homework ii

Ayush Kumar Shah

2020-09-11

Introduction

In this report, we will create different visualizations on the nyc311 data using the open-source R package `ggplot2`, from CRAN. `ggplot2` is a written implementation of the layered grammar of graphics concept by Wickham. We will explore the concepts of layered grammar of graphics through various examples of the use of ggplot2 library to build higher level tools for data analysis. We will be mainly dealing with the following major components of the layered grammar:

1. Layers

- Data and mapping
- Statistical Transformation
- Geometric Object
- Position Adjustment

2. Scales
3. Coordinate System
4. Faceting

These components allow us to completely and explicitly describe a wide range of graphics.

Initialization

Here we load the tidyverse packages and the `data.table` package and load the nyc311 data set. Then we fix the column names of the nyc311 data so that they have no spaces.

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------------
```

```
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
```

```
## -- Conflicts ------------------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(data.table)
```

```
##
## Attaching package: 'data.table'
```

```
## The following objects are masked from 'package:dplyr':
##
##     between, first, last
```

```
## The following object is masked from 'package:purrr':
##
##     transpose
```

```
# fast for when you are starting out:
# nyc311<-fread("311_Service_Requests_from_2010_to_Present.csv",nrow=10000)
# after you get going:
nyc311<-fread("311_Service_Requests_from_2010_to_Present.csv")
names(nyc311)<-names(nyc311) %>%
  stringr::str_replace_all("\\s", ".")
```

Description

Here we describe the data, showing both a sample and a data dictionary.

The head of the table

Here we produce a table of just some relevant columns of data.

```
library(xtable)
options(xtable.comment=FALSE)
options(xtable.booktabs=TRUE)
narrow<-nyc311 %>%
  select(Agency,
     Complaint.Type,
     Descriptor,
     Incident.Zip,
     Status,
     Borough)
xtable(head(narrow))
```

	Agency	Complaint.Type	Descriptor	Incident.Zip	Status	Borough
1	NYPD	Vending	In Prohibited Area	10465	Closed	BRONX
2	NYPD	Blocked Driveway	No Access	11234	Open	BROOKLYN
3	NYPD	Noise - Street/Sidewalk	Loud Music/Party	11204	Open	BROOKLYN
4	NYPD	Noise - Street/Sidewalk	Loud Talking	11211	Assigned	BROOKLYN
5	NYPD	Noise - Street/Sidewalk	Loud Talking	10025	Closed	MANHATTAN
6	NYPD	Noise - Street/Sidewalk	Loud Talking	11205	Closed	BROOKLYN

Data Dictionary

First dropping the irrelevant columns

```
nyc311 <- subset(nyc311, select = c(1:4, 6:9, 20, 24:26, 40, 50:52 ))
```

Building the data dictionary

```
var_desc <- c("Unique identifier of a Service Request (SR) in the open data set",
              "Date Service Request (SR) was created",
              "Date SR was closed by responding agency",
              "Acronym of responding City Government Agency",
              " This is the first level of a hierarchy identifying the topic \n of the incident or condi
              "This is associated to the Complaint Type, and provides further detail on the incident or
              "Describes the type of location used in the address information",
              "Incident location zip code, provided by geo validation.",
              "Status of SR submitted",
              "Provided by the submitter and confirmed by geovalidation.",
              "Geo validated, X coordinate of the incident location.",
              "Geo validated, Y coordinate of the incident location.",
              "If the incident is a taxi, this field describes the type of TLC vehicle.",
              "Geo based Lat of the incident location",
              "Geo based Long of the incident location",
              "Combination of the geo based lat & long of the incident location")

text <- "Plain text"
num <- "Number"
date <- "Date and Time"
var_type <- c(text, date, date, text, text, text, text, text, text, text,
              num, num, text, num, num, "Location")
data_dict <- data.frame("Field Name" = names(nyc311),
                        "Variable Description" = var_desc,
                        "Variable Type" = var_type)
xtable(data_dict)
```

Exploration

Here we explore the columns in the data set.

Largest Responding City Government Agencies

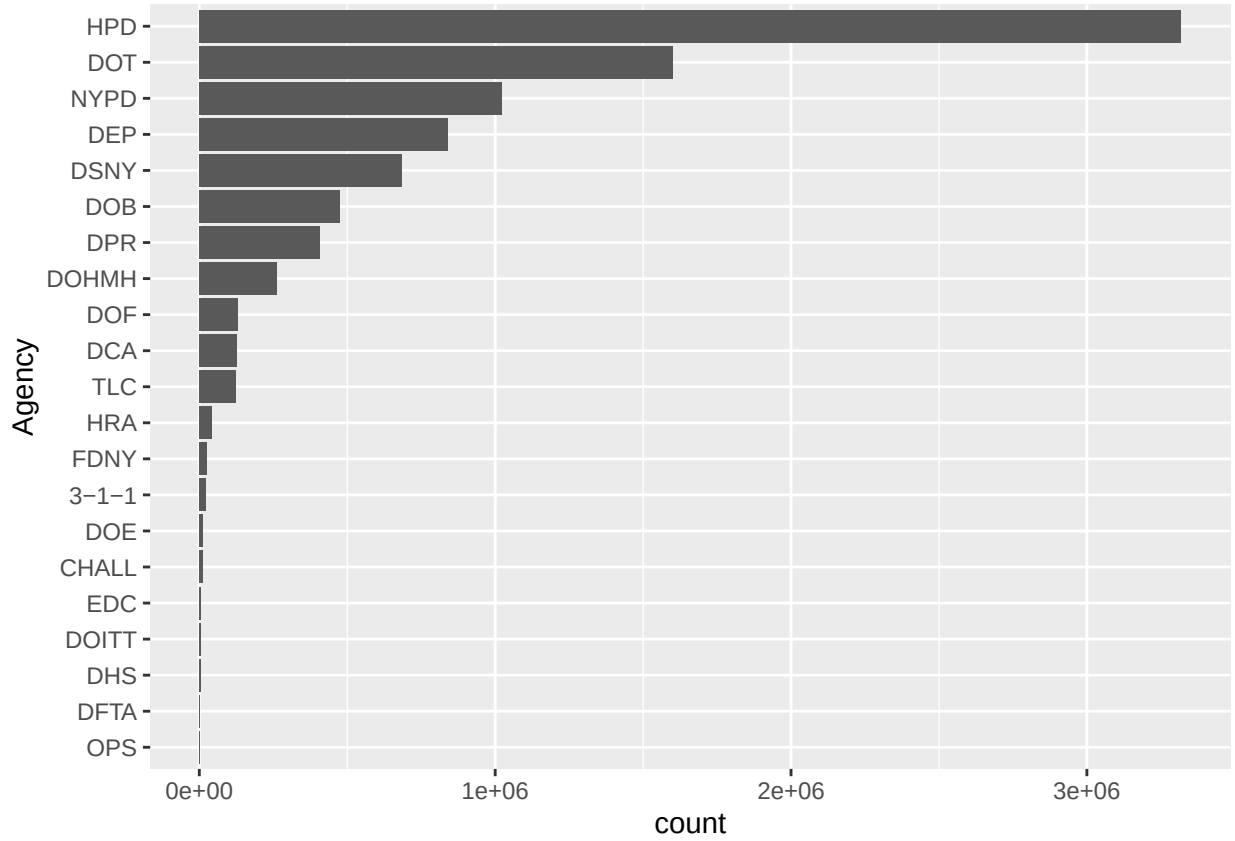
We find out the city government agencies that have responded to more than 1000 Service Requests (SR) with the number of SR in an increasing order using bar chart.

```
bigAgency <- narrow %>%
  group_by(Agency) %>%
  summarize(count=n()) %>%
  filter(count>1000)
```

	Field.Name	Variable.Description
1	Unique.Key	Unique identifier of a Service Request (SR) in the open data set
2	Created.Date	Date Service Request (SR) was created
3	Closed.Date	Date SR was closed by responding agency
4	Agency	Acronym of responding City Government Agency
5	Complaint.Type	This is the first level of a hierarchy identifying the topic of the incident or condition. C
6	Descriptor	This is associated to the Complaint Type, and provides further detail on the incident o
7	Location.Type	Describes the type of location used in the address information
8	Incident.Zip	Incident location zip code, provided by geo validation.
9	Status	Status of SR submitted
10	Borough	Provided by the submitter and confirmed by geovalidation.
11	X.Coordinate.(State.Plane)	Geo validated, X coordinate of the incident location.
12	Y.Coordinate.(State.Plane)	Geo validated, Y coordinate of the incident location.
13	Vehicle.Type	If the incident is a taxi, this field describes the type of TLC vehicle.
14	Latitude	Geo based Lat of the incident location
15	Longitude	Geo based Long of the incident location
16	Location	Combination of the geo based lat & long of the incident location

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
bigAgency$Agency<-factor(bigAgency$Agency,
  levels=bigAgency$Agency[order(bigAgency$count)])
p<-ggplot(bigAgency,aes(x=Agency,y=count)) +
   geom_bar(stat="identity") +
   coord_flip()
p
```



Status of Complaints across Boroughs

We plot the different status of complaints and how it vaires across different Boroughs.

```
s1 <- ggplot(narrow) +
      geom_bar(mapping = aes(Status, fill = Borough), position="fill") +
      coord_flip() +
      ggtitle("Status of SR Submitted across Boroughs")

s2 <- ggplot(narrow) +
  geom_bar(mapping = aes(Status, fill = Borough), show.legend = FALSE) +
  coord_flip()

if (!require(gridExtra)) {
install.packages("gridExtra",dependencies=TRUE)
library(gridExtra)
}
```

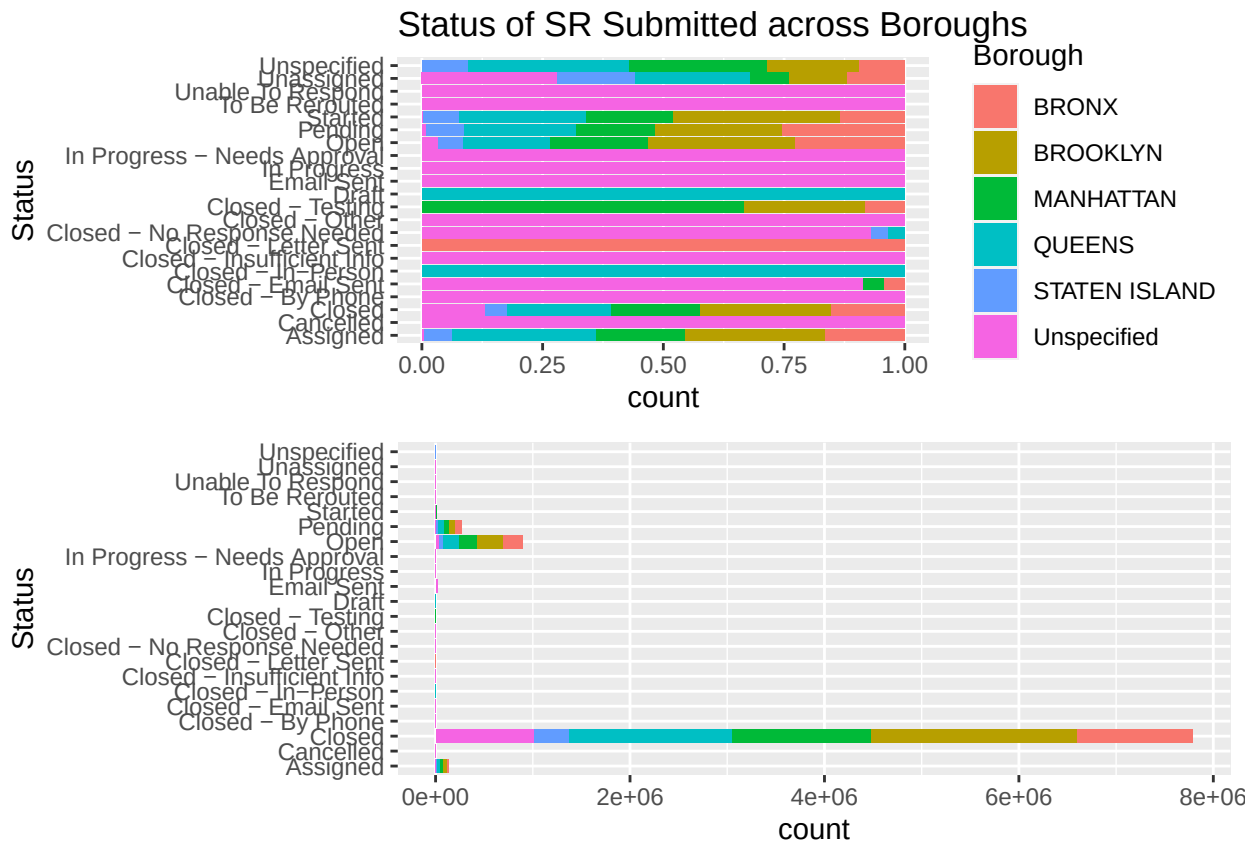
```
## Loading required package: gridExtra
```

```
##
## Attaching package: 'gridExtra'
```

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

```
##
##     combine
```

```
grid.arrange(s1, s2, nrow = 2)
```



Top complaint types

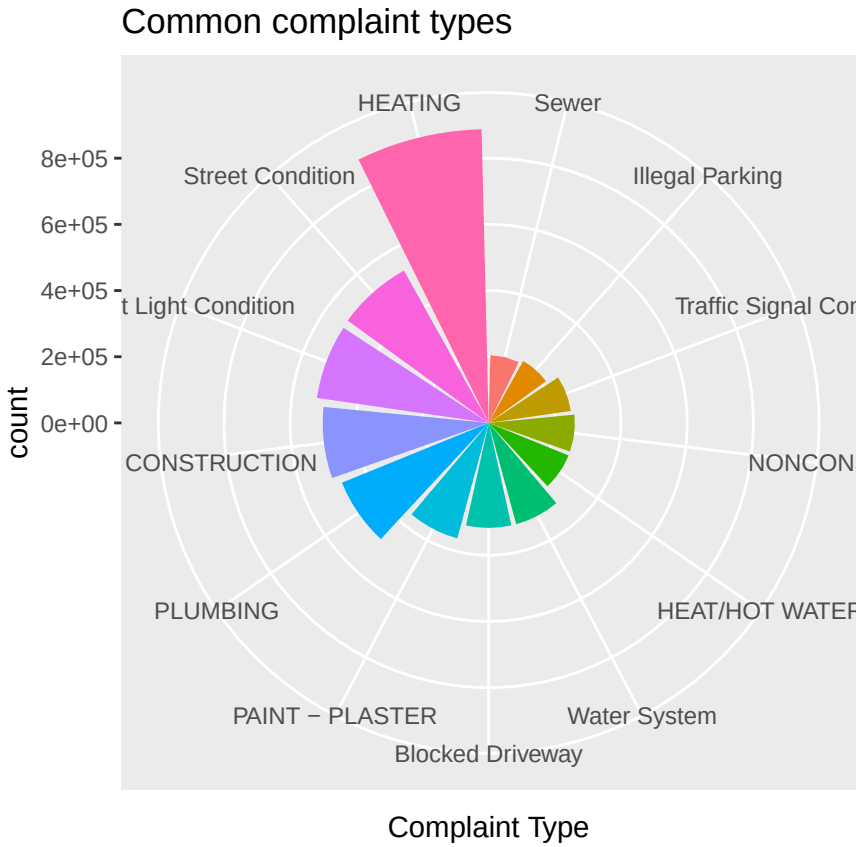
```
top_complaints <- narrow %>%
  group_by(Complaint.Type) %>%
  summarize(count=n()) %>%
  filter(count>200000)
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
top_complaints$Complaint.Type <- factor(top_complaints$Complaint.Type,
          levels=top_complaints$Complaint.Type[order(top_complaints$count)])

t <- ggplot(top_complaints,aes(Complaint.Type,count, fill=Complaint.Type)) +
  geom_bar(stat="identity", show.legend = FALSE) +
  coord_polar() +
  xlab("Complaint Type") +
  ggtitle("Common complaint types")

t
```



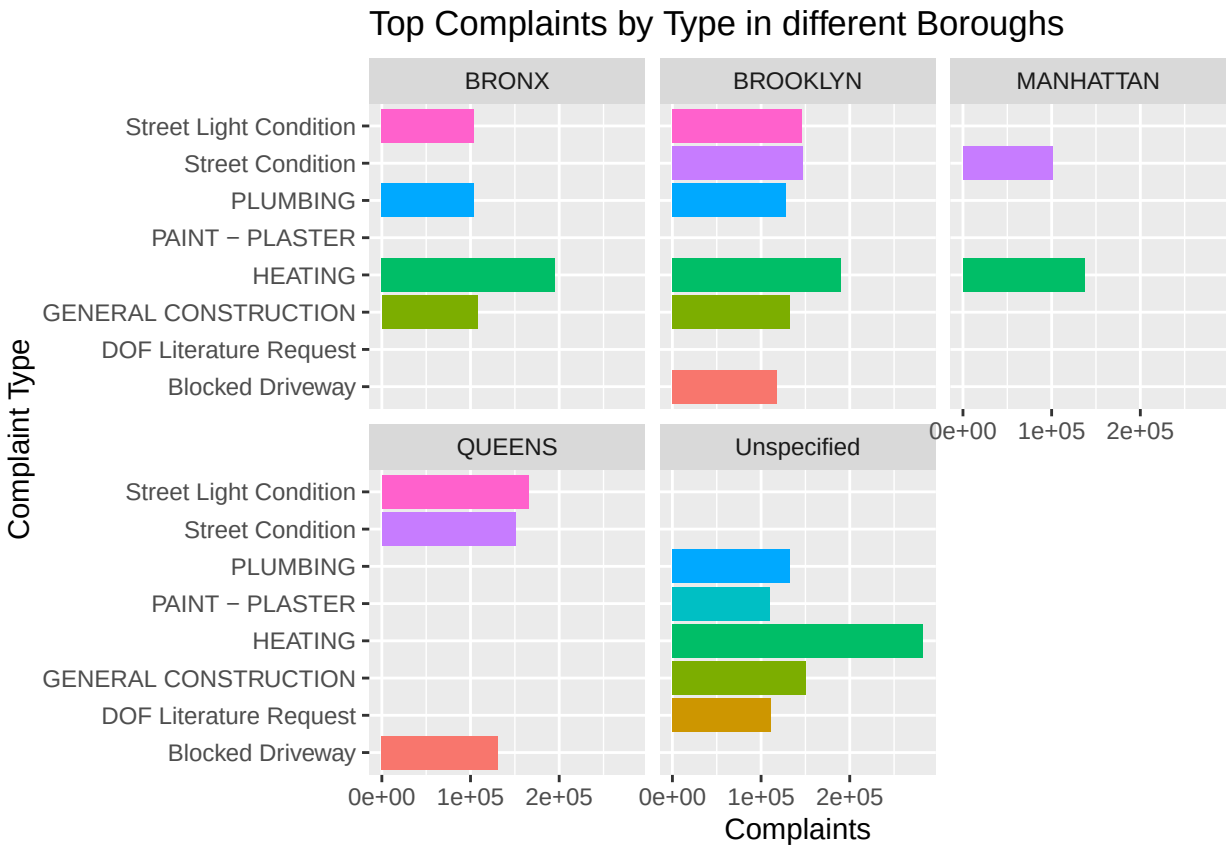
Complaint types

```
complaint_types <- narrow %>%
  group_by(Complaint.Type, Borough) %>%
  summarize(Complaints = length(Complaint.Type)) %>%
  filter(Complaints>100000)
```

```
## `summarise()` regrouping output by 'Complaint.Type' (override with `.groups` argument)
```

```
ct <- ggplot(complaint_types) +
  geom_bar(stat="identity", aes(x=Complaint.Type, y=Complaints,
                                fill=Complaint.Type),
           show.legend = FALSE) +
  facet_wrap(~ Borough) +
  coord_flip() +
  xlab("Complaint Type") +
  ggtitle("Top Complaints by Type in different Boroughs")

ct
```



Response time

```
df<-nyc311 %>%
  select(Agency,
         Complaint.Type,
         Descriptor,
         Incident.Zip,
         Status,
         Borough,
         Created.Date,
         Closed.Date)
df$Response.Time.hrs <- as.numeric(difftime(as.Date(as.character(df$Closed.Date),
                                                    format="%m/%d/%Y %H:%M:%S %p"),
                                            as.Date(as.character(df$Created.Date),
                                                    format="%m/%d/%Y %H:%M:%S %p")
                                            , units = "hours"))

df <- df[df$Response.Time.hrs > 0]
cat("Average response time = ", mean(df$Response.Time.hrs) %/% 24 , "days")
```

```
## Average response time =  22 days
```

Cross tabulations

Next we include a crosstabulation.

1. Borough and Complaint Type

```
xtabA<-dplyr::filter(narrow,
  Complaint.Type=='HEATING' |
  Complaint.Type=='GENERAL CONSTRUCTION' |
  Complaint.Type=='PLUMBING'
)
xtabB<-select(xtabA,Borough,"Complaint.Type")
library(gmodels)
CrossTable(xtabB$Borough,xtabB$'Complaint.Type')
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  1868064
##
##
##                | xtabB$Complaint.Type
## xtabB$Borough | GENERAL CONSTRUCTION |              HEATING |             PLUMBING |            Row Total |
## --------------|----------------------|----------------------|----------------------|----------------------|
##         BRONX |               107626 |               195246 |               103964 |               406836 |
##               |               23.326 |               19.145 |                1.030 |                      |
##               |                0.265 |                0.480 |                0.256 |                0.218 |
##               |                0.215 |                0.220 |                0.217 |                      |
##               |                0.058 |                0.105 |                0.056 |                      |
## --------------|----------------------|----------------------|----------------------|----------------------|
##      BROOKLYN |               132552 |               190268 |               128383 |               451203 |
##               |             1076.405 |             2717.190 |             1398.387 |                      |
##               |                0.294 |                0.422 |                0.285 |                0.242 |
##               |                0.264 |                0.214 |                0.268 |                      |
##               |                0.071 |                0.102 |                0.069 |                      |
## --------------|----------------------|----------------------|----------------------|----------------------|
##     MANHATTAN |                61453 |               137458 |                63103 |               262014 |
##               |             1123.330 |             1347.582 |              245.877 |                      |
##               |                0.235 |                0.525 |                0.241 |                0.140 |
##               |                0.123 |                0.155 |                0.132 |                      |
##               |                0.033 |                0.074 |                0.034 |                      |
## --------------|----------------------|----------------------|----------------------|----------------------|
##        QUEENS |                41277 |                75776 |                43604 |               160657 |
##               |               79.707 |                4.192 |              142.183 |                      |
##               |                0.257 |                0.472 |                0.271 |                0.086 |
##               |                0.082 |                0.085 |                0.091 |                      |
##               |                0.022 |                0.041 |                0.023 |                      |
## --------------|----------------------|----------------------|----------------------|----------------------|
## STATEN ISLAND |                 8329 |                 6011 |                 7525 |                21865 |
##               |             1030.062 |             1845.525 |              657.654 |                      |
##               |                0.381 |                0.275 |                0.344 |                0.012 |
##               |                0.017 |                0.007 |                0.016 |                      |
##               |                0.004 |                0.003 |                0.004 |                      |
## --------------|----------------------|----------------------|----------------------|----------------------|
##   Unspecified |               150277 |               282916 |               132296 |               565489 |
```

```
##               |              15.587 |             750.862 |            1106.709 |                     |
##               |               0.266 |               0.500 |               0.234 |               0.303 |
##               |               0.300 |               0.319 |               0.276 |                     |
##               |               0.080 |               0.151 |               0.071 |                     |
## --------------|---------------------|---------------------|---------------------|---------------------|
##  Column Total |              501514 |              887675 |              478875 |             1868064 |
##               |               0.268 |               0.475 |               0.256 |                     |
## --------------|---------------------|---------------------|---------------------|---------------------|
##
##
```

A summary table will be generated with cell row, column and table proportions and marginal totals and proportions.

2. Borough and Status

```
filtered_status<-dplyr::filter(narrow,
                       Status=='Closed' |
                       Status=='Open' |
                       Status=='Pending'
)
CrossTable(filtered_status$Borough, filtered_status$Status)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  8953711
##
##
##                          | filtered_status$Status
## filtered_status$Borough |    Closed |      Open |   Pending | Row Total |
## ------------------------|-----------|-----------|-----------|-----------|
##                   BRONX |   1190472 |    203951 |     68633 |   1463056 |
##                         |  5259.123 | 22240.719 | 13825.306 |           |
##                         |     0.814 |     0.139 |     0.047 |     0.163 |
##                         |     0.153 |     0.227 |     0.255 |           |
##                         |     0.133 |     0.023 |     0.008 |           |
## ------------------------|-----------|-----------|-----------|-----------|
##                BROOKLYN |   2119906 |    274308 |     71078 |   2465292 |
##                         |   266.607 |  2931.575 |   123.283 |           |
##                         |     0.860 |     0.111 |     0.029 |     0.275 |
##                         |     0.272 |     0.305 |     0.264 |           |
##                         |     0.237 |     0.031 |     0.008 |           |
## ------------------------|-----------|-----------|-----------|-----------|
##                MANHATTAN |   1429165 |    182086 |     43515 |   1654766 |
##                         |    66.965 |  1549.349 |   778.617 |           |
##                         |     0.864 |     0.110 |     0.026 |     0.185 |
##                         |     0.184 |     0.203 |     0.162 |           |
##                         |     0.160 |     0.020 |     0.005 |           |
## ------------------------|-----------|-----------|-----------|-----------|
##                  QUEENS |   1673100 |    161352 |     62725 |   1897177 |
##                         |   329.591 |  4423.519 |   569.884 |           |
##                         |     0.882 |     0.085 |     0.033 |     0.212 |
##                         |     0.215 |     0.180 |     0.233 |           |
```

```
##                         |     0.187 |     0.018 |     0.007 |           |
## ------------------------|-----------|-----------|-----------|-----------|
##           STATEN ISLAND |    367007 |     47382 |     21006 |    435395 |
##                         |   356.112 |   312.085 |  4792.015 |           |
##                         |     0.843 |     0.109 |     0.048 |     0.049 |
##                         |     0.047 |     0.053 |     0.078 |           |
##                         |     0.041 |     0.005 |     0.002 |           |
## ------------------------|-----------|-----------|-----------|-----------|
##             Unspecified |   1006480 |     29376 |      2169 |   1038025 |
##                         | 11939.813 | 53692.866 | 27013.202 |           |
##                         |     0.970 |     0.028 |     0.002 |     0.116 |
##                         |     0.129 |     0.033 |     0.008 |           |
##                         |     0.112 |     0.003 |     0.000 |           |
## ------------------------|-----------|-----------|-----------|-----------|
##            Column Total |   7786130 |    898455 |    269126 |   8953711 |
##                         |     0.870 |     0.100 |     0.030 |           |
## ------------------------|-----------|-----------|-----------|-----------|
##
##
```

3. Status and Complaint Type

```
xtabC<-select(xtabA,Status,Complaint.Type)
CrossTable(xtabC$Status, xtabC$Complaint.Type)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  1868064
##
##
##              | xtabC$Complaint.Type
## xtabC$Status | GENERAL CONSTRUCTION |              HEATING |             PLUMBING |            Row Total |
## -------------|----------------------|----------------------|----------------------|----------------------|
##       Closed |               470206 |               876269 |               406114 |              1752589 |
##              |                0.200 |             2268.588 |             4146.063 |                      |
##              |                0.268 |                0.500 |                0.232 |                0.938 |
##              |                0.938 |                0.987 |                0.848 |                      |
##              |                0.252 |                0.469 |                0.217 |                      |
## -------------|----------------------|----------------------|----------------------|----------------------|
##         Open |                31308 |                11404 |                72761 |               115473 |
##              |                3.046 |            34433.104 |            62928.280 |                      |
##              |                0.271 |                0.099 |                0.630 |                0.062 |
##              |                0.062 |                0.013 |                0.152 |                      |
##              |                0.017 |                0.006 |                0.039 |                      |
## -------------|----------------------|----------------------|----------------------|----------------------|
##      Pending |                    0 |                    2 |                    0 |                    2 |
##              |                0.537 |                1.159 |                0.513 |                      |
##              |                0.000 |                1.000 |                0.000 |                0.000 |
##              |                0.000 |                0.000 |                0.000 |                      |
##              |                0.000 |                0.000 |                0.000 |                      |
## -------------|----------------------|----------------------|----------------------|----------------------|
## Column Total |               501514 |               887675 |               478875 |              1868064 |
##              |                0.268 |                0.475 |                0.256 |                      |
## -------------|----------------------|----------------------|----------------------|----------------------|
```

```
##
##
```

Conclusion

So, we learned how to visualize the `nyc311` data using ggplot2. We applied and understood different components of layered grammar of graphics like data and aesthetic mappings, facets, geometric objects, statistical transformations, position arguments and coordinates transformation using the `nyc311` data as an example.