Project

Monica Puerto

3/10/2019

Introduction

The number 311 is s Smith cited[@Smith2009]

```
suppressPackageStartupMessages(library(tidyverse))
```

```
## Warning: package 'ggplot2' was built under R version 3.4.4
```

```
## Warning: package 'tibble' was built under R version 3.4.4
```

```
## Warning: package 'tidyr' was built under R version 3.4.4
```

```
## Warning: package 'purrr' was built under R version 3.4.4
```

```
## Warning: package 'dplyr' was built under R version 3.4.4
```

```
## Warning: package 'forcats' was built under R version 3.4.3
```

```
#library(tidyverse)
requests <- read_csv('City_Service_Requests_in_2018.csv')
```

```
## Parsed with column specification:
## cols(
##   .default = col_character(),
##   X = col_double(),
##   Y = col_double(),
##   OBJECTID = col_integer(),
##   SERVICECALLCOUNT = col_integer(),
##   ADDDATE = col_datetime(format = ""),
##   RESOLUTIONDATE = col_datetime(format = ""),
##   SERVICEDUEDATE = col_datetime(format = ""),
##   SERVICEORDERDATE = col_datetime(format = ""),
##   INSPECTIONDATE = col_datetime(format = ""),
##   XCOORD = col_double(),
##   YCOORD = col_double(),
##   LATITUDE = col_double(),
##   LONGITUDE = col_double(),
##   ZIPCODE = col_integer(),
##   MARADDRESSREPOSITORYID = col_integer()
## )
```

```
## See spec(...) for full column specifications.
```

Data observations

There were 333,105 city requests made in DC across a population of X. (verify if you have to be a DC resident to do a 311 request) This dataset had 30 variables of information which of consisted of timestamps, numeric, and categorical such as service codes. Of the 30 columns, there were 12 variables with NAs. The columns with the most NAs were about the Inspector Name and Inspection Date which were 100% and 92% comprised of NAs respectively. That would have been interesting to analyze if certain Inspectors were assigned to certain

Wards, how many inspectors per Ward and the ratio of the population to inspectors in each Ward. A column that I do care about is resolution date and I was alarmed when I saw about ~53K rows blank but that just comprises of 16% of the data; which is still a hefty chunk of the data but at least it was not over 1/2 of the data and just 1/6.

The

```
summary(requests)
```

```
## Warning in as.POSIXlt.POSIXct(x, tz): unknown timezone 'zone/tz/2018i.1.0/
## zoneinfo/America/New_York'
```

```
##        X                Y             OBJECTID      SERVICECODE
##  Min.   :-77.11   Min.   :38.81   Min.   :306864   Length:333105
##  1st Qu.:-77.04   1st Qu.:38.89   1st Qu.:420161   Class :character
##  Median :-77.02   Median :38.91   Median :518266   Mode  :character
##  Mean   :-77.01   Mean   :38.91   Mean   :516828
##  3rd Qu.:-76.99   3rd Qu.:38.93   3rd Qu.:614640
##  Max.   :-76.91   Max.   :39.00   Max.   :734178
##
##  SERVICECODEDESCRIPTION SERVICETYPECODEDESCRIPTION ORGANIZATIONACRONYM
##  Length:333105          Length:333105              Length:333105
##  Class :character       Class :character           Class :character
##  Mode  :character       Mode  :character           Mode  :character
##
##
##
##
##  SERVICECALLCOUNT    ADDDATE
##  Min.   :1        Min.   :2018-01-01 00:05:35
##  1st Qu.:1        1st Qu.:2018-04-09 07:50:35
##  Median :1        Median :2018-07-02 12:27:45
##  Mean   :1        Mean   :2018-07-01 15:31:45
##  3rd Qu.:1        3rd Qu.:2018-09-22 13:50:43
##  Max.   :1        Max.   :2018-12-31 23:32:09
##
##  RESOLUTIONDATE                SERVICEDUEDATE
##  Min.   :2018-01-01 00:19:26   Min.   :2018-01-02 01:00:08
##  1st Qu.:2018-04-06 16:55:46   1st Qu.:2018-04-27 17:00:00
##  Median :2018-06-26 15:21:08   Median :2018-07-30 17:00:00
##  Mean   :2018-06-26 14:35:54   Mean   :2018-08-07 03:14:08
##  3rd Qu.:2018-09-06 13:42:49   3rd Qu.:2018-10-24 17:00:00
##  Max.   :2019-01-09 15:01:37   Max.   :2020-12-10 14:18:00
##  NA's   :52756                 NA's   :9
##  SERVICEORDERDATE              INSPECTIONFLAG
##  Min.   :2018-01-01 00:05:35   Length:333105
##  1st Qu.:2018-04-09 07:50:35   Class :character
##  Median :2018-07-02 12:27:45   Mode  :character
##  Mean   :2018-07-01 15:31:45
##  3rd Qu.:2018-09-22 13:50:43
##  Max.   :2018-12-31 23:32:09
##
##  INSPECTIONDATE                INSPECTORNAME      SERVICEORDERSTATUS
##  Min.   :2018-01-02 14:50:00   Length:333105      Length:333105
##  1st Qu.:2018-04-18 19:03:15   Class :character   Class :character
##  Median :2018-07-13 15:00:30   Mode  :character   Mode  :character
```

```
##  Mean   :2018-07-09 18:03:37
##  3rd Qu.:2018-10-01 11:06:30
##  Max.   :2019-01-09 10:48:00
##  NA's   :307183
##  STATUS_CODE        SERVICEREQUESTID     PRIORITY
##  Length:333105      Length:333105      Length:333105
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  STREETADDRESS          XCOORD           YCOORD          LATITUDE
##  Length:333105      Min.   :390090   Min.   :127297   Min.   :38.81
##  Class :character   1st Qu.:396661   1st Qu.:136286   1st Qu.:38.89
##  Mode  :character   Median :398427   Median :137961   Median :38.91
##                     Mean   :398786   Mean   :138199   Mean   :38.91
##                     3rd Qu.:400936   3rd Qu.:140493   3rd Qu.:38.93
##                     Max.   :407848   Max.   :147500   Max.   :39.00
##
##    LONGITUDE          CITY              STATE              ZIPCODE
##  Min.   :-77.11   Length:333105      Length:333105      Min.   :-4453
##  1st Qu.:-77.04   Class :character   Class :character   1st Qu.:20004
##  Median :-77.02   Mode  :character   Mode  :character   Median :20011
##  Mean   :-77.01                                         Mean   :20016
##  3rd Qu.:-76.99                                         3rd Qu.:20019
##  Max.   :-76.91                                         Max.   :83127
##                                                         NA's   :14
##  MARADDRESSREPOSITORYID     WARD             DETAILS
##  Min.   :     1         Length:333105      Length:333105
##  1st Qu.: 72198         Class :character   Class :character
##  Median :243465         Mode  :character   Mode  :character
##  Mean   :322083
##  3rd Qu.:307138
##  Max.   :914266
##  NA's   :2629
```

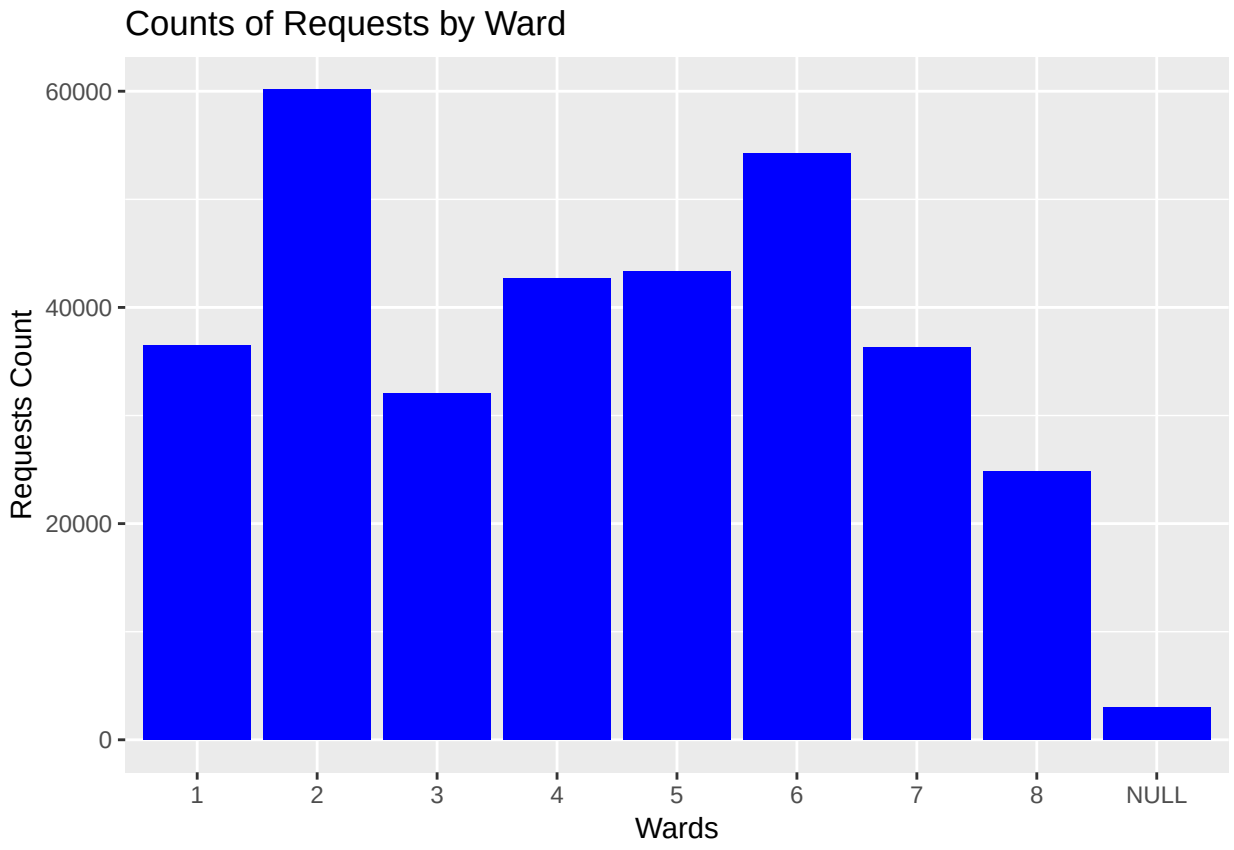
```
column_names = c(names(requests))

requests %>%
  summarize_all(funs(sum(is.na(.)))) %>%
  gather(key=column_names,value="NA_COUNT") %>%
  arrange(desc(NA_COUNT)) %>%
  filter(NA_COUNT > 0) %>%
  mutate(NA_PCT = round((NA_COUNT/nrow(requests)*100)))
```

```
## Warning: package 'bindrcpp' was built under R version 3.4.4
```

```
## # A tibble: 12 x 3
##    column_names               NA_COUNT NA_PCT
##    <chr>                         <int>  <dbl>
##  1 INSPECTORNAME                333105    100
##  2 INSPECTIONDATE               307183     92
##  3 DETAILS                       86131     26
##  4 RESOLUTIONDATE                52756     16
```

```
##  5 STREETADDRESS                 22540      7
##  6 CITY                          22539      7
##  7 STATE                         22539      7
##  8 MARADDRESSREPOSITORYID         2629      1
##  9 ZIPCODE                          14      0
## 10 SERVICEDUEDATE                    9      0
## 11 SERVICETYPECODEDESCRIPTION        1      0
## 12 ORGANIZATIONACRONYM               1      0
```

```
ggplot(requests,aes(x=WARD)) +
  geom_bar(fill='blue') +
  xlab("Wards") +
  ylab("Requests Count") +
  ggtitle("Counts of Requests by Ward")
```




```
requests %>%
  group_by(WARD,SERVICECODEDESCRIPTION) %>%
  count() %>%
  arrange(desc(n))
```

```
## # A tibble: 810 x 3
## # Groups:   WARD, SERVICECODEDESCRIPTION [810]
##    WARD  SERVICECODEDESCRIPTION     n
##    <chr> <chr>                  <int>
##  1 2     Parking Meter Repair   26532
##  2 4     Bulk Collection        10294
##  3 7     Bulk Collection         9174
##  4 5     Bulk Collection         9023
```

```
##  5 6     Parking Enforcement     8835
##  6 6     Bulk Collection         8114
##  7 2     Parking Enforcement     7420
##  8 1     Parking Enforcement     6389
##  9 8     Bulk Collection         5726
## 10 5     Parking Enforcement     5616
## # ... with 800 more rows
```

```
length(unique(requests$SERVICECODEDESCRIPTION))
```

```
## [1] 109
```

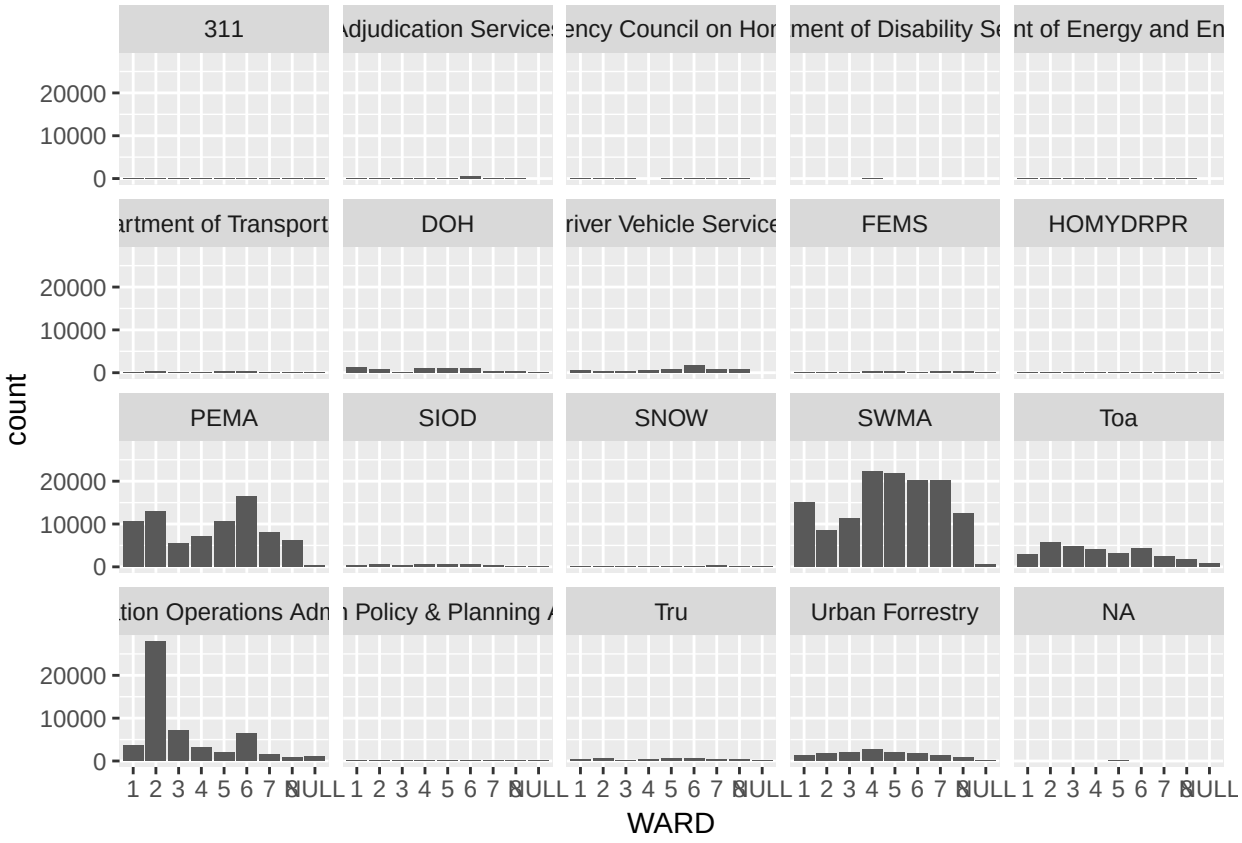
```
length(unique(requests$SERVICECODE))
```

```
## [1] 106
```

```
requests %>%
separate(SERVICETYPECODEDESCRIPTION,into=c('CODE','DESCODE'),sep='-') %>%
  ggplot(aes(x=WARD)) +
  geom_bar() +
  facet_wrap(~CODE)
```

```
## Warning: Expected 2 pieces. Additional pieces discarded in 11518 rows [44,
## 184, 208, 461, 472, 487, 492, 653, 681, 783, 860, 864, 1090, 1240, 1420,
## 1870, 2138, 2223, 2379, 2398, ...].
```

```
## Warning: Expected 2 pieces. Missing pieces filled with `NA` in 81050 rows
## [4, 14, 26, 42, 50, 72, 76, 79, 83, 85, 93, 94, 98, 102, 116, 131, 138,
## 141, 142, 145, ...].
```



```
wards_demo <- read_csv('Ward_from_2012.csv')
```

```
## Parsed with column specification:
## cols(
##   .default = col_integer(),
##   NAME = col_character(),
##   REP_NAME = col_character(),
##   WEB_URL = col_character(),
##   REP_PHONE = col_character(),
##   REP_EMAIL = col_character(),
##   REP_OFFICE = col_character(),
##   LABEL = col_character(),
##   AREASQMI = col_double(),
##   Shape_Length = col_double(),
##   Shape_Area = col_double(),
##   MEDIAN_AGE = col_double(),
##   UNEMPLOYMENT_RATE = col_double(),
##   PCT_FAMILY_HH = col_double(),
##   PCT_NONFAMILY_HH = col_double(),
##   PCT_BELOW_POV = col_double(),
##   PCT_BELOW_POV_FAM = col_double(),
##   PCT_BELOW_POV_WHITE = col_double(),
##   PCT_BELOW_POV_BLACK = col_double(),
##   PCT_BELOW_POV_NAT_AMER = col_double(),
##   PCT_BELOW_POV_ASIAN = col_double()
##   # ... with 8 more columns
## )
```

```
## See spec(...) for full column specifications.
```

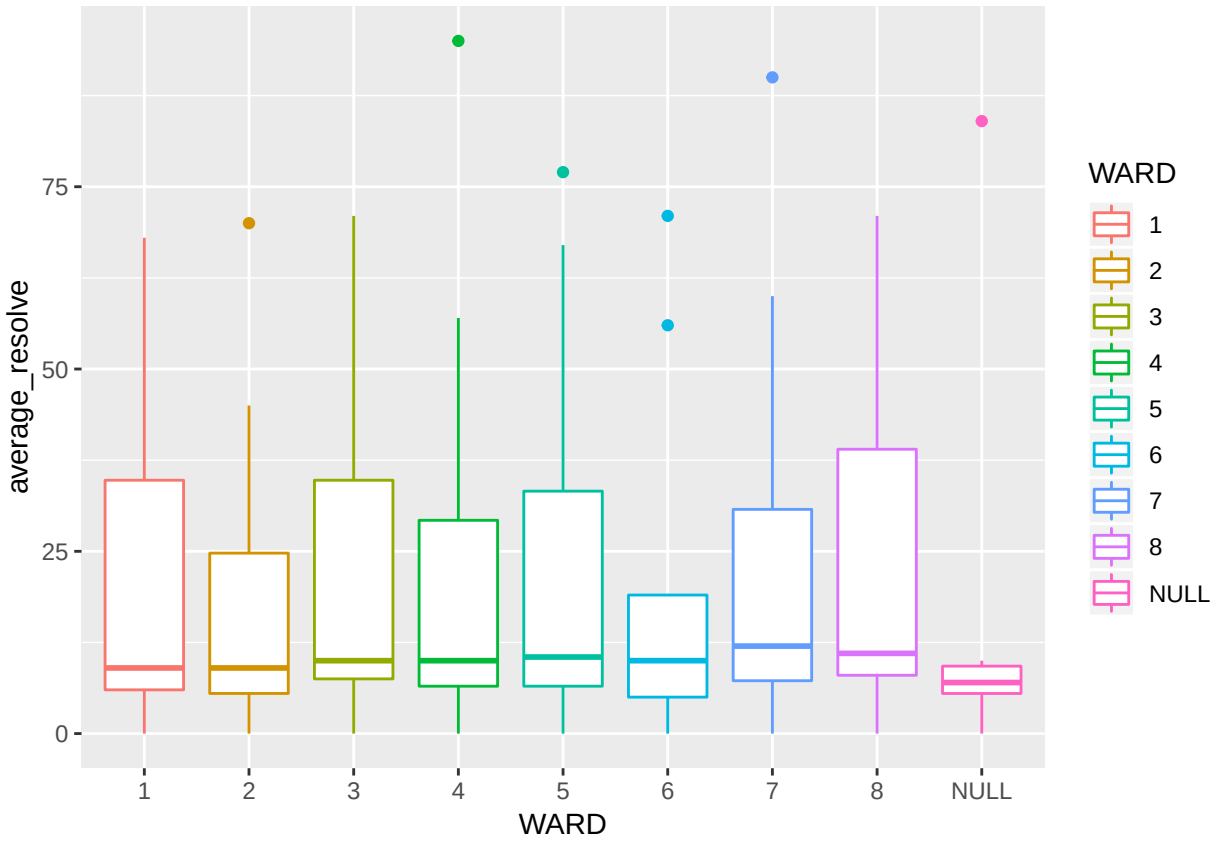
```
col_names = c(names(wards_demo))

wards_demo %>%
  summarize_all(funs(sum(is.na(.)))) %>%
  gather(key=column_names,value="NA_COUNT") %>%
  arrange(desc(NA_COUNT)) %>%
  filter(NA_COUNT > 0) %>%
  mutate(NA_PCT = round((NA_COUNT/nrow(wards_demo)*100)))
```

```
## # A tibble: 2 x 3
##   column_names          NA_COUNT NA_PCT
##   <chr>                    <int>  <dbl>
## 1 POP_25_PLUS                  8    100
## 2 MARRIED_COUPLE_FAMILY        8    100
```

```
requests %>%
  mutate(resolve_time_days = round(difftime(RESOLUTIONDATE,ADDDATE,units = "days"))) %>%
  group_by(WARD,ORGANIZATIONACRONYM) %>%
  summarise(average_resolve = round(mean(resolve_time_days,na.rm = TRUE))) %>%
  arrange(desc(average_resolve)) %>%
  ggplot(aes(WARD,average_resolve,color=WARD)) +
  geom_boxplot()
```

```
## Don't know how to automatically pick scale for object of type difftime. Defaulting to continuous.
```

```
## Warning: Removed 7 rows containing non-finite values (stat_boxplot).
```


average_resolve
75
50
25
0
1
2
3
4
5
6
7
8
NULL
WARD
WARD
1
2
3
4
5
6
7
8
NULL