Homework VI: Communicating Findings from the 311 Dataset

Heather Akers-Healy

2019-12-09

Introduction

This homework assignment is the final assignment in R, where we communicate our findings. Our datasets are the 311 call data from New York City and another dataset of social indicators that can be matched with the 311 call data by the name of the borough.

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)
library(data.table)
library(formatR)
library(kableExtra)
library(knitr)
```

```
nyc311<-fread("/Users/heather/Downloads/311_Service_Requests_from_2010_to_Present.csv")
names(nyc311)<-names(nyc311) %>%
  stringr::str_replace_all("\\s", ".")
```

Tidying up and joining the datasets

Both datasets must be cleaned and reformatted. Below I re-create the steps I took for previous assignments and complete the join.

Reformatting steps from previous assignments

```
# nyc311 data

# fixing date columns
nyc311 <- nyc311 %>% separate(Created.Date, c("Created.Date", "Created.Time"),
    sep = "\\s", extra = "merge")

nyc311 <- nyc311 %>% separate(Closed.Date, c("Closed.Date", "Closed.Time"),
    sep = "\\s", extra = "merge")
nyc311 <- nyc311 %>% separate(Resolution.Action.Updated.Date, c("Resolution.Action.Updated.Date",
    "Resolution.Action.Updated.Time"), sep = "\\s", extra = "merge") %>% separate(Due.Date,
```

```r
    c("Due.Date", "Due.Time"), sep = "\\s", extra = "merge")
# remove location column with duplicate data
nyc311 <- nyc311 %>% select(-c(Location))
# remove the city column with unreliable data
nyc311 <- nyc311 %>% select(-c(City))
# formatting NAs
nyc311[nyc311 == ""] <- NA
nyc311[nyc311 == "N/A"] <- NA
nyc311[nyc311 == "Unspecified"] <- NA
```

```r
# social indicator data

social_ind <- fread("https://data.cityofnewyork.us/api/views/8ek7-jxw6/rows.csv",
    header = TRUE)

# select borough data
borough_ind <- filter(social_ind, `Dimension Category` == "Borough")
# rename borough column
colnames(borough_ind)[colnames(borough_ind) == "Dimension"] <- "Borough"
# create reference dataframe from columns that are not necessary for main
# dataset
reference_df <- data.frame(`Dimension Category` = borough_ind$`Dimension Category`,
    Domain = borough_ind$Domain, Indicator = borough_ind$Indicator, Defintion = borough_ind$Defintion,
    Source = borough_ind$Source, Notes = borough_ind$Notes)
reference_no_dupe <- distinct(reference_df)
# delete reference columns from main dataset
drops <- c("Dimension Category", "Domain", "Defintion", "Source", "Notes")
borough_ind <- borough_ind[, !(names(borough_ind) %in% drops)]
# gather year columns
borough_ind <- borough_ind %>% gather(`2000`, `2001`, `2002`, `2003`, `2004`,
    `2005`, `2006`, `2007`, `2008`, `2009`, `2010`, `2011`, `2012`, `2013`,
    `2014`, `2015`, `2016`, `2017`, key = "Year", value = "percent")
# spread indicator values into columns
borough_ind <- spread(borough_ind, key = Indicator, value = percent)
names(borough_ind) <- names(borough_ind) %>% stringr::str_replace_all("\\s",
    ".")
# reformat NAs
borough_ind[borough_ind == "*"] <- NA
```

```r
#final steps for join
borough_ind <- borough_ind %>%
      mutate(Borough = toupper(Borough))
#add a year column from the Created.Date
nyc311 <- nyc311 %>%
    mutate(Year = substring(Created.Date,7,10))
```

```r
# remove all % and $ signs
borough_ind[] <- lapply(borough_ind, gsub, pattern = "$", replacement = "")
borough_ind[] <- lapply(borough_ind, gsub, pattern = "%", replacement = "")
# reformat numbers columns to numbers
borough_ind$Admissions.to.Department.of.Correction.per.1000 <- as.numeric(borough_ind$Admissions.to.Depa
borough_ind$Average.Weekly.Earnings <- as.numeric(borough_ind$Average.Weekly.Earnings)
borough_ind$Curbside.and.Containerized.Diversion.Rate <- as.numeric(borough_ind$Curbside.and.Containeri
```

```r
borough_ind$Department.of.Probation.Population <- as.numeric(borough_ind$Department.of.Probation.Popula
borough_ind$Department.of.Probation.Population.per.1000 <- as.numeric(borough_ind$Department.of.Probati
borough_ind$`Disconnected.Youth,.16-24.Not.at.Work.&.Not.in.School` <- as.numeric(borough_ind$`Disconne
borough_ind$Eligible.Voter.Registration.Rate <- as.numeric(borough_ind$Eligible.Voter.Registration.Rate
borough_ind$`Enrolled.Full.Day.Pre-K.Students` <- as.numeric(borough_ind$`Enrolled.Full.Day.Pre-K.Studer
borough_ind$`Four-Year.High.School.Graduation.Rate` <- as.numeric(borough_ind$`Four-Year.High.School.Gra
borough_ind$Labor.Force.Participation.Rate <- as.numeric(borough_ind$Labor.Force.Participation.Rate)
borough_ind$Mean.Travel.Time.to.Work <- as.numeric(borough_ind$Mean.Travel.Time.to.Work)
borough_ind$New.York.City.Households.with.Internet.Access <- as.numeric(borough_ind$New.York.City.House
borough_ind$New.Yorkers.Living.Within.Walking.Distance.of.a.Park <- as.numeric(borough_ind$New.Yorkers.
borough_ind$`Notices.of.Foreclosure.Rate.per.1,000.(1-4.family.&.condo.properties)` <- as.numeric(borou
borough_ind$Number.of.Jobs.in.the.City <- as.numeric(borough_ind$Number.of.Jobs.in.the.City)
borough_ind$`Outdoor.Air.Pollution/Fine.Particulate.Matter.(PM2.5).Levels` <- as.numeric(borough_ind$`Ou
borough_ind$Percent.of.Adults.with.Serious.Psychological.Distress.Who.Received.Mental.Health.Treatment
borough_ind$Rental.Housing.Vacancy.Rate <- as.numeric(borough_ind$Rental.Housing.Vacancy.Rate)
borough_ind$`Serious.Housing.Code.Violations.(per.1,000.privately.owned.rental.units)` <- as.numeric(bo
borough_ind$`Severely.Rent-Burdened.Households` <- as.numeric(borough_ind$`Severely.Rent-Burdened.House
borough_ind$Total.Supplemental.Nutrition.Assistance.Program.Recipients <- as.numeric(borough_ind$Total.
borough_ind$`Turnout.Among.Registered.Voters.(percent)` <- as.numeric(borough_ind$`Turnout.Among.Registe
borough_ind$`Turnout.Among.Voting.Age.Population.(percent)` <- as.numeric(borough_ind$`Turnout.Among.Vo
borough_ind$Unemployment.Rate <- as.numeric(borough_ind$Unemployment.Rate)
borough_ind$Violent.Crime.per.1000 <- as.numeric(borough_ind$Violent.Crime.per.1000)
borough_ind$Violent.Victimization.of.Youth.per.1000 <- as.numeric(borough_ind$Violent.Victimization.of.
```

```r
#shortening the names of some columns for displaying the data dictionary and tables
colnames(borough_ind)[15] <- 'Living.Walking.Distance.Park'
colnames(borough_ind)[18] <- 'Outdoor.Air.Pollution'
colnames(borough_ind)[21] <- "Srs.Housing.Code.Violations"
colnames(borough_ind)[19] <- "Pct.Adults.Psych.Distress"
colnames(borough_ind)[16] <- "Foreclosure.Rate.per.1000"
colnames(borough_ind)[23] <- "Total.SNAP.Recipients"
colnames(borough_ind)[25] <- "Turnout.Voting.Age.Pop.(pct)"
colnames(borough_ind)[14] <- "Households.w.Internet.Access"
colnames(borough_ind)[3] <- "Corrections.Pop.per.1000"
colnames(borough_ind)[6] <- "Probation.Population"
colnames(borough_ind)[7] <- "Probation.Pop.per.1000"
colnames(borough_ind)[8] <- "Disconnected.Youth"
colnames(borough_ind)[28] <- "Violent.Victimization.Youth.per.1000"
#colnames(borough_ind)
```

Creating a clean join with a summary dataset

As discussed previously, some caution would need to be used in doing analysis with a set joined between the 311 call data and the social indicators data because it would contain two kinds data of per row (values for the specific 311 call, and generalized social data for the borough by year.) In order to do analysis on a clean joined set, I will join the social indicators set with a summary 311 dataset with the top 5 complaints and agencies, grouped by the borough and year. A joined dataset with the social indicators will allow me to look for relationships between the call data and possible causes for some of the patterns seen there.

```r
nyc311_summary <- nyc311 %>% group_by(Borough, Year) %>% summarize(Calls.Total = n(),
    Calls.HPD = sum(str_count(Agency, "HPD")), Calls.DOT = sum(str_count(Agency,
        "DOT")), Calls.NYPD = sum(str_count(Agency, "NYPD")), Calls.DEP = sum(str_count(Agency,
```

```r
        "DEP")), Calls.DSNY = sum(str_count(Agency, "DSNY")), Calls.Closed = sum(str_count(Status,
        "Closed")), Calls.Open = sum(str_count(Status, "Open")), Calls.Pending = sum(str_count(Status,
        "Pending")), Calls.Assigned = sum(str_count(Status, "Assigned")), Cmplts.Heating = sum(str_coun
        "HEATING")), Cmplts.StreetCondition = sum(str_count(Complaint.Type,
        "Street Condition")), Cmplts.StreetLight = sum(str_count(Complaint.Type,
        "Street Light Condition")), Cmplts.GenConstruction = sum(str_count(Complaint.Type,
        "GENERAL CONSTRUCTION")), Cmplts.Plumbing = sum(str_count(Complaint.Type,
        "PLUMBING")))
```

Creating the joined dataset with the nyc311 summary data and the social indicators data matched on Borough and Year:

```r
nyc_summary_joined <- left_join(nyc311_summary, borough_ind, by = c("Borough", "Year"))
```

```r
head(nyc_summary_joined)
```

```
## # A tibble: 6 x 43
## # Groups:   Borough [1]
##   Borough Year  Calls.Total Calls.HPD Calls.DOT Calls.NYPD Calls.DEP
##   <chr>   <chr>       <int>     <int>     <int>      <int>     <int>
## 1 BRONX   2003         1907         0      1877          0         0
## 2 BRONX   2004          808         0       705          2         0
## 3 BRONX   2005            7         0         0          0         0
## 4 BRONX   2006          374         0         1          0         0
## 5 BRONX   2007          434         0         0          1         0
## 6 BRONX   2008          631         0         7         10         0
## # ... with 36 more variables: Calls.DSNY <int>, Calls.Closed <int>,
## #   Calls.Open <int>, Calls.Pending <int>, Calls.Assigned <int>,
## #   Cmplts.Heating <int>, Cmplts.StreetCondition <int>,
## #   Cmplts.StreetLight <int>, Cmplts.GenConstruction <int>,
## #   Cmplts.Plumbing <int>, Corrections.Pop.per.1000 <dbl>,
## #   Average.Weekly.Earnings <dbl>,
## #   Curbside.and.Containerized.Diversion.Rate <dbl>,
## #   Probation.Population <dbl>, Probation.Pop.per.1000 <dbl>,
## #   Disconnected.Youth <dbl>, Eligible.Voter.Registration.Rate <dbl>,
## #   `Enrolled.Full.Day.Pre-K.Students` <dbl>,
## #   `Four-Year.High.School.Graduation.Rate` <dbl>,
## #   Labor.Force.Participation.Rate <dbl>, Mean.Travel.Time.to.Work <dbl>,
## #   Households.w.Internet.Access <dbl>,
## #   Living.Walking.Distance.Park <dbl>, Foreclosure.Rate.per.1000 <dbl>,
## #   Number.of.Jobs.in.the.City <dbl>, Outdoor.Air.Pollution <dbl>,
## #   Pct.Adults.Psych.Distress <dbl>, Rental.Housing.Vacancy.Rate <dbl>,
## #   Srs.Housing.Code.Violations <dbl>,
## #   `Severely.Rent-Burdened.Households` <dbl>,
## #   Total.SNAP.Recipients <dbl>,
## #   `Turnout.Among.Registered.Voters.(percent)` <dbl>,
## #   `Turnout.Voting.Age.Pop.(pct)` <dbl>, Unemployment.Rate <dbl>,
## #   Violent.Crime.per.1000 <dbl>,
## #   Violent.Victimization.Youth.per.1000 <dbl>
```

Findings

Findings from the NYC311 call data

What are the top complaints?

A sorted table of complaint types with over 100,000 calls. Heating issues generate the most complaints, followed by street condition and street light condition.

```
top_complaints <- narrow %>%
  group_by(Complaint.Type) %>%
  summarize(count=n()) %>% filter(count>100000)
knitr::kable((arrange(top_complaints, desc(count))), format = "markdown")
```

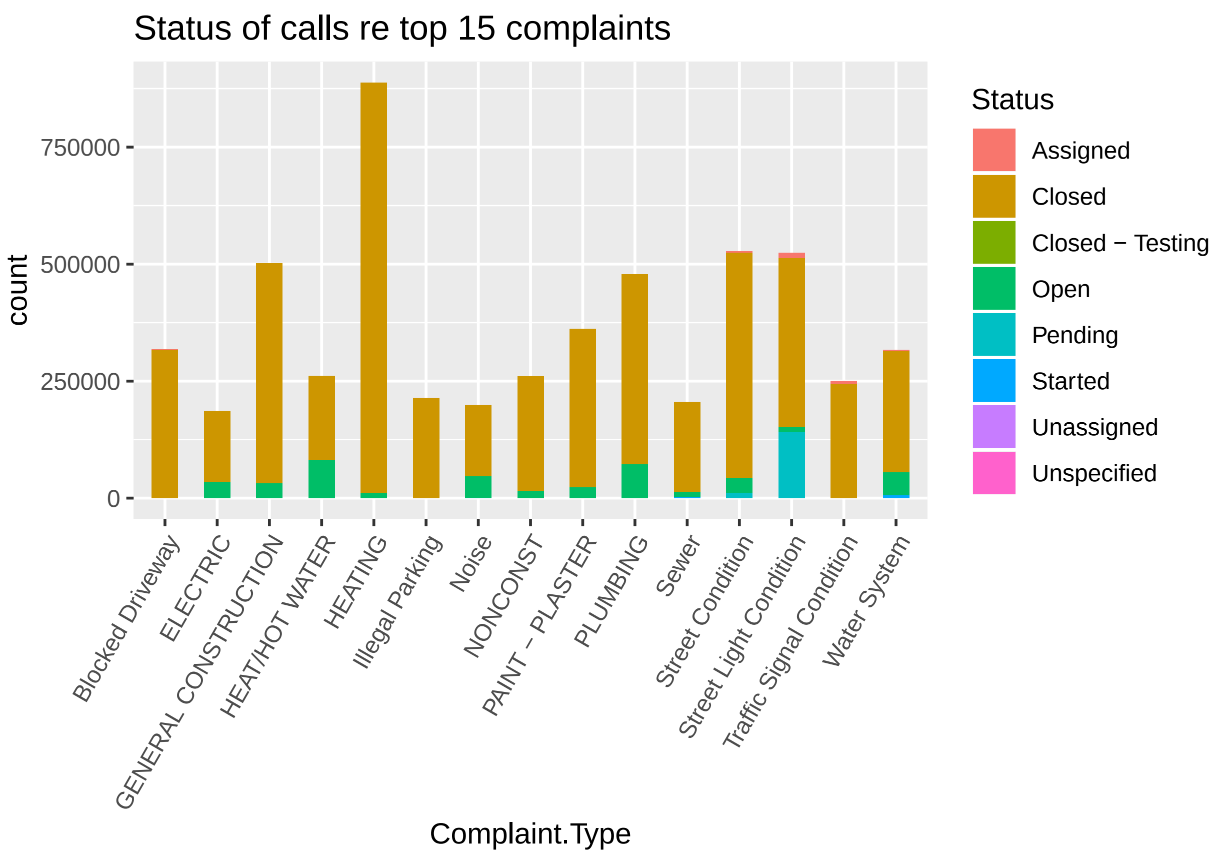
Complaint.Type	count
HEATING	887675
Street Condition	526797
Street Light Condition	524501
GENERAL CONSTRUCTION	501514
PLUMBING	478875
PAINT - PLASTER	361449
Blocked Driveway	317163
Water System	317075
HEAT/HOT WATER	260936
NONCONST	260405
Traffic Signal Condition	250781
Illegal Parking	214134
Sewer	205106
Noise	198803
ELECTRIC	185966
Dirty Conditions	178920
Damaged Tree	156007
General Construction/Plumbing	155307
Building/Use	140001
Sanitation Condition	135026
Noise - Commercial	134222
Rodent	124020
Broken Muni Meter	121629
Noise - Street/Sidewalk	119540
DOF Literature Request	110859
Consumer Complaint	104675
Taxi Complaint	100767

What is the status of calls in the top 15 complaints?

```
top_15compl <- narrow[which(narrow$Complaint.Type == "HEATING" | narrow$Complaint.Type ==
    "Street Condition" | narrow$Complaint.Type == "Street Light Condition" |
    narrow$Complaint.Type == "GENERAL CONSTRUCTION" | narrow$Complaint.Type ==
    "PLUMBING" | narrow$Complaint.Type == "PAINT - PLASTER" | narrow$Complaint.Type ==
    "Blocked Driveway" | narrow$Complaint.Type == "Water System" | narrow$Complaint.Type ==
```

```
    "HEAT/HOT WATER" | narrow$Complaint.Type == "NONCONST" | narrow$Complaint.Type ==
    "Traffic Signal Condition" | narrow$Complaint.Type == "Illegal Parking" |
    narrow$Complaint.Type == "Sewer" | narrow$Complaint.Type == "Noise" | narrow$Complaint.Type ==
    "ELECTRIC"), ]
```

```
top_15compl_plot <- ggplot(top_15compl, aes(Complaint.Type))
top_15compl_plot + geom_bar(aes(fill = Status), width = 0.5) + labs(title = "Status of calls re top 15 c
    theme(axis.text.x = element_text(angle = 60, hjust = 1))
```



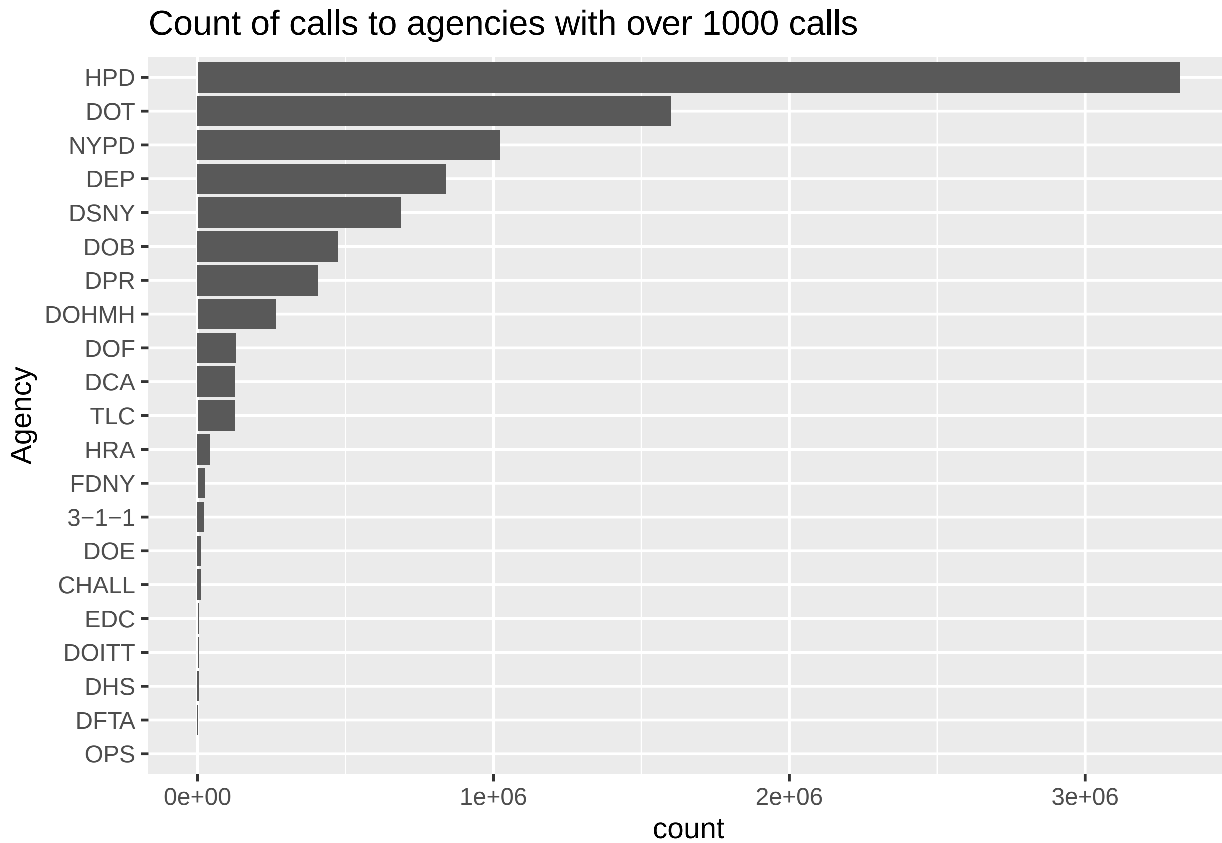
Most calls regarding the top 15 complaint categories have been closed.

Which top complaints are most associated with which agencies?

From previous homework, we determined that the Department of Housing Preservation and Development ("HPD"), Department of Transportation ("DOT"), the New York Police Department ("NYPD") and the Department of Environmental Protection ("DEP"") had the top 4 most calls per agency.

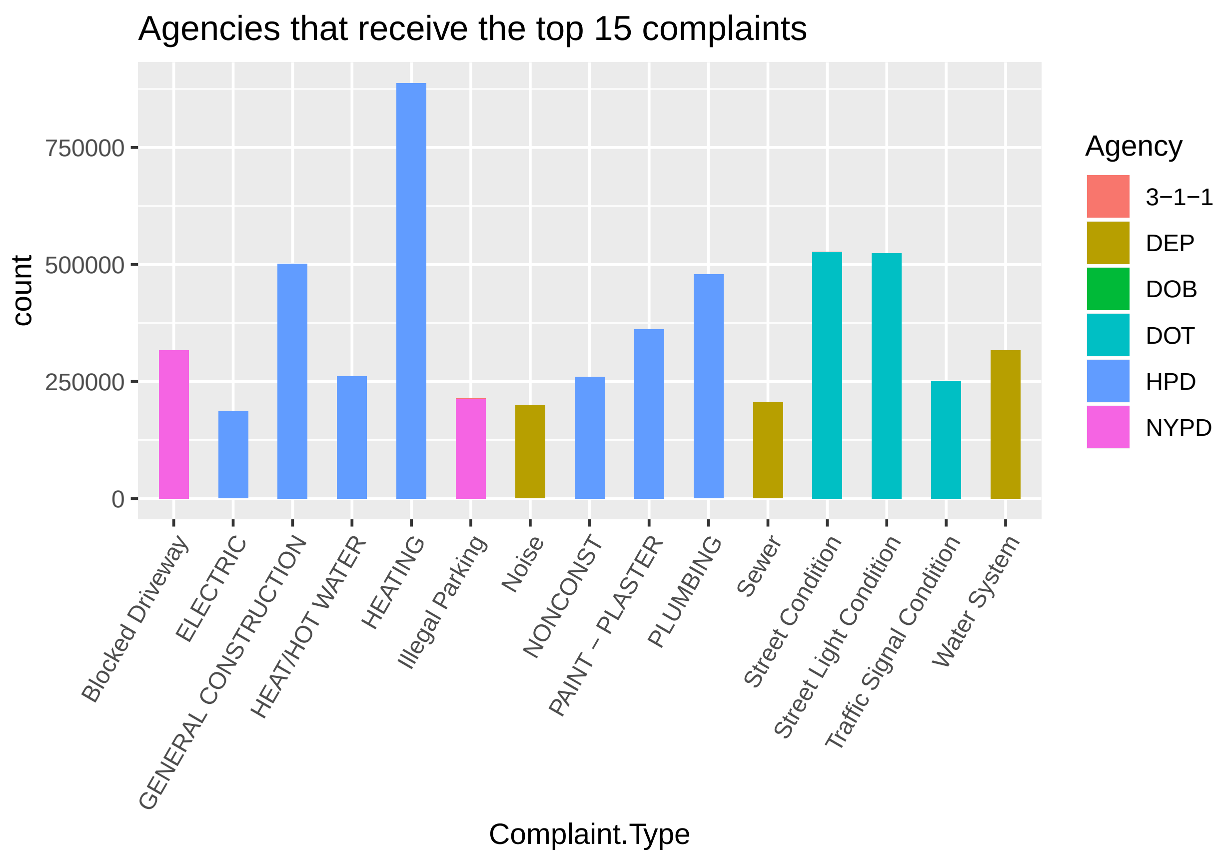
Here again is the plot of agencies that received over 1000 calls.

```
bigAgency <- narrow %>% group_by(Agency) %>% summarize(count = n()) %>% filter(count >
    1000)
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() + labs(title = "Count of calls to agencies with over 1000 calls")
p
```



Of the top fifteen complaint types, let's see which are routed most often to which agency.

```
top_15compl_ag<-ggplot(top_15compl, aes(Complaint.Type))
top_15compl_ag + geom_bar(aes(fill=Agency), width = 0.5) + labs(title="Agencies that receive the top 15
```



DOT gets routed complaints about traffic light, street and street light conditions. HPD receives complaints regarding housing conditions (plumbing, heating, electric, etc.) NYPD receives a large amount of complaints about blocked driveways and illegal parking. DEP handles water system and sewer complaints. Most of the top 15 complaint types are routed to these top 4 agencies.

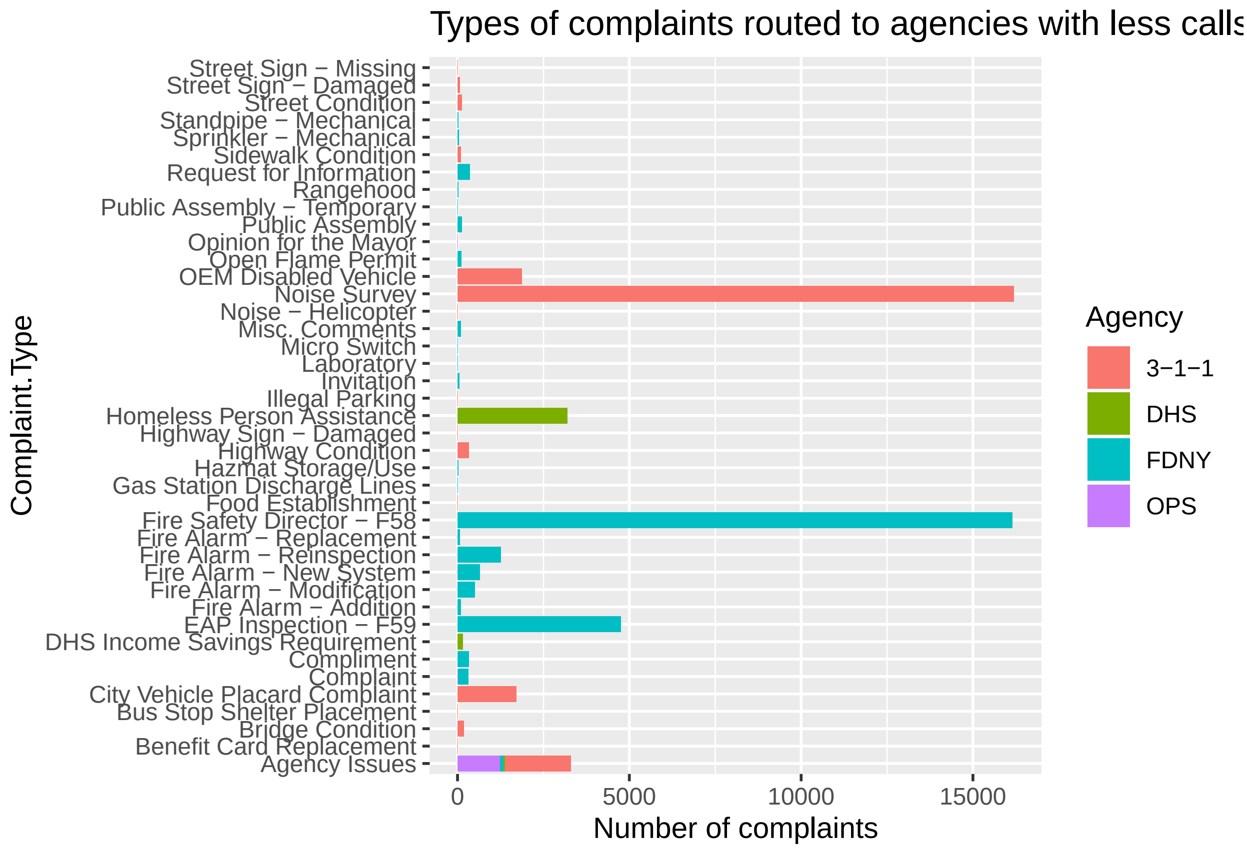
What about agencies with less calls?

Four of the big agencies that took the least amount of calls were the Fire Department ("FDNY"), 3—1—1, Department of Homeless Services ("DHS") and Mayor's Office of Operations ("OPS"). What kinds of calls did they take?

```
less_call_agencies <- narrow[which(narrow$Agency == "FDNY" | narrow$Agency ==
    "3-1-1" | narrow$Agency == "DHS" | narrow$Agency == "OPS"), ]

compl_less_call_agencies_sum <- less_call_agencies %>% group_by(Complaint.Type,
    Agency) %>% tally()
```

```
less_calls<-ggplot(compl_less_call_agencies_sum, aes(x = Complaint.Type, y = n, fill=Agency)) +
      geom_bar(stat="identity") +
      labs(title="Types of complaints routed to agencies with less calls") +
      coord_flip() + ylab("Number of complaints")
less_calls
```

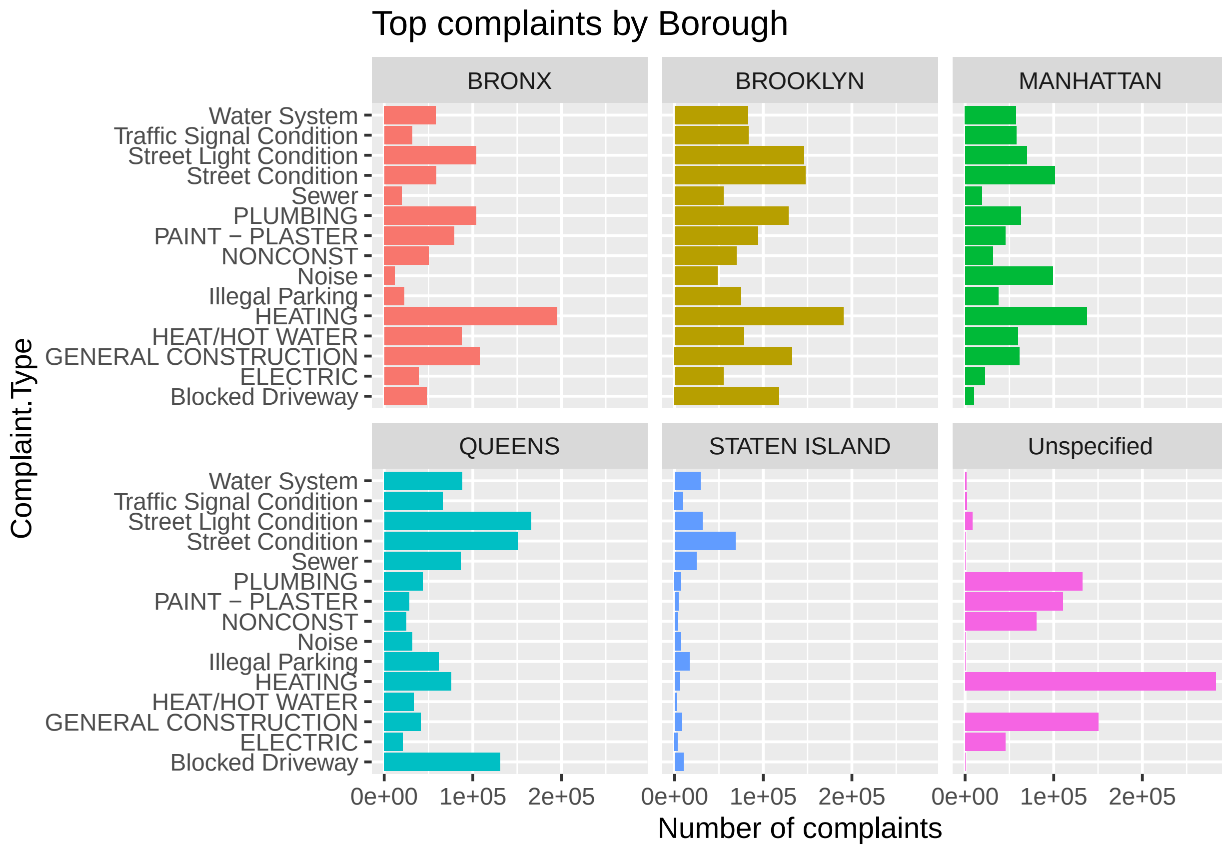


3-1-1 tends to mostly get calls about noise surveys. DHS receives calls regarding assistance with homeless persons, and OPS regarding agency issues. The bulk of fire department calls are about building certifications from the fire safety director.

Which Boroughs are most associated with which top fifteen complaints?

```
top15_compl_borough_sum <- top_15compl %>%
      group_by(Complaint.Type, Borough) %>%
      tally()
```

```
grid<-ggplot(top15_compl_borough_sum,aes(x = Complaint.Type, y = n, fill=Borough)) + geom_bar(stat="ider
grid + facet_wrap(~Borough) + coord_flip() + theme(legend.position="none") + ylab("Number of complaints
```



Heating is the top complaint in Brooklyn, the Bronx, and Manhattan. Street light condition is the top complaint type from Queens, while street condition is confirmed as the most common complaint from Staten Island. The majority of heating and general construction complaints were not assigned to a borough.

Relationships between social indicators and top complaint types

Using the scale at http://www.statstutor.ac.uk/resources/uploaded/pearsons.pdf to evaluate the strength of correlations, I will examine the relationships between the top complaint types and social indicators (in related groups where possible).

A corrplot of social indicators and the complaint types gives a quick overview of relationships. The corrplot is split into two graphs due to the number of social indicators available.

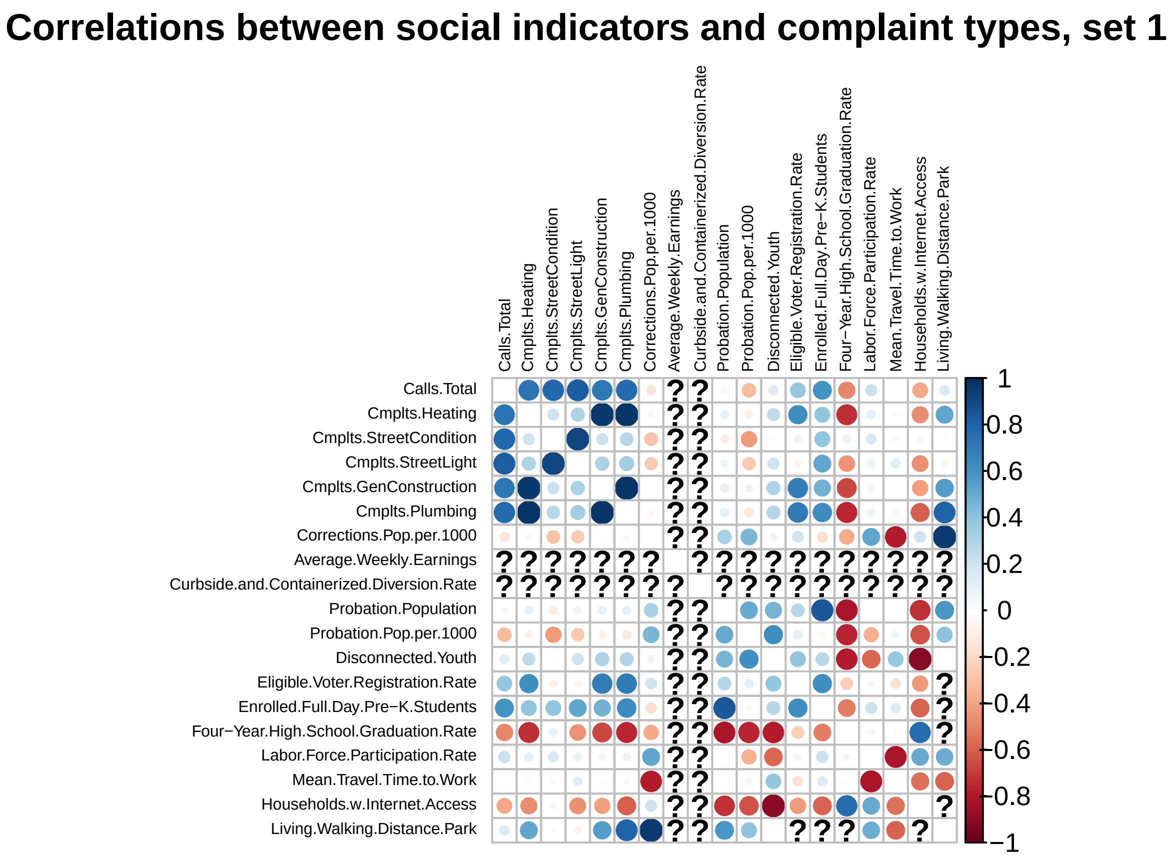
```
numeric_data_cols_1 <- subset(nyc_summary_joined, select = c("Calls.Total",
    "Cmplts.Heating", "Cmplts.StreetCondition", "Cmplts.StreetLight", "Cmplts.GenConstruction",
    "Cmplts.Plumbing", "Corrections.Pop.per.1000", "Average.Weekly.Earnings",
    "Curbside.and.Containerized.Diversion.Rate", "Probation.Population", "Probation.Pop.per.1000",
    "Disconnected.Youth", "Eligible.Voter.Registration.Rate", "Enrolled.Full.Day.Pre-K.Students",
    "Four-Year.High.School.Graduation.Rate", "Labor.Force.Participation.Rate",
    "Mean.Travel.Time.to.Work", "Households.w.Internet.Access", "Living.Walking.Distance.Park"))



numeric_data_1 <- cor(numeric_data_cols_1, use = "pairwise.complete.obs")

library(corrplot)
```

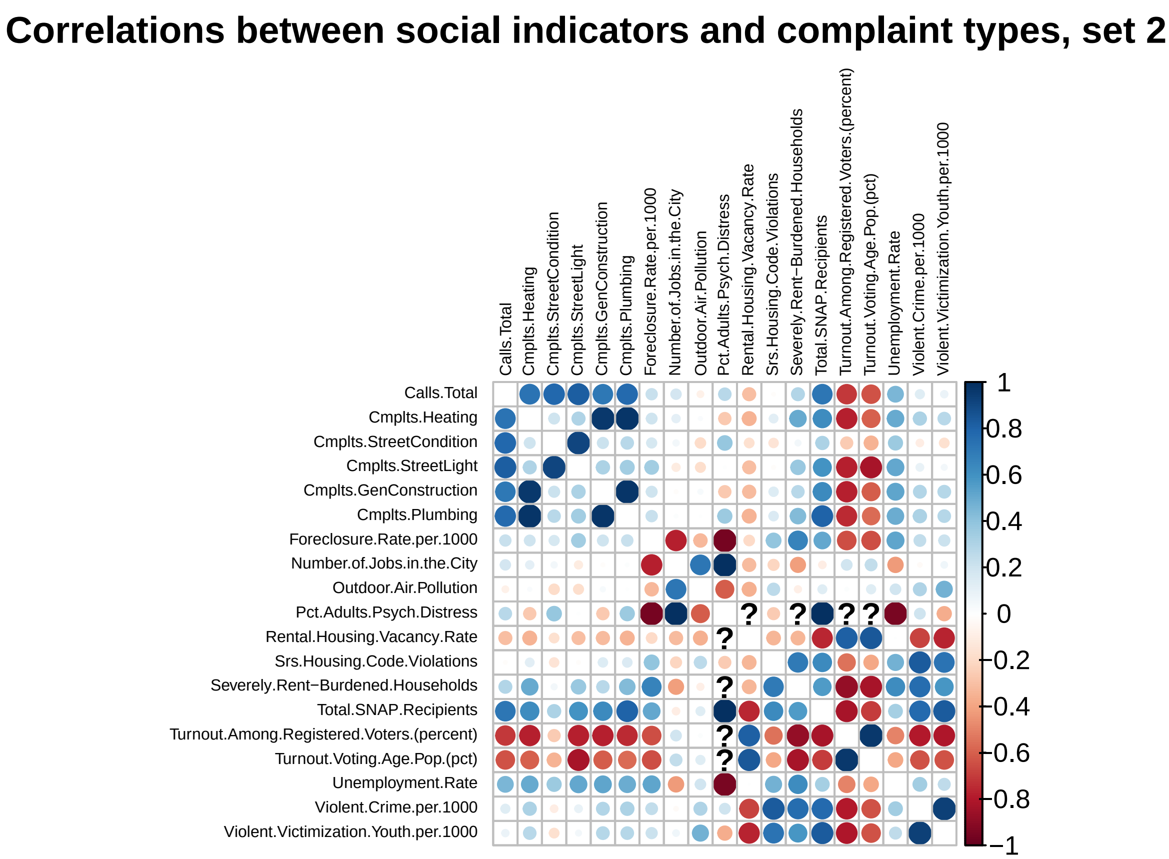
```r
plot.new()
corrplot(numeric_data_1, method = "circle", tl.cex = 0.5, tl.col = "black",
    diag = FALSE, title = "Correlations between social indicators and complaint types, set 1",
    mar = c(0, 0, 2, 0))
```




```r
numeric_data_cols_2 <- subset(nyc_summary_joined, select = c("Calls.Total",
    "Cmplts.Heating", "Cmplts.StreetCondition", "Cmplts.StreetLight", "Cmplts.GenConstruction",
    "Cmplts.Plumbing", "Foreclosure.Rate.per.1000", "Number.of.Jobs.in.the.City",
    "Outdoor.Air.Pollution", "Pct.Adults.Psych.Distress", "Rental.Housing.Vacancy.Rate",
    "Srs.Housing.Code.Violations", "Severely.Rent-Burdened.Households", "Total.SNAP.Recipients",
    "Turnout.Among.Registered.Voters.(percent)", "Turnout.Voting.Age.Pop.(pct)",
    "Unemployment.Rate", "Violent.Crime.per.1000", "Violent.Victimization.Youth.per.1000"))

numeric_data_2 <- cor(numeric_data_cols_2, use = "pairwise.complete.obs")

plot.new()
corrplot(numeric_data_2, method = "circle", tl.cex = 0.5, tl.col = "black",
    diag = FALSE, title = "Correlations between social indicators and complaint types, set 2",
    mar = c(0, 0, 2, 0))
```



There are quite a few strong relationships between the top complaints and the social indicators. For completeness I will examine all of the relationships below.

Environmental indicators and top complaint types:

```
env_cols <- subset(nyc_summary_joined, select = c("Calls.Total", "Cmplts.Heating",
    "Cmplts.StreetCondition", "Cmplts.StreetLight", "Cmplts.GenConstruction",
    "Cmplts.Plumbing", "Mean.Travel.Time.to.Work", "Living.Walking.Distance.Park",
    "Outdoor.Air.Pollution"))
env_cor <- cor(env_cols, use = "pairwise.complete.obs")
knitr::kable(env_cor[1:6, 7:9], format = "markdown", caption = "Correlations between complaint types and environmental factors") %>%
    kable_styling(font_size = 7) %>% landscape()
```

	Mean.Travel.Time.to.Work	Living.Walking.Distance.Park	Outdoor.Air.Pollution
Calls.Total	-0.0093387	0.1506594	-0.0702156
Cmplts.Heating	-0.0298254	0.5211594	0.0207513
Cmplts.StreetCondition	0.0351557	0.0201730	-0.1879323
Cmplts.StreetLight	0.1231961	-0.0643121	-0.1716714
Cmplts.GenConstruction	0.0228220	0.5566144	0.0370391
Cmplts.Plumbing	0.0373031	0.8095501	0.0056554

Mean travel time to work and the level of pollution do not have a signification relationship with any of the top complaint types. The amount of people living within walking distance of a park has a very strong, positive correlation with plumbing complaints, a moderate relationship to heating and general construction complaints, and a weak correlation with the total calls.

Housing indicators:

```
housing_cols <- subset(nyc_summary_joined, select = c("Calls.Total", "Cmplts.Heating",
    "Cmplts.StreetCondition", "Cmplts.StreetLight", "Cmplts.GenConstruction",
    "Cmplts.Plumbing", "Foreclosure.Rate.per.1000", "Rental.Housing.Vacancy.Rate",
    "Srs.Housing.Code.Violations"))
housing_cor <- cor(housing_cols, use = "pairwise.complete.obs")
knitr::kable(t(housing_cor[7:9, 1:6]), format = "markdown", caption = "Correlations between complaint types and housing factors") %>%
    kable_styling(font_size = 7) %>% landscape()
```

	Foreclosure.Rate.per.1000	Rental.Housing.Vacancy.Rate	Srs.Housing.Code.Violations
Calls.Total	0.2212228	-0.3031887	-0.0161389
Cmplts.Heating	0.2020596	-0.3479525	0.1267841
Cmplts.StreetCondition	0.1766502	-0.1604710	-0.1495396
Cmplts.StreetLight	0.3410168	-0.3005602	-0.0178389
Cmplts.GenConstruction	0.2087250	-0.3100611	0.1491176
Cmplts.Plumbing	0.2251227	-0.3478510	0.1540689

The foreclosure rate has only a weak to very weak positive correlation with the total calls and the top complaints. All of the top complaint types have a weak to very weak, negative correlation to the vacancy rate, with plumbing and heating the strongest.

Somewhat surprisingly, none of the top complaint types appear to have any significant correlation with serious housing code violations. It's possible that not all housing code violations are being reported correctly from the 311 calls. If I were working for the city, I might recommend reviewing that process.

Psychiatric distress: (Note that for correlation tables with only one social indicator, only the last row and column are relevant.)

```
mnt_hth_cols <- subset(nyc_summary_joined, select = c("Calls.Total", "Cmplts.Heating",
    "Cmplts.StreetCondition", "Cmplts.StreetLight", "Cmplts.GenConstruction",
    "Cmplts.Plumbing", "Pct.Adults.Psych.Distress"))
mnt_hth_cor <- cor(mnt_hth_cols, use = "pairwise.complete.obs")
knitr::kable((mnt_hth_cor), format = "markdown", caption = "Correlations between complaint types and psychological distress") %>%
    kable_styling(font_size = 7) %>% landscape()
```

	Calls.Total	Cmplts.Heating	Cmplts.StreetCondition	Cmplts.StreetLight	Cmplts.GenConstruction	Cmplts.Plumbing	Pct.Adults.Psych.Distress
Calls.Total	1.0000000	0.7304898	0.7872922	0.8216617	0.7128808	0.7724521	0.2756771
Cmplts.Heating	0.7304898	1.0000000	0.2049223	0.3050547	0.9691566	0.9761679	-0.2673310
Cmplts.StreetCondition	0.7872922	0.2049223	1.0000000	0.9131864	0.2132996	0.2739035	0.3829542
Cmplts.StreetLight	0.8216617	0.3050547	0.9131864	1.0000000	0.3174865	0.3498641	0.0099146
Cmplts.GenConstruction	0.7128808	0.9691566	0.2132996	0.3174865	1.0000000	0.9784977	-0.2673310
Cmplts.Plumbing	0.7724521	0.9761679	0.2739035	0.3498641	0.9784977	1.0000000	0.3642663
Pct.Adults.Psych.Distress	0.2756771	-0.2673310	0.3829542	0.0099146	-0.2673310	0.3642663	1.0000000

Heating and general construction complaints have a weak negative relationship with the percent of adults seeking treatment for psychiatric distress, while total calls, street condition and plumbing have a weak positive relationship.

Crime factors:

```
crime_cols <- subset(nyc_summary_joined, select = c("Calls.Total", "Cmplts.Heating",
    "Cmplts.StreetCondition", "Cmplts.StreetLight", "Cmplts.GenConstruction",
    "Cmplts.Plumbing", "Violent.Crime.per.1000", "Violent.Victimization.Youth.per.1000",
    "Corrections.Pop.per.1000", "Probation.Population", "Probation.Pop.per.1000"))
crime_cor <- cor(crime_cols, use = "pairwise.complete.obs")
knitr::kable(t(crime_cor[7:11, 1:6]), format = "markdown", caption = "Correlations between complaint types and crime factors") %>%
    kable_styling(font_size = 7) %>% landscape()
```

	Violent.Crime.per.1000	Violent.Victimization.Youth.per.1000	Corrections.Pop.per.1000	Probation.Population	Probation.Pop.per.1000
Calls.Total	0.1382545	0.0857901	-0.1379565	0.0451072	-0.3194636
Cmplts.Heating	0.3107015	0.2792147	0.0404659	0.1016801	-0.0724468
Cmplts.StreetCondition	-0.0929067	-0.1614418	-0.2832668	-0.0946958	-0.4238912
Cmplts.StreetLight	0.0975791	0.0523063	-0.2573954	0.0770618	-0.2668317
Cmplts.GenConstruction	0.2941138	0.2870008	-0.0140354	0.1048845	-0.0754074
Cmplts.Plumbing	0.3191177	0.2804206	-0.0474153	0.1078841	-0.1220950

The violent crime and violent youth victimization rates have only a weak correlation to plumbing, heating, and general construction complaints. The rate of people per 1000 sent to prison has a weak, negative correlation to street condition and street light condition complaints. The rate of the population per 1000 on probation has a weak, negative correlation to the total calls and complaints about street condition and street light condition.

Disconnected youth:

```
dis_youth_cols <- subset(nyc_summary_joined, select = c("Calls.Total", "Cmplts.Heating",
    "Cmplts.StreetCondition", "Cmplts.StreetLight", "Cmplts.GenConstruction",
    "Cmplts.Plumbing", "Disconnected.Youth"))
dis_youth_cor <- cor(dis_youth_cols, use = "pairwise.complete.obs")
knitr::kable(dis_youth_cor, format = "markdown", caption = "Correlations between complaint types and disconnected youth") %>%
    kable_styling(font_size = 7) %>% landscape()
```

	Calls.Total	Cmplts.Heating	Cmplts.StreetCondition	Cmplts.StreetLight	Cmplts.GenConstruction	Cmplts.Plumbing	Disconnected.Youth
Calls.Total	1.0000000	0.7304898	0.7872922	0.8216617	0.7128808	0.7724521	0.1358809
Cmplts.Heating	0.7304898	1.0000000	0.2049223	0.3050547	0.9691566	0.9761679	0.2558621
Cmplts.StreetCondition	0.7872922	0.2049223	1.0000000	0.9131864	0.2132996	0.2739035	0.0154292
Cmplts.StreetLight	0.8216617	0.3050547	0.9131864	1.0000000	0.3174865	0.3498641	0.2097736
Cmplts.GenConstruction	0.7128808	0.9691566	0.2132996	0.3174865	1.0000000	0.9784977	0.3083259
Cmplts.Plumbing	0.7724521	0.9761679	0.2739035	0.3498641	0.9784977	1.0000000	0.2943788
Disconnected.Youth	0.1358809	0.2558621	0.0154292	0.2097736	0.3083259	0.2943788	1.0000000

The amount of youth categorized as disconnected has a weak, positive correlation with plumbing, general construction, street light condition, and heating complaints.

Internet access:

```r
internet_acc_cols <- subset(nyc_summary_joined, select = c("Calls.Total", "Cmplts.Heating",
    "Cmplts.StreetCondition", "Cmplts.StreetLight", "Cmplts.GenConstruction",
    "Cmplts.Plumbing", "Households.w.Internet.Access"))
internet_acc_cor <- cor(internet_acc_cols, use = "pairwise.complete.obs")
knitr::kable(internet_acc_cor, format = "markdown", caption = "Correlations between complaint types and internet access") %>%
    kable_styling(font_size = 7) %>% landscape()
```

	Calls.Total	Cmplts.Heating	Cmplts.StreetCondition	Cmplts.StreetLight	Cmplts.GenConstruction	Cmplts.Plumbing	Households.w.Internet.Access
Calls.Total	1.0000000	0.7304898	0.7872922	0.8216617	0.7128808	0.7724521	-0.3881881
Cmplts.Heating	0.7304898	1.0000000	0.2049223	0.3050547	0.9691566	0.9761679	-0.4669426
Cmplts.StreetCondition	0.7872922	0.2049223	1.0000000	0.9131864	0.2132996	0.2739035	-0.0418516
Cmplts.StreetLight	0.8216617	0.3050547	0.9131864	1.0000000	0.3174865	0.3498641	-0.4540284
Cmplts.GenConstruction	0.7128808	0.9691566	0.2132996	0.3174865	1.0000000	0.9784977	-0.4189878
Cmplts.Plumbing	0.7724521	0.9761679	0.2739035	0.3498641	0.9784977	1.0000000	-0.5970603
Households.w.Internet.Access	-0.3881881	-0.4669426	-0.0418516	-0.4540284	-0.4189878	-0.5970603	1.0000000

Total calls, heating, street light condition, general construction, and plumbing complaints all have a moderate, negative correlation with internet access. Of those, plumbing complaints seem to have the strongest relationship. Internet access may be considered an indicator of income, so this might suggest that those with higher income make less complaints.

Struggling households:

```r
low_inc_cols <- subset(nyc_summary_joined, select = c("Calls.Total", "Cmplts.Heating",
    "Cmplts.StreetCondition", "Cmplts.StreetLight", "Cmplts.GenConstruction",
    "Cmplts.Plumbing", "Severely.Rent-Burdened.Households", "Total.SNAP.Recipients"))
low_inc_cor <- cor(low_inc_cols, use = "pairwise.complete.obs")
knitr::kable(low_inc_cor[7:8, 1:6], format = "markdown", caption = "Correlations between complaint types and struggling households") %>%
    kable_styling(font_size = 7) %>% landscape()
```

	Calls.Total	Cmplts.Heating	Cmplts.StreetCondition	Cmplts.StreetLight	Cmplts.GenConstruction	Cmplts.Plumbing
Severely.Rent-Burdened.Households	0.2977214	0.5002023	0.0540981	0.3781051	0.2708738	0.4396465
Total.SNAP.Recipients	0.7204625	0.6249998	0.3129439	0.5962434	0.6300092	0.8076121

All of the top complaint types as well as total calls are weakly to moderately correlated with rent-burdened households, except for the street condition. The number of Supplemental Nutrition Assistance Program (SNAP) recipients has a strong positive correlation with the total calls and almost all of the top complaint types. The weakest relationship is with complaints about street condition, and the strongest relationship is with plumbing complaints. It may be that struggling economically is generally associated with making a complaint of any type, possibly due to these households being unable to afford the cost of better living conditions.

Voting-related indicators:

```r
voter_turn_age_cols <- subset(nyc_summary_joined, select = c("Calls.Total",
    "Cmplts.Heating", "Cmplts.StreetCondition", "Cmplts.StreetLight", "Cmplts.GenConstruction",
    "Cmplts.Plumbing", "Turnout.Among.Registered.Voters.(percent)", "Turnout.Voting.Age.Pop.(pct)",
    "Eligible.Voter.Registration.Rate"))
voter_turn_age_cor <- cor(voter_turn_age_cols, use = "pairwise.complete.obs")
knitr::kable(voter_turn_age_cor[7:9, 1:6], format = "markdown", caption = "Correlations between complaint types and voting") %>%
    kable_styling(font_size = 7) %>% landscape()
```

	Calls.Total	Cmplts.Heating	Cmplts.StreetCondition	Cmplts.StreetLight	Cmplts.GenConstruction	Cmplts.Plumbing
Turnout.Among.Registered.Voters.(percent)	-0.7080163	-0.7739984	-0.2593741	-0.7767826	-0.7777395	-0.7456381

	Calls.Total	Cmplts.Heating	Cmplts.StreetCondition	Cmplts.StreetLight	Cmplts.GenConstruction	Cmplts.Plumbing
Turnout.Voting.Age.Pop.(pct)	-0.6364371	-0.5949009	-0.3498165	-0.8290802	-0.6027809	-0.5655395
Eligible.Voter.Registration.Rate	0.3866116	0.6127693	-0.0824938	-0.0572532	0.6979481	0.7066592

Voter turnout among registered voters and the eligible population has a moderate to strong negative correlation with the total calls and every top complaint type except for street condition. The voter registration rate is somewhat inverse to turnout, with strong, positive correlations with plumbing, general construction, and heating complaints, and a weak positive relationship with the total calls.

Education indicators:

```
education_cols <- subset(nyc_summary_joined, select = c("Calls.Total", "Cmplts.Heating",
    "Cmplts.StreetCondition", "Cmplts.StreetLight", "Cmplts.GenConstruction",
    "Cmplts.Plumbing", "Enrolled.Full.Day.Pre-K.Students", "Four-Year.High.School.Graduation.Rate"))
education_cor <- cor(education_cols, use = "pairwise.complete.obs")
knitr::kable(education_cor[7:8, 1:6], format = "markdown", caption = "Correlations between complaint types and pre-college education") %>%
    kable_styling(font_size = 7) %>% landscape()
```

	Calls.Total	Cmplts.Heating	Cmplts.StreetCondition	Cmplts.StreetLight	Cmplts.GenConstruction	Cmplts.Plumbing
Enrolled.Full.Day.Pre-K.Students	0.5927185	0.3964040	0.3983968	0.5200397	0.4763672	0.6257577
Four-Year.High.School.Graduation.Rate	-0.4860981	-0.7383153	0.0961329	-0.4404035	-0.6617817	-0.7506323

Pre-college education generally seems to have at least a moderately strong relationship with the total calls and top complaints. The high school graduation rate has a strong, negative correlation with heating, general construction, and plumbing complaints, and a more moderate negative correlation with the total calls and street light condition. The amount of pre-kindergarten students has a strong, positive correlation with plumbing complaints, and a more moderate positive correlation with the total calls, heating, street condition, street light condition, and general construction complaints.

Labor indicators:

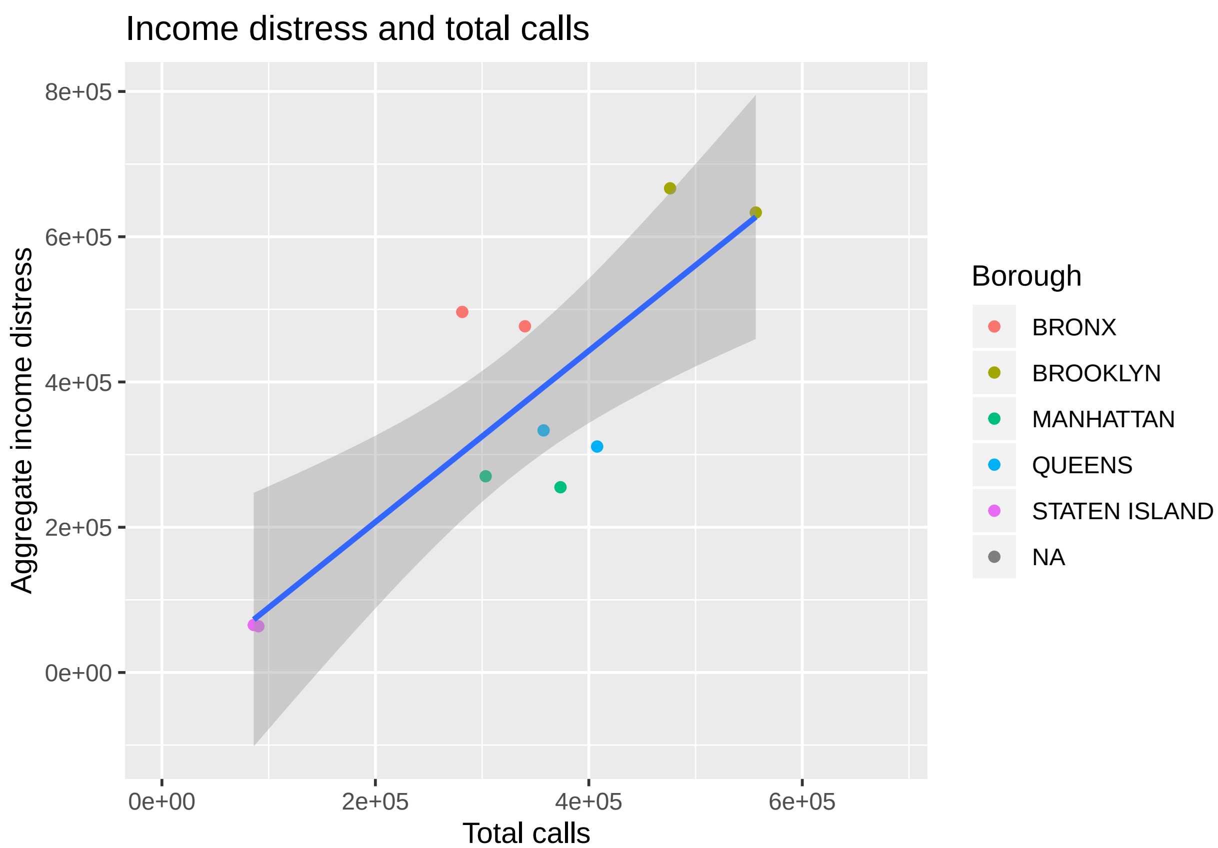
```
earnings_unemp_cols <- subset(nyc_summary_joined, select = c("Calls.Total",
    "Cmplts.Heating", "Cmplts.StreetCondition", "Cmplts.StreetLight", "Cmplts.GenConstruction",
    "Cmplts.Plumbing", "Unemployment.Rate", "Labor.Force.Participation.Rate",
    "Number.of.Jobs.in.the.City", "Average.Weekly.Earnings"))
earnings_unemp_cor <- cor(earnings_unemp_cols, use = "pairwise.complete.obs")
knitr::kable(earnings_unemp_cor[7:9, 1:6], format = "markdown", caption = "Correlations between complaint types and employment factors") %>%
    kable_styling(font_size = 7) %>% landscape()
```

	Calls.Total	Cmplts.Heating	Cmplts.StreetCondition	Cmplts.StreetLight	Cmplts.GenConstruction	Cmplts.Plumbing
Unemployment.Rate	0.4549602	0.5047034	0.3602387	0.5174704	0.5234889	0.4963890
Labor.Force.Participation.Rate	0.2142126	0.1177766	0.1630658	0.0939138	0.0612362	0.0841232
Number.of.Jobs.in.the.City	0.1848851	0.1139791	0.0587847	-0.1094688	-0.0188217	0.0158384

The unemployment rate has a moderate positive correlation with total calls and each top complaint type, another case for low income possibly being the driver of complaints. The number of jobs does not appear to have a significant relationship with any of the complaint types. Unfortunately we do not have data on weekly earnings for the years in the dataset.

Plotting the unemployment rate, the number of SNAP recipients, and the number of rent-burdened households together provides strong visual proof of the relationship between distressed/low income households and the number of calls.

```
ggplot(data = nyc_summary_joined, aes(Calls.Total, (`Severely.Rent-Burdened.Households` +
    Total.SNAP.Recipients + Unemployment.Rate))) + geom_point(aes(color = Borough)) +
    geom_smooth(method = "lm") + labs(title = "Income distress and total calls") +
    ylab("Aggregate income distress") + xlab("Total calls")
```




```r
# correlation coeff of total calls with sum of
# Severely.Rent-Burdened.Households,Total.SNAP.Recipients,Unemployment.Rate
cor(nyc_summary_joined$Calls.Total, nyc_summary_joined$`Severely.Rent-Burdened.Households` +
    nyc_summary_joined$Total.SNAP.Recipients + nyc_summary_joined$Unemployment.Rate,
    use = "pairwise.complete.obs")
```

```
## [1] 0.838795
```

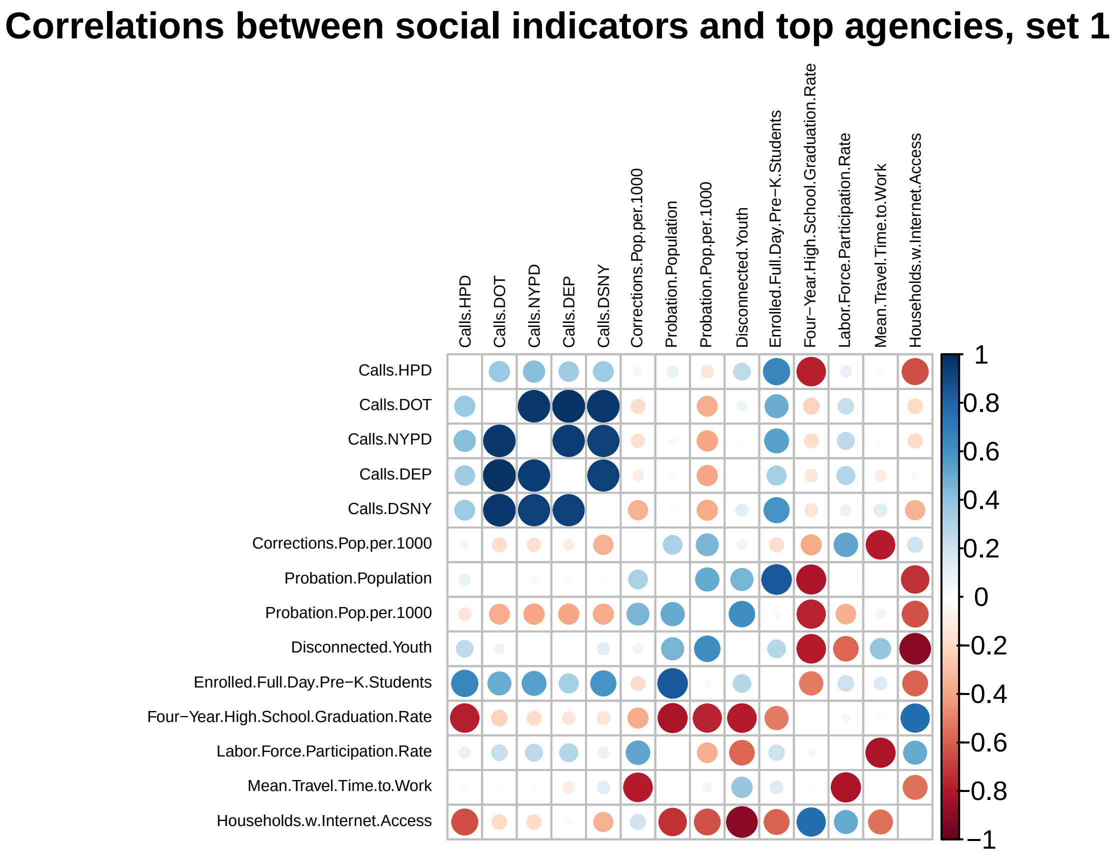
A combined income distress metric of severely rent-burdened households, total SNAP recipients, and the unemployment rate correlates very strongly with the total number of calls, with a correlation coefficient of about .84.

Relationships between agencies and social indicators

Corrplots are employed again to get an overview of relationships, and split into two graphs due to the number of social indicators. 'Average.Weekly.Earnings','Curbside.and.Containerized.Diversion.Rate', 'Living.Walking.Distance.Park' and 'Eligible.Voter.Registration.Rate' are omitted due to lack of data.

```r
agency_cols <- subset(nyc_summary_joined, select = c("Calls.HPD", "Calls.DOT",
    "Calls.NYPD", "Calls.DEP", "Calls.DSNY", "Corrections.Pop.per.1000", "Probation.Population",
    "Probation.Pop.per.1000", "Disconnected.Youth", "Enrolled.Full.Day.Pre-K.Students",
    "Four-Year.High.School.Graduation.Rate", "Labor.Force.Participation.Rate",
    "Mean.Travel.Time.to.Work", "Households.w.Internet.Access"))
```

```r
agency_cor <- cor(agency_cols, use = "pairwise.complete.obs")
plot.new()
corrplot(agency_cor, method = "circle", tl.cex = 0.5, tl.col = "black", diag = FALSE,
    title = "Correlations between social indicators and top agencies, set 1",
    mar = c(0, 0, 2, 0))
```



Enrolled pre-K students, the high school graduation rate, and internet access all seem to have significant relationships with some of the top agencies.

```r
library(dplyr)
agency1_breakout <- subset(nyc_summary_joined, select = c("Calls.HPD", "Calls.DOT",
    "Calls.NYPD", "Calls.DEP", "Calls.DSNY", "Enrolled.Full.Day.Pre-K.Students",
    "Four-Year.High.School.Graduation.Rate", "Households.w.Internet.Access"))

agency1_breakout_cor <- cor(agency1_breakout, use = "pairwise.complete.obs")
knitr::kable(agency1_breakout_cor[6:8, 1:5], format = "markdown", caption = "Correlations between agencies, education, and internet access") %>%
    kable_styling(font_size = 7) %>% landscape()
```

	Calls.HPD	Calls.DOT	Calls.NYPD	Calls.DEP	Calls.DSNY
Enrolled.Full.Day.Pre.K.Students	0.6565872	0.4989928	0.5402678	0.3358951	0.5965520
Four.Year.High.School.Graduation.Rate	-0.7764363	-0.2209960	-0.1726854	-0.1395269	-0.1474461
Households.w.Internet.Access	-0.6439616	-0.1912869	-0.1866136	0.0338975	-0.3445193

Both education indicators as well as internet access are strongly correlated with calls that were routed to Department of Housing Preservation and Development (HPD). The high school graduation rate has the strongest correlation at negative .77. Recall that HPD calls are mostly complaints about heating, plumbing, and other housing related issues. Higher numbers of high school graduates are associated with less calls about housing issues.

The number of pre-K students is also moderately to strongly correlated with complaints routed to the

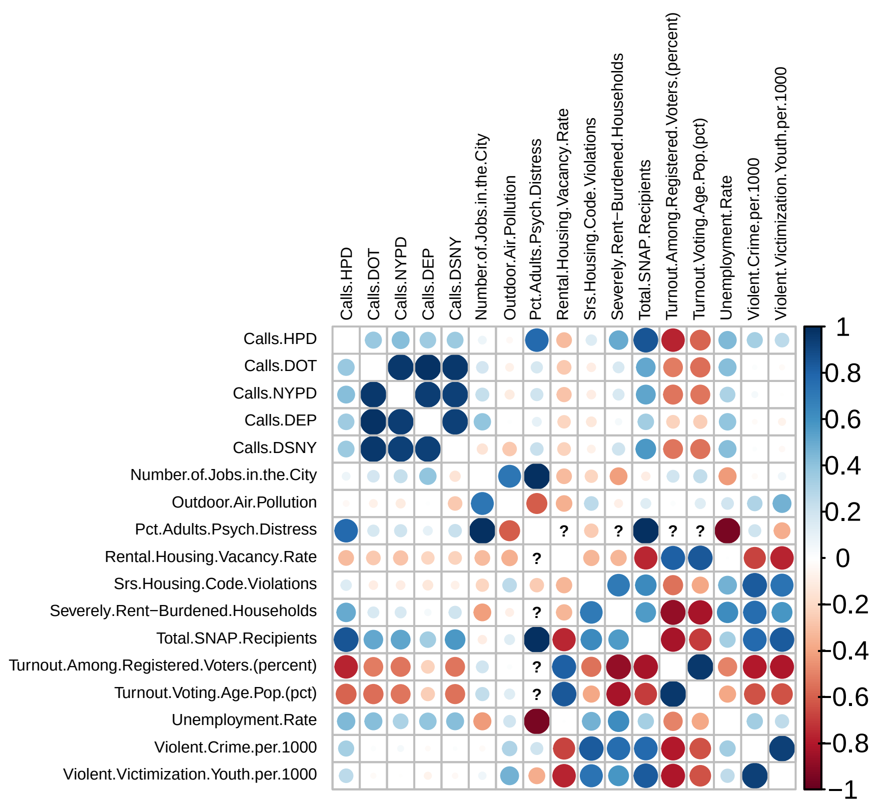
New York Police Department (NYPD), the Department of Transportation (DOT) and the Department of Sanitation (DSNY).

```r
agency_cols_2 <- subset(nyc_summary_joined, select = c("Calls.HPD", "Calls.DOT",
    "Calls.NYPD", "Calls.DEP", "Calls.DSNY", "Number.of.Jobs.in.the.City", "Outdoor.Air.Pollution",
    "Pct.Adults.Psych.Distress", "Rental.Housing.Vacancy.Rate", "Srs.Housing.Code.Violations",
    "Severely.Rent-Burdened.Households", "Total.SNAP.Recipients", "Turnout.Among.Registered.Voters.(per
    "Turnout.Voting.Age.Pop.(pct)", "Unemployment.Rate", "Violent.Crime.per.1000",
    "Violent.Victimization.Youth.per.1000"))

agency_cor_2 <- cor(agency_cols_2, use = "pairwise.complete.obs")

plot.new()
corrplot(agency_cor_2, method = "circle", tl.cex = 0.5, tl.col = "black", diag = FALSE,
    number.cex = 0.5, title = "Correlations between social indicators and top agencies, set 2",
    mar = c(0, 0, 2, 0))
```

Correlations between social indicators and top agencies, set 2



The percent of adults experiencing psychiatric distress, severely rented burdened households, the total SNAP recipients, voter turnout, and the unemployment rate all have a moderate to strong association with calls to the top 5 agencies.

```r
library(dplyr)
agency2_breakout <- subset(nyc_summary_joined, select = c("Calls.HPD", "Calls.DOT",
    "Calls.NYPD", "Calls.DEP", "Calls.DSNY", "Pct.Adults.Psych.Distress", "Severely.Rent-Burdened.Households",
    "Total.SNAP.Recipients", "Turnout.Among.Registered.Voters.(percent)", "Turnout.Voting.Age.Pop.(pct)",
    "Unemployment.Rate"))

agency2_breakout_cor <- cor(agency2_breakout, use = "pairwise.complete.obs")
knitr::kable(agency2_breakout_cor[6:11, 1:5], format = "markdown", caption = "Correlations between agencies, voter turnout, income distress factors and psychological distress") %>%
    kable_styling(font_size = 7) %>% landscape()
```

	Calls.HPD	Calls.DOT	Calls.NYPD	Calls.DEP	Calls.DSNY
Pct.Adults.Psych.Distress	0.7703865	0.1734177	0.2049296	0.1029719	0.2208998
Severely.Rent-Burdened.Households	0.5071962	0.1608509	0.1641184	0.0479054	0.2099028
Total.SNAP.Recipients	0.8593708	0.5137546	0.5289404	0.3493334	0.5754234
Turnout.Among.Registered.Voters.(percent)	-0.7687408	-0.5184310	-0.5389098	-0.2278963	-0.5305899
Turnout.Voting.Age.Pop.(pct)	-0.5891138	-0.5554787	-0.5326802	-0.2465933	-0.5443573
Unemployment.Rate	0.4432978	0.4266585	0.3131394	0.3966764	0.4214835

All of the indicators charted have moderate to strong relationship with calls to HPD, with the number of SNAP recipients most strongly correlated at .85. SNAP recipients are also fairly strongly correlated with calls to DSNY.

```
cor(nyc_summary_joined$Calls.HPD, nyc_summary_joined$`Severely.Rent-Burdened.Households` +
    nyc_summary_joined$Total.SNAP.Recipients + nyc_summary_joined$Unemployment.Rate,
    use = "pairwise.complete.obs")
```

```
## [1] 0.9263153
```

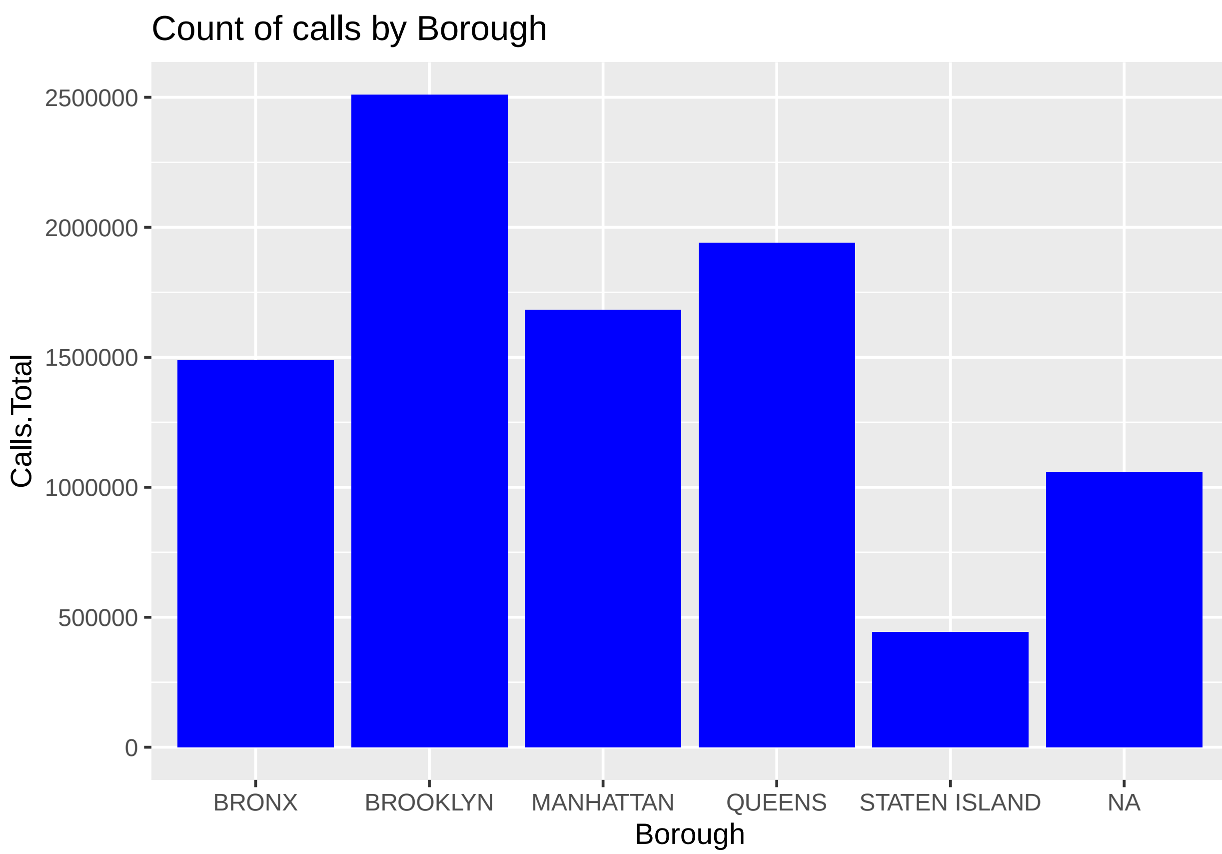
From a previous section, I concluded that income distress indicators, including the number of SNAP recipients, are very strongly correlated with the total number of calls. They are also very strongly associated with calls to HPD with a positive correlation coefficient of about .93. Recall that HPD overwhelmingly receives the most 311 calls, so the correlation here may just be a result of income distress correlating to the total number of calls. We would also expect that economically-struggling households have trouble affording better housing conditions.

Relationships between boroughs and social indicators

As stated in an earlier section, heating is the top complaint in Brooklyn, the Bronx, and Manhattan. Street light condition is the top complaint type from Queens, while street condition is confirmed as the most common complaint from Staten Island. Are there social indicators that might correlate with a high number of certain kinds of calls in the boroughs?

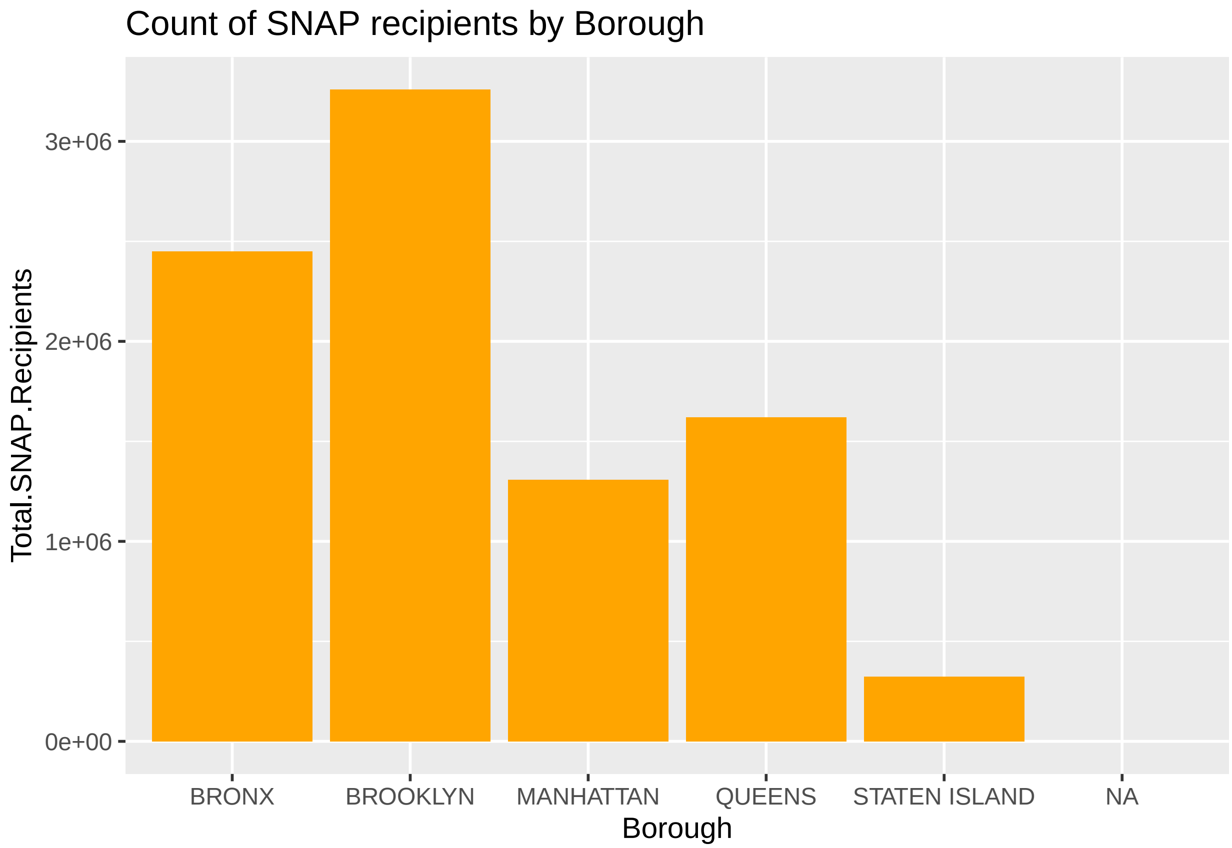
Recall that Brooklyn had the most total calls:

```
b <- ggplot(nyc_summary_joined, aes(x = Borough, y = Calls.Total)) + geom_bar(stat = "identity",
    fill = "blue") + labs(title = "Count of calls by Borough")
b
```



Of all the social indicators, total calls is most closely associated with the number of SNAP recipients, with a .72 correlation coefficient. Are there a large amount of SNAP recipients in Brooklyn?

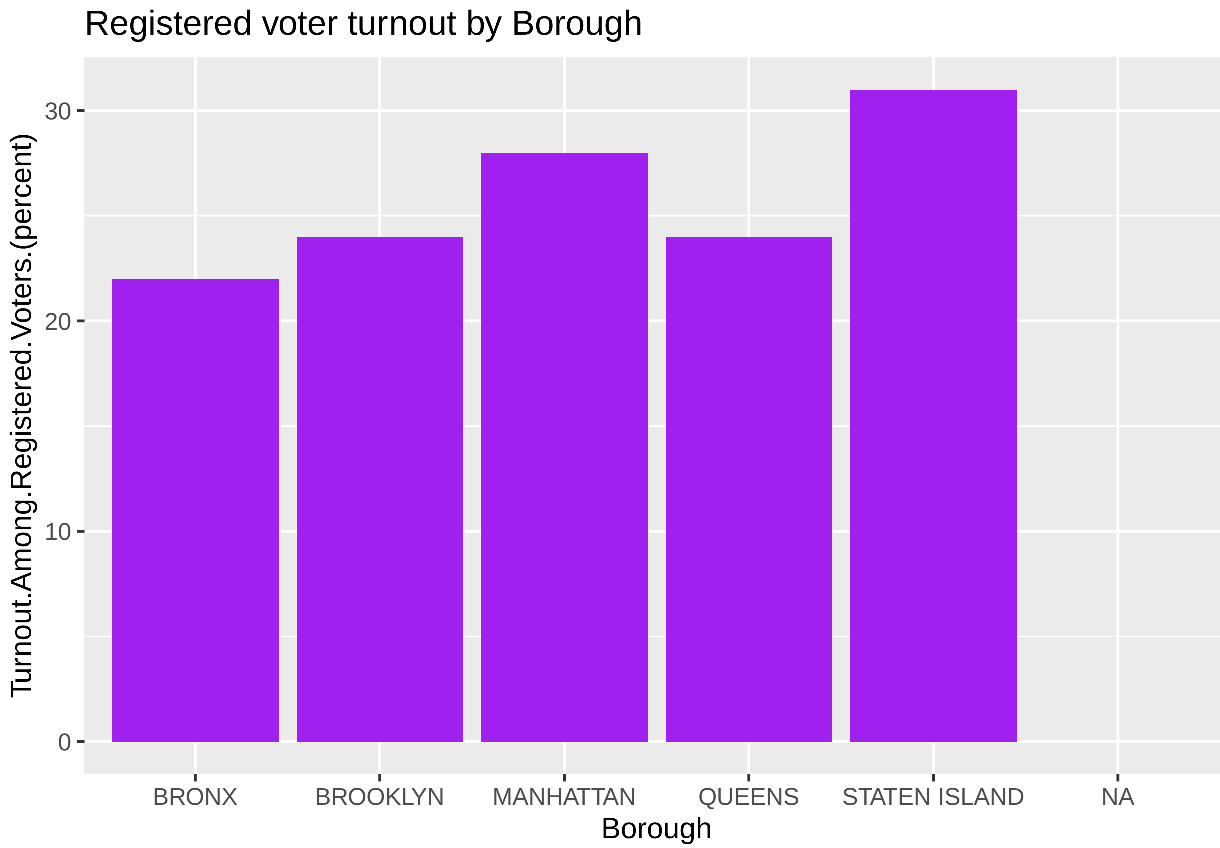
```
snap <- ggplot(nyc_summary_joined, aes(x = Borough, y = Total.SNAP.Recipients)) +
    geom_bar(stat = "identity", fill = "orange") + labs(title = "Count of SNAP recipients by Borough")
snap
```



Brooklyn has by far the highest total SNAP recipients, which corresponds with having the most 311 calls.

Heating complaints have a strong negative correlation with turnout among registered voters. Heating is also the top complaint in the Bronx, Brooklyn, and Manhattan, in that order. Does a lesser degree of voter turnout correspond to high numbers of heating complaints if we look at the boroughs?

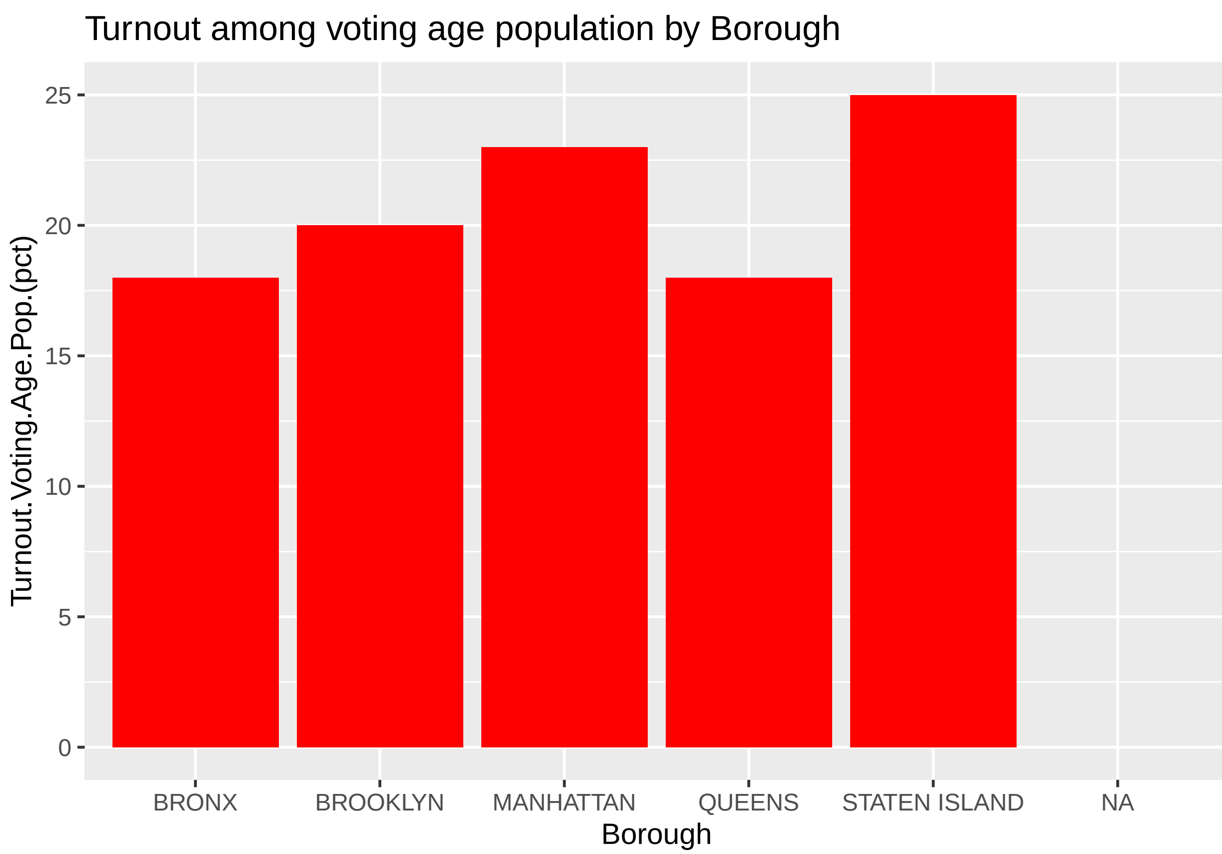
```
reg_voter <- ggplot(nyc_summary_joined, aes(x = Borough, y = `Turnout.Among.Registered.Voters.(percent)
    geom_bar(method = "mean", stat = "identity", fill = "purple") + labs(title = "Registered voter turn
reg_voter
```



It appears that lower voter turnout does correspond roughly with high numbers of complaints about heating by borough, however Staten Island defies the trend.

Turnout among voting age population has a very strong negative correlation with complaints about street light condition, which as we saw in a previous section is the top complaint in Queens. Does Queens have the lowest turnout of all boroughs by the voting age population?

```
voting_age_turnout <- ggplot(nyc_summary_joined, aes(x = Borough, y = `Turnout.Voting.Age.Pop.(pct)`)) +
    geom_bar(method = "mean", stat = "identity", fill = "red") + labs(title = "Turnout among voting age
voting_age_turnout
```



Queens is tied with the Bronx for least amount of voter turnout by the voting age population. As we've discovered with the other social indicators, the turnout among the voting age population is a fairly good predictor of the type of complaint that dominates each borough. The Bronx is an outlier here, with a lesser amount of complaints about the street light condition not corresponding with its tie for the lowest turnout.

Relationships between call status and social indicators

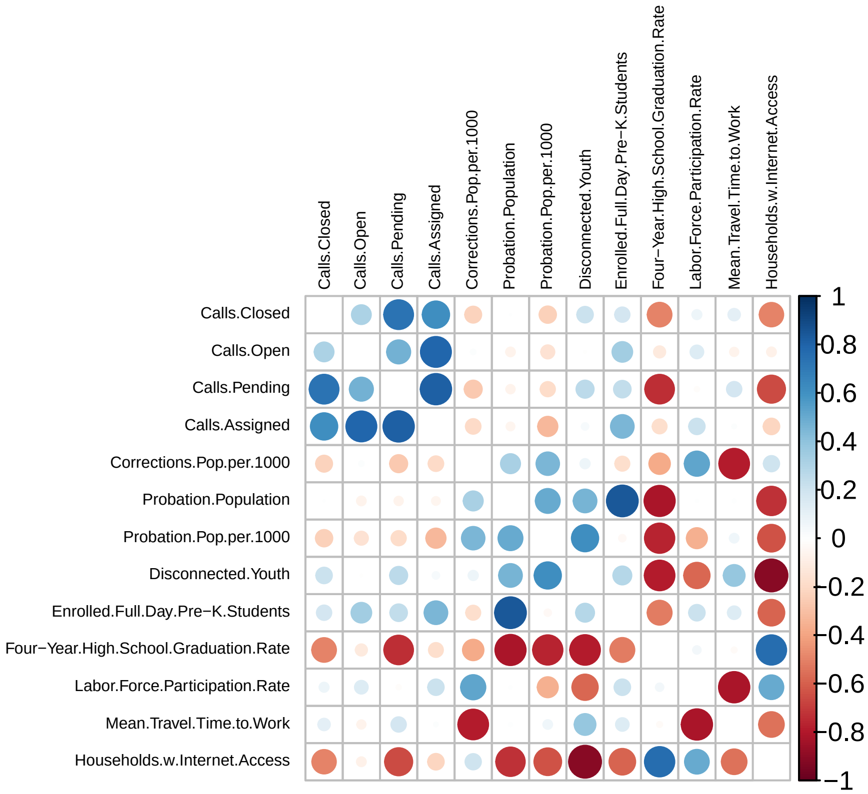
Corrplots of the call status and social indicators again give a quick overview of the relationships between the columns. 'Average.Weekly.Earnings','Curbside.and.Containerized.Diversion.Rate', 'Living.Walking.Distance.Park' and 'Eligible.Voter.Registration.Rate' are omitted due to lack of data.

```
status_cols <- subset(nyc_summary_joined, select = c("Calls.Closed", "Calls.Open",
    "Calls.Pending", "Calls.Assigned", "Corrections.Pop.per.1000", "Probation.Population",
    "Probation.Pop.per.1000", "Disconnected.Youth", "Enrolled.Full.Day.Pre-K.Students",
    "Four-Year.High.School.Graduation.Rate", "Labor.Force.Participation.Rate",
    "Mean.Travel.Time.to.Work", "Households.w.Internet.Access"))



status_cor <- cor(status_cols, use = "pairwise.complete.obs")
plot.new()
corrplot(status_cor, method = "circle", tl.cex = 0.5, tl.col = "black", diag = FALSE,
    title = "Correlations between social indicators and call status, set 1",
    mar = c(0, 0, 2, 0))
```

Correlations between social indicators and call status, set 1



From this corrplot, it appears there is possibly a strong negative relationship between calls pending and the high school graduation rate, and households with internet access. Looking closer at those relationships:

```
library(dplyr)
pending_cols <- subset(nyc_summary_joined, select = c("Calls.Pending", "Four-Year.High.School.Graduation.Rate",
    "Households.w.Internet.Access"))

pending_cor <- cor(pending_cols, use = "pairwise.complete.obs")
knitr::kable(pending_cor[2:3, 1:3], format = "markdown", caption = "Correlations between calls pending, high school graduation rate, and internet access") %>%
    kable_styling(font_size = 7) %>% landscape()
```

	Calls.Pending	Four-Year.High.School.Graduation.Rate	Households.w.Internet.Access
Four-Year.High.School.Graduation.Rate	-0.7308674	1.0000000	0.7651981
Households.w.Internet.Access	-0.6542204	0.7651981	1.0000000

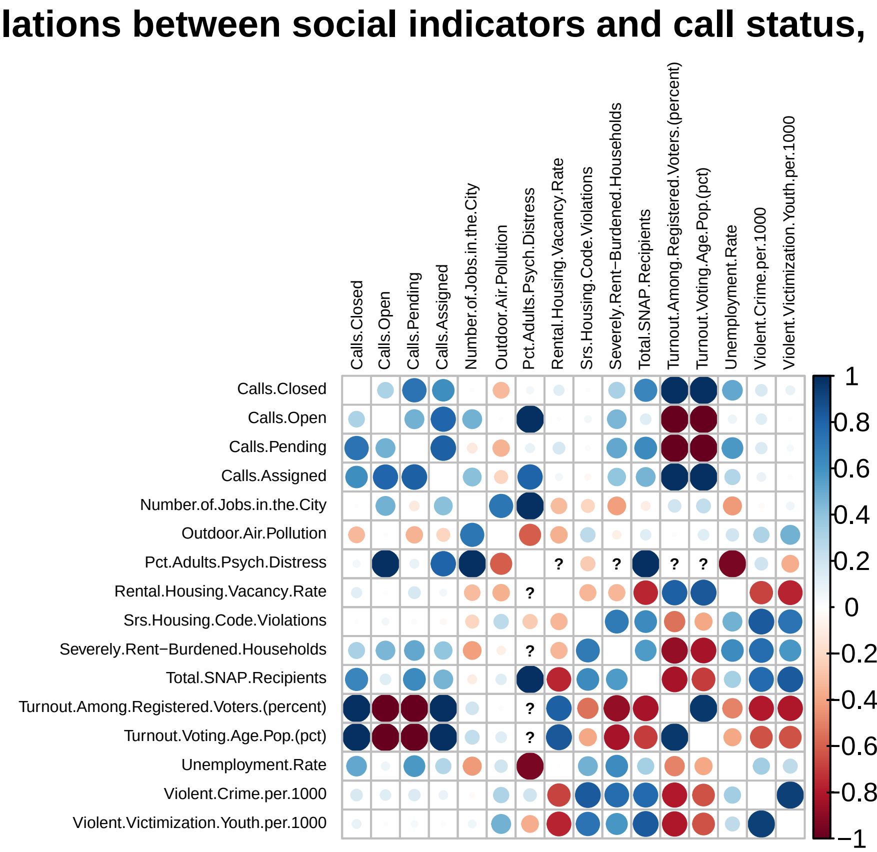
Calls pending, the high school graduation rate, and internet access are all strongly correlated with each other. A larger number of high school graduates and households with internet access seems to be a predictor of less calls, which again points to income factors driving complaints.

```
status_cols_2 <- subset(nyc_summary_joined, select = c("Calls.Closed", "Calls.Open",
    "Calls.Pending", "Calls.Assigned", "Number.of.Jobs.in.the.City", "Outdoor.Air.Pollution",
    "Pct.Adults.Psych.Distress", "Rental.Housing.Vacancy.Rate", "Srs.Housing.Code.Violations",
    "Severely.Rent-Burdened.Households", "Total.SNAP.Recipients", "Turnout.Among.Registered.Voters.(per
    "Turnout.Voting.Age.Pop.(pct)", "Unemployment.Rate", "Violent.Crime.per.1000",
    "Violent.Victimization.Youth.per.1000"))

status_cor_2 <- cor(status_cols_2, use = "pairwise.complete.obs")

plot.new()
```

```r
corrplot(status_cor_2, method = "circle", tl.cex = 0.5, tl.col = "black", diag = FALSE,
    number.cex = 0.5, title = "Correlations between social indicators and call status, set 2",
    mar = c(0, 0, 2, 0))
```



Taking the second half of the social indicators with the status of calls, the two turnout indicators have the strongest relationships with each of the status types.

```r
status_turnout <- subset(nyc_summary_joined, select = c("Calls.Closed", "Calls.Open",
    "Calls.Pending", "Calls.Assigned", "Turnout.Among.Registered.Voters.(percent)",
    "Turnout.Voting.Age.Pop.(pct)"))

status_turnout_cor <- cor(status_turnout, use = "pairwise.complete.obs")
knitr::kable(status_turnout_cor[5:6, 1:4], format = "markdown", caption = "Correlations between call status and voter turnout") %>%
    kable_styling(font_size = 7) %>% landscape()
```

	Calls.Closed	Calls.Open	Calls.Pending	Calls.Assigned
Turnout.Among.Registered.Voters.(percent)	1	-1	-1	1
Turnout.Voting.Age.Pop.(pct)	1	-1	-1	1

I believe the 1:1 relationships shown above with the two turnout indicators are due to having only single years of data available for each borough. The relationships may be very strong, but more data is needed before making that conclusion.

```r
status_psych <- subset(nyc_summary_joined, select = c("Calls.Closed", "Calls.Open",
    "Calls.Pending", "Calls.Assigned", "Pct.Adults.Psych.Distress"))

status_psych_cor <- cor(status_psych, use = "pairwise.complete.obs")
knitr::kable(status_psych_cor, format = "markdown", caption = "Correlations between call status and psychological distress") %>%
    kable_styling(font_size = 7) %>% landscape()
```

	Calls.Closed	Calls.Open	Calls.Pending	Calls.Assigned	Pct.Adults.Psych.Distress
Calls.Closed	1.0000000	0.3162352	0.7381017	0.6185213	0.0551449

	Calls.Closed	Calls.Open	Calls.Pending	Calls.Assigned	Pct.Adults.Psych.Distress
Calls.Open	0.3162352	1.0000000	0.4734141	0.7968686	0.9945867
Calls.Pending	0.7381017	0.4734141	1.0000000	0.8199100	0.0923722
Calls.Assigned	0.6185213	0.7968686	0.8199100	1.0000000	0.8070141
Pct.Adults.Psych.Distress	0.0551449	0.9945867	0.0923722	0.8070141	1.0000000

Calls open and calls assigned correlate very strongly with the percent of adults in psychiatric distress. This is expected, as 311 is advertised as one of the numbers to call in case of mental health emergencies.

The previously identified income distress indicators also seem to strongly correlate with call status, as shown below.

```
income_distress <- subset(nyc_summary_joined, select = c("Calls.Closed", "Calls.Open",
    "Calls.Pending", "Calls.Assigned", "Severely.Rent-Burdened.Households",
    "Total.SNAP.Recipients", "Unemployment.Rate"))

income_distress_cor <- cor(income_distress, use = "pairwise.complete.obs")
knitr::kable(income_distress_cor[5:7, 1:4], format = "markdown", caption = "Correlations between call status and income distress indicators") %>%
    kable_styling(font_size = 7) %>% landscape()
```

	Calls.Closed	Calls.Open	Calls.Pending	Calls.Assigned
Severely.Rent-Burdened.Households	0.3297177	0.4599207	0.5117555	0.3966277
Total.SNAP.Recipients	0.6590262	0.1372137	0.6389223	0.4674983
Unemployment.Rate	0.5132366	0.0778088	0.5742735	0.2948578

Calls pending has a strong, positive correlation with the total SNAP recipients, which would correspond to the number of SNAP recipients having a strong correlation with the total calls.

```
cor(nyc_summary_joined$Calls.Pending, (nyc_summary_joined$`Severely.Rent-Burdened.Households` +
    nyc_summary_joined$Total.SNAP.Recipients + nyc_summary_joined$Unemployment.Rate)/3,
    use = "pairwise.complete.obs")
```

```
## [1] 0.9859281
```

Taken together as a cluster, the income distress indicators (unemployment rate, severely rent burdened households, and SNAP recipients), have a very strong positive correlation of about .99 to the calls pending.

```
cor(nyc_summary_joined$Calls.Open, (nyc_summary_joined$`Severely.Rent-Burdened.Households` +
    nyc_summary_joined$Total.SNAP.Recipients + nyc_summary_joined$Unemployment.Rate)/3,
    use = "pairwise.complete.obs")
```

```
## [1] 0.8665656
```

```
cor(nyc_summary_joined$Calls.Assigned, (nyc_summary_joined$`Severely.Rent-Burdened.Households` +
    nyc_summary_joined$Total.SNAP.Recipients + nyc_summary_joined$Unemployment.Rate)/3,
    use = "pairwise.complete.obs")
```

```
## [1] 0.8035085
```

The income distress cluster also has very strong, positive correlation coefficients of about .87 and .80 to calls open and calls assigned.

However, less of these calls from income distressed households are resolved, with a still strong but lesser correlation of .70 to calls closed.

```r
cor(nyc_summary_joined$Calls.Closed, (nyc_summary_joined$`Severely.Rent-Burdened.Households` +
    nyc_summary_joined$Total.SNAP.Recipients + nyc_summary_joined$Unemployment.Rate)/3,
    use = "pairwise.complete.obs")
```

```
## [1] 0.7067364
```

Conclusion

This was our final R exercise: Communicating our findings from the NYC 311 dataset.

Summary of Findings

Heating issues generate the most complaints, followed by street condition and street light condition. Most calls in the top complaint categories have been closed. DOT gets routed complaints about traffic light, street and street light conditions. HPD receives complaints regarding housing conditions such as heating, NYPD receives a large amount of complaints about blocked driveways and illegal parking, and DEP handles complaints related to sanitation. Most of the top 15 complaint types are routed to these top 4 agencies. Heating is the top complaint in Brooklyn, the Bronx, and Manhattan. Street light condition is the top complaint type from Queens, while street condition is confirmed as the most common complaint from Staten Island.

Perhaps unsurprisingly, social factors relating to income levels, and related living conditions, seem to be associated with calls to 311. Income distress indicators (severely rent-burdened households, total SNAP recipients, and the unemployment rate) correlate very strongly with the number of complaints. The number of SNAP recipients, both education indicators, as well as internet access, are strongly correlated with calls that were routed to HPD. SNAP recipients are also fairly strongly correlated with calls to DSNY. It's likely that as a result of income distress indicators correlating highly with the number of calls, they also generally correlate with calls made to the top agencies.

The social indicators carry over into the types of complaints that are mostly strongly associated with each borough and call status. Brooklyn has the highest number of calls, and also the highest number of SNAP recipients, which strongly correlates with the total number of calls. The two kinds of voter turnout metrics also appear to generally correspond with complaints about heating and street light conditions in the boroughs. The income distress indicators very strongly correlate with open, assigned, and pending calls, but less so with calls that have been closed.

Caveats

Additional years of data are necessary in order to draw actionable conclusions and make recommendations. For some of the social indicators like voter turnout, very little data was available. There are also some odd discrepancies in the data, such as the fact that heating complaints are being recorded at a high rate but the borough is often not noted. This makes it more difficult to rely on social indicators as a predictor of complaints because the data itself may not be reliable.

APPENDIX

Joint data dictionary for the joined summary dataset

For columns derived from the NYC 311 call data, I've assumed that data such as the borough, date and complaint type were recorded by the 311 call operator who took the call, and that status fields were updated by the agency to which it was routed. The rest of the columns derived from the social indicators data have clear sources defined in the reference dataframe. This data dictionary is split into two tables due to the length of the descriptions and data source text.

Note to instructor: For reasons I don't understand, the formatting I used previously to prevent table columns from overlapping in the pdf output is not working in this file. I spent over 15 hours trying to debug this and other issues with tables, including printing them using different libraries, and was never able to fix it. The data dictionary below is readable and usable when viewed through RStudio. For comparison, please see my previous homework assignment where I was able to customize the width of the columns using the same column_spec option.

```r
library(webshot)
column_names <- names(nyc_summary_joined)
data_dict <- data.frame(matrix(ncol = 43, nrow = 0))
colnames(data_dict) <- column_names
data_dict <- as.data.frame(t(as.matrix(data_dict)))

# data_dict['Columns'] <- names(nyc_summary_joined)
data_dict["Description"] <- c("Name of Borough", "Year the call was created and social indicator data recorded",
    "Total 311 Calls", "Number of calls to HPD", "Number of calls to DOT", "Number of calls to NYPD",
    "Number of calls to DEP", "Number of calls to DSNY", "Number of calls closed",
    "Number of calls open", "Number of calls pending resolution", "Number of calls assigned to an agency",
    "Number of complaints about heating", "Number of complaints about street condition",
    "Number of complaints about street light condition", "Number of complaints about general construction",
    "Number of complaints about plumbing", "Admissions to the Department of Corrections per 1000 people in the borough",
    "Average weekly earnings per person in the borough($)", "Textile & Used Materials Donations, E-Waste, Hazardous Waste, and other Organics collections. Does not include Redeemed Bottle & Can Deposit containers",
    "Number of individuals supervised by the Dept. of Probation during the fiscal year",
    "Number of individuals per 1,000 people supervised by the Dept. of Probation during the fiscal year",
    "The percentage of all person 16 to 24 years old who are not at work (unemployed or not in labor force) and not enrolled in school--% as total of all 16 to 24 year-olds",
    "Percent of NYC residents who meet the eligibility requirements to vote and registered to vote",
    "The percentage of students enrolled in Pre-K", "The percentage of students who graduated with a diploma within four years in August out of the cohort of all students who entered ninth grade",
    "The percentage of working age New Yorkers who are either employed or looking for work",
    "Mean travel time for workers 16 years and over who did not work at home - Group quarters not included",
    "Percent of persons in NYC households with Internet access. Households with access include those with and without a subscription",
    "Percentage of New Yorkers who live within a quarter-mile walk of a small park (under 6 acres) or a half-mile walk of a larger park (over 6 acres)",
    "Total number of residential properties (single- and multi-family buildings and condominium apartment units) that had mortgage foreclosure actions initiated against them",
    "Number of job openings compiled from the Quarterly Employer Survey of Employers by the Bureau of Labor Statisitcs",
    "Average concentration of PM2.5 (fine particulate matter with a diameter smaller than 2.5 microns), micrograms (one-millionth of a gram) of PM2.5 per cubic meter of air",
    "Age-adjusted percent of adults with serious psychological distress who reported receiving mental health counseling or treatment in the last 12 months",
    "The net rental vacancy rate is calculated by dividing the number of vacant, habitable, and available-for-rent units by the number of renter-occupied units plus vacant, habitable, and available for-rent units. This calculation excludes housing units in group quarters,
    "Serious Housing Code Violations (class C) recorded by the Dept. of Housing Preservation and Development per 1000 rental units",
    "The share of renter-occupied households whose gross rent (rent plus electricity and heating fuel costs) equaled at least 50 percent of their monthly pre-tax income, excluding those living in public housing or renting with the use of a voucher",
    "Total low-income New Yorkers who receive critical nutrition assistance",
    "Turnout among registered voters (Active + Inactive)", "Turnout among NYC population age 18+",
    "Percentage of the total labor force that is unemployed, but actively seeking employment and willing to work",
    "Violent crime reports (murder, rape and sexual assault, robbery, and aggravated assault) per 1000 people.",
    "NYPD complaints for violent crimes (murder, rape, robbery, and felony assault) where the victim was a youth or young adult between the ages of 16 and 24")
data_dict["Data Source"] <- c("nyc311 call data - 311 operator", "nyc311 call data - 311 operator",
    "nyc311 call data - 311 operator", "nyc311 call data - 311 operator", "nyc311 call data - 311 operator",
    "nyc311 call data - 311 operator", "nyc311 call data - 311 operator", "nyc311 call data - 311 operator",
    "nyc311 call data - Agency assigned", "nyc311 call data - Agency assigned",
    "nyc311 call data - Agency assigned", "nyc311 call data - Agency assigned",
    "nyc311 call data - 311 operator", "nyc311 call data - 311 operator", "nyc311 call data - 311 operator",
    "nyc311 call data - 311 operator", "nyc311 call data - 311 operator", "Dept. of Corrections Population Research",
    "2005-2016 American Community Survey 1-yr PUMS data, analyzed by NYC Opportunity",
    "NYC DSNY Annual Report: FY 2016 http://www1.nyc.gov/assets/dsny/docs/about_dsny-curbside-collections-FY2016.pdf",
    "Snapshot of Department of Probation population as of Septermber 6, 2017",
    "Snapshot of Department of Probation population as of Septermber 6, 2017",
    "2005-2016 American Community Survey PUMS, U.S. Bureau of the Census", "2016 Source: NYC Board of Elections Annual Report 2016, Voting Age Population Data from the U.S. Census Bureau 2015; 2014 Source: NYC Board of Elections voter file data, Population data from U.S.
    "NYC Department of Education", "NYC Department of Education; http://schools.nyc.gov/Accountability/data/GraduationDropoutReports/default.htm",
    "American Community Survey 1-yr PUMS data, analyzed by NYC Opportunity",
    "2005-2016 American Community Survey PUMS as Augmented by NYC Opportunity",
    "2013-2016 American Community Survey PUMS, U.S. Bureau of the Census", "Department of Parks and Recreation; data derived from Walk to a Park Analysis run in July 2017 using 2010 Census Block Data",
    "NYU Furman Center Core Data http://coredata.nyc/", "U.S. Census Bureau, Quarterly Workforce Indicators (QWI) via NYC Economic Development Corporation",
    "New York City Community Air Survey (NYCCAS)", "NYC Community Health Survey 2009-10, 2012-13, 2015. CHS 2002-2008 data are weighted to the NYC adult population per Census 2000; starting in 2009, data are weighted to the 2008 HVS for phone usage and the Census 2000.CHS
    "U.S. Census Bureau: NYC Housing and Vacancy Survey 2005, 2008, 2011, 2014) , HPD : Selected Initial Findings of the 2017 New York City Housing and Vacancy Survey",
```

```r
    "New York City Department of Housing Preservation and Development, New York City Department of Finance Final Tax Roll File, New York City Housing Authority, NYU Furman Center",
    "U.S. Census Bureau: NYC Housing and Vacancy Survey (2005, 2008, 2011, 2014)",
    "NYC figures from HRA administrative data; NYS figures from NYS Office of Temporary and Disability Assistance; U.S. figures from the U.S. Department of Agriculture for 2012-2015 and from the Food Research and Action Center",
    "2016 Source: NYC Board of Elections Annual Report 2016, Voting Age Population Data from the U.S. Census Bureau 2015; 2014 Source: NYC Board of Elections voter file data, Population data from U.S. Census Bureau, 2009-2013 American Community Survey Five-Year Estimates"
    "2016 Source: NYC Board of Elections Annual Report 2016, Voting Age Population Data from the U.S. Census Bureau 2015; 2014 Source: NYC Board of Elections voter file data, Population data from U.S. Census Bureau, 2009-2013 American Community Survey Five-Year Estimates"
    "2005-2016 American Community Survey 1-yr PUMS data, analyzed by NYC Opportunity",
    "NYPD Compstat for violent crimes (murder, rape, robbery, and felony assault)",
    "NYPD Compstat for violent crimes (murder, rape, robbery, and felony assault), Population Estimates from the American Community Survey")

# renders table without column formatting when knit to pdf
knitr::kable(data_dict, format = "markdown") %>% column_spec(1, width = "25em") %>%
    column_spec(2, width = "35em") %>% column_spec(3, width = "40em") %>% landscape()
```

	Description	Data Source
Borough	Name of Borough	nyc311 call data - 311 operator
Year	Year the call was created and social indicator data recorded	nyc311 call data - 311 operator
Calls.Total	Total 311 Calls	nyc311 call data - 311 operator
Calls.HPD	Number of calls to HPD	nyc311 call data - 311 operator
Calls.DOT	Number of calls to DOT	nyc311 call data - 311 operator
Calls.NYPD	Number of calls to NYPD	nyc311 call data - 311 operator
Calls.DEP	Number of calls to DEP	nyc311 call data - 311 operator
Calls.DSNY	Number of calls to DSNY	nyc311 call data - 311 operator
Calls.Closed	Number of calls closed	nyc311 call data - Agency assigned
Calls.Open	Number of calls open	nyc311 call data - Agency assigned
Calls.Pending	Number of calls pending resolution	nyc311 call data - Agency assigned
Calls.Assigned	Number of calls assigned to an agency	nyc311 call data - Agency assigned
Cmplts.Heating	Number of complaints about heating	nyc311 call data - 311 operator
Cmplts.Street.Condition	Number of complaints about street condition	nyc311 call data - 311 operator
Cmplts.Street.Light	Number of complaints about street light condition	nyc311 call data - 311 operator
Cmplts.Gen.Construction	Number of complaints about general construction	nyc311 call data - 311 operator
Cmplts.Plumbing	Number of complaints about plumbing	nyc311 call data - 311 operator
Corrections.Pop.per.1000	Admissions to the Department of Corrections per 1000 people in the borough	Dept. of Corrections Population Research
Average.Weekly.Earnings	Average weekly earnings per person in the borough($)	2005-2016 American Community Survey 1-yr PUMS data, analyzed by NYC Opportunity
Curbside.and.Containerized.Diversion.Rate	Textile & Used Materials Donations, E-Waste, Hazardous Waste, and other Organics collections. Does not include Redeemed Bottle & Can Deposit containers	NYC DSNY Annual Report: FY 2016 http://www1.nyc.gov/assets/dsny/docs/about_dsny-curbside-collections-FY2016.pdf
Probation.Population	Number of individuals supervised by the Dept. of Probation during the fiscal year	Snapshot of Department of Probation population as of Septermber 6, 2017
Probation.Pop.per.1000	Number of individuals per 1,000 people supervised by the Dept. of Probation during the fiscal year	Snapshot of Department of Probation population as of Septermber 6, 2017
Disconnected.Youth	The percentage of all person 16 to 24 years old who are not at work (unemployed or not in labor force) and not enrolled in school–% as total of all 16 to 24 year-olds	2005-2016 American Community Survey PUMS, U.S. Bureau of the Census
Eligible.Voter.Registration.Rate	Percent of NYC residents who meet the eligibility requirements to vote and registered to vote	2016 Source: NYC Board of Elections Annual Report 2016, Voting Age Population Data from the U.S. Census Bureau 2015; 2014 Source: NYC Board of Elections voter file data, Population data from U.S. Census Bureau, 2009-2013 American Community Survey Five-Year Estimates
Enrolled.Full.Day.Pre-K.Students	The percentage of students enrolled in Pre-K	NYC Department of Education
Four-Year.High.School.Graduation.Rate	The percentage of students who graduated with a diploma within four years in August out of the cohort of all students who entered ninth grade	NYC Department of Education; http://schools.nyc.gov/Accountability/data/GraduationDropoutReports/default.htm
Labor.Force.Participation.Rate	The percentage of working age New Yorkers who are either employed or looking for work	American Community Survey 1-yr PUMS data, analyzed by NYC Opportunity
Mean.Travel.Time.to.Work	Mean travel time for workers 16 years and over who did not work at home - Group quarters not included	2005-2016 American Community Survey PUMS as Augmented by NYC Opportunity
Households.w.Internet.Access	Percent of persons in NYC households with Internet access. Households with access include those with and without a subscription	2013-2016 American Community Survey PUMS, U.S. Bureau of the Census
Living.Walking.Distance.Park	Percentage of New Yorkers who live within a quarter-mile walk of a small park (under 6 acres) or a half-mile walk of a larger park (over 6 acres)	Department of Parks and Recreation; data derived from Walk to a Park Analysis run in July 2017 using 2010 Census Block Data
Foreclosure.Rate.per.1000	Total number of residential properties (single- and multi-family buildings and condominium apartment units) that had mortgage foreclosure actions initiated against them	NYU Furman Center Core Data http://coredata.nyc/
Number.of.Jobs.in.the.City	Number of job openings compiled from the Quarterly Employer Survey of Employers by the Bureau of Labor Statisitcs	U.S. Census Bureau, Quarterly Workforce Indicators (QWI) via NYC Economic Development Corporation
Outdoor.Air.Pollution	Average concentration of PM2.5 (fine particulate matter with a diameter smaller than 2.5 microns), micrograms (one-millionth of a gram) of PM2.5 per cubic meter of air	New York City Community Air Survey (NYCCAS)
Pct.Adults.Psych.Distress	Age-adjusted percent of adults with serious psychological distress who reported receiving mental health counseling or treatment in the last 12 months	NYC Community Health Survey 2009-10, 2012-13, 2015. CHS 2002-2008 data are weighted to the NYC adult population per Census 2000; starting in 2009, data are weighted to the 2008 HVS for phone usage and the Census 2000.CHS 2012 data are weighted to the adult residential population per the American Community Survey, 2011.CHS 2013 data are weighted to the adult residential population per the American Community Survey, 2012. CHS 2015 data are weighted to the adult residential population per the American Community Survey, 2014
Rental.Housing.Vacancy.Rate	The net rental vacancy rate is calculated by dividing the number of vacant, habitable, and available-for-rent units by the number of renter-occupied units plus vacant, habitable, and available for-rent units. This calculation excludes housing units in group quarters, such as hospitals, jails, mental institutions, and college dormitories, as well as units that are rented but not occupied and vacant units that are in such poor condition that they are not habitable	U.S. Census Bureau: NYC Housing and Vacancy Survey 2005, 2008, 2011, 2014) , HPD : Selected Initial Findings of the 2017 New York City Housing and Vacancy Survey
Srs.Housing.Code.Violations	Serious Housing Code Violations (class C) recorded by the Dept. of Housing Preservation and Development per 1000 rental units	New York City Department of Housing Preservation and Development, New York City Department of Finance Final Tax Roll File, New York City Housing Authority, NYU Furman Center

	Description	Data Source
Severely.Rent.Burdened.Households	The share of renter-occupied households whose gross rent (rent plus electricity and heating fuel costs) equaled at least 50 percent of their monthly pre-tax income, excluding those living in public housing or renting with the use of a voucher	U.S. Census Bureau: NYC Housing and Vacancy Survey (2005, 2008, 2011, 2014)
Total.SNAP.Recipients	Total low-income New Yorkers who receive critical nutrition assistance	NYC figures from HRA administrative data; NYS figures from NYS Office of Temporary and Disability Assistance; U.S. figures from the U.S. Department of Agriculture for 2012-2015 and from the Food Research and Action Center
Turnout.Among.Registered.Voters (percent)	Turnout among registered voters (Active + Inactive)	2016 Source: NYC Board of Elections Annual Report 2016, Voting Age Population Data from the U.S. Census Bureau 2015; 2014 Source: NYC Board of Elections voter file data, Population data from U.S. Census Bureau, 2009-2013 American Community Survey Five-Year Estimates
Turnout.Voting.Age.Pop (pct)	Turnout among NYC population age 18+	2016 Source: NYC Board of Elections Annual Report 2016, Voting Age Population Data from the U.S. Census Bureau 2015; 2014 Source: NYC Board of Elections voter file data, Population data from U.S. Census Bureau, 2009-2013 American Community Survey Five-Year Estimates
Unemployment.Rate	Percentage of the total labor force that is unemployed, but actively seeking employment and willing to work	2005-2016 American Community Survey 1-yr PUMS data, analyzed by NYC Opportunity
Violent.Crime.per.1000	Violent crime reports (murder, rape and sexual assault, robbery, and aggravated assault) per 1000 people.	NYPD Compstat for violent crimes (murder, rape, robbery, and felony assault)
Violent.Victimization.Youth.per.1000	NYPD complaints for violent crimes (murder, rape, robbery, and felony assault) where the victim was a youth or young adult between the ages of 16 and 24	NYPD Compstat for violent crimes (murder, rape, robbery, and felony assault), Population Estimates from the American Community Survey

```r
data_dict_pt2 <- data.frame(matrix(ncol = 43, nrow = 0))
colnames(data_dict_pt2) <- names(nyc_summary_joined)
data_dict_pt2 <- as.data.frame(t(as.matrix(data_dict_pt2)))
data_dict_pt2["Data Type"]<- c( "character", "character" , "integer","integer" , "integer" ,"integer", "integer", "integer","integer","integer","integer","integer","integer","integer", "integer","integer","integer","double", "double","double" ,"double", "double","double",
data_dict_pt2["Data Size"] <- c("CHAR(13)", "CHAR(4)" , "LONG INT","LONG INT" , "LONG INT" ,"LONG INT", "LONG INT", "LONG INT","LONG INT","LONG INT","LONG INT","LONG INT","LONG INT","LONG INT", "LONG INT","LONG INT","LONG INT","DOUBLE", "DOUBLE","DOUBLE" ,"DOUBLE", "DOUBL
knitr::kable(data_dict_pt2, format = 'markdown')
```

	Data Type	Data Size
Borough	character	CHAR(13)
Year	character	CHAR(4)
Calls.Total	integer	LONG INT
Calls.HPD	integer	LONG INT
Calls.DOT	integer	LONG INT
Calls.NYPD	integer	LONG INT
Calls.DEP	integer	LONG INT
Calls.DSNY	integer	LONG INT
Calls.Closed	integer	LONG INT
Calls.Open	integer	LONG INT
Calls.Pending	integer	LONG INT
Calls.Assigned	integer	LONG INT
Cmplts.Heating	integer	LONG INT
Cmplts.StreetCondition	integer	LONG INT
Cmplts.StreetLight	integer	LONG INT
Cmplts.GenConstruction	integer	LONG INT
Cmplts.Plumbing	integer	LONG INT
Corrections.Pop.per.1000	double	DOUBLE
Average.Weekly.Earnings	double	DOUBLE
Curbside.and.Containerized.Diversion.Rate	double	DOUBLE
Probation.Population	double	DOUBLE
Probation.Pop.per.1000	double	DOUBLE
Disconnected.Youth	double	DOUBLE
Eligible.Voter.Registration.Rate	double	DOUBLE
Enrolled.Full.Day.Pre-K.Students	double	DOUBLE
Four-Year.High.School.Graduation.Rate	double	DOUBLE
Labor.Force.Participation.Rate	double	DOUBLE
Mean.Travel.Time.to.Work	double	DOUBLE
Households.w.Internet.Access	double	DOUBLE
Living.Walking.Distance.Park	double	DOUBLE
Foreclosure.Rate.per.1000	double	DOUBLE
Number.of.Jobs.in.the.City	double	DOUBLE
Outdoor.Air.Pollution	double	DOUBLE
Pct.Adults.Psych.Distress	double	DOUBLE
Rental.Housing.Vacancy.Rate	double	DOUBLE
Srs.Housing.Code.Violations	double	DOUBLE
Severely.Rent-Burdened.Households	double	DOUBLE
Total.SNAP.Recipients	double	DOUBLE
Turnout.Among.Registered.Voters.(percent)	double	DOUBLE
Turnout.Voting.Age.Pop.(pct)	double	DOUBLE
Unemployment.Rate	double	DOUBLE
Violent.Crime.per.1000	double	DOUBLE
Violent.Victimization.Youth.per.1000	double	DOUBLE

Data dictionary for the NYC 311 call dataset

```r
col_names <- names(nyc311)
data_dict2 <- data.frame(matrix(ncol = 55, nrow = 0))
colnames(data_dict2) <- col_names
data_dict2 <- as.data.frame(t(as.matrix(data_dict2)))
data_dict2["Description"] <- c("Unique call ID", "Date of call", "Time of call",
    "Date the case was closed", "Time the case was closed", "Agency assigned to call (abbr.)",
```

```
    "Name of agency", "Type of complaint", "Description of scene", "Type of location",
    "Zip code for incident", "Address for incident", "Street name", "First cross street",
    "Second cross street", "First street intersection", "Second street intersection",
    "Type of Address", "Nearest landmark", "Type of facility", "Current status of case",
    "Assigned due date", "Assigned due time", "Date last updated", "Time last updated",
    "Community board assigned", "Borough of incident", "X coordinate in state plane",
    "Y coordinate in state plane", "Park facility name", "Park borough", "School name",
    "School ID number", "School region", "School code", "School phone number",
    "School address", "City of school location", "State of school location",
    "Zip code of school", "School not found", "School or city-wide complaint",
    "Type of vehicle", "Borough of taxi company", "Location of taxi pick-up",
    "Bridge or highway name", "Bridge or highway direction", "Road onramp",
    "Segment of bridge or highway", "Name of garage or lot", "Direction of ferry",
    "Name of ferry terminal", "Latitude of incident location", "Longitude of incident location",
    "Year")
data_dict2["Data Source"] <- c("311 call database", "311 operator", "311 operator",
    "Agency assigned", "Agency assigned", "311 operator", "311 operator", "311 operator",
    "Agency assigned", "311 operator", "311 operator", "311 operator", "311 operator",
    "311 operator", "311 operator", "311 operator", "311 operator", "311 operator",
    "311 operator", "311 operator", "311 operator", "Agency assigned", "Agency assigned",
    "Agency assigned", "Agency assigned", "311 operator", "311 operator", "311 call database",
    "311 call database", "311 operator", "311 operator", "311 operator", "311 operator",
    "311 operator", "311 operator", "311 operator", "311 operator", "311 operator",
    "311 operator", "311 operator", "311 operator", "311 operator", "311 operator",
    "311 operator", "311 operator", "311 operator", "311 operator", "311 operator",
    "311 operator", "311 operator", "311 operator", "311 operator", "311 call database",
    "311 call database", "311 operator")
data_dict2["Data Type"] <- c("integer", "character", "character", "character",
    "character", "character", "character", "character", "character", "character",
    "character", "character", "character", "character", "character", "character",
    "character", "character", "character", "character", "character", "character",
    "character", "character", "character", "character", "character", "integer",
    "integer", "character", "character", "character", "character", "character",
    "character", "character", "character", "character", "character", "character",
    "character", "character", "character", "character", "character", "character",
    "character", "character", "character", "character", "character", "character",
    "real", "real", "character")
data_dict2["Data Size"] <- c("LONG INT", "CHAR(10)", "CHAR(11)", "CHAR(10)",
```

```
    "CHAR(11)", "CHAR(10)", "CHAR(91)", "CHAR(41)", "CHAR(104)", "CHAR(36)",
    "CHAR(10)", "CHAR(80)", "CHAR(80)", "CHAR(36)", "CHAR(36)", "CHAR(38)",
    "CHAR(47)", "CHAR(12)", "CHAR(48)", "CHAR(15)", "CHAR(28)", "CHAR(10)",
    "CHAR(11)", "CHAR(10)", "CHAR(11)", "CHAR(25)", "CHAR(13)", "LONG INT",
    "LONG INT", "CHAR(95)", "CHAR(13)", "CHAR(95)", "CHAR(8)", "CHAR(27)", "CHAR(6)",
    "CHAR(10)", "CHAR(120)", "CHAR(19)", "CHAR(2)", "CHAR(5)", "CHAR(1)", "CHAR(18)",
    "CHAR(23)", "CHAR(13)", "CHAR(27)", "CHAR(42)", "CHAR(30)", "CHAR(7)", "CHAR(100)",
    "CHAR(27)", "CHAR(19)", "CHAR(95)", "DOUBLE", "DOUBLE", "CHAR(4)")
knitr::kable(data_dict2, format = "markdown")
```

	Description	Data Source	Data Type	Data Size
Unique.Key	Unique call ID	311 call database	integer	LONG INT
Created.Date	Date of call	311 operator	character	CHAR(10)
Created.Time	Time of call	311 operator	character	CHAR(11)
Closed.Date	Date the case was closed	Agency assigned	character	CHAR(10)
Closed.Time	Time the case was closed	Agency assigned	character	CHAR(11)
Agency	Agency assigned to call (abbr.)	311 operator	character	CHAR(10)
Agency.Name	Name of agency	311 operator	character	CHAR(91)
Complaint.Type	Type of complaint	311 operator	character	CHAR(41)
Descriptor	Description of scene	Agency assigned	character	CHAR(104)
Location.Type	Type of location	311 operator	character	CHAR(36)
Incident.Zip	Zip code for incident	311 operator	character	CHAR(10)
Incident.Address	Address for incident	311 operator	character	CHAR(80)
Street.Name	Street name	311 operator	character	CHAR(80)
Cross.Street.1	First cross street	311 operator	character	CHAR(36)
Cross.Street.2	Second cross street	311 operator	character	CHAR(36)
Intersection.Street.1	First street intersection	311 operator	character	CHAR(38)
Intersection.Street.2	Second street intersection	311 operator	character	CHAR(47)
Address.Type	Type of Address	311 operator	character	CHAR(12)
Landmark	Nearest landmark	311 operator	character	CHAR(48)
Facility.Type	Type of facility	311 operator	character	CHAR(15)
Status	Current status of case	311 operator	character	CHAR(28)
Due.Date	Assigned due date	Agency assigned	character	CHAR(10)
Due.Time	Assigned due time	Agency assigned	character	CHAR(11)
Resolution.Action.Updated.Date	Date last updated	Agency assigned	character	CHAR(10)
Resolution.Action.Updated.Time	Time last updated	Agency assigned	character	CHAR(11)
Community.Board	Community board assigned	311 operator	character	CHAR(25)
Borough	Borough of incident	311 operator	character	CHAR(13)

	Description	Data Source	Data Type	Data Size
X.Coordinate.(State.Plane)	X coordinate in state plane	311 call database	integer	LONG INT
Y.Coordinate.(State.Plane)	Y coordinate in state plane	311 call database	integer	LONG INT
Park.Facility.Name	Park facility name	311 operator	character	CHAR(95)
Park.Borough	Park borough	311 operator	character	CHAR(13)
School.Name	School name	311 operator	character	CHAR(95)
School.Number	School ID number	311 operator	character	CHAR(8)
School.Region	School region	311 operator	character	CHAR(27)
School.Code	School code	311 operator	character	CHAR(6)
School.Phone.Number	School phone number	311 operator	character	CHAR(10)
School.Address	School address	311 operator	character	CHAR(120)
School.City	City of school location	311 operator	character	CHAR(19)
School.State	State of school location	311 operator	character	CHAR(2)
School.Zip	Zip code of school	311 operator	character	CHAR(5)
School.Not.Found	School not found	311 operator	character	CHAR(1)
School.or.Citywide.Complaint	School or city-wide complaint	311 operator	character	CHAR(18)
Vehicle.Type	Type of vehicle	311 operator	character	CHAR(23)
Taxi.Company.Borough	Borough of taxi company	311 operator	character	CHAR(13)
Taxi.Pick.Up.Location	Location of taxi pick-up	311 operator	character	CHAR(27)
Bridge.Highway.Name	Bridge or highway name	311 operator	character	CHAR(42)
Bridge.Highway.Direction	Bridge or highway direction	311 operator	character	CHAR(30)
Road.Ramp	Road onramp	311 operator	character	CHAR(7)
Bridge.Highway.Segment	Segment of bridge or highway	311 operator	character	CHAR(100)
Garage.Lot.Name	Name of garage or lot	311 operator	character	CHAR(27)
Ferry.Direction	Direction of ferry	311 operator	character	CHAR(19)
Ferry.Terminal.Name	Name of ferry terminal	311 operator	character	CHAR(95)
Latitude	Latitude of incident location	311 call database	real	DOUBLE
Longitude	Longitude of incident location	311 call database	real	DOUBLE
Year	Year	311 operator	character	CHAR(4)