Topos Internship Assignment-Meghna Das

Meghna Das

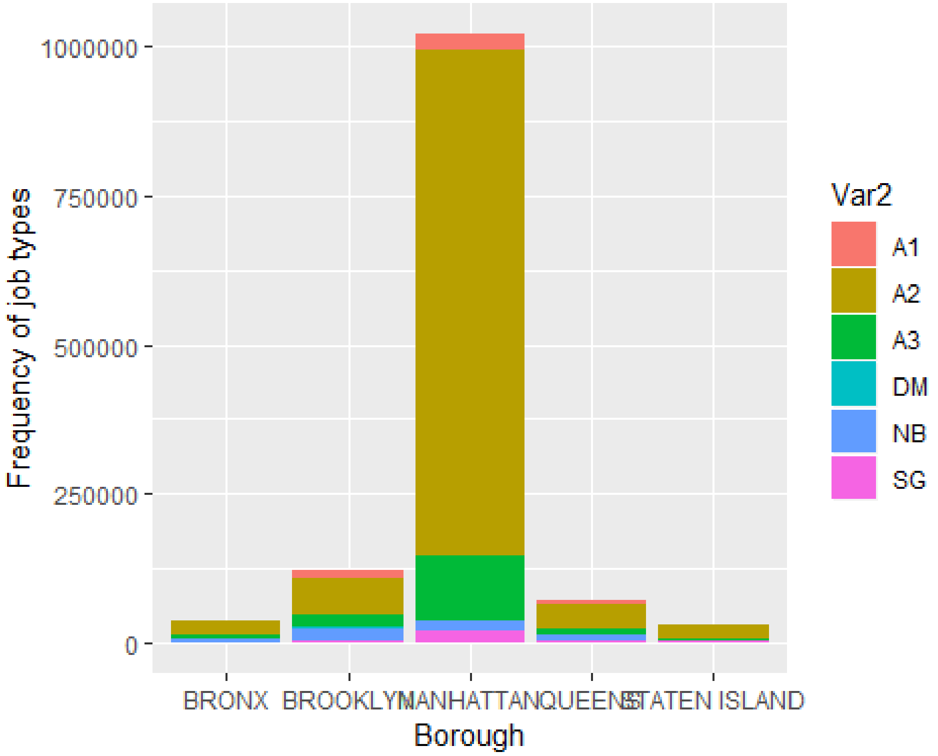
April 3, 2019

```
#install.packages("ggplot2")
#install.packages("dplyr")
#install.packages("plyr")

setwd("C:\\Users\\Meghna Das\\Desktop\\DOB")
DoB<-read.csv("C:\\Users\\Meghna Das\\Desktop\\DOB\\DOB_Permit_Issuance (1).c
sv")
DoBApproved<-read.csv("C:\\Users\\Meghna Das\\Desktop\\DOB\\DOB_Approved_Perm
its.csv")

#Change column names in DoB to make the join easier
names(DoB)[2]<-paste("Bin")
names(DoB)[3]<-paste("House.No")

#merge
DoBTotal<-merge(DoB, DoBApproved, by=c("House.No", "Bin"), no.dups= TRUE)

## Warning in `[<-.factor`(`*tmp*`, ri, value = c(4617966L, 5067707L,
## 1001362L, : invalid factor level, NA generated

na.omit(DoBTotal)

#relation between borough and job type(Frequency of each job type in a boroug
h)
(count1<-data.frame(table(DoBTotal$ï..BOROUGH,DoBTotal$Job.Type)))

##              Var1 Var2   Freq
## 1           BRONX   A1   1839
## 2        BROOKLYN   A1  13737
## 3       MANHATTAN   A1  28189
## 4          QUEENS   A1   6125
## 5   STATEN ISLAND   A1    889
## 6           BRONX   A2  22323
## 7        BROOKLYN   A2  63343
## 8       MANHATTAN   A2 846984
## 9          QUEENS   A2  41102
## 10  STATEN ISLAND   A2  24057
## 11          BRONX   A3   7871
## 12       BROOKLYN   A3  19345
## 13      MANHATTAN   A3 108232
## 14         QUEENS   A3  10608
## 15  STATEN ISLAND   A3   1577
## 16          BRONX   DM    768
## 17       BROOKLYN   DM   3335
## 18      MANHATTAN   DM   2407
## 19         QUEENS   DM   1728
## 20  STATEN ISLAND   DM    553
## 21          BRONX   NB   4022
## 22       BROOKLYN   NB  20116
## 23      MANHATTAN   NB  16045
## 24         QUEENS   NB   9397
## 25  STATEN ISLAND   NB   2041
## 26          BRONX   SG    891
## 27       BROOKLYN   SG   3238
## 28      MANHATTAN   SG  20156
## 29         QUEENS   SG   2533
## 30  STATEN ISLAND   SG   1642
```

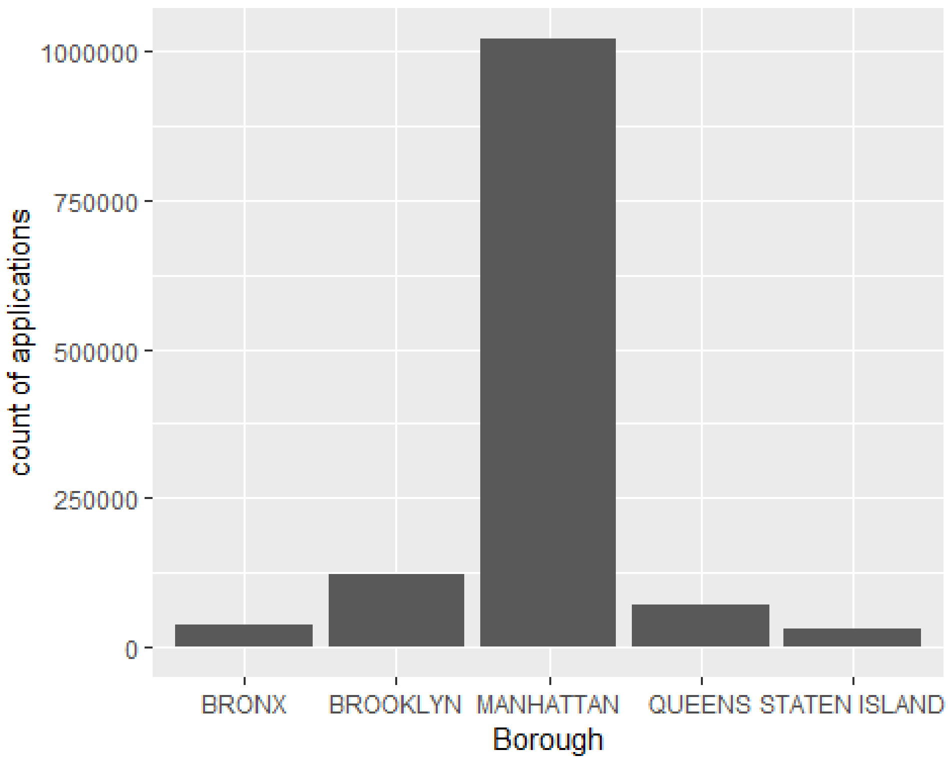
```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.5.3
```

```
ggplot(data=count1 , aes(x=Var1, y=Freq, fill=Var2)) + geom_bar(stat="identit
y")+xlab("Borough")+ylab("Frequency of job types")
```



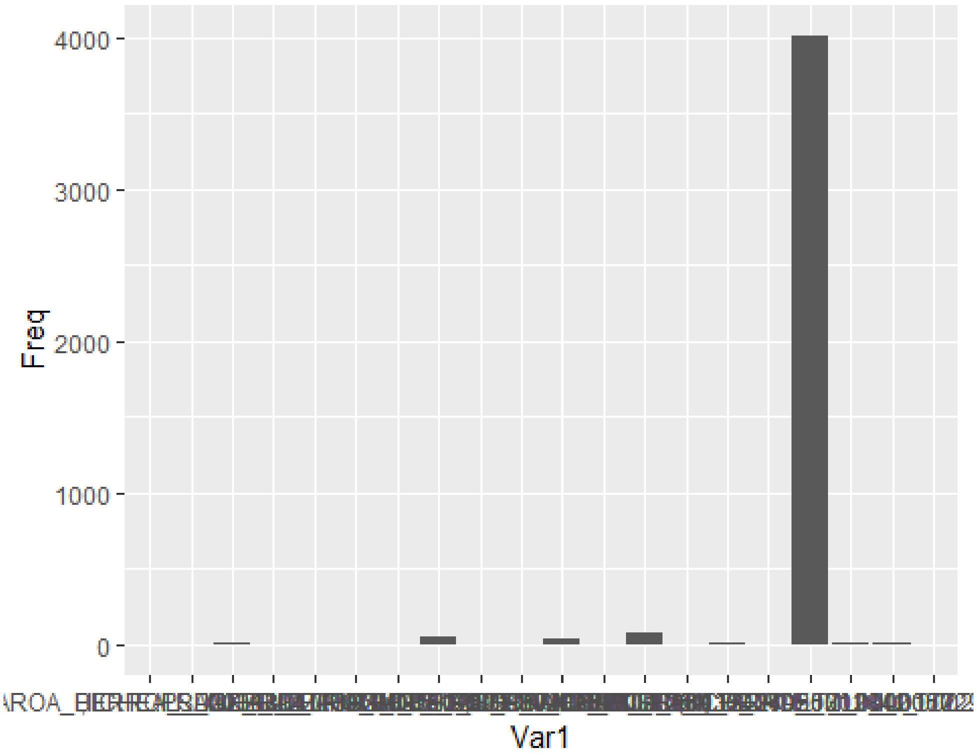
```
#count of applications for each borough
library(ggplot2)
ggplot(DoBTotal, aes(x=ï..BOROUGH)) + geom_bar() + xlab("Borough") +ylab("cou
nt of applications")
```



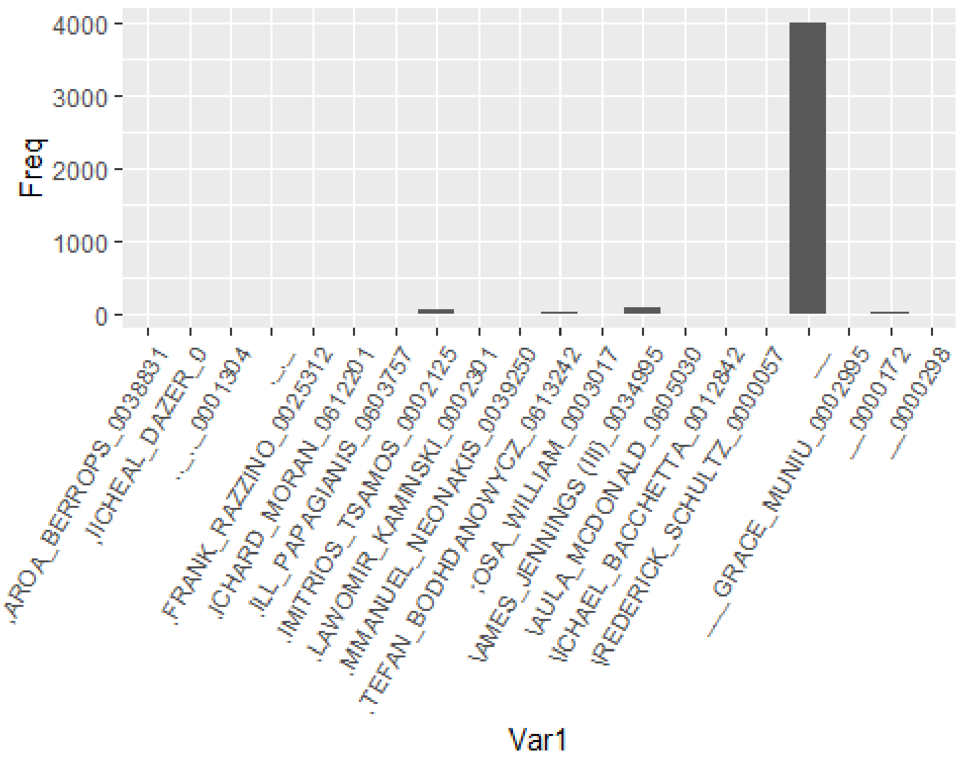
```
#combining permittee first name and last name

library(dplyr)

## Warning: package 'dplyr' was built under R version 3.5.3

##
## 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

DoBTotal<- mutate(DoBTotal,
             Permittee = paste(Permittee.s.First.Name, Permittee.s.Last.Name,
sep = '_'))

#However, we need to include the license number as well in order to identify
the unique permittees
DoBTotal<- mutate(DoBTotal,
            Permittee = paste(Permittee.s.First.Name, Permittee.s.Last.Name,
Permittee.s.License..,sep = '_'))
```

```
#Frequency of each Permittee
(count2<-data.frame(table(DoBTotal$Permittee)))

#order in descending order

count2[order(-count2$Freq),]

##                                                Var1  Freq
## 1                            ,AROA_BERROPS_0038831     4
## 2                                 ,IICHEAL_DAZER_0     3
## 3                                    .._.._0001304     6
## 4                                             ._._     1
## 5                           .FRANK_RAZZINO_0025312     2
## 6                            .ICHARD_MORAN_0612201     3
## 7                          .ILL_PAPAGIANIS_0603757     4
## 8                         .IMITRIOS_TSAMOS_0002125    48
## 9                        .LAWOMIR_KAMINSKI_0002301     2
## 10                       .MMANUEL_NEONAKIS_0039250     5
## 11                     .TEFAN_BODHDANOWYCZ_0613242    40
## 12                            ;OSA_WILLIAM_0003017     4
## 13                   \\AMES_JENNINGS (III)_0034995    81
## 14                         \\AULA_MCDONALD_0605030     1
## 15                      \\ICHAEL_BACCHETTA_0012842     6
## 16                      \\REDERICK_SCHULTZ_0000057     1
## 17                                              __  4011
## 18                          ___GRACE_MUNIU_0002995     8
## 19                                       __0000172    14
## 20                                       __0000298     3
## 22107                                JOEY_MARRONE_     2
## 22108                         JOEY_PROCIDA_0001863     2
## 22109                     JOFSEPH_CONRETTA_0009086     3
## 22110                         JOGN_KRISTIS_0000502     2
## 22111                           JOGN_WHITE_0002660     3
## 22112                                  JOH_CURLEY_     1
## 22113                            JOH_KELLY_0001529     4
## 22114                            JOH_WHITE_0002660     1
## 22115                    JOHAH_FINKELSTEIN_0036699     1
## 22116                                   JOHAN_LAM_     4
## 22117                               JOHAN_PALONE_0     2
## 22118                               JOHAN_SOLTANI_     2
## 22119                             JOHANN_APKARIAN_     4
## 22120                      JOHANN_APKARIAN_0032781     6
## 22121                      JOHANNA_ESCOBAR_0617316   145
## 22122                     JOHANNA_POLLEMAN_0003715     4
## 22123                         JOHANNA_VEGA_0617819    19
## 22124                      JOHANNES_SANZIN_0019776     4
## 22125                     JOHANTHAN_DISICK_0023324     2
## 22126                      JOHBN_SALLUSTIO_0001791     3
## 22127                     JOHGN_CIARMAGLIA_0008569     6
## 22128                       JOHGN_MCNAMARA_0009666     4
## 22129                              JOHH_BOCCIERI_0001649     1
## 22130                                JOHH_KRITIS_0000502     2
## 22131                               JOHHNY_NANHU_0001504     2
## 22132                                JOHMN_WHITE_0002660     6
```

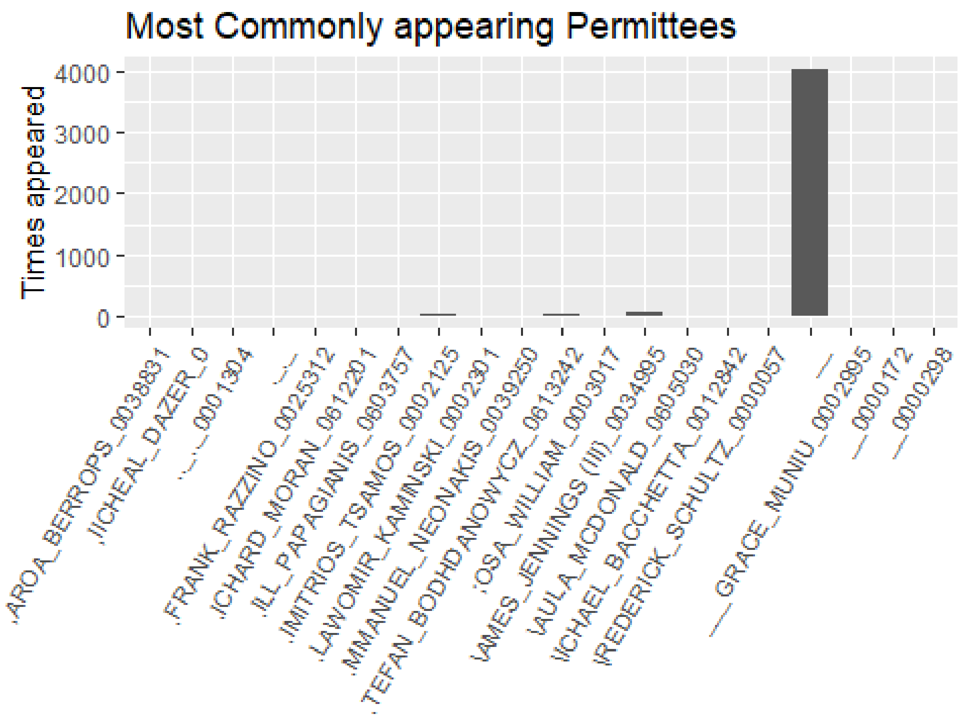
```
#Let's create a bar chart to capture the same information for the 20 most com
monly appearing Permmittees.

barchart.fig = ggplot(data = count2[1:20, ], mapping = aes(x = Var1, y = Freq
))
barchart.fig + geom_bar(stat = "identity")
```



```
#There's an issue with this plot: the labels along the x-axis have all blende
d together and are incomprehensible.To adjust the text, we can use the theme
command.
barchart.fig + geom_bar(stat = "identity") +
               theme(axis.text.x = element_text(angle = 60, vjust = 1, hjust
= 1))
```



```
#Number of job filings per permittee
#We should probably also change the axis labels and title to something more m
eaningful

barchart.fig + geom_bar(stat = "identity") +
               theme(axis.text.x = element_text(angle = 60, vjust = 1, hjust
= 1)) +
               labs(x = "", y = "Times appeared", title = "Most Commonly appe
aring Permittees")
```



```
#Testing difference in means of estimated job costs broken down by borough
#How can we assess whether this difference is statistically significant?
#Let's compute a summary table
library(plyr)

## Warning: package 'plyr' was built under R version 3.5.3

## ------------------------------------------------------------------------

## You have loaded plyr after dplyr - this is likely to cause problems.
## If you need functions from both plyr and dplyr, please load plyr first, th
en dplyr:
## library(plyr); library(dplyr)

## ------------------------------------------------------------------------

##
## Attaching package: 'plyr'

## The following objects are masked from 'package:dplyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize

ddply(DoBTotal, ~ ï..BOROUGH, summarize,
      mean.estimated.job.costs = mean(Estimated.Job.Costs),
      sd.estimated.job.costs = sd(Estimated.Job.Costs)
      )
```

```
##      ï..BOROUGH mean.estimated.job.costs sd.estimated.job.costs
## 1         BRONX                15673.268               34195.91
## 2      BROOKLYN              4400959.900            22992174.59
## 3     MANHATTAN                29331.744              120371.89
## 4        QUEENS                 9380.724               33600.70
## 5 STATEN ISLAND                 9542.832               16455.39
```

```
#The standard deviation is good to have, but to assess statistical significan
ce we really want to have the standard error (which the standard deviation ad
justed by the group size).

ddply(DoBTotal, ~ ï..BOROUGH, summarize,
      group.size = length(Estimated.Job.Costs),
      mean.estimated.job.costs = mean(Estimated.Job.Costs),
      sd.estimated.job.costs = sd(Estimated.Job.Costs),
      se.estimated.job.costs = sd.estimated.job.costs / sqrt(group.size)
      )
```

```
##      ï..BOROUGH group.size mean.estimated.job.costs sd.estimated.job.costs
## 1         BRONX      37714                15673.268               34195.91
## 2      BROOKLYN     123114              4400959.900            22992174.59
## 3     MANHATTAN    1022013                29331.744              120371.89
## 4        QUEENS      71493                 9380.724               33600.70
## 5 STATEN ISLAND      30759                 9542.832               16455.39
##   se.estimated.job.costs
## 1              176.08523
## 2            65527.91169
## 3              119.06849
## 4              125.66563
## 5               93.82575
```

```
#Let's look at the average estimated job costs and proportion of residential
projects broken down by boroughs and job types

ddply(DoBTotal, ~ ï..BOROUGH+Job.Type, summarize,
      mean.Estimated.Job.Costs = mean(Estimated.Job.Costs),
      Residential.prop = mean(Residential == "YES"))
```

```
##       ï..BOROUGH Job.Type mean.Estimated.Job.Costs Residential.prop
## 1          BRONX       A1                 7803.972       0.50462208
## 2          BRONX       A2                15725.515       0.29960131
## 3          BRONX       A3                19745.746       0.58137467
## 4          BRONX       DM                 6923.747       0.00000000
## 5          BRONX       NB                14626.107       0.61834908
## 6          BRONX       SG                 6899.030       0.00000000
## 7       BROOKLYN       A1                 5508.650       0.64883162
## 8       BROOKLYN       A2              6438613.479       0.27254787
## 9       BROOKLYN       A3              6909076.891       0.45407082
## 10      BROOKLYN       DM                 3833.690       0.00000000
## 11      BROOKLYN       NB                10110.363       0.72608869
## 12      BROOKLYN       SG                 9497.643       0.00000000
## 13     MANHATTAN       A1                22053.232       0.42250523
## 14     MANHATTAN       A2                29423.388       0.13051368
## 15     MANHATTAN       A3                31864.858       0.28566413
## 16     MANHATTAN       DM                 6955.232       0.00000000
## 17     MANHATTAN       NB                20357.932       0.51841695
## 18     MANHATTAN       SG                31873.597       0.00000000
## 19        QUEENS       A1                11101.468       0.61191837
## 20        QUEENS       A2                 8398.577       0.20814072
## 21        QUEENS       A3                11619.897       0.43193816
## 22        QUEENS       DM                 8741.766       0.00000000
## 23        QUEENS       NB                11043.547       0.59359370
## 24        QUEENS       SG                 6046.364       0.00000000
## 25 STATEN ISLAND       A1                 5357.936       0.41619798
## 26 STATEN ISLAND       A2                 9662.518       0.02053456
## 27 STATEN ISLAND       A3                14094.667       0.12111604
## 28 STATEN ISLAND       DM                 8439.380       0.00000000
## 29 STATEN ISLAND       NB                 6515.662       0.42969133
## 30 STATEN ISLAND       SG                 9817.803       0.00000000
```

```
#Permit Status broken down by borough
Permit.borough.tbl<-with(DoBTotal, table(ï..BOROUGH,Permit.Status))
Permit.borough.tbl
```

```
##                Permit.Status
## ï..BOROUGH              IN PROCESS  ISSUED RE-ISSUED REVOKED
##   BRONX              56        316   36902       440       0
##   BROOKLYN          559       1003  119454      2098       0
##   MANHATTAN         498       4169 1003654     13692       0
##   QUEENS            199        456   69951       887       0
##   STATEN ISLAND      43        125   30295       296       0
```

```
#To test for significance, we just need to perform chi-square test
#chisq.test(Permit.borough.tbl)

#Permit.borough.test<-fisher.test(Permit.borough.tbl)
#Permit.borough.test
```