MA20277 2022 - Coursework 2

23011

```
knitr::opts_chunk$set(echo = TRUE)
```

```
x = c('dplyr','ggplot2','tidytext','wordcloud','widyr','patchwork','tidyr',
      'sf','ggmap','sp','gstat','spatstat','maptools','spdep')
invisible(lapply(x, library, character.only = TRUE))
```

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]**

Loading in `Anne of Green Gables UTF8.txt` and `Blue Castle UTF8.txt`.

```
AnneGreen_raw=readLines( "Anne of Green Gables UTF8.txt")
BlueCastle_raw = readLines("Blue Castle UTF8.txt")
AnneGreen_raw = data.frame(text=AnneGreen_raw)
BlueCastle_raw = data.frame(text=BlueCastle_raw)
```

Removing stop words and correcting italicized words.

```
data( "stop_words" )
AnneGreen = AnneGreen_raw %>% unnest_tokens( word, text) %>%
  mutate( word = gsub( "\\_", "", word ) ) %>% anti_join(stop_words)
BlueCastle = BlueCastle_raw %>% unnest_tokens( word, text ) %>%
  mutate( word = gsub( "\\_", "", word ) ) %>% anti_join(stop_words)
```

Removing words with count less than 5 for each book. Then calculating word frequency in both books. Removing the words "I'm", "don't", "it's", "didn't", "I've" and "I'll". Plotting a ordered bar graph to visualize the 10 most frequent words.

```
AnneGreen_Count = AnneGreen %>% count( word, sort=TRUE ) %>% filter(n >4)
BlueCastle_Count = BlueCastle %>% count( word, sort=TRUE ) %>% filter(n >4)
inner_join(AnneGreen_Count,BlueCastle_Count, by = ('word' = 'word')) %>%
  group_by(word) %>% summarise('Word frequency' = n.x+n.y ,n.x=n.x, n.y=n.y) %>%
  arrange(desc(`Word frequency`)) %>%
  filter(!word%in%c('i'm','don't','it's','didn't','i've','i'll')) %>%
  mutate(word = reorder(word,`Word frequency`)) %>% slice(1:10) %>%
  pivot_longer(cols = c('n.x','n.y'), names_to = 'Book count') %>%
  ggplot( aes( x=value, y=word , fill = `Book count`) ) +
  geom_bar(position="stack", stat="identity") +
  labs( x="Count", y="Word" ) +
  labs(title = "A graph to visualise the most frequent words that
```

```
occur in Lucy Montgommery's books combined",
x = 'Number of times word is used in book') +
  scale_fill_discrete(name = 'Book', labels = c('Anne of Green Gables', 'Blue Castle'))
```

A graph to visualise the most frequent words that
occur in Lucy Montgommery's books combined
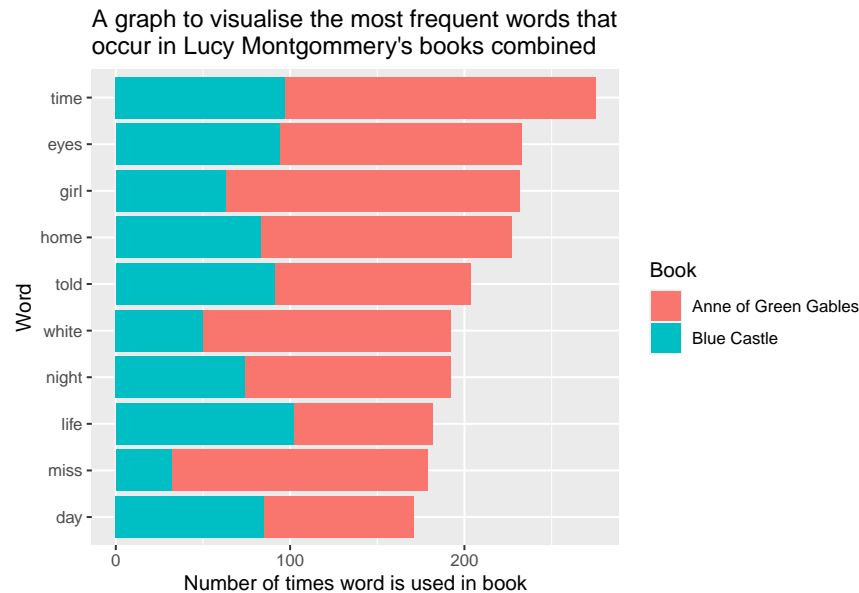
b) *Some scholars say that "Anne of Green Gables" is patterned after the book "Rebecca of Sunnybrook Farm" by Kate Douglas Wiggin. The text for "Rebecca of Sunnybrook Farm" is provided in the file "Rebecca of Sunnybrook Farm.txt". Extract the top two words with the highest term frequency-inverse document 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**]

Data frame with top 2 and 5 words with highest TFIDF for Anne of Green Gables and Rebecca of Sunnybrook Farm respectively.

```
#Creating corpus
AnneGreen_raw=AnneGreen_raw%>% mutate(title = 'Anne of Green Gables')
RebeccaSunny_raw=readLines( "Rebecca of Sunnybrook Farm UTF8.txt")
RebeccaSunny_raw = data.frame(text=RebeccaSunny_raw)
RebeccaSunny_raw=RebeccaSunny_raw  %>% mutate(title = 'Rebecca of Sunnybrook Farm')
LucyM_tf.idf=rbind(RebeccaSunny_raw,AnneGreen_raw) %>%
  unnest_tokens( word, text ) %>%
  mutate( word = gsub( "\\_", "", word ) ) %>% count( title, word, sort=TRUE ) %>%
  bind_tf_idf( word, title, n ) %>%
  arrange( desc(tf_idf) )
#calculating IFIDF
rbind(LucyM_tf.idf %>% filter(title == 'Anne of Green Gables' ) %>%
  arrange(desc(tf_idf)) %>% slice(1:2), LucyM_tf.idf %>% filter(
  title == 'Rebecca of Sunnybrook Farm' ) %>% arrange(desc(tf_idf)) %>% slice(1:5))
```

```
##                          title      word    n          tf       idf       tf_idf
## 1       Anne of Green Gables      anne 1102 0.010670437 0.6931472 0.0073961830
## 2       Anne of Green Gables   marilla  795 0.007697819 0.6931472 0.0053357218
## 3 Rebecca of Sunnybrook Farm   rebecca  572 0.007727850 0.6931472 0.0053565374
## 4 Rebecca of Sunnybrook Farm     don't  147 0.001986003 0.6931472 0.0013765927
## 5 Rebecca of Sunnybrook Farm      it's  105 0.001418574 0.6931472 0.0009832805
## 6 Rebecca of Sunnybrook Farm rebecca's  105 0.001418574 0.6931472 0.0009832805
```

```
## 7 Rebecca of Sunnybrook Farm     cobb   90 0.001215920 0.6931472 0.0008428118
```

Choose to ignore the words 'don't', 'it's' and 'rebecca's' from data frame above. Reasoning is as follows

- 'don't' and 'it's' are written with a different apostrophes in `Anne of Green Gables.txt`, and since 'don't' and 'it's' have a high occurrence in `Anne of Green Gables.txt`, they should have a low TFIDF.
- 'rebecca's' means the same as 'rebecca has/(an equivalent stop word given context)', and 'has/(an equivalent stop word given context)' will have a high usage in both texts, so should have a low TFIDF.

Therefore the two words with the highest IDF for Rebecca of Sunnybrook farm and Anne of Green Gables are 'rebecca' and 'cobb', and 'anne' and 'marilla' respectively. Interestingly the top words are names for each text. This provides evidence for plagiarism since it implies that Lucy used similar language to Rebecca of Sunnybrook Farm when writing Anne of Green Gables.

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]**

Since the location of the measurement stations is known, and the variable of interest is the PM10 measurements, choose a point-referenced data visualization.

```r
#Choose gradient with no white point colour, and opposite colours to aid
#visualization
Manchester = read.csv( "Manchester.csv" )
PlotDim = 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=12) ) +
  geom_point( data=Manchester, size=3, aes( x=Lon, y=Lat, color=Level ) ) +
  scale_color_gradient(low="blue", high="orange") +
  labs( color="Level", x="Longitude", y="Latitude" ) +
  labs(title = 'A graph to visualise the PM10 levels from
60 measurement stations in Machester', color = 'PM10 Level')
```

A graph to visualise the PM10 levels from 60 measurement stations in Machester
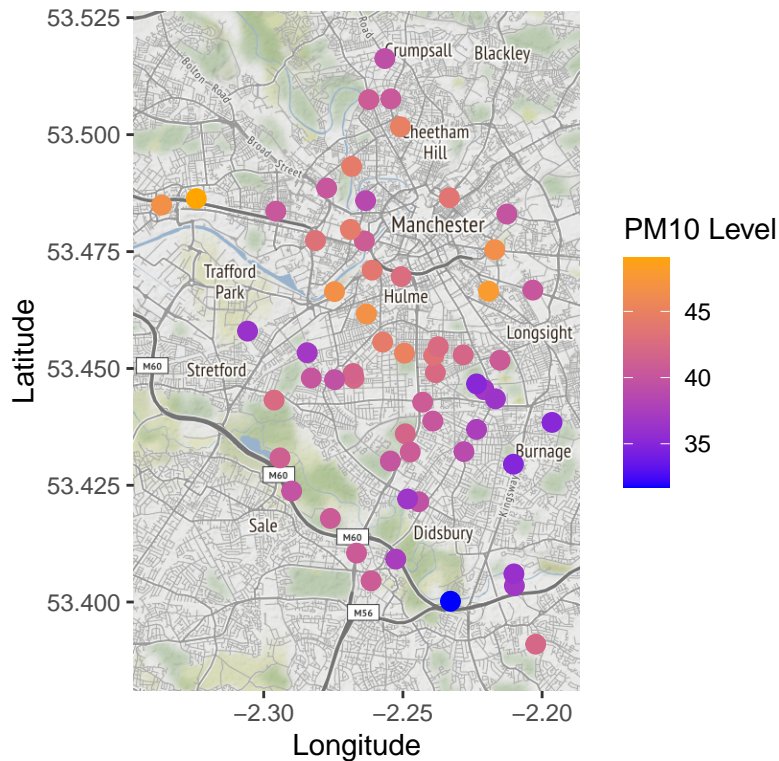
The data graphic above shows the PM10 levels for the urban background sites in the center tend to be significantly higher than for regional areas outside the center of Manchester. PM10 Measurements taken at green spaces, such as parks, are significantly lower than urban areas in the center of Manchester.

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

```
# change the variogram width to reduce noise
Manchester = read.csv( "Manchester.csv" )
coordinates(Manchester ) <- ~Lon+Lat
gamma_hat <- variogram( Level~1, Manchester, width = 0.0065)
ggplot( gamma_hat, aes( x=dist, y=gamma/2 ) ) + geom_point( size=2 ) +
  labs( x="Distance", y="Semi-variogram",
        title = 'Semi-Variogram for PM10 meausurements in Manchester')
```

Semi−Variogram for PM10 meausurements in Manchester
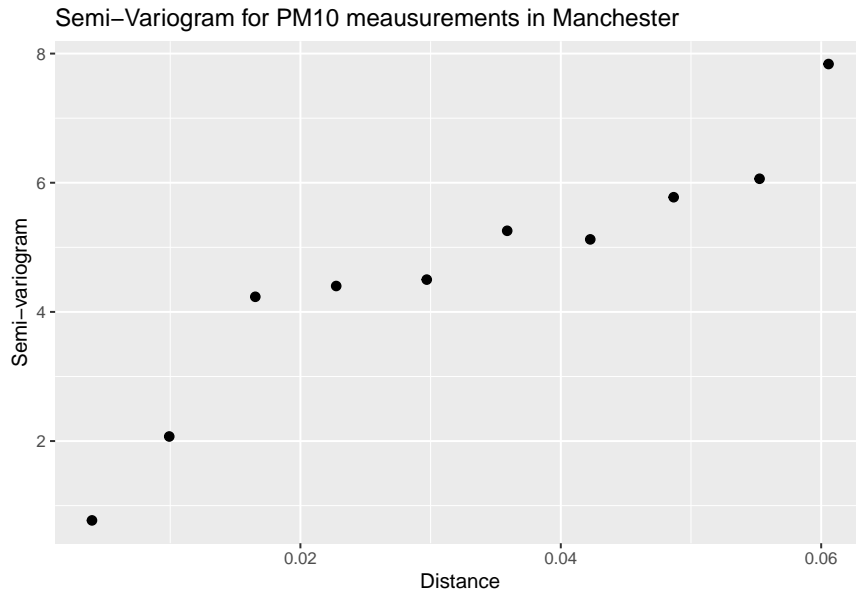
The variogram indicates that spatial dependence decreases with increasing spatial distance between the measurement sites. There is positive correlation of spatially close measurement sites. There is some variation for measurement sites with distances 0.02 and above. For distances above 0.02, the points show weaker positive correlation. This may be due to differences in PM10 measurements from urban to green backgrounds (discussed more in 2(a)).

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] )
 }
IDW( Manchester$Level,cbind( Manchester$Lon, Manchester$Lat),c(-2.275, 53.354), p=2 )
```

```
## [1] 40.14383
```

```
IDW( Manchester$Level,cbind( Manchester$Lon, Manchester$Lat),c(-2.250, 53.471), p=2 )
```

```
## [1] 42.68654
```

Estimate PM10 level for location (1) is 40.14383. Estimate PM10 level for location (2) is 42.68654.

```
Manchester = read.csv( "Manchester.csv" )
PlotDim = c( left=min(Manchester$Lon)-0.01, right=max(Manchester$Lon)+0.01,
             top=max(Manchester$Lat)+0.01, bottom=min(Manchester$Lat)-0.05 )
ggmap( get_stamenmap(PlotDim, maptype="terrain", zoom=11) ) +
  geom_point( data=Manchester, size=3, aes( x=Lon, y=Lat, fill = 'black')) +
  labs(x="Longitude", y="Latitude" ) +
  geom_point(aes(x=-2.275,y=53.354,color="red"), size = 3) +
  geom_point(aes(x=-2.250,y=53.471,color="blue"), size = 3) +
  scale_color_discrete(name = 'Estimates', labels = c('(1)','(2)')) +
  scale_fill_discrete(name = 'PH10 Measurement\nStations', labels = c('Stations')) +
```

```
  labs( title="A graph to show location of the estiamtes
in relation to the measurement stations
in Manchester")
```

A graph to show location of the estiamtes in relation to the measurement stations in Manchester
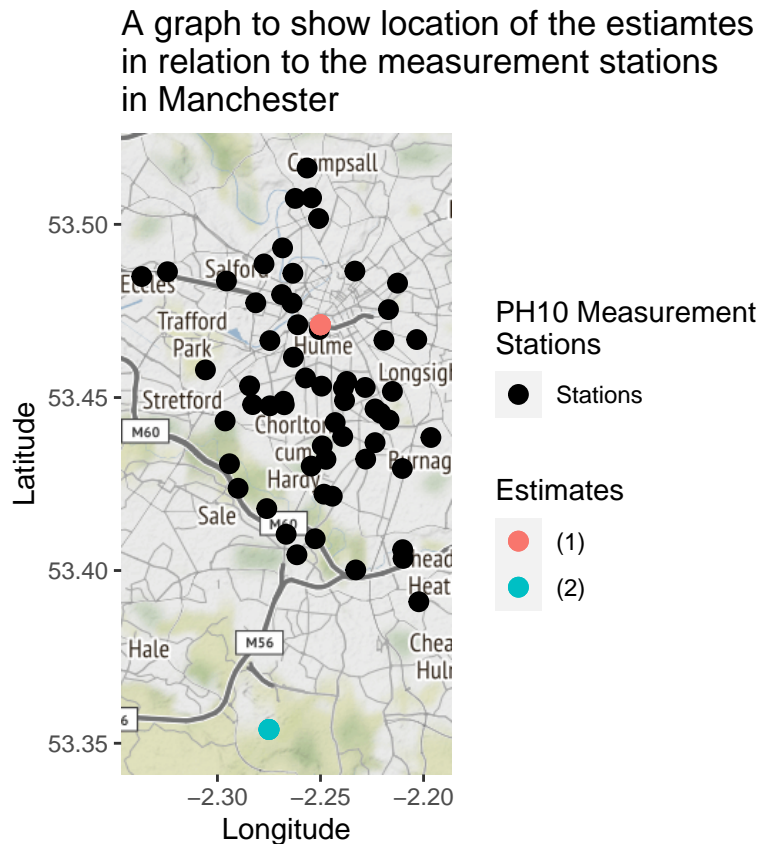
Notice that estimate (1) lies in the center of a cluster of measurement stations all within close proximity and strictly in urban locations. This implies that estimate (1) is a good estimate of the PM10 level.

Estimate (2) is a bad estimate of the PM10 Level. This is because the variogram shows that spatial dependence decreases as spatial distance increases. So, naturally there will be more uncertainty with the estimate. Also, an observation from plot (2a), green spaces tend to have lower PM10 levels. However, the closest points to estimate (2), although outside Manchester city center, are not entirely in green spaces. The inverse distance weighing function does not take this into account when calculating the estimate (2) (provides evidence for higher inaccuracies in estimate).

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**]

```
UtopiaCrimes=read.csv('UtopiaCrimes.csv')
UtopiaPopulation=read.csv('UtopiaPopulation.csv')
UtopiaCrimes%>%group_by(Category)%>%summarize('Number of crimes comitted'=n())%>%
  arrange(desc(`Number of crimes comitted`))%>%slice(1:3)
```

```
## # A tibble: 3 x 2
##   Category         `Number of crimes comitted`
##   <chr>                                  <int>
## 1 Burglary                               16513
## 2 Drug Possession                        10551
## 3 Assault                                10169
```

The 3 most common crimes in Utopia are Burglary, Drug Possession and Assault.

```
UtopiaShape = read_sf("UtopiaShapefile.shp" )
UtopiaShape$NAME_1 = as.numeric(gsub('District','', UtopiaShape$NAME_1))
DistrictBurglary=UtopiaCrimes %>% filter(Category == 'Burglary') %>%
  group_by(District_ID) %>% summarise('Burglary per 1000 pop' = n())
DistrictBurglary$`Burglary per 1000 pop` =
  (DistrictBurglary$`Burglary per 1000 pop`)/(UtopiaPopulation$Population/1000)
Utopia = inner_join(UtopiaShape, DistrictBurglary, by=c("NAME_1"="District_ID"))
ggplot() + geom_sf( data=Utopia, aes(fill=`Burglary per 1000 pop`) ) + theme_bw()+
  labs(
  title="A graph to visualise the number of burglary incidences
per 1000 population across Utopia",
  x="Longitude", y="Latitude", fill="Burglary incidences\nper 1000 population")+
  scale_fill_distiller( palette="Reds", trans="reverse" )
```


A graph to visualise the number of burglary incidences per 1000 population across Utopia
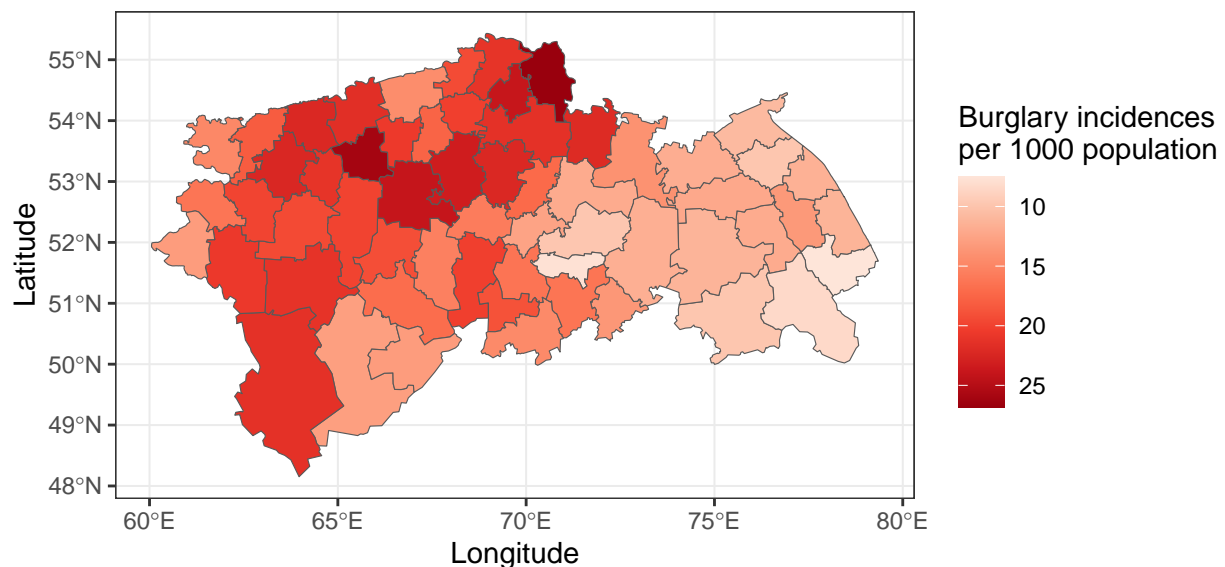
The map clearly visualizes the districts worst effected by burglaries in Utopia. The worst effected districts are the regions with the highest number of burglary incidences per 1000 population; these regions are prominently in the north west of Utopia.

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**]

Since the police are choosing to conduct a raid in the near future, intuitively, they should use information which is fairly recent, and on drug possession incidences which have no arrests. This is based on the reasonable assumption that when criminals get arrested, they will not commit the same crime in the same area, and if criminals have not been caught, they will still be committing crimes in the same area. In real life, police raid operations can take up to 2 years of planning, and gathering of evidence to implement a raid (particularly if they are high profile cases); using information gathered over an appropriate timeframe is important when making the correct decision.

```
par(mai=c(0.1,0.1,0.5,0.5) )
DistrictShape44=UtopiaShape %>% filter(NAME_1 == 44)
DistrictDrug44=UtopiaCrimes %>% filter(District_ID == 44, Category ==
    'Drug Possession', Arrest == 'No', Year > 2019)
DistrictShape44_sp=as(DistrictShape44 , "Spatial" )
DistrictShape44_sp=slot( DistrictShape44_sp, "polygons" )
DistrictShape44_win=lapply(DistrictShape44_sp,function(z){SpatialPolygons(list(z))})
DistrictShape44_win = lapply( DistrictShape44_win, as.owin )[[1]]
District44_ppp = ppp( x=DistrictDrug44$Longitude, y=DistrictDrug44$Latitude,
                      window = DistrictShape44_win )
plot( quadratcount( District44_ppp, nx=5 ),
main="Quadrat Counting in district 44 for incidences of drug
possession with criminals not arrested over 2020-2021")
```
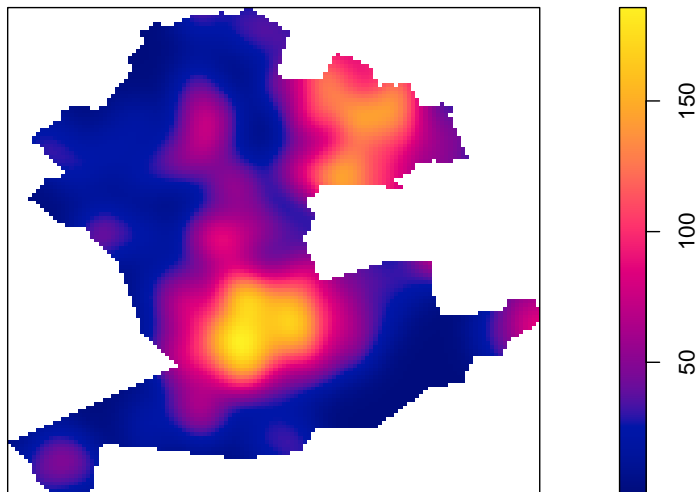
Quadrat Counting in district 44 for incidences of drug possession with criminals not arrested over 2020–2021
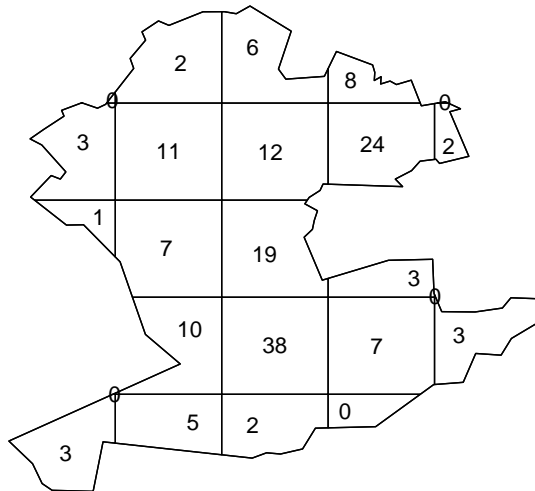

```
plot(density.ppp(District44_ppp,edge = TRUE,sigma = 0.1),
    main="Kernel Smoothed Intensity Function in District 44 for incidences
of drug possession with criminals not arrested over 2020-2021")
```

**Kernel Smoothed Intensity Function in District 44 for incidences
of drug possession with criminals not arrested over 2020–2021**

The quadrat count plot shows that there are two locations where there are a high number of drug possession incidences with no arrests in 2020-2021. These locations are in the north east and southern middle of district 44. There are large differences between number of incidences of drug possession in parts of district 44, therefore non-homogeneity is a reasonable assumption.

The kernel intensity plot shows that there are distinct regions in the north east and south of district 44 that have a significantly higher intensity than the rest of district 44. These locations are considered as drug possession hot-spots, and the kernel intensity plot suggests significant clustering in these regions, particularly the south middle region of district 44. Therefore, the police should conduct a raid in the high intensity location at the southern middle region of district 44.

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 "Crimes.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**]

```
burglary=UtopiaCrimes %>% filter(Category == 'Burglary')
TotalBurglary = burglary %>% summarise(n=n())
PeopleGroups = c('young single','young couple','middle-aged single',
                'middle-aged couple','elderly single','elderly couple')
PeopleGroupsN = data.frame()
for( i in 1:6){
PeopleGroupsN = rbind(PeopleGroupsN, burglary %>%
    filter((grepl(PeopleGroups[i], .$Description)))%>%
    summarise('Number of people burgled'=n()))}
PeopleGroupsN = cbind(PeopleGroups,PeopleGroupsN) %>%
  arrange(desc(`Number of people burgled`))
PeopleGroupsN
```

```
##         PeopleGroups Number of people burgled
## 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
```
```
burglary %>% filter(!(grepl('One criminal', .$Description) | grepl(
  'Two criminals', .$Description))) %>% summarise('Proportion of burglaries that
involved more than two criminals' = n()/TotalBurglary$n)
```

```
##   Proportion of burglaries that\ninvolved more than two criminals
## 1                                                      0.2440501
```

The group most at risk of a burglary in Utopia is 'elderly single'. 24% of burglaries in Utopia involved 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**]

By investigating regional arrest rates for assault, and spatial dependence for assault incidences across Utopia, is there anything concerning that Utopia's police department should be aware of?

```
neighbours_utopia=poly2nb(UtopiaShape)
neighbours_utopia=nb2listw(neighbours_utopia, style="B")
UtopiaCrimes = UtopiaCrimes %>%
  mutate(Arrest= case_when(Arrest == 'No' ~ 0, Arrest == 'Yes' ~ 1))
Arrestation=UtopiaCrimes %>% filter(Category == 'Assault') %>%
  group_by(District_ID) %>% summarise(n = mean(Arrest))
inner_join(UtopiaShape, Arrestation, by=c("NAME_1"="District_ID")) %>% ggplot()+
  geom_sf(aes(fill=n) ) + theme_bw() +
  labs( title="A graph to show the assault arrest rate across the districts
of Utopia",
  x="Longitude", y="Latitude", fill="Assult arrest rate") +
  scale_fill_distiller( palette="Reds", trans="reverse")
```
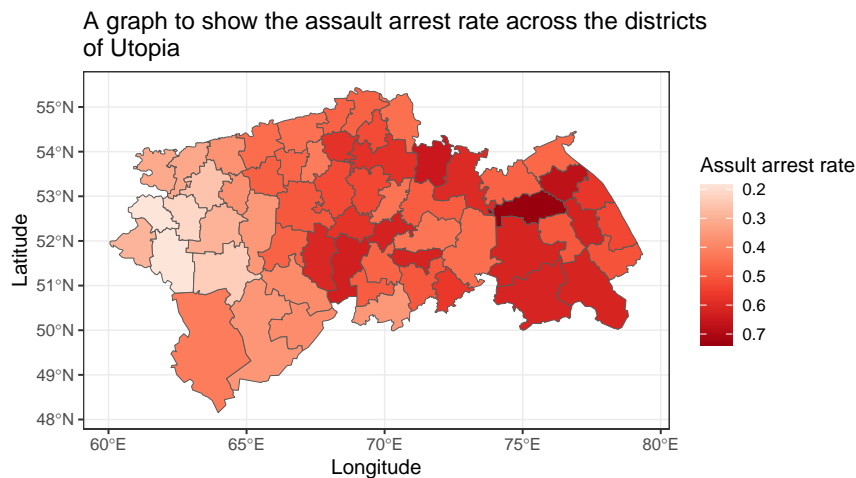

Note that incidences of assault may be driven by differences in population density per district; since, it is assumed that a high population density provides more opportunities for assaults to happen. Therefore, choose to measure geographical spatial dependence of neighboring districts on assault incidences per 1000 population.

```
AssaultRate=UtopiaCrimes %>% filter(Category =='Assault') %>%
  group_by(District_ID)%>% summarise(n = n())
```

```
AssaultRate$n=
  (AssaultRate$n)/(UtopiaPopulation$Population[AssaultRate$District_ID]/1000)
MoranLocal <- localmoran( x=AssaultRate$n, listw=neighbours_utopia)
AssaultRate$Moran = MoranLocal[,"Ii"]
inner_join(UtopiaShape, AssaultRate, by=c("NAME_1"="District_ID")) %>% ggplot()+
  geom_sf(aes(fill = Moran, style = 'quantile') ) + theme_bw() +
  labs( title="A graph to show simularities between the district's
assault incidences per 1000 population in Utopia",
  x="Longitude", y="Latitude", fill="Local Moran's I")+
  geom_point( aes(color = n, geometry = geometry), size = 4, shape = 18,
  stat="sf_coordinates")+labs(color='Assault incidences per\n1000 population') +
  scale_colour_distiller( palette="Reds",trans="reverse")
```


A graph to show simularities between the district's assault incidences per 1000 population in Utopia
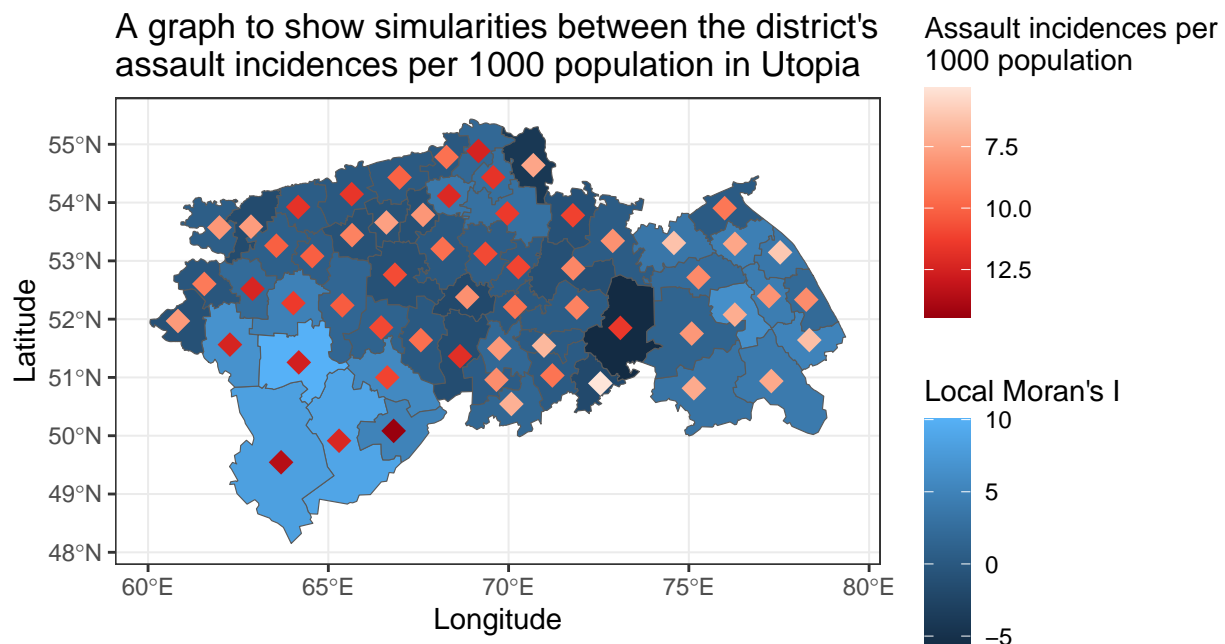
Utopias police department should definitely be concerned that the southern west region of Utopia have high assault incidences per 1000 population, but low arrest rates compared to the rest of Utopia. This contrasts the east region of Utopia, which has low assault incidences per 1000 population, but high assault arrest rates. Intuitively, it is suspected that socio-economic differences between districts could be a reason for the observed constrasting differences described above.

There is also geographical clustering of spatially dependent neighboring districts with similar assault incidences per 1000 population. Observe that the values for local Moran's I in the South west and east regions of Utopia are positive and similar in value. This is concerning to Utopia's Police Department. This is because we observe that increased assault incidences per 1000 population in the south west side of Utopia are not random between districts (neighboring districts share similar high values). Therefore, the high assault incidences per 1000 population in the south west of Utopia is a regional issue; Utopia's Police department should definitely be aware of this.

Note that there is also geographical clustering of spatial dependent districts with similar assault incidences per 1000 population in the northern middle of Utopia. These districts are of less concern to Utopia's police department. This is because the districts have a high arrest rate for assault.

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**]

A report was conducted, following the request of Utopia's police department, to analyse regional characteristics of certain crimes in order to provide useful recommendations. The information used was provided by Utopia's police department over a timeframe of approximately 7 years (January 2015 to December 2021). The report showed that from January 2015, to December 2021, at least 73,561 crimes were committed in Utopia. During this timeframe, the crime with the most incidences in Utopia was burglary, with at least 16,513 incidences; this was followed by drug possession and assault, with number of incidences of at least 10,551 and 10,169 respectively. From January 2015 to December 2021, the north-west region of Utopia was effected the most by burglary in terms of number of incidents per 1,000 population, with the maximum burglary incidences per 1,000 population at approximately 25 in a single district. Elderly singles were found to be most at risk of being burgled in Utopia, with at least 4410 incidences occurring from January 2015 to December 2021. Young couples were found to be least at risk of being burgled in Utopia, with 1488 incidences occurring from January 2015 to December 2021. 24% of burglaries from January 2015 to December 2021 in Utopia were committed by more than two criminals. In general, significant differences in burglary incidences per 1,000 population, between the east and west regions of Utopia, were found during the full length of the timeframe; the east was found to be significantly lower than the north west of Utopia, with incidences per 1,000 population in the range of approximately 5-15 compared to 15-25 respectively. From January 2020 to December 2021, the report found evidence to suggest that drug possession incidences which did not result in arrestation, were not uniformly spread around district 44; there was significant clustering present in the north-east and southern middle of the district 44, with the southern middle obtaining the highest intensity of drug possession incidences which did not result in arrestation. From January 2015 to December 2021, the report highlighted contrasting regions over arrestation for assault. During this time, the south west region of Utopia had a high number of assault incidences per 1000 population and a low arrest rate, whereas, the east region of Utopia had a low number of assault incidences per 1000 population and a high arrest rate. Geographical clustering was present in south west and east regions, with neighbouring districts all having a similar number of assault incidences per 1000 population.

The report recommends that the police conduct a raid in the southern middle region of district 44. The report also recommends that Utopia's police department should focus on protecting the elderly from being burgled. Some suitable suggestions are, making sure elderly people are more 'up to date' with modern burglary methods, introducing more police presence in neighbourhoods where the average age is higher, and educating people on how to protect their homes from intrusion. It may be interesting to investigate if there are any socio-economic differences between the north west and east region of Utopia; this could explain the differences in number of burglary incidences per 1000 population. This information would certainly be useful when implementing burglary mitigation strategies. The report recommends an investigation into the south west region of Utopia; it is deemed necessary to investigate if more police presence is required to keep assault incidences down, and increase arrestation of assault. It would certainly be in Utopia's police department's interest to increase arrestation of assault in the south west region of Utopia.