Text & Spatial Data Analysis

Will Harrison

```
library(ggplot2)
library(dplyr)
library(tidytext)
library(ggmap)
library(sp)
library(gstat)
library(sf)
library(spatstat)
library(maptools)
library(spdep)
```

Question 1 [9 marks]

We want to analyze the books "Anne of Green Gables" and "Blue Castle" by Lucy Maud Montgomery. The two books are provided in the files "Anne of Green Gables.txt" and "Blue Castle.txt".

a) *Visualize the frequency of the 10 most frequent words that satisfy the following three criteria: (1) The word occurs at least five times in each book, (2) The word is not a stop word according to the usual stop list considered in the lectures, (3) The word is not "I'm", "don't", "it's", "didn't", "I've" or "I'll".* **[6 marks]**

```
data("stop_words")

AnneGreen_raw <- readLines("Anne of Green Gables.txt")
AnneGreen_raw <- data.frame(text = AnneGreen_raw)
AnneGreen <- AnneGreen_raw %>%
  unnest_tokens(word, text)

AnneGreen$word <- gsub("\\_" , "", AnneGreen$word)

AnneGreen_count <- AnneGreen %>%
  count(word, sort = TRUE)

AnneGreen_count <- AnneGreen_count %>%
  anti_join(stop_words) %>%
  filter(n >= 5) %>%
  filter(!(word %in% c("i'm", "don't", "it's", "didn't", "i've","i'll")))

BlueCastle_raw <- readLines("Blue Castle.txt")
BlueCastle_raw <- data.frame(text = BlueCastle_raw)
BlueCastle <- BlueCastle_raw %>%
  unnest_tokens(word, text)
```

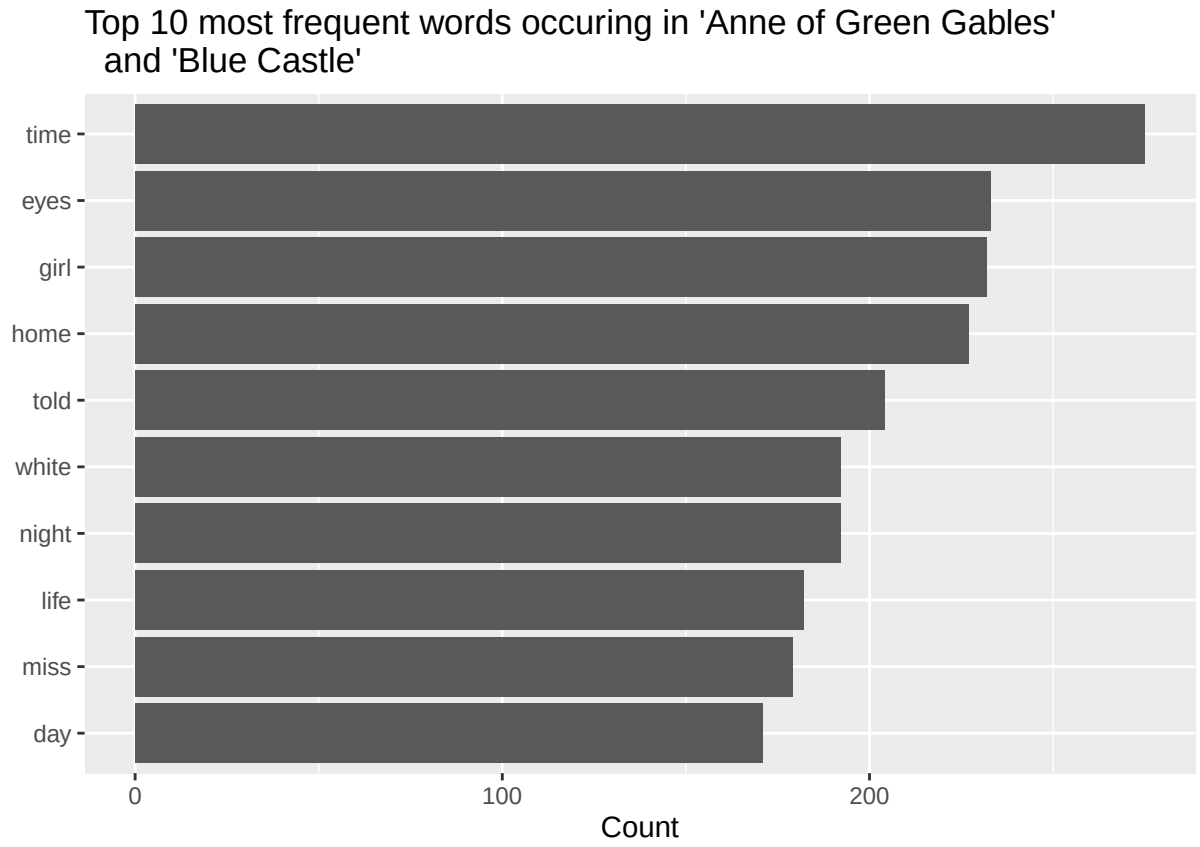
```
BlueCastle$word <- gsub("\\_", "", BlueCastle$word)

BlueCastle_count <- BlueCastle %>%
  count(word, sort = TRUE)

BlueCastle_count <- BlueCastle_count %>%
  anti_join(stop_words) %>%
  filter(n >= 5) %>%
  filter(!(word %in% c("i'm", "don't", "it's", "didn't", "i've","i'll")))

AnneCastle <- inner_join(AnneGreen_count, BlueCastle_count, by = "word",
                         keep = TRUE) %>%
  mutate("word" = word.x, "count" = n.x + n.y, .keep = "none") %>%
  arrange(desc(count)) %>%
  head(10)

AnneCastle %>% mutate(word = reorder(word, count)) %>%
  ggplot(aes(x = count, y =word)) +
  geom_col() +
  labs(title = "Top 10 most frequent words occuring in 'Anne of Green Gables'
  and 'Blue Castle'", x = "Count", y = "")
```



b) *Some scholars say that "Anne of Green Gables" is patterned after the book "Rebecca of Sunnybrook Farm" by Kate Douglas Wiggin. The text for "Rebecca of Sunnybrook Farm" is provided in the file "Rebecca of Sunnybrook Farm.txt". Extract the top two words with the highest term frequency-inverse*

document frequency for each of the two books, "Anne of Green Gables" and "Rebecca of Sunnybrook Farm", with the corpus only containing these books. **[3 marks]**

```
Rebecca_raw <- readLines("Rebecca of Sunnybrook Farm.txt")
Rebecca_raw <- data.frame(text = Rebecca_raw)
Rebecca <- Rebecca_raw %>%
  unnest_tokens(word, text)

Rebecca$word <- gsub("\\_" , "", Rebecca$word)

Rebecca$title <- "Rebecca of Sunnybrook Farm"

AnneGreen$title <- "Anne of Green Gables"

AnneRebecca <- full_join(AnneGreen,Rebecca)

AnneRebecca_count <- AnneRebecca %>% count(title, word, sort = TRUE)

AnneRebecca_tf_idf <- AnneRebecca_count %>%
  bind_tf_idf(word, title, n)

AnneRebecca_tf_idf %>%
  group_by(title) %>%
  arrange(desc(tf_idf)) %>%
  slice(1:2) %>%
  select(title, word, tf_idf)
```

```
## # A tibble: 4 x 3
## # Groups:   title [2]
##   title                      word     tf_idf
##   <chr>                      <chr>     <dbl>
## 1 Anne of Green Gables       anne    0.00740
## 2 Anne of Green Gables       marilla 0.00534
## 3 Rebecca of Sunnybrook Farm rebecca 0.00536
## 4 Rebecca of Sunnybrook Farm don't   0.00138
```

Question 2 [**9 marks**]

We were given PM10 measurements from 60 measurement stations in the Greater Manchester area, including the locations of the stations. The data can be found in the file "Manchester.csv". A detailed description of the variables is provided in the file "DataDescriptions.pdf".

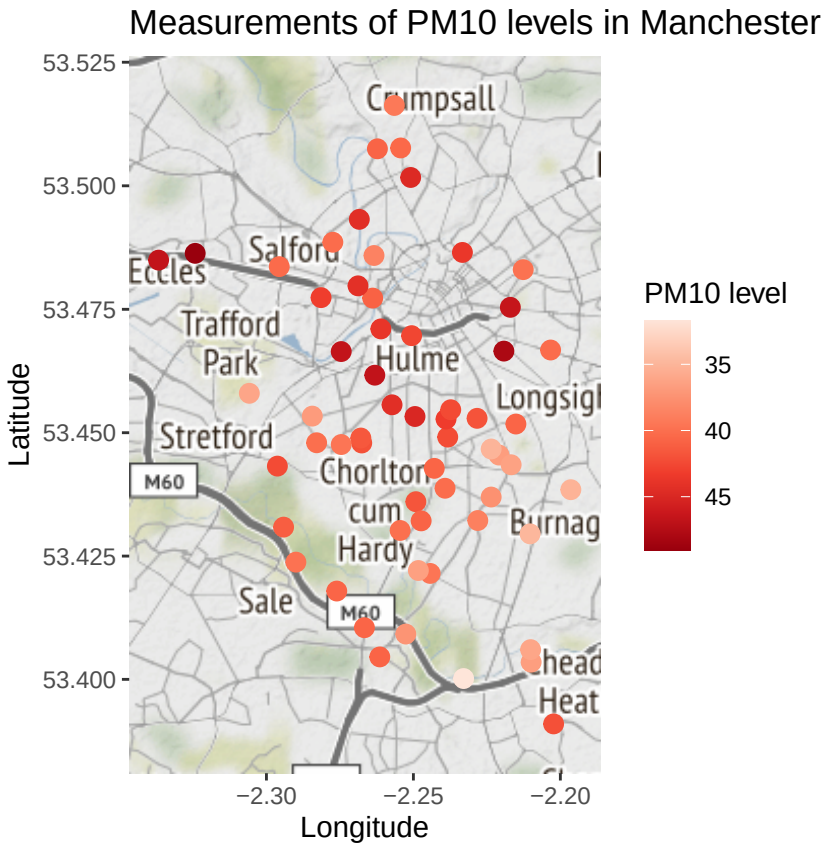
a) *Visualize the data in an informative way and provide an interpretation of your data graphic.* **[3 marks]**

```
Manchester <- read.csv("Manchester.csv")

PlotDim <- c(c(left = min(Manchester$Lon) - 0.01,
               right = max(Manchester$Lon) + 0.01,
               top = max(Manchester$Lat) + 0.01,
               bottom = min(Manchester$Lat) - 0.01))

ggmap(get_stamenmap(PlotDim, maptype = "terrain", zoom = 11)) +
```

```
  geom_point(data = Manchester, aes(x = Lon, y =  Lat, color = Level),
  size = 3) +
  scale_color_distiller(palette = "Reds", trans = "reverse") +
  labs(title = "Measurements of PM10 levels in Manchester", x = "Longitude",
       y = "Latitude", color = "PM10 level")
```

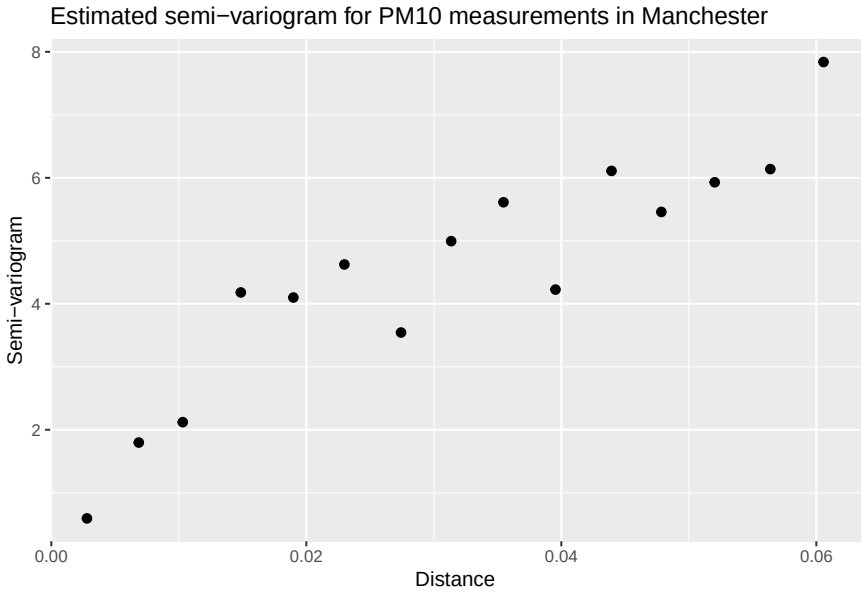


From this plot we can see that higher measurements of PM10 look to be nearer to the centre of Manchester, with lower measurements appearing on the southern side of the city.

b) *Explore the spatial dependence of the PM10 measurements.* [**3 marks**]

```
Manchester_Spatial <- Manchester
coordinates(Manchester_Spatial) <- ~ Lon + Lat
gamma_hat <- variogram(Level ~ 1, data = Manchester_Spatial)

ggplot(gamma_hat, aes(x = dist, y = gamma/2)) +
  geom_point(size = 2) +
  labs(x = "Distance", y = "Semi-variogram",
       title = "Estimated semi-variogram for PM10 measurements in Manchester")
```



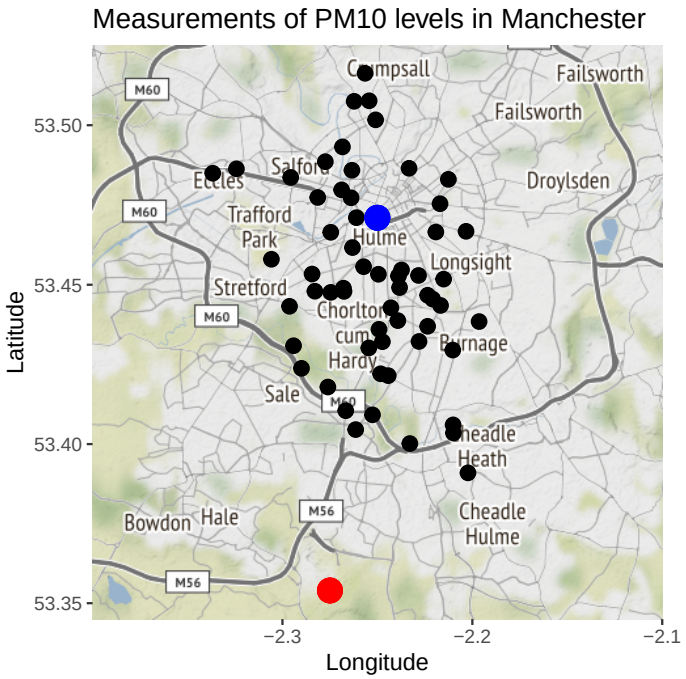
We can see from the variogram that the spatial dependence of two observations looks to decrease as distance increases between the measurement stations in Manchester.

c) *Provide estimates of PM10 levels for two locations: (1) Latitude=53.354, Longitude=-2.275 and (2) Latitude=53.471, Longitude=-2.250. Comment on the reliability of your estimates.* **[3 marks]**

```
IDW <- function( X, S, s_star, p){
  d <- sqrt( (S[,1]-s_star[1])^2 + (S[,2]-s_star[2])^2 )
  w <- d^(-p)
  if( min(d) > 0 )
    return( sum( X * w ) / sum( w ) )
  else
    return( X[d==0] )
}
```

```
PlotDim2 <- c(c(left = -2.4,
                right = -2.1,
                top = 53.525,
                bottom = 53.345))

ggmap(get_stamenmap(PlotDim2, maptype = "terrain", zoom = 11)) +
  geom_point(data = Manchester, aes(x = Lon, y =  Lat),   size = 3) +
  geom_point(aes(y = 53.354, x = -2.275), size = 5, color = "red") +
  geom_point(aes(y = 53.471, x = -2.250), size = 5, color = "blue") +
  labs(title = "Measurements of PM10 levels in Manchester",
         x = "Longitude", y = "Latitude")
```



By looking at this plot, we can see that point 1 (shown in red) is very far from any of the observations, therefore the method of inverse-distance weighting used for prediction will be unreliable. Point 2 (shown in blue) is very central and has a lot of points in the data surrounding it in all directions - therefore the prediction for this point will be a lot more reliable.

```
coord <- cbind(Manchester_Spatial$Lon, Manchester_Spatial$Lat)
s_1 <- c(-2.275,53.354)
s_2 <- c(-2.250,53.471)
p_1 <- IDW(X = Manchester_Spatial$Level, S = coord, s_1, p = 2)
p_2 <- IDW(X = Manchester_Spatial$Level, S = coord, s_2, p = 2)
sprintf("Prediction for (1): %s", round(p_1,5))
```

```
## [1] "Prediction for (1): 40.14383"
```

```
sprintf("Prediction for (2): %s", round(p_2,5))
```

```
## [1] "Prediction for (2): 42.68654"
```

Question 3 [28 marks]

After hearing about the work you did for Utopia's health department, the country's police department got in touch. They need help with analyzing their 2015-2021 data regarding certain crimes. The data is provided in the file "UtopiaCrimes.csv" and a detailed explanation of the variables is provided in the file "Data Descriptions.pdf".

Utopia consists of 59 districts and a shapefile of Utopia is provided together with the other files. To hide Utopia's location, the latitude and longitude coordinates have been manipulated, but the provided shapes are correct. The districts vary in terms of their population and the population for each district is provided in the file "UtopiaPopulation.csv".

a) *What are the three most common crimes in Utopia? Create a map that visualizes the districts worst affected by the most common crime in terms of number of incidents per 1,000 population.* **[5 marks]**

```r
Utopia <- read_sf("UtopiaShapefile.shp")
Utopia <- Utopia %>%
  rename("District_ID" = NAME_1) %>%
  mutate(District_ID = as.integer(gsub("District ","", District_ID)))
Crimes <- read.csv("UtopiaCrimes.csv")
Pop <- read.csv("UtopiaPopulation.csv")

Crimes %>%
  count(Category) %>%
  arrange(desc(n)) %>%
  head(3)
```

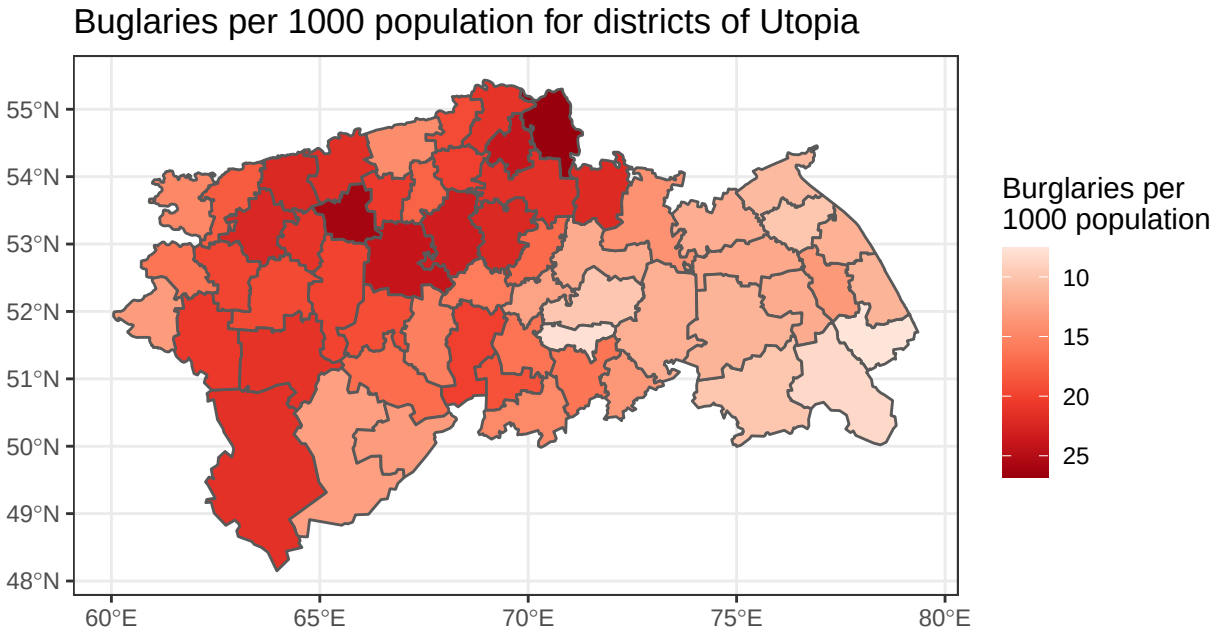
```
##          Category     n
## 1        Burglary 16513
## 2 Drug Possession 10551
## 3         Assault 10169
```

We see that the three most common crimes are Burglary, Drug Possession and Assault.

```r
Crimes_numbers <- Crimes %>%
  filter(Category == "Burglary") %>%
  count(District_ID) %>%
  inner_join(Pop, by = "District_ID") %>%
  mutate("BurglaryPer1000" = n * 1000 / Population)

CrimesPlotBurglary <- inner_join(Utopia, Crimes_numbers, by = "District_ID")

ggplot(CrimesPlotBurglary, aes(fill = BurglaryPer1000)) +
  geom_sf() +
  theme_bw() +
  scale_fill_distiller(palette = "Reds", trans = "reverse") +
  labs(title = "Buglaries per 1000 population for districts of Utopia",
       fill = "Burglaries per \n1000 population")
```

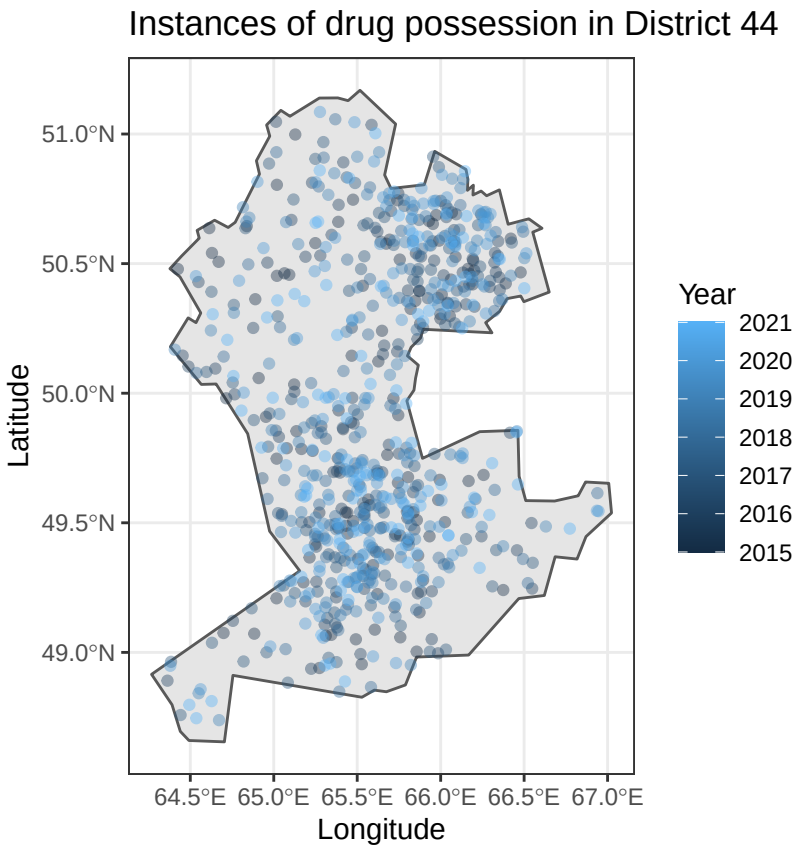


b) *You are told that District 44 is notorious for drug possession. The police is planning to conduct a raid to tackle the issue, but they are unsure on which area of the district they should focus on. Help them make the correct decision.* [**5 marks**]

```
District44Drugs <- Crimes %>%
  filter(Category == "Drug Possession" & District_ID == 44)


District44Map <- Utopia %>%
  filter(District_ID == 44)


ggplot(District44Map) + geom_sf() + theme_bw() +
  geom_point(data = District44Drugs,
             aes(x = Longitude, y = Latitude, color = Year), alpha = 0.4) +
  labs(title = "Instances of drug possession in District 44")
```



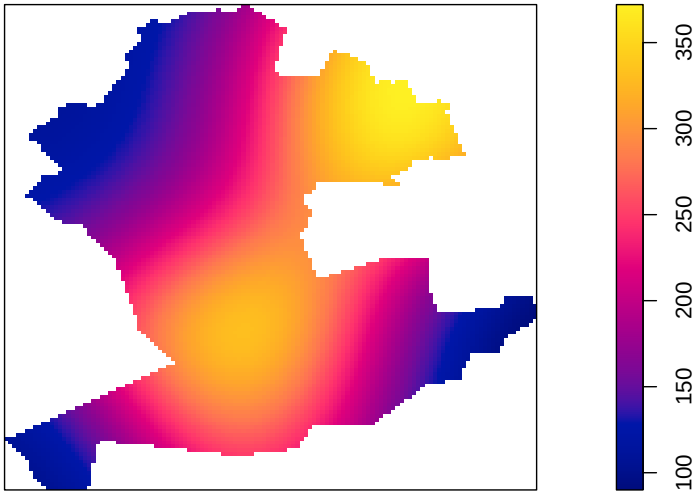
The crimes look to be spread across the district with two larger clusters in the slightly south of the centre and the north-east. Further, the crimes do not seem to show temporal dependence - as a simplification for analysis, all points will be treated with equal weighting, but it could be argued that more recent crimes should have a higher weighting.
Which of the two clusters should the police prioritise? Look to model the intensity function, use the uniformly corrected smoothed kernel intensity function to correct for edge effect bias. To do this, we must first convert our data on district 44 to a ppp object.

```
District44_sp <- as(District44Map, "Spatial")
District44_sp <- slot(District44_sp, "polygons")
District44_win <- lapply(District44_sp, function(z) {SpatialPolygons(list(z))})
District44_win <- lapply(District44_win, as.owin)[[1]]

District44_ppp <- ppp(x = District44Drugs$Longitude,
                      y = District44Drugs$Latitude,
                      window = District44_win)

par(mai = c(0.1,0.1,0.5,0.1))
lambdaC <- density.ppp(District44_ppp, edge = TRUE, sigma = 0.4)
plot(lambdaC, main =
       "Intensity function for instances of drug possession in District 44")
```

Intensity function for instances of drug possession in District 44



This plot reveals that the intensity function is largest in the north west of District 44. Sigma has been chosen so that the intensity function is still quite concentrated close to the observations, while remaining a good fit. Therefore it is recommended that the police focus on the north-east area of the district in their raid. Since the southern region still has quite a high intensity, police should use any remaining resources on this area, however they should prioritise the north east.

c) *The police would also like to understand which group of people is most at risk of a burglary. The possible victims are: "young single", "young couple", "middle-aged single", "middle-aged couple", "elderly single" and "elderly couple". Use the short description provided in "UtopiaCrimes.csv" to extract which group of people is suffering from the highest number of burglaries. What is the proportion of burglaries that involved more than two criminals?* **[4 marks]**

```
Crimes_burglary <- Crimes %>%
  filter(Category == "Burglary") %>%
  mutate(id = row_number(), .keep = "all") %>%
  select(id, Description)  %>%
  unnest_tokens(Description, Description, token = "regex", pattern = " ; ")


Crimes_burglary_vict <- Crimes_burglary %>%
  group_by(id) %>%
  filter(row_number() == 2) # second entry corresponds to victim

Crimes_burglary_vict %>%
  group_by(Description) %>%
  summarise("Occurences" = n()) %>%
  arrange(desc(Occurences))
```

```
## # A tibble: 6 x 2
##   Description        Occurences
##   <chr>                   <int>
## 1 elderly single           4410
## 2 elderly couple           3429
## 3 middle-aged single       3043
```

```
## 4 young single             2126
## 5 middle-aged couple       2017
## 6 young couple             1488
```

We see that elderly single people are most at risk of burglary. We also see that elderly couple are the second most common, so overall the elderly are most vulnerable.

```
Crimes_burglary_perp <- Crimes_burglary %>%
  filter(Description %in% c("three criminals", "more than 3 criminals"))

Crimes_burglary_perp %>%
  summarise("Proportion" = round(n()/max(Crimes_burglary$id), 5))
```

```
##   Proportion
## 1    0.24405
```

See that just under 25% of burglaries involve more than two criminals.

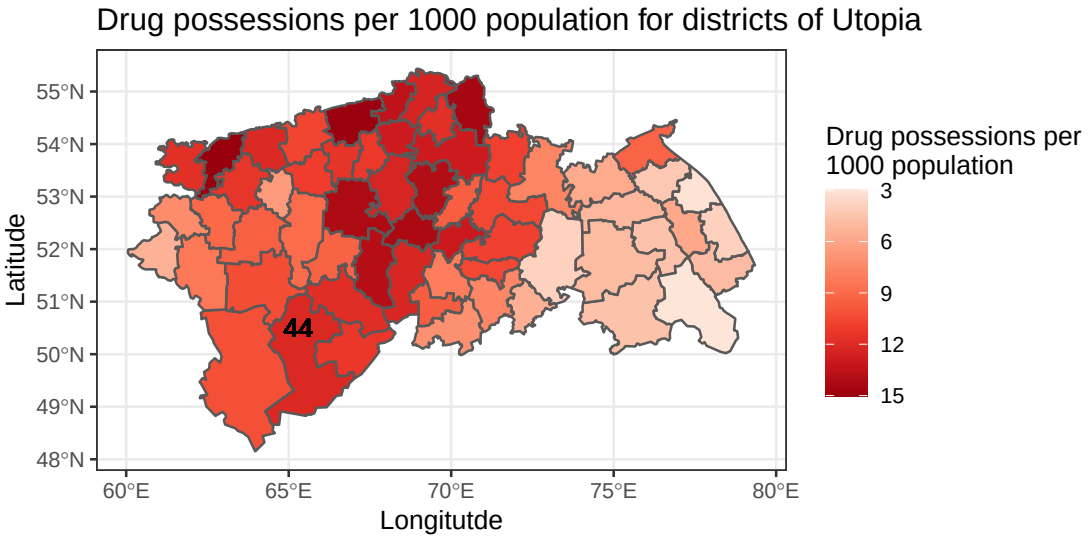
d) *Make up your own question and answer it. Your question should consider 1-2 aspects different to that in parts 3a)-3c). Originality will be rewarded.* [**7 marks**]
Is district 44's notoriety for drug possession justified? Does it really have a significantly higher number of cases of drug possession per 1000 population?

To determine this, use local Moran's I, if district 44 deserves its' reputation, then it will have negative spatial dependence with its' neighbours, i.e. significantly more drug possessions than its surroundings.

```
Drugs_numbers <- Crimes %>%
  filter(Category == "Drug Possession") %>%
  count(District_ID) %>%
  inner_join(Pop, by = "District_ID") %>%
  mutate("DrugPer1000" = n * 1000 / Population)

CrimesPlotDrug <- inner_join(Utopia, Drugs_numbers, by = "District_ID")

ggplot(CrimesPlotDrug, aes(fill = DrugPer1000)) +
  geom_sf() +
  theme_bw() +
  scale_fill_distiller(palette = "Reds", trans = "reverse") +
  labs(title = "Drug possessions per 1000 population for districts of Utopia",
       fill = "Drug possessions per \n1000 population",
       x = "Longitutde", y = "Latitude") +
  geom_text(label = "44", aes(x = 65.3, y = 50.5))
```



From this plot we see that district 44 seems to have a similar number of drug possessions as the districts near it, in fact it does not even have the highest rate of drug possessions. Calculate local Moran's I to check this interpretation:

```
neighbours_Utopia <- poly2nb(CrimesPlotDrug)
neighbours_Utopia <- nb2listw(neighbours_Utopia, style= "B")
MoranLocal <- localmoran(x = CrimesPlotDrug$DrugPer1000,
                         listw = neighbours_Utopia)
print(c(as.character(CrimesPlotDrug$District_ID[44]),MoranLocal[[44]]))
```

```
## [1] "44"                "1.25657437797221"
```

District 44 has a positive spatial dependence, so it isn't any worse than its neighbours for drug possession, and perhaps its reputation is not justified.

e) *Write a short (two paragraphs) report about the findings of your analysis in parts a-d. The report should be readable for people without data science knowledge. Make it sound interesting and state possible recommendations that may be of interest to Utopia's police department.* [**7 marks**]

From the analysis, it was found that burglaries were the most common crime in Utopia (followed by drug possession and assault), with burglaries affecting the western and northern regions of the country the most. The single elderly are the most common victims of this crime and roughly a quarter of burglaries are carried out by more than two criminals. We recommend that the police department send out a warning to the elderly and prioritise them in any anti-burglary initiatives.

District 44 has built a reputation for drug possession, however, after looking holistically and locally at drug possessions in Utopia, it may not deserve this reputation. By looking at the locations of these crimes, it is found that most of these crimes occur in the north east and south of the district. Further analysis recommends that the police begin their raids in the north east, and if possible conduct raids in the south.