Project 1

Zechen Wang (zew20@pitt.edu)

January 28, 2019

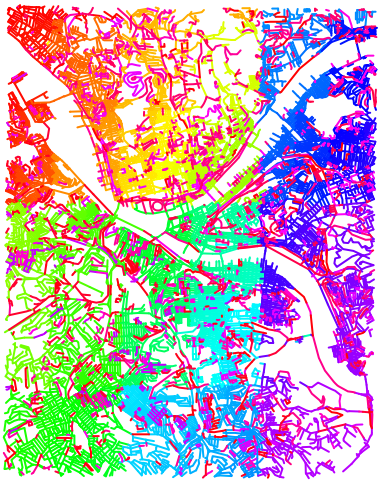
Load libraries and data sets

```
setwd("D:/workspace/r/INFSCI 2809 Spatial Data Analytics/Project 1")

library(sp)
library(spdep)
library(classInt)
library(rgeos)
library(maptools)
library(rgdal)
library(ggplot2)
library(weights)
library(ape)
library(GISTools)
library(maps)
library(raster)
library(dplyr)
library(plyr)

filename <- "pgh_streets/pgh_streets/pgh_streets.shp"
s <- shapefile(filename, stringsAsFactors=T)
summary(s)
```

```
## Object of class SpatialLinesDataFrame
## Coordinates:
##         min       max
## x -80.06888 -79.93147
## y  40.37077  40.50327
## Is projected: FALSE
## proj4string :
## [+proj=longlat +datum=NAD83 +no_defs +ellps=GRS80 +towgs84=0,0,0]
## Data attributes:
##        TLID           FNODE           TNODE           LENGTH
##  51603411:    1   Min.   :11708   Min.   :11674   Min.   :0.00030
##  51616344:    1   1st Qu.:18767   1st Qu.:18695   1st Qu.:0.02762
##  51619508:    1   Median :24873   Median :24940   Median :0.04275
##  51619527:    1   Mean   :24743   Mean   :24762   Mean   :0.05980
##  51619529:    1   3rd Qu.:30607   3rd Qu.:30704   3rd Qu.:0.07596
##  51619535:    1   Max.   :37414   Max.   :37473   Max.   :1.46654
##  (Other) :22216
##   FEDIRP           FENAME          FETYPE      FEDIRS           CFCC
##  E   :  294   Liberty :  132   St     :9712   E   :    5   A41    :20522
##  N   :  281   Penn    :   91   Ave    :5472   S   :    3   A31    :  296
##  S   :  472   5th     :   85   Way    :2288   W   :    5   A73    :  293
##  W   :  229   Carson  :   82   Rd     :1195   NA's:22209   A35    :  289
##  NA's:20946   Brighton:   81   Dr     : 891                A25    :  243
##               (Other) :20682   (Other):1263                A15    :  181
```

```
##                NA's    : 1069   NA's   :1401                (Other):  398
##      FRADDL           TOADDL           FRADDR           TOADDR
##  1       :  862   199     :  632   2       :  856   198     :  625
##  101     :  646   99      :  603   100     :  634   98      :  564
##  201     :  534   299     :  517   200     :  536   298     :  515
##  301     :  430   399     :  444   300     :  435   398     :  454
##  401     :  373   499     :  360   400     :  370   498     :  341
##  (Other):14079   (Other):14368   (Other):14076   (Other):14408
##  NA's    : 5298   NA's    : 5298   NA's    : 5315   NA's    : 5315
##       ZIPL             ZIPR        CENSUS1   CENSUS2   CFCC1     CFCC2
##  15212  : 2081   15212  : 2064   0:22222   0:22222   A:22219   A1:  204
##  15210  : 1757   15210  : 1751                       P:    3   A2:  302
##  15216  :  945   15216  :  931                                 A3:  585
##  15219  :  896   15219  :  899                                 A4:20534
##  15201  :  869   15201  :  850                                 A6:  180
##  (Other):10376   (Other):10412                                 A7:  414
##  NA's   : 5298   NA's   : 5315                                 P4:    3
##  SOURCE
##  A:19635
##  J:  903
##  K: 1114
##  L:  105
##  M:  122
##  N:  342
##  O:    1
```

```
n <- length(s)
plot(s, col=rainbow(n))
```



Spatial Data Manipulation

- Find the total number of road segments

Total number of road segments is 1195.

```
s_Rd <- s[s$FETYPE=="Rd",]
nrow(s_Rd)
```

```
## [1] 1195
```

- Calculate the minimum, maximum and mean segment lengths

Min: 0.00084, Max: 0.90640, Mean: 0.08383

```
summary(s_Rd)
```

```
## Object of class SpatialLinesDataFrame
## Coordinates:
##         min       max
## x -80.06888 -79.93317
## y  40.37091  40.50238
## Is projected: FALSE
## proj4string :
## [+proj=longlat +datum=NAD83 +no_defs +ellps=GRS80 +towgs84=0,0,0]
## Data attributes:
##        TLID          FNODE           TNODE           LENGTH
##  51619568:   1   Min.   :11708   Min.   :11675   Min.   :0.00084
##  51619614:   1   1st Qu.:15762   1st Qu.:15690   1st Qu.:0.03441
```

```
##  51619626:   1   Median :22063   Median :22067   Median :0.05853
##  51620423:   1   Mean   :22825   Mean   :22819   Mean   :0.08383
##  51620425:   1   3rd Qu.:29359   3rd Qu.:29426   3rd Qu.:0.10789
##  51620736:   1   Max.   :37208   Max.   :37211   Max.   :0.90640
##  (Other) :1189
##   FEDIRP             FENAME        FETYPE      FEDIRS          CFCC
##  E   :  20   Brownsville: 78   Rd     :1195   E   :   0   A41    :1021
##  N   :   0   Brighton   : 75   Aly    :   0   S   :   0   A25    :  71
##  S   :   0   Greentree  : 52   Ave    :   0   W   :   0   A31    :  49
##  W   :  13   Mount Troy : 42   Blvd   :   0   NA's:1195   A35    :  48
##  NA's:1162   Cochran    : 35   Brg    :   0               A73    :   3
##              Noblestown : 34   Byp    :   0               A21    :   1
##              (Other)    :879   (Other):   0               (Other):   2
##      FRADDL        TOADDL        FRADDR        TOADDR         ZIPL
##  1      : 38   99     : 28   2      : 38   98     : 27   15212  :132
##  101    : 26   299    : 27   400    : 24   298    : 26   15227  :118
##  401    : 23   599    : 22   100    : 23   598    : 24   15220  : 82
##  201    : 22   799    : 20   200    : 21   798    : 21   15210  : 80
##  501    : 20   199    : 19   800    : 20   198    : 19   15214  : 80
##  (Other):875   (Other):888   (Other):887   (Other):896   (Other):512
##  NA's   :191   NA's   :191   NA's   :182   NA's   :182   NA's   :191
##       ZIPR     CENSUS1  CENSUS2  CFCC1    CFCC2     SOURCE
##  15212  :127   0:1195   0:1195   A:1194   A1:   0   A:1066
##  15227  :115                     P:   1   A2:  72   J:  11
##  15220  : 87                              A3:  97   K:  85
##  15210  : 81                              A4:1021   L:   4
##  15228  : 79                              A6:   0   M:   6
##  (Other):524                              A7:   4   N:  23
##  NA's   :182                              P4:   1   O:   0
```

- Filter out the segments that are below the mean length that you calculated in (b) and then create a map showing the remaining segments.

```
s_Rd_filtered <- s_Rd[s_Rd$LENGTH >= mean(s_Rd$LENGTH),]
plot(s_Rd_filtered)
```

Spatial Data Aggregation

- Aggregate the data based on the mean of the point values. Create a map and prepare a report on the result.

Explore the data set first.

```
load("lnd.RData")
load("stations.RData")

plot(lnd)
points(stations)
```

Aggregate the data based on the mean of the point values.

```
stations_agg <- over(lnd, stations[c("coords.x1", "coords.x2", "Value")], fn = mean)
plot(lnd)
points(stations_agg)
```

```
lnd_stations <- lnd
lnd_stations@data <- cbind(lnd@data, stations_agg)
colnames(lnd_stations@data)[1] <- "id"
head(lnd_stations@data)
```

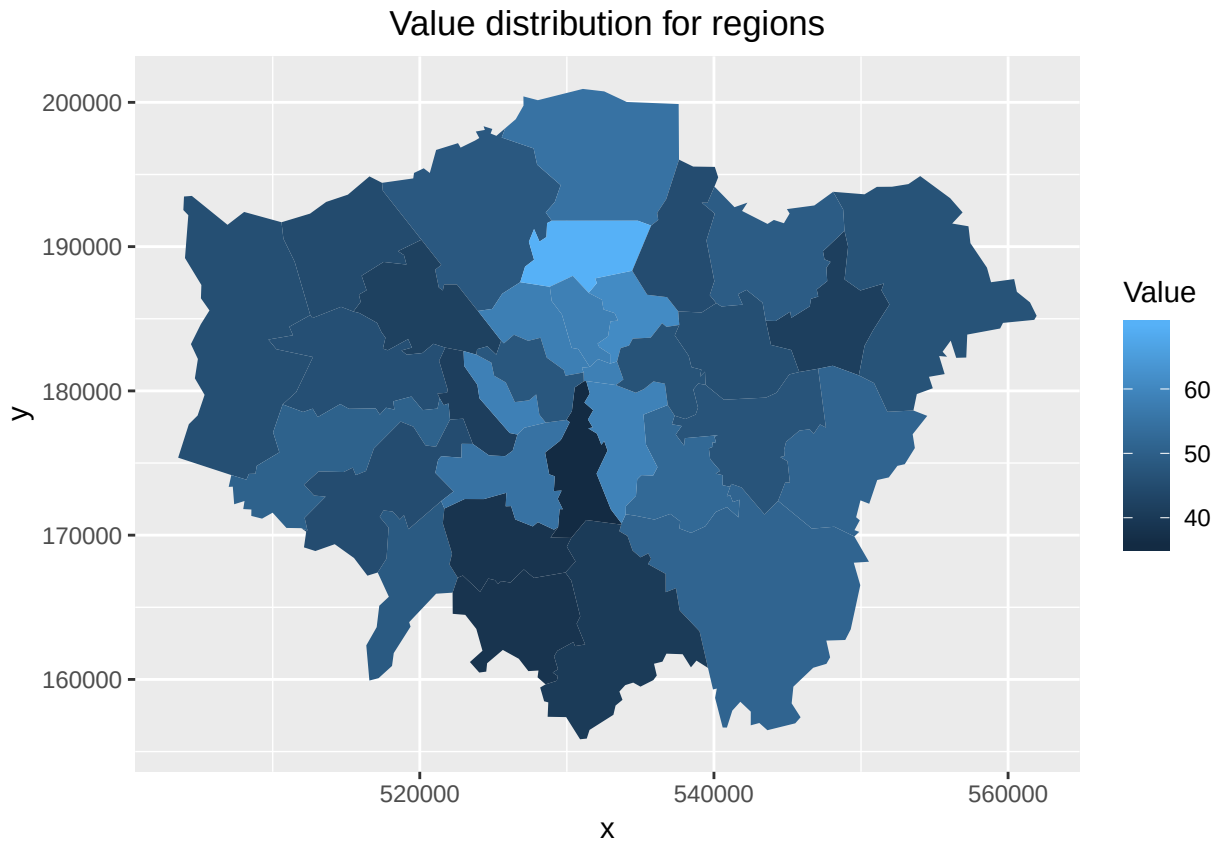
```
##     id                 name Partic_Per Pop_2001 CrimeCount station.count
## 0 00AF              Bromley       21.7   295535      15172            48
## 1 00BD Richmond upon Thames       26.6   172330       9715            22
## 2 00AS           Hillingdon       21.5   243006      15302            43
## 3 00AR             Havering       17.9   224262      12611            18
## 4 00AX Kingston upon Thames       24.4   147271       9023            12
## 5 00BF               Sutton       19.3   179767       8810            13
##   coords.x1 coords.x2    Value
## 0  540217.1  167562.6 51.22460
## 1  517046.3  173448.1 45.08194
## 2  508002.0  182491.3 45.70525
## 3  552512.5  187192.4 46.72895
## 4  519322.7  166832.4 48.53035
## 5  527114.6  164839.9 38.35842
```

```
lnd_stations.points <- fortify(lnd_stations, region = "id")
lnd_stations.points <- join(lnd_stations.points, lnd_stations@data, by="id")
head(lnd_stations.points)
```

```
##       long      lat order  hole piece   id  group           name
## 1 531026.9 181611.1     1 FALSE     1 00AA 00AA.1 City of London
## 2 531554.9 181659.3     2 FALSE     1 00AA 00AA.1 City of London
```

```
## 3 532135.6 182198.4     3 FALSE     1 00AA 00AA.1 City of London
## 4 532946.1 181894.8     4 FALSE     1 00AA 00AA.1 City of London
## 5 533410.7 182037.9     5 FALSE     1 00AA 00AA.1 City of London
## 6 533842.7 180793.6     6 FALSE     1 00AA 00AA.1 City of London
##   Partic_Per Pop_2001 CrimeCount station.count coords.x1 coords.x2
## 1        9.1     7181       3735             5  532820.7  181089.2
## 2        9.1     7181       3735             5  532820.7  181089.2
## 3        9.1     7181       3735             5  532820.7  181089.2
## 4        9.1     7181       3735             5  532820.7  181089.2
## 5        9.1     7181       3735             5  532820.7  181089.2
## 6        9.1     7181       3735             5  532820.7  181089.2
##      Value
## 1 58.21482
## 2 58.21482
## 3 58.21482
## 4 58.21482
## 5 58.21482
## 6 58.21482
```

```
ggplot(data = lnd_stations.points, aes(x = long, y = lat, group = id, fill = Value)) + geom_polygon() +
    labs(x = "x", y = "y") + theme(plot.title = element_text(hjust = 0.5))
```



- Run regression on the point values before and after aggregation.

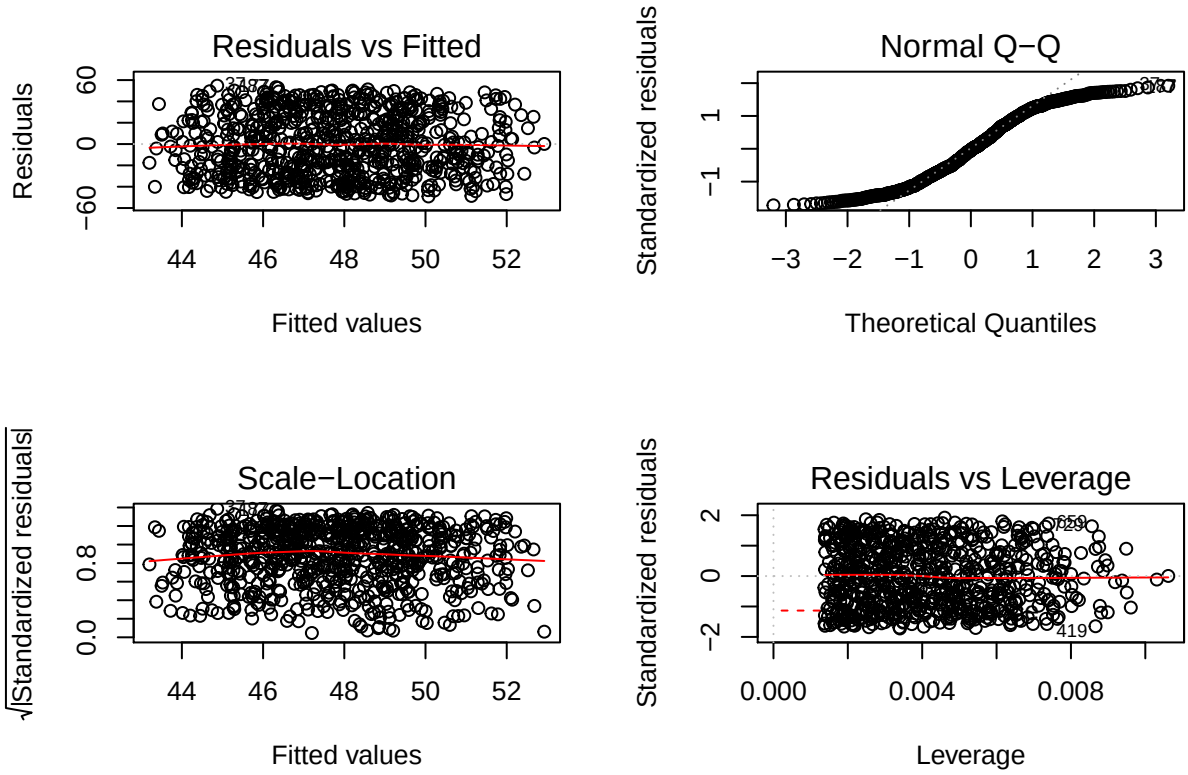
We can see that both p-value is larger than 0.05, which indicates that they are bad models, the coordinates and value may not have relationship at all. Even though, we can still find out that after aggregation, the R-sqaured has a huge increase, which means that the aggregation of the data may exaggerate the relationship

between the variables.

```
par(mfrow=c(2,2))

lm_before <- lm(Value ~ coords.x1 + coords.x2, data = stations)
summary(lm_before)
```

```
##
## Call:
## lm(formula = Value ~ coords.x1 + coords.x2, data = stations)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -49.138 -24.226  -1.232  25.363  54.759
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.419e+01  5.422e+01  -0.815   0.4154
## coords.x1    1.033e-04  9.003e-05   1.147   0.2517
## coords.x2    2.075e-04  1.220e-04   1.701   0.0893 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 28.59 on 728 degrees of freedom
## Multiple R-squared:  0.00533,    Adjusted R-squared:  0.002597
## F-statistic: 1.951 on 2 and 728 DF,  p-value: 0.1429
```

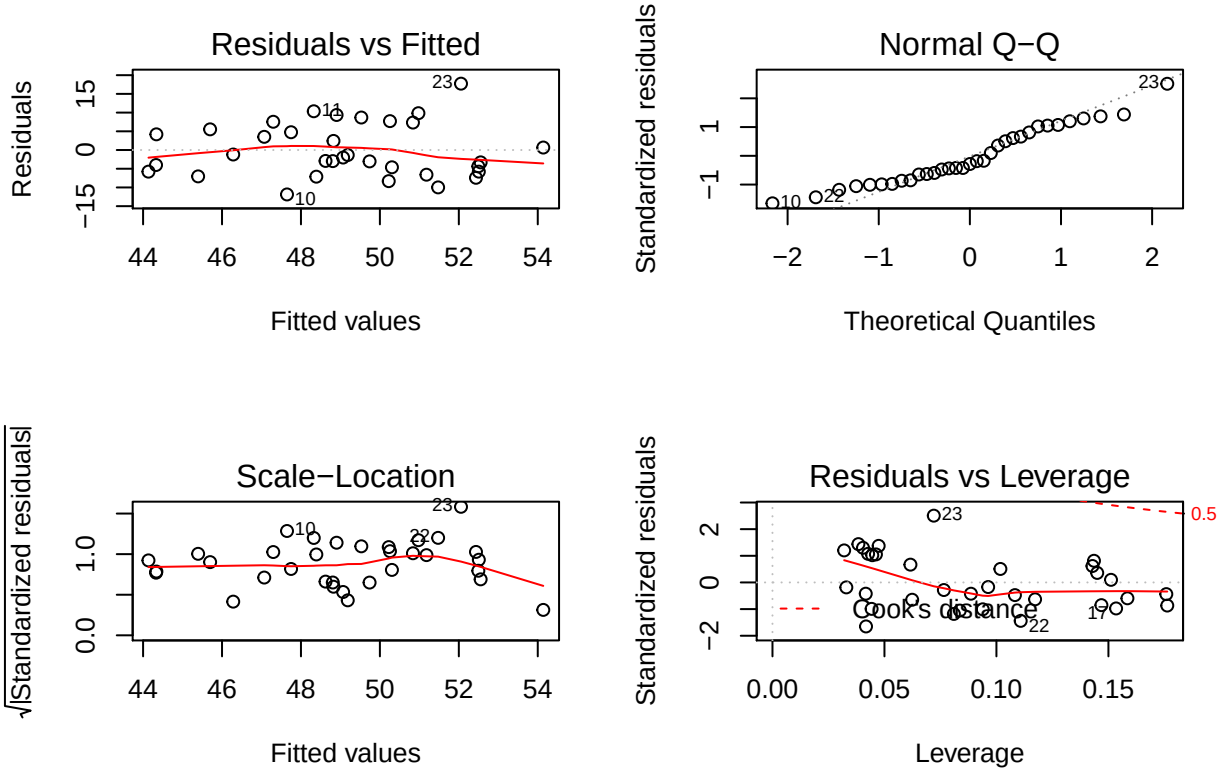
```
plot(lm_before)
```




```
lm_after <- lm(Value ~ coords.x1 + coords.x2, data = stations_agg)
summary(lm_after)
```

```
##
## Call:
## lm(formula = Value ~ coords.x1 + coords.x2, data = stations_agg)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.871  -5.778  -2.010   5.526  17.729
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.601e+01  6.833e+01  -0.527    0.602
## coords.x1    5.421e-05  1.215e-04   0.446    0.659
## coords.x2    3.129e-04  1.640e-04   1.908    0.066 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.343 on 30 degrees of freedom
## Multiple R-squared:  0.1186, Adjusted R-squared:  0.05981
## F-statistic: 2.018 on 2 and 30 DF,  p-value: 0.1506
```

```
plot(lm_after)
```


Residuals vs Fitted
Residuals
Fitted values
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Cook's distance
0.5
Leverage