+03  -1.207  0.22752
## GenderFemale           2.885e-01  4.492e-01  5.044e+03   0.642  0.52082
## GenderMale             3.752e-02  1.270e-02  5.047e+03   2.953  0.00316 **
## Temp                  -2.537e-03  8.368e-04  3.649e+03  -3.032  0.00245 **
## Humidity               7.793e-02  2.924e-02  4.943e+03   2.665  0.00771 **
## Facility_TypeU         2.428e-02  6.160e-02  4.562e+01   0.394  0.69531
## log(Pop_Dens)          2.638e-02  6.572e-02  6.618e+01   0.401  0.68946
## log(E_UNEMP)           5.319e-01  2.946e-01  1.185e+01   1.805  0.09649 .
## RPL_THEMES            -1.099e+00  4.305e-01  1.150e+01  -2.552  0.02611 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) Dy_TyW tm_tO_ Nmb__P GndrFm GndrMl Temp   Humdty Fcl_TU
## Day_TypWknd -0.018
## tm_typOffw_ -0.013 -0.612
## Numbr_f_Ppl  0.002 -0.004  0.001
## GenderFemal -0.029  0.005  0.003 -0.021
## GenderMale  -0.001  0.010 -0.014  0.024  0.013
## Temp         0.009 -0.166  0.081  0.032 -0.009 -0.004
## Humidity    -0.027  0.040 -0.043  0.046  0.006  0.006  0.385
## Faclty_TypU -0.136  0.072 -0.022 -0.007 -0.001 -0.003  0.016  0.037
## log(Pp_Dns) -0.360  0.030 -0.004 -0.010  0.093 -0.009 -0.045  0.036  0.048
## lg(E_UNEMP) -0.967  0.012  0.013 -0.006  0.006  0.000 -0.014  0.010  0.124
## RPL_THEMES   0.552  0.005  0.006  0.006  0.019  0.005 -0.002 -0.006 -0.085
##             l(P_D) l(E_UN
## Day_TypWknd
## tm_typOffw_
## Numbr_f_Ppl
## GenderFemal
## GenderMale
## Temp
## Humidity
## Faclty_TypU
## log(Pp_Dns)
## lg(E_UNEMP)  0.127
## RPL_THEMES   0.129 -0.680
## convergence code: 0
## boundary (singular) fit: see ?isSingular
```

```
anova(Model_individual,Model_time_type,refit=F)
```

```
## Data: Master_reg
## Models:
## Model_individual: Touch_Binary ~ Day_Type + time_type + Number_of_People + Gender +
## Model_individual:     Temp + Humidity + Facility_Type + log(Pop_Dens) + log(E_UNEMP) +
## Model_individual:     RPL_THEMES + (1 | Borough) + (1 | Observer)
## Model_time_type: Touch_Binary ~ Day_Type + time_type + Number_of_People + Gender +
## Model_time_type:     Temp + Humidity + Facility_Type + log(Pop_Dens) + log(E_UNEMP) +
## Model_time_type:     RPL_THEMES + (1 | Borough) + (time_type | Observer)
##                  Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
```

```
## Model_individual 15 6326.0 6423.9 -3148.0   6296.0
## Model_time_type  17 6327.3 6438.4 -3146.7   6293.3 2.6451      2     0.2665
```

```
Model_Number_of_People<-lmer(Touch_Binary~Day_Type+time_type+Number_of_People+Gender+Temp+Humidity+Facil
```

```
## boundary (singular) fit: see ?isSingular
```

```
summary(Model_Number_of_People)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: Touch_Binary ~ Day_Type + time_type + Number_of_People + Gender +
##     Temp + Humidity + Facility_Type + log(Pop_Dens) + log(E_UNEMP) +
##     RPL_THEMES + (1 | Borough) + (Number_of_People | Observer)
##    Data: Master_reg
##
## REML criterion at convergence: 6275.8
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.1096 -0.9425  0.2568  0.7853  2.0617
##
## Random effects:
##  Groups   Name             Variance Std.Dev. Corr
##  Observer (Intercept)      0.050434 0.22458
##           Number_of_People 0.003234 0.05687  -0.93
##  Borough  (Intercept)      0.000000 0.00000
##  Residual                  0.197677 0.44461
## Number of obs: 5070, groups:  Observer, 16; Borough, 4
##
## Fixed effects:
##                          Estimate Std. Error         df t value Pr(>|t|)
## (Intercept)           -5.338e+00  2.580e+00  1.071e+01  -2.069  0.06354 .
## Day_TypeWeekend        4.867e-02  2.310e-02  5.025e+03   2.107  0.03517 *
## time_typeOffwork_hour -4.325e-02  1.934e-02  5.052e+03  -2.236  0.02540 *
## Number_of_People      -1.728e-02  1.995e-02  1.395e+01  -0.866  0.40107
## GenderFemale           3.328e-01  4.481e-01  4.925e+03   0.743  0.45766
## GenderMale             3.533e-02  1.269e-02  5.052e+03   2.784  0.00539 **
## Temp                  -2.535e-03  8.306e-04  5.024e+03  -3.052  0.00229 **
## Humidity               7.646e-02  2.916e-02  5.054e+03   2.622  0.00876 **
## Facility_TypeU         3.914e-02  5.498e-02  3.537e+01   0.712  0.48121
## log(Pop_Dens)          7.003e-02  5.827e-02  3.567e+01   1.202  0.23731
## log(E_UNEMP)           5.318e-01  2.334e-01  1.034e+01   2.279  0.04506 *
## RPL_THEMES            -8.191e-01  3.476e-01  1.208e+01  -2.356  0.03617 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) Dy_TyW tm_tO_ Nmb__P GndrFm GndrMl Temp   Humdty Fcl_TU
## Day_TypWknd -0.013
## tm_typOffw_ -0.004 -0.689
## Numbr_f_Ppl  0.000 -0.002 -0.005
## GenderFemal -0.016  0.004  0.003 -0.018
```

```
## GenderMale  -0.002  0.008 -0.015  0.013  0.012
## Temp         0.008 -0.169  0.087  0.018 -0.008 -0.005
## Humidity    -0.017  0.037 -0.046  0.025  0.006  0.006  0.385
## Faclty_TypU -0.174  0.035  0.013  0.080 -0.012  0.005  0.003  0.021
## log(Pp_Dns) -0.349  0.033 -0.014 -0.009  0.085 -0.013 -0.035  0.037  0.087
## lg(E_UNEMP) -0.959  0.006  0.004 -0.015 -0.009  0.002 -0.018 -0.005  0.146
## RPL_THEMES   0.532  0.014  0.005 -0.021  0.036  0.006 -0.002  0.021 -0.048
##             l(P_D) l(E_UN
## Day_TypWknd
## tm_typOffw_
## Numbr_f_Ppl
## GenderFemal
## GenderMale
## Temp
## Humidity
## Faclty_TypU
## log(Pp_Dns)
## lg(E_UNEMP)  0.083
## RPL_THEMES   0.184 -0.678
## convergence code: 0
## boundary (singular) fit: see ?isSingular
```

```
anova(Model_individual,Model_Number_of_People,refit=F)
```

```
## Data: Master_reg
## Models:
## Model_individual: Touch_Binary ~ Day_Type + time_type + Number_of_People + Gender +
## Model_individual:     Temp + Humidity + Facility_Type + log(Pop_Dens) + log(E_UNEMP) +
## Model_individual:     RPL_THEMES + (1 | Borough) + (1 | Observer)
## Model_Number_of_People: Touch_Binary ~ Day_Type + time_type + Number_of_People + Gender +
## Model_Number_of_People:     Temp + Humidity + Facility_Type + log(Pop_Dens) + log(E_UNEMP) +
## Model_Number_of_People:     RPL_THEMES + (1 | Borough) + (Number_of_People | Observer)
##                         Df    AIC    BIC  logLik deviance  Chisq Chi Df
## Model_individual        15 6326.0 6423.9 -3148.0   6296.0
## Model_Number_of_People 17 6309.8 6420.8 -3137.9   6275.8 20.153      2
##                         Pr(>Chisq)
## Model_individual
## Model_Number_of_People  4.206e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Model_Gender<- lmer(Touch_Binary~Day_Type+time_type+Number_of_People+Gender+Temp+Humidity+Facility_Type
```

```
## boundary (singular) fit: see ?isSingular
```

```
## Warning: Model failed to converge with 2 negative eigenvalues: -2.0e-07 -7.9e-02
```

```
summary(Model_Gender)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
```

```
## Formula: Touch_Binary ~ Day_Type + time_type + Number_of_People + Gender +
##     Temp + Humidity + Facility_Type + log(Pop_Dens) + log(E_UNEMP) +
##     RPL_THEMES + (1 | Borough) + (Gender | Observer)
##    Data: Master_reg
##
## REML criterion at convergence: 6285
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.0596 -0.9416  0.2761  0.7724  2.0513
##
## Random effects:
##  Groups   Name          Variance Std.Dev. Corr
##  Observer (Intercept)   0.032045 0.17901
##           GenderFemale  0.239810 0.48970   0.05
##           GenderMale    0.004698 0.06854  -0.36  0.74
##  Borough  (Intercept)   0.000000 0.00000
##  Residual               0.197663 0.44459
## Number of obs: 5070, groups:  Observer, 16; Borough, 4
##
## Fixed effects:
##                           Estimate Std. Error         df t value Pr(>|t|)
## (Intercept)           -4.702e+00  3.352e+00  1.406e+01  -1.403  0.18238
## Day_TypeWeekend        4.946e-02  2.312e-02  5.038e+03   2.139  0.03245 *
## time_typeOffwork_hour -4.465e-02  1.935e-02  5.049e+03  -2.307  0.02109 *
## Number_of_People      -1.560e-02  1.289e-02  5.049e+03  -1.211  0.22602
## GenderFemale          -4.655e-02  6.028e-01  1.744e-01  -0.077  0.97130
## GenderMale             2.987e-02  2.241e-02  1.266e+01   1.333  0.20610
## Temp                  -2.379e-03  8.316e-04  5.048e+03  -2.861  0.00424 **
## Humidity               8.037e-02  2.916e-02  5.050e+03   2.756  0.00587 **
## Facility_TypeU         2.335e-02  6.254e-02  5.273e+01   0.373  0.71044
## log(Pop_Dens)          1.457e-02  6.651e-02  7.134e+01   0.219  0.82720
## log(E_UNEMP)           5.446e-01  3.044e-01  1.215e+01   1.789  0.09849 .
## RPL_THEMES            -1.040e+00  4.458e-01  1.178e+01  -2.332  0.03830 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) Dy_TyW tm_tO_ Nmb__P GndrFm GndrMl Temp   Humdty Fcl_TU
## Day_TypWknd -0.026
## tm_typOffw_  0.008 -0.688
## Numbr_f_Ppl -0.003 -0.001  0.002
## GenderFemal -0.067  0.009 -0.004 -0.021
## GenderMale  -0.010  0.003 -0.008  0.016  0.092
## Temp         0.012 -0.169  0.088  0.030 -0.004 -0.006
## Humidity    -0.028  0.037 -0.045  0.046  0.004  0.002  0.384
## Faclty_TypU -0.114  0.045  0.014 -0.006  0.048  0.010  0.014  0.036
## log(Pp_Dns) -0.351  0.034 -0.021 -0.007  0.085 -0.011 -0.046  0.037  0.042
## lg(E_UNEMP) -0.968  0.020 -0.007 -0.001  0.053  0.008 -0.016  0.011  0.103
## RPL_THEMES   0.558 -0.003  0.009  0.005 -0.042 -0.009  0.000  0.002 -0.076
##             l(P_D) l(E_UN
## Day_TypWknd
## tm_typOffw_
## Numbr_f_Ppl
```

```
## GenderFemal
## GenderMale
## Temp
## Humidity
## Faclty_TypU
## log(Pp_Dns)
## lg(E_UNEMP)  0.120
## RPL_THEMES   0.133 -0.684
## convergence code: 0
## boundary (singular) fit: see ?isSingular
```

```
anova(Model_individual, Model_Gender,refit=F)
```

```
## Data: Master_reg
## Models:
## Model_individual: Touch_Binary ~ Day_Type + time_type + Number_of_People + Gender +
## Model_individual:     Temp + Humidity + Facility_Type + log(Pop_Dens) + log(E_UNEMP) +
## Model_individual:     RPL_THEMES + (1 | Borough) + (1 | Observer)
## Model_Gender: Touch_Binary ~ Day_Type + time_type + Number_of_People + Gender +
## Model_Gender:     Temp + Humidity + Facility_Type + log(Pop_Dens) + log(E_UNEMP) +
## Model_Gender:     RPL_THEMES + (1 | Borough) + (Gender | Observer)
##                  Df  AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## Model_individual 15 6326 6423.9 -3148.0     6296
## Model_Gender     20 6325 6455.6 -3142.5     6285 10.992      5    0.05154 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Model_Temp<- lmer(Touch_Binary~Day_Type+time_type+Number_of_People+Gender+Temp+Humidity+Facility_Type+l
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
## Model failed to converge with max|grad| = 5.09882 (tol = 0.002, component 1)
```

```
## Warning: Model failed to converge with 1 negative eigenvalue: -1.9e-01
```

```
summary(Model_Temp)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: Touch_Binary ~ Day_Type + time_type + Number_of_People + Gender +
##     Temp + Humidity + Facility_Type + log(Pop_Dens) + log(E_UNEMP) +
##     RPL_THEMES + (1 | Borough) + (Temp | Observer)
##    Data: Master_reg
##
## REML criterion at convergence: 6296.7
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.1102 -0.9526  0.2806  0.7786  2.0562
##
## Random effects:
##  Groups   Name        Variance  Std.Dev. Corr
```

```
##  Observer (Intercept) 3.632e-02 0.190586
##           Temp        6.585e-06 0.002566 -0.49
##  Borough  (Intercept) 1.997e+00 1.412997
##  Residual             1.982e-01 0.445172
## Number of obs: 5070, groups:  Observer, 16; Borough, 4
##
## Fixed effects:
##                          Estimate Std. Error         df t value Pr(>|t|)
## (Intercept)           -5.954e+00  5.188e+01  4.993e+03  -0.115  0.90865
## Day_TypeWeekend        5.265e-02  2.322e-02  4.996e+03   2.267  0.02343 *
## time_typeOffwork_hour -4.604e-02  1.942e-02  5.029e+03  -2.371  0.01779 *
## Number_of_People      -1.617e-02  1.288e-02  5.047e+03  -1.256  0.20935
## GenderFemale           2.812e-01  4.496e-01  5.043e+03   0.625  0.53168
## GenderMale             3.769e-02  1.270e-02  5.046e+03   2.968  0.00301 **
## Temp                  -2.323e-03  1.112e-03  1.945e+01  -2.090  0.05001 .
## Humidity               7.845e-02  2.933e-02  4.868e+03   2.675  0.00750 **
## Facility_TypeU         1.107e-03  6.475e-02  6.119e+01   0.017  0.98641
## log(Pop_Dens)          2.356e-02  7.410e-02  1.203e+02   0.318  0.75113
## log(E_UNEMP)           6.622e-01  4.928e+00  4.995e+03   0.134  0.89311
## RPL_THEMES            -1.217e+00  6.434e+00  5.022e+03  -0.189  0.84997
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) Dy_TyW tm_tO_ Nmb__P GndrFm GndrMl Temp   Humdty Fcl_TU
## Day_TypWknd -0.002
## tm_typOffw_  0.001 -0.684
## Numbr_f_Ppl  0.000 -0.003  0.002
## GenderFemal -0.003  0.004  0.001 -0.021
## GenderMale   0.000  0.007 -0.014  0.024  0.013
## Temp         0.001 -0.123  0.067  0.028 -0.011 -0.008
## Humidity    -0.002  0.037 -0.046  0.047  0.007  0.007  0.285
## Faclty_TypU -0.004  0.042  0.018 -0.007 -0.005 -0.001  0.015  0.036
## log(Pp_Dns) -0.029  0.026 -0.021 -0.005  0.105 -0.005 -0.052  0.041 -0.036
## lg(E_UNEMP) -0.997  0.001 -0.001  0.000  0.001  0.000 -0.001  0.001  0.004
## RPL_THEMES   0.573 -0.001  0.001  0.000  0.001  0.000  0.000  0.000 -0.005
##             l(P_D) l(E_UN
## Day_TypWknd
## tm_typOffw_
## Numbr_f_Ppl
## GenderFemal
## GenderMale
## Temp
## Humidity
## Faclty_TypU
## log(Pp_Dns)
## lg(E_UNEMP)  0.012
## RPL_THEMES   0.008 -0.636
## convergence code: 0
## Model failed to converge with max|grad| = 5.09882 (tol = 0.002, component 1)
```

```
anova(Model_individual,Model_Temp)
```

```
## refitting model(s) with ML (instead of REML)
```

```
## Data: Master_reg
## Models:
## Model_individual: Touch_Binary ~ Day_Type + time_type + Number_of_People + Gender +
## Model_individual:     Temp + Humidity + Facility_Type + log(Pop_Dens) + log(E_UNEMP) +
## Model_individual:     RPL_THEMES + (1 | Borough) + (1 | Observer)
## Model_Temp: Touch_Binary ~ Day_Type + time_type + Number_of_People + Gender +
## Model_Temp:     Temp + Humidity + Facility_Type + log(Pop_Dens) + log(E_UNEMP) +
## Model_Temp:     RPL_THEMES + (1 | Borough) + (Temp | Observer)
##                   Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## Model_individual 15 6269.0 6366.9 -3119.5   6239.0
## Model_Temp       17 6269.3 6380.3 -3117.6   6235.3 3.7094      2     0.1565
```

After including the number of people being observed varying at the observer level, we account for the random effect of the touch_binary outcome that varies by different observers. This is a more complex model that is justified by conducting the LRT test with the simple model(P=0.00004).

6 Limitations and Drawbacks

Overall, we fitted multiple regression and multi-level models on our outcome variable Touch_Binary. Instead of dealing with correlated structures is to treat clustering as a nuisance, multi-level modeling treats hierarchical structures as a feature of the population that is of interest. However, there still exist some limitations on the analysis and model fitting. First of all, although the multi-level model accounts for random effect among different observers, it did not identify the specific effect and impact among groups. Also, from the regression model, we can learn that the variance is still considerably large. Therefore, more predictors might need to be included in the regression model.

7 Conclusion

In general, we discovered the pattern of distribution of our outcome variable Touch_Binary(Whether people touch any object or not), PPE_Use(Whether subject wears Personal Protective Equipment or not) among different facilities, and observers. We have also created a 3D bubble graph on PPE_Use and Touch_Binary with the percentage of male and record size as they, z-axis, and colored by borough. Finally, we fitted the evaluate different models. We find out that the logistic regression(binomial) is a better-fitted model for the touch_binary outcome variable. It shows that touch_binary is correlated with Day_Type, time_type, gender, borough, facility type, an interaction term between temperature and humidity, and the logarithm of population density. For the multi-level model, it seems to perform better when allowing the number of subject varying at the observer's(data collectors) level.