homework iii

Ayush Kumar Shah

2020-10-22

Introduction

In this report, we try to explore the nyc311 data and find answers about the data. Some of the questions we try to answer in this exploration are:

- What are the most frequent categories of complaints?
- How is the frequent complaint categories distributed across different Boroughs?
- How does the overall status of the complaints vary across different categories?
- What are the status of the most frequent complaint categories?
- Which agencies are the top 10 largest responding City Government Agencies?

We perform various transformations to the data using the `dplyr` package to answer these queries. We also use `ggplot` to visualize the results and also the distributions of different variables of the data.

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 ------------------------------------------------------------------------------------ ti
## 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
```

```
nyc311<-fread("311_Service_Requests_from_2010_to_Present.csv")
names(nyc311)<-names(nyc311) %>%
stringr::str_replace_all("\\s", ".")
mini311<-nyc311[sample(nrow(nyc311),10000),]
write.csv(mini311,"mini311.csv")
```

Working with maps

Reading the saved shorter sample of the data

This is done since original data has too many points to visualize.

```
sample<-fread("mini311.csv")
```

Selecting a single complaint type "Noise"

```
complaintlocs <- sample %>%
  select(Complaint.Type,
    Longitude,
    Latitude
  )
noisecompl <- complaintlocs %>%
  filter(Complaint.Type == "Noise")
```

Displaying the map

```
library(ggmap)
```

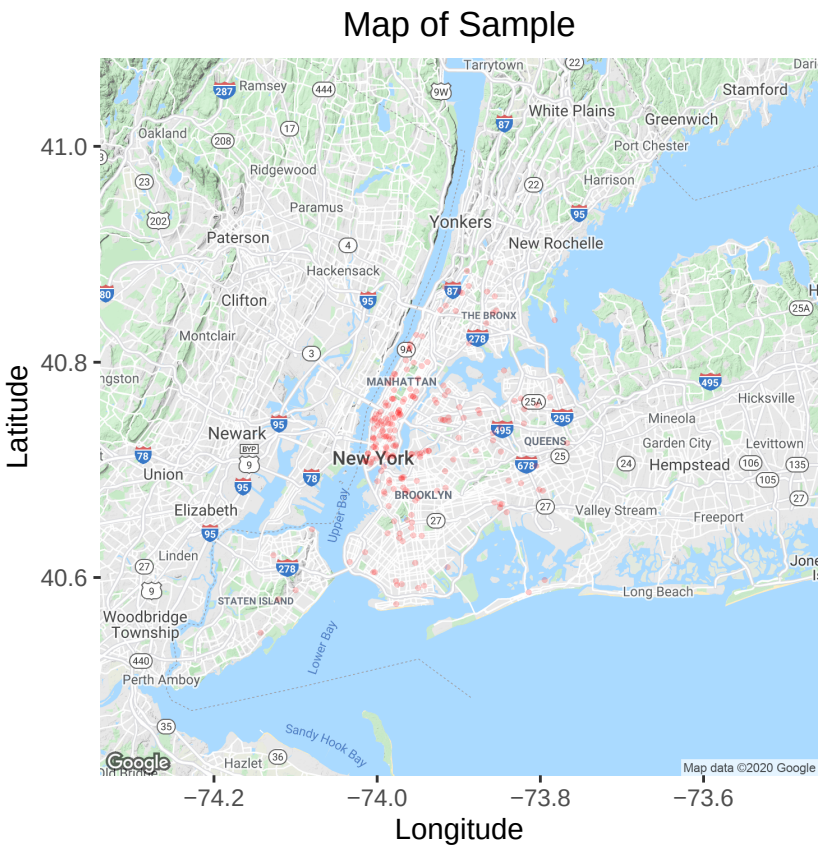
```
## Google's Terms of Service: https://cloud.google.com/maps-platform/terms/.

## Please cite ggmap if you use it! See citation("ggmap") for details.
```

```
key <- "AIzaSyC5GbsILImvioh5il6W3HQaIgKR3ra5v1o"
register_google(key=key)
nyc_map <- get_map(location=c(lon=-73.9,lat=40.75),
           maptype="terrain",zoom=10)
```

```
## Source : https://maps.googleapis.com/maps/api/staticmap?center=40.75,-73.9&zoom=10&size=640x640&scal
```

```
map <- ggmap(nyc_map) +
  geom_point(data=noisecompl,aes(x=Longitude,y=Latitude),
         size=0.4,alpha=0.2,color="red") +
  ggtitle("Map of Sample") +
  theme(plot.title=element_text(hjust=0.5)) +
  xlab("Longitude") + ylab("Latitude")
map
```

```
## Warning: Removed 4 rows containing missing values (geom_point).
```



Most frequent Complaint Categories

Let's view the Top 10 most frequent categories of the complaints registered along with the count and count %.

```
top10_complaints <-
  nyc311 %>%
  group_by(Complaint.Type) %>%
  summarize(count = n()) %>%
  mutate(rank = min_rank(desc(count)),
         'proportion in %' = count / sum(count) * 100) %>%
  filter(rank <= 10) %>%
  arrange(rank) %>%
  select(-rank)
```

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

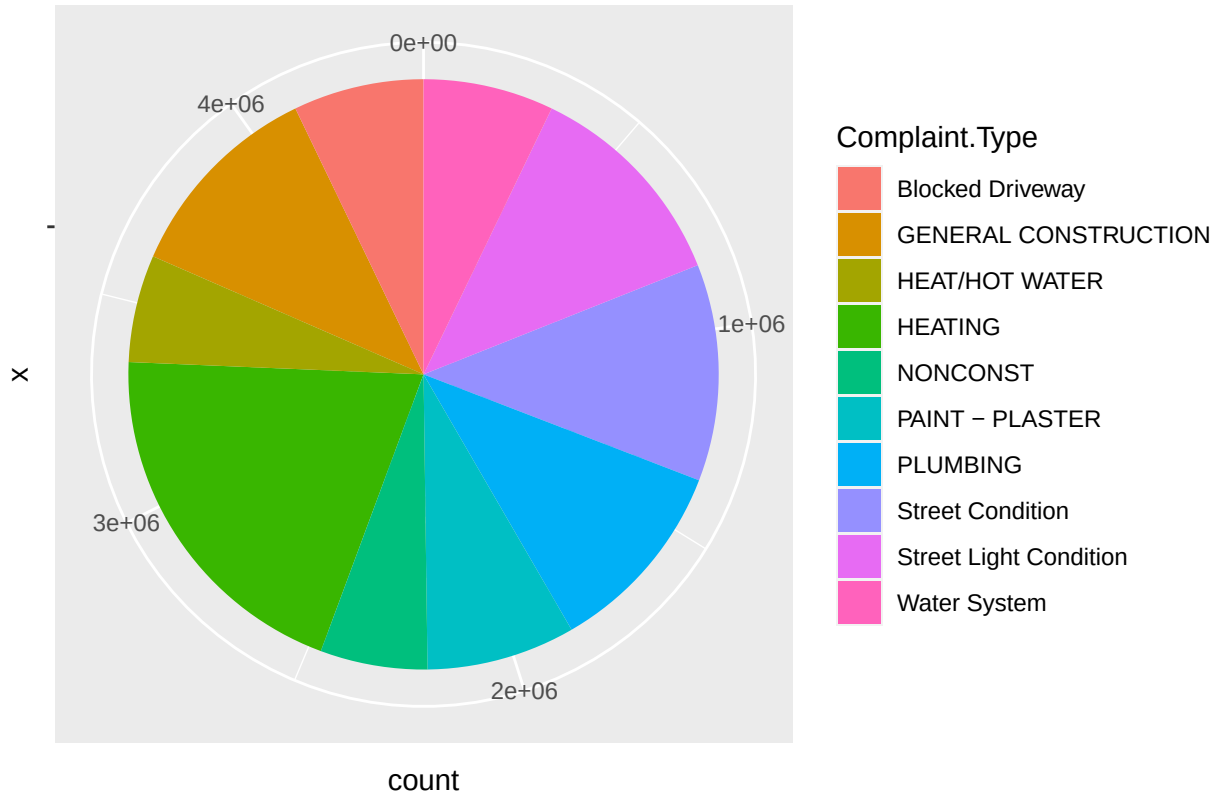
```
top10_complaints
```

```
## # A tibble: 10 x 3
##    Complaint.Type          count `proportion in %`
##    <chr>                   <int>             <dbl>
##  1 HEATING                 887675              9.73
##  2 Street Condition        526797              5.77
##  3 Street Light Condition  524501              5.75
##  4 GENERAL CONSTRUCTION    501514              5.50
##  5 PLUMBING                478875              5.25
##  6 PAINT - PLASTER         361449              3.96
##  7 Blocked Driveway        317163              3.48
##  8 Water System            317075              3.47
##  9 HEAT/HOT WATER          260936              2.86
## 10 NONCONST                260405              2.85
```

Pie chart of the most frequent complaint categiories

Let's plot the counts generated above in a pie chart.

```
top10_complaints_plot <-
  ggplot(top10_complaints, aes(x="", y=count, fill=Complaint.Type)) +
  geom_bar(stat = "identity") +
  coord_polar("y", start=0)

top10_complaints_plot
```



Most freqeuent Complaint Categories Count across different Boroughs

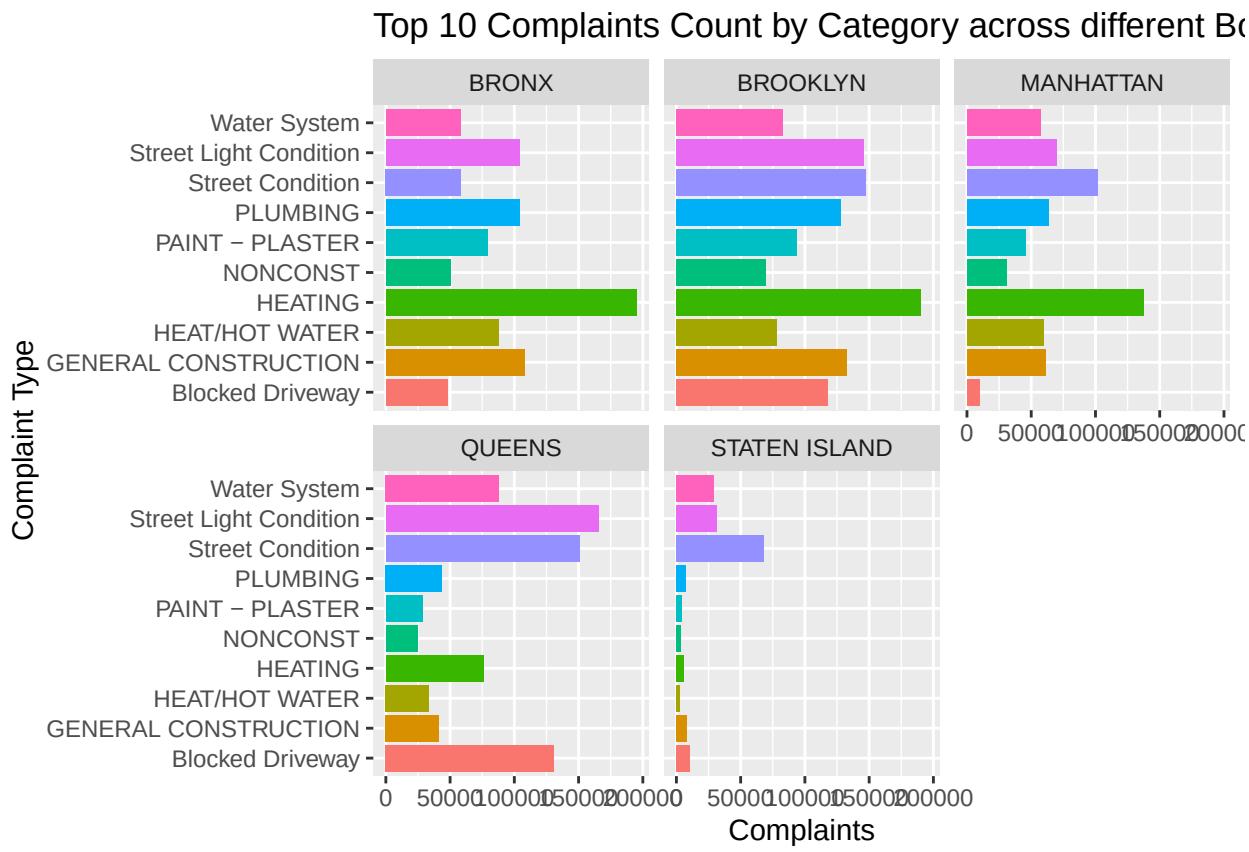
Now, let's view the counts of the top 10 frequent complain categories across different Boroughs using a facet plot.

```
complaint_types <- nyc311 %>%
  group_by(Complaint.Type, Borough) %>%
  summarize(Complaints = n()) %>%
  filter(Complaint.Type %in% top10_complaints$Complaint.Type,
         Borough != 'Unspecified')
```

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

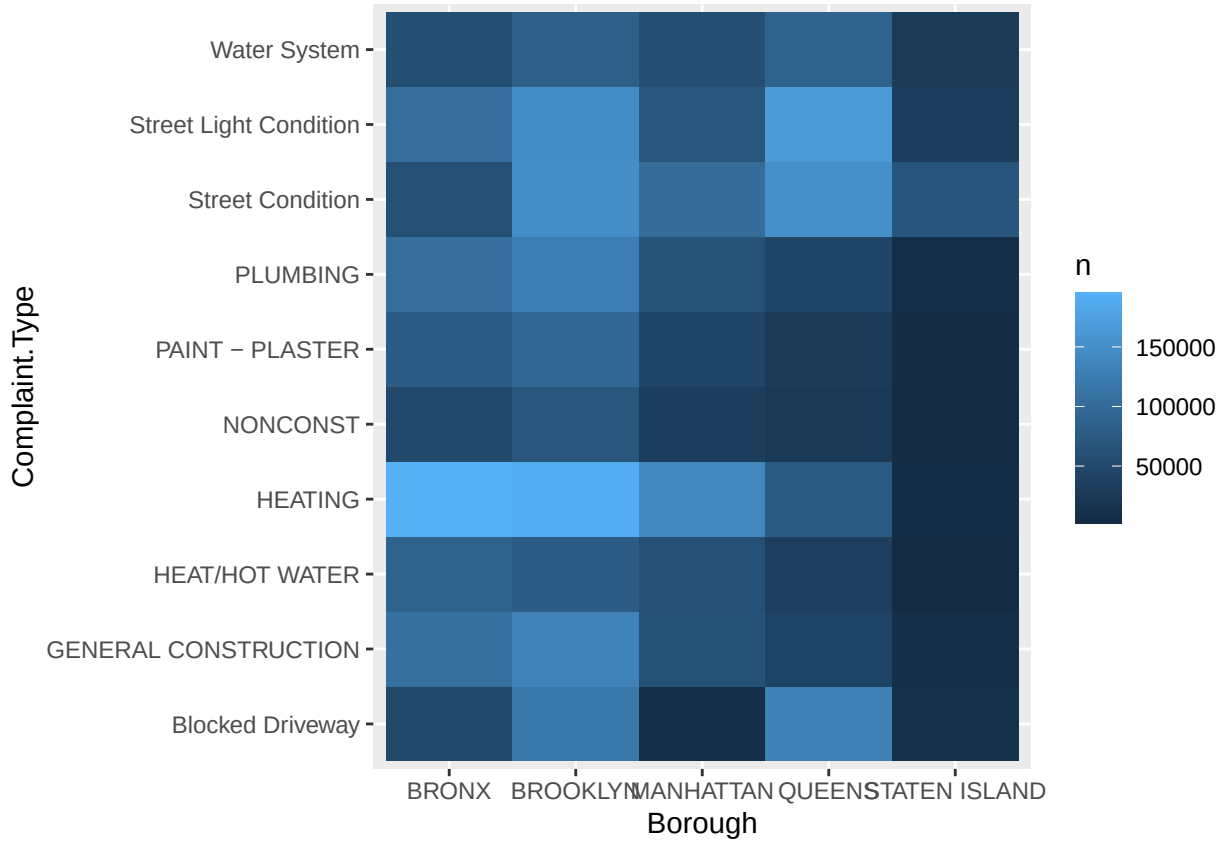
```
ct_borough <- ggplot(complaint_types) +
  geom_bar(stat="identity",
           aes(x=Complaint.Type, y=Complaints, fill=Complaint.Type),
           show.legend = FALSE) +
  facet_wrap(~ Borough) +
  coord_flip() +
  xlab("Complaint Type") +
  ggtitle("Top 10 Complaints Count by Category across different Boroughs")
```

```
ct_borough
```



Visualizing using geom_tile

```
nyc311 %>%
  filter(Complaint.Type %in% top10_complaints$Complaint.Type,
         Borough != 'Unspecified') %>%
  count(Complaint.Type, Borough) %>%
  ggplot(mapping = aes(x = Borough, y = Complaint.Type)) +
  geom_tile(mapping = aes(fill = n))
```



Status of complaints

Overall status of the complaints

```
nyc311 %>%
  group_by(Status)%>%
  summarize(count = n()) %>%
  arrange(desc(count))
```

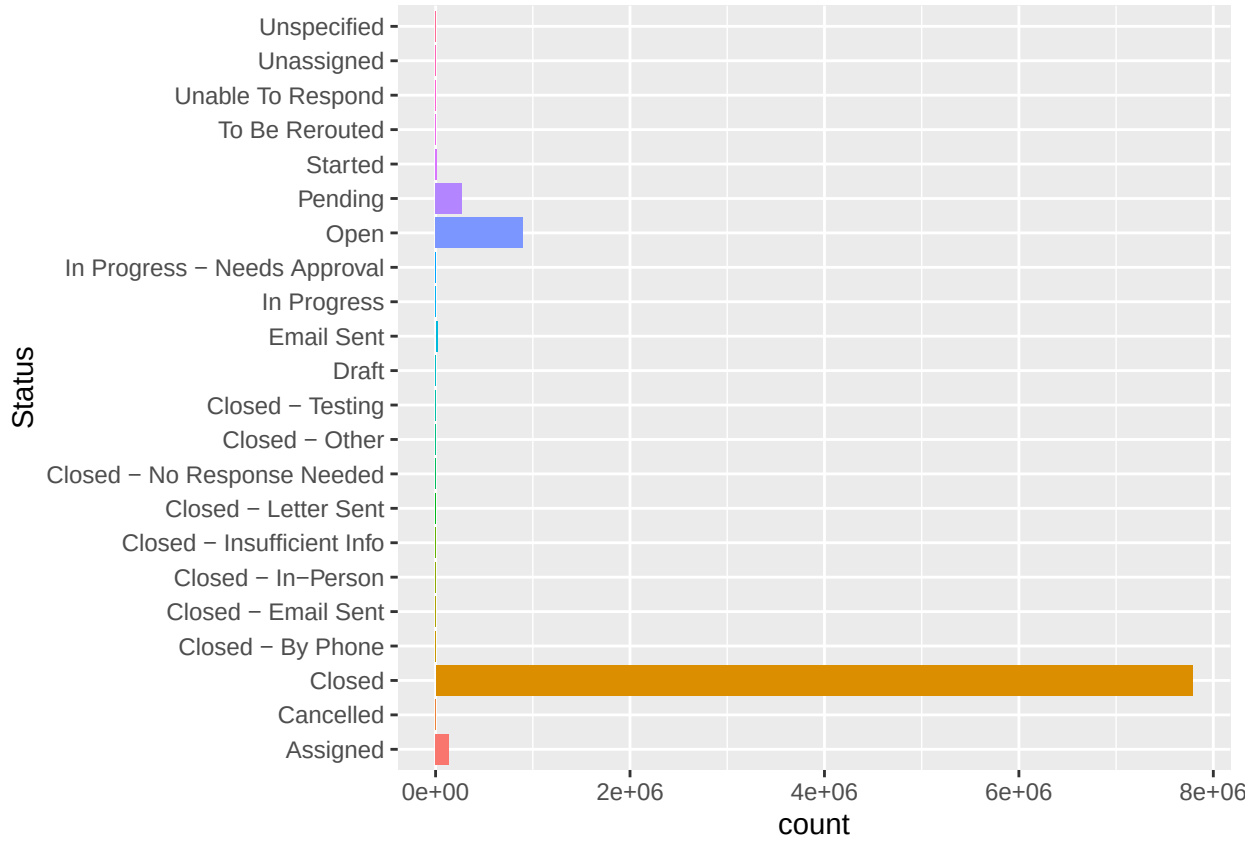
```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 22 x 2
##    Status                        count
##    <chr>                         <int>
##  1 Closed                      7786130
##  2 Open                         898455
##  3 Pending                      269126
##  4 Assigned                     138769
##  5 Email Sent                    20670
##  6 Started                       11639
##  7 Closed - No Response Needed      28
##  8 Unassigned                       25
```

```
##  9 Closed - Email Sent              23
## 10 Unspecified                      21
## # ... with 12 more rows
```

```
ggplot(data = nyc311, aes(x = Status, fill=Status)) +
  geom_bar(show.legend = FALSE) +
  coord_cartesian(ylim = c(0, 1)) +
  coord_flip()
```

```
## Coordinate system already present. Adding new coordinate system, which will replace the existing one
```



We can see that most of the status categories have very few count. So, we only consider the 3 major categories for analysis further.

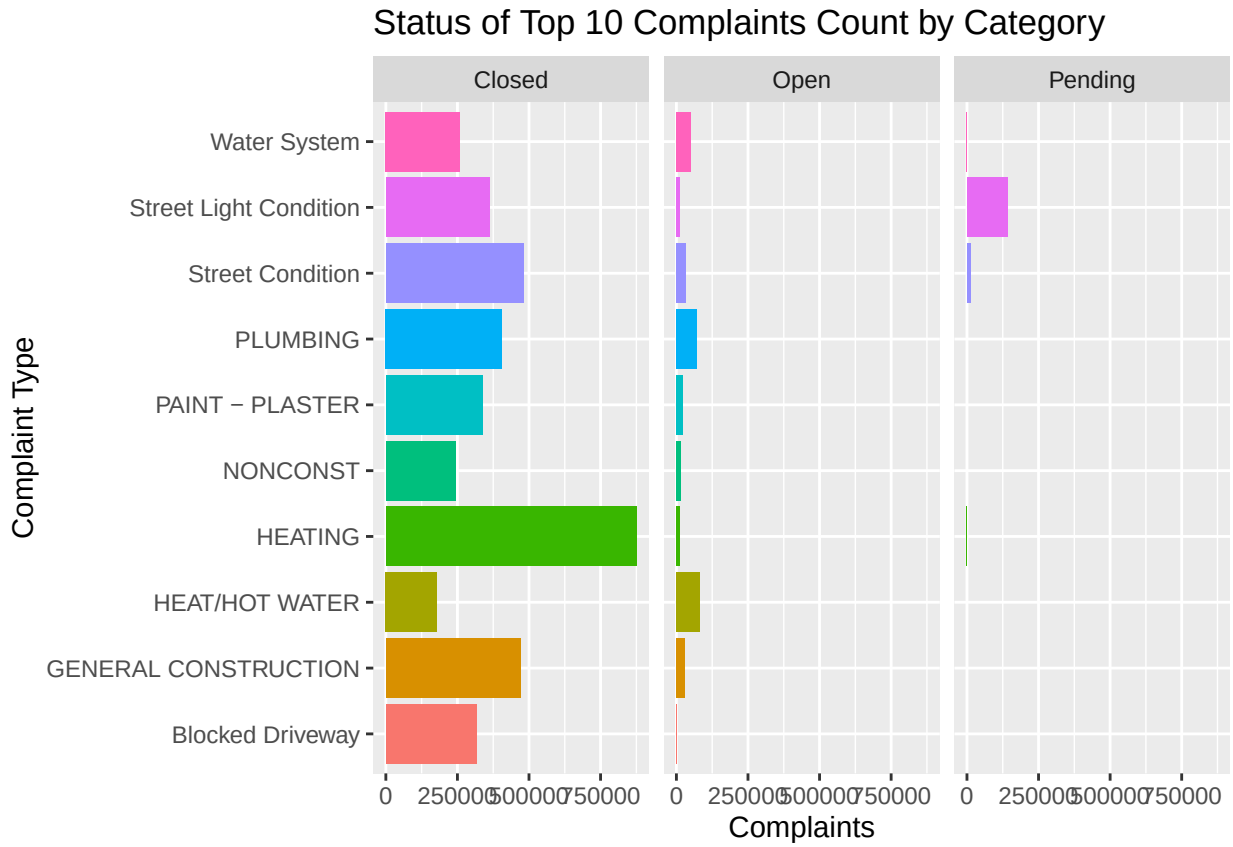
Status of the top 10 frequent complaint categories

We are only interested in the 3 major status of the most frequent complaint categories.

```
status_complaints <-
  nyc311 %>%
  group_by(Complaint.Type, Status) %>%
  summarize(Complaints = n()) %>%
  filter(Complaint.Type %in% top10_complaints$Complaint.Type,
         Status %in% c('Open', 'Closed', 'Pending'))
```

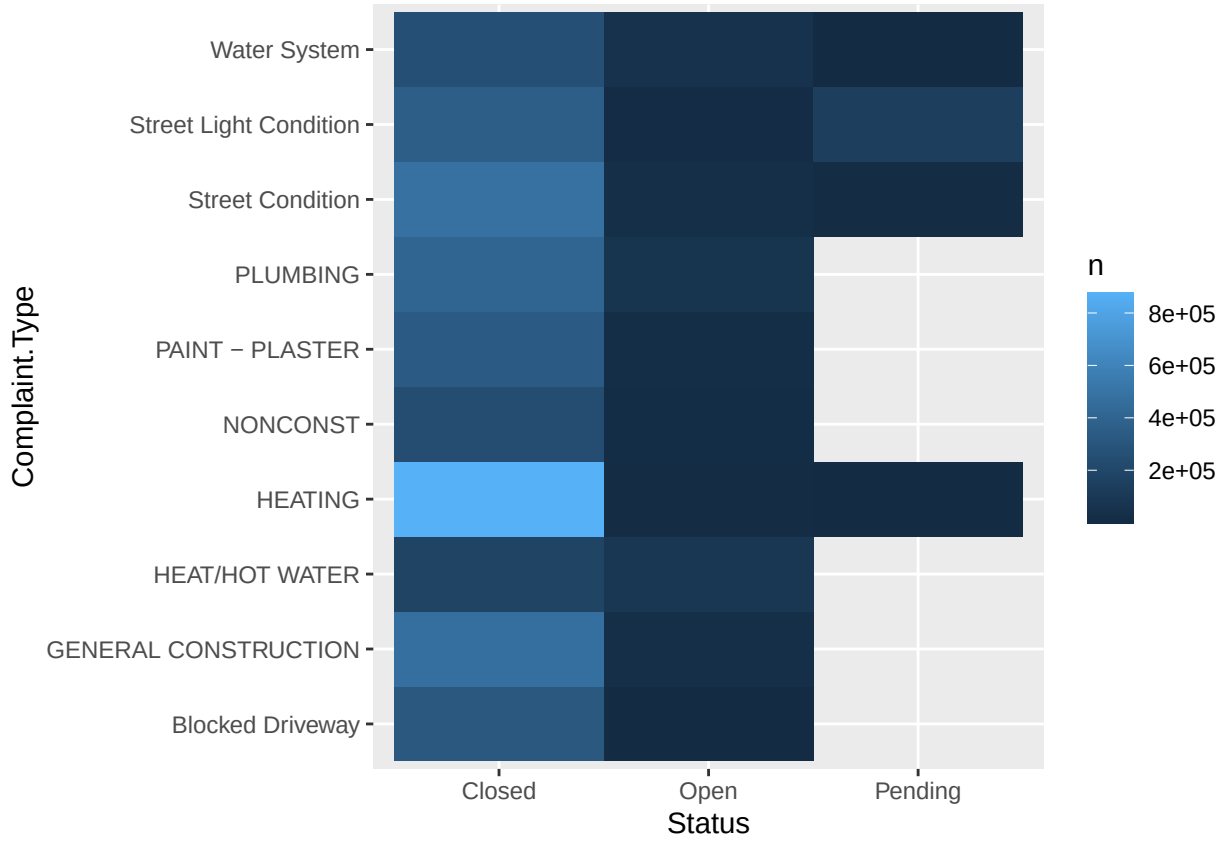
```
## `summarise()` regrouping output by 'Complaint.Type' (override with `.groups` argument)
```

```
## Plotting the status of complaints
status_complaints_plot <- ggplot(status_complaints) +
  geom_bar(stat="identity",
           aes(x=Complaint.Type, y=Complaints, fill=Complaint.Type),
           show.legend = FALSE) +
  facet_wrap(~ Status) +
  coord_flip() +
  xlab("Complaint Type") +
  ggtitle("Status of Top 10 Complaints Count by Category")

status_complaints_plot
```



Visualizing using geom_tile

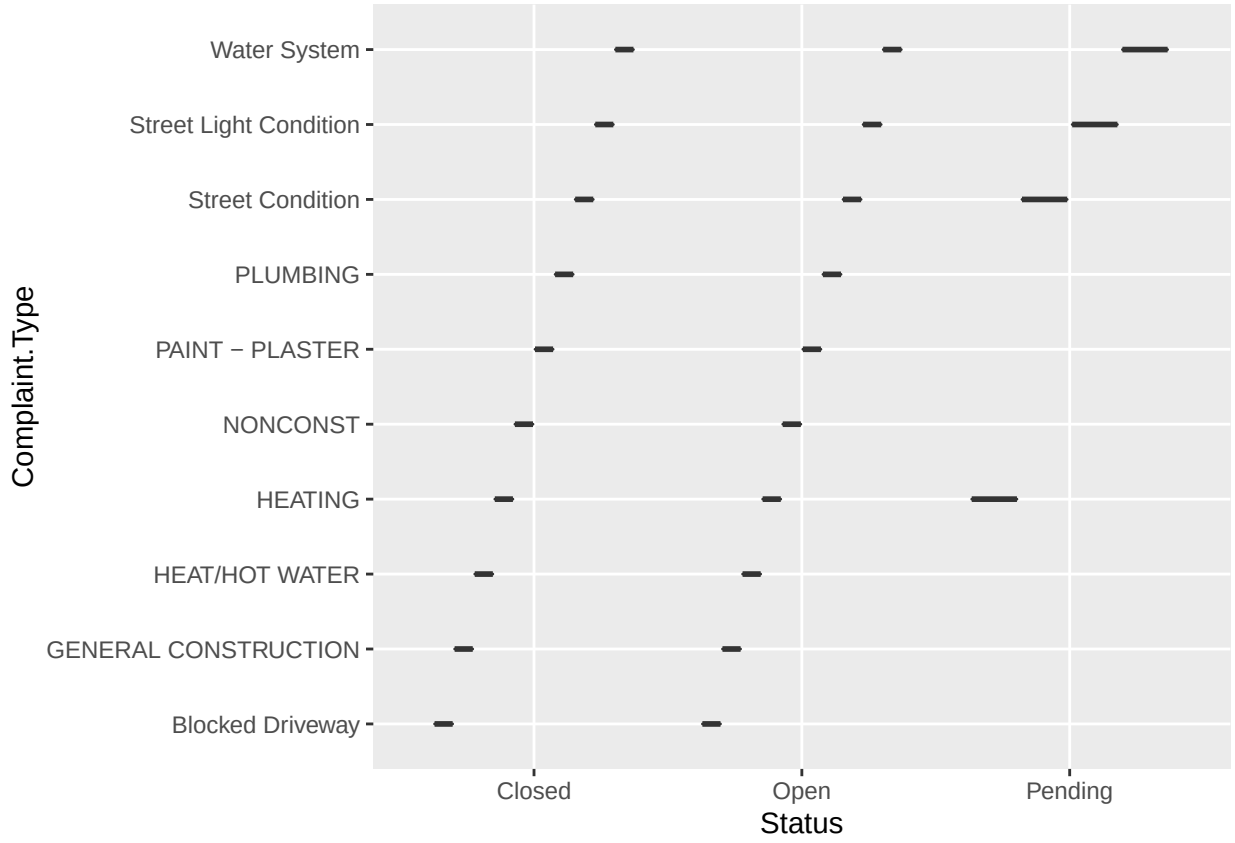
```
nyc311 %>%
  filter(Complaint.Type %in% top10_complaints$Complaint.Type,
         Status %in% c('Open', 'Closed', 'Pending')) %>%
  count(Complaint.Type, Status) %>%
  ggplot(mapping = aes(x = Status, y = Complaint.Type)) +
  geom_tile(mapping = aes(fill = n))
```



Hence, we can see that among the top 10 complaints category, most of them are closed with a very few pending cases.

Visualizing using box_plot

```
nyc311 %>%
  filter(Complaint.Type %in% top10_complaints$Complaint.Type,
         Status %in% c('Open', 'Closed', 'Pending')) %>%
  ggplot(mapping = aes(x = Status, y = Complaint.Type)) +
  geom_boxplot()
```



Top 10 Largest Responding City Government Agencies

We find out the top 10 city government agencies in terms of the largest Service Requests (SR) with the count and proportion count (in percentage).

```
bigAgency <- nyc311 %>%
  group_by(Agency) %>%
  summarize(complaints_count=n()) %>%
  mutate(rank = min_rank(desc(complaints_count)),
         'proportion in %' = complaints_count / sum(complaints_count) * 100) %>%
  filter(rank <= 10) %>%
  arrange(rank) %>%
  select(-rank)
```

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

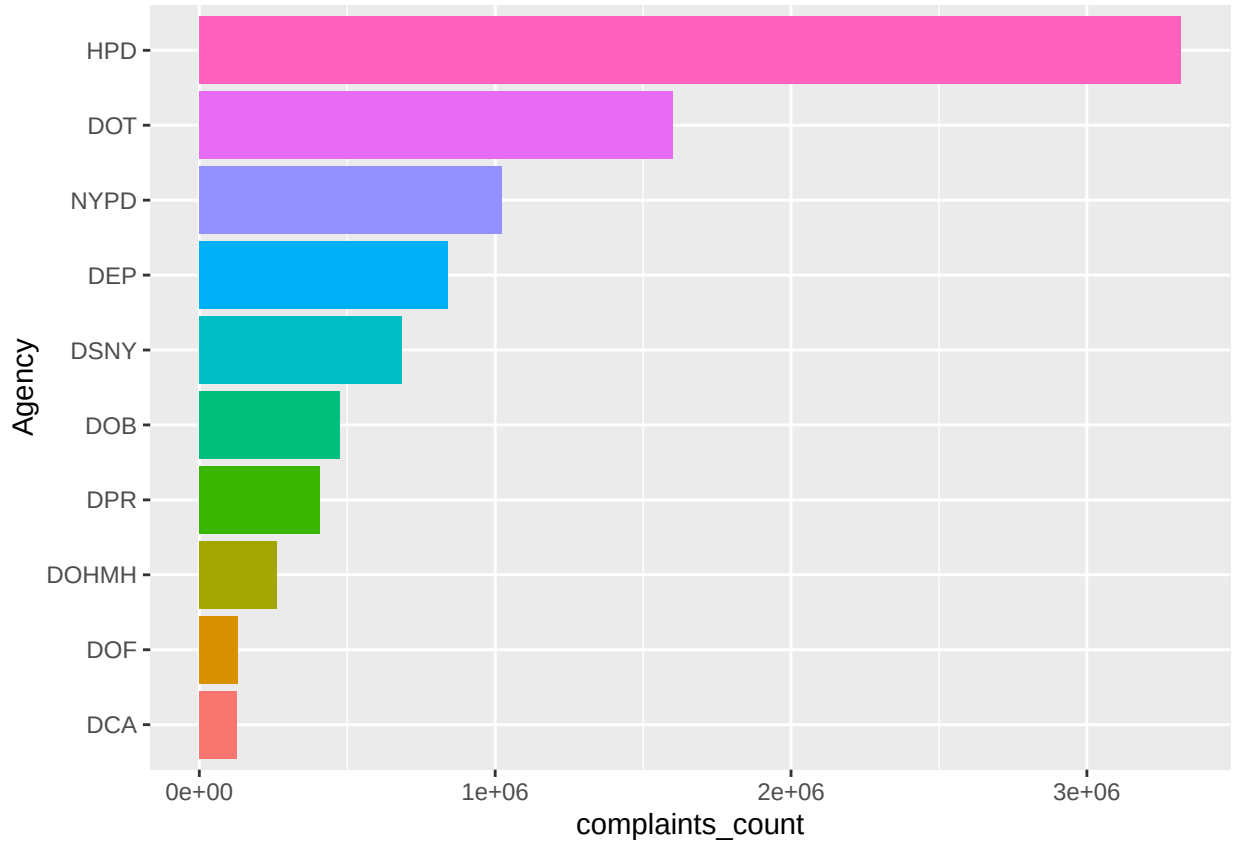
```
bigAgency
```

```
## # A tibble: 10 x 3
##    Agency complaints_count `proportion in %`
##    <chr>             <int>             <dbl>
##  1 HPD             3319073              36.4
##  2 DOT             1601604              17.6
```

```
##  3 NYPD            1023154            11.2
##  4 DEP              838855             9.19
##  5 DSNY             686032             7.52
##  6 DOB              475674             5.21
##  7 DPR              405939             4.45
##  8 DOHMH            263568             2.89
##  9 DOF              129615             1.42
## 10 DCA              125554             1.38
```

Then we visualize it using a bar chart.

```
bigAgency$Agency<-factor(bigAgency$Agency,
  levels=bigAgency$Agency[order(bigAgency$complaints_count)])

p<-ggplot(bigAgency,aes(x=Agency,y=complaints_count, fill=Agency)) +
   geom_bar(stat="identity", show.legend = FALSE) +
   coord_flip()
p
```



Conclusion

Hence, we applied exploratory data analysis on the `nyc311` data using various data transformation and visualization techniques. We were able to answer several questions about the data like:

- There are only 3 important status of complaints to consider in the data.
- There is a wide variance of the top complaint categories across different Boroughs.
- Among the top 10 complaints category, most of them are closed with a very few pending cases.