MA20277 2022 - Coursework 2

Ariana Vass 23111

```
library( dplyr, warn.conflicts = F, quietly = T )
library( ggplot2 )
library( tidytext )
library( widyr )
library( patchwork )
library( ggmap )
library( sf )
library(tidyr)
library(gstat)
library(sp)
```

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]**

First we need to turn both texts into tokens. I am going use “anti_join()” to remove the stop words and the words “I’m”, “don’t”, “it’s”, “didn’t”, “I’ve” or “I’ll”, to satisfy criteria (2) & (3):

```
data("stop_words")
Anne_raw <- readLines("Anne of Green Gables.txt")
Blue_raw <- readLines("Blue Castle.txt")
contractions <- data.frame("word" = c("I’m", "don’t", "it’s", "didn’t", "I’ve", "I’ll"))

Anne <- tibble(text = Anne_raw)
Anne <- Anne %>% unnest_tokens(word, text)
Anne$word <- gsub("\\_", "", Anne$word)
Anne <- Anne %>% anti_join(stop_words, copy = TRUE) %>%
  anti_join(contractions, copy = TRUE)
```

```
## Joining, by = "word"
## Joining, by = "word"
```

```
Blue <- tibble(text = Blue_raw)
Blue <- Blue %>% unnest_tokens(word, text)
Blue$word <- gsub("\\_", "", Blue$word)
Blue <- Blue %>% anti_join(stop_words, copy = TRUE) %>%
  anti_join(contractions, copy = TRUE)
```

```
## Joining, by = "word"
## Joining, by = "word"
```

Now we want to make Anne_count and Blue_Count, and filter them by count >= 5 to satisfy criterion (1):

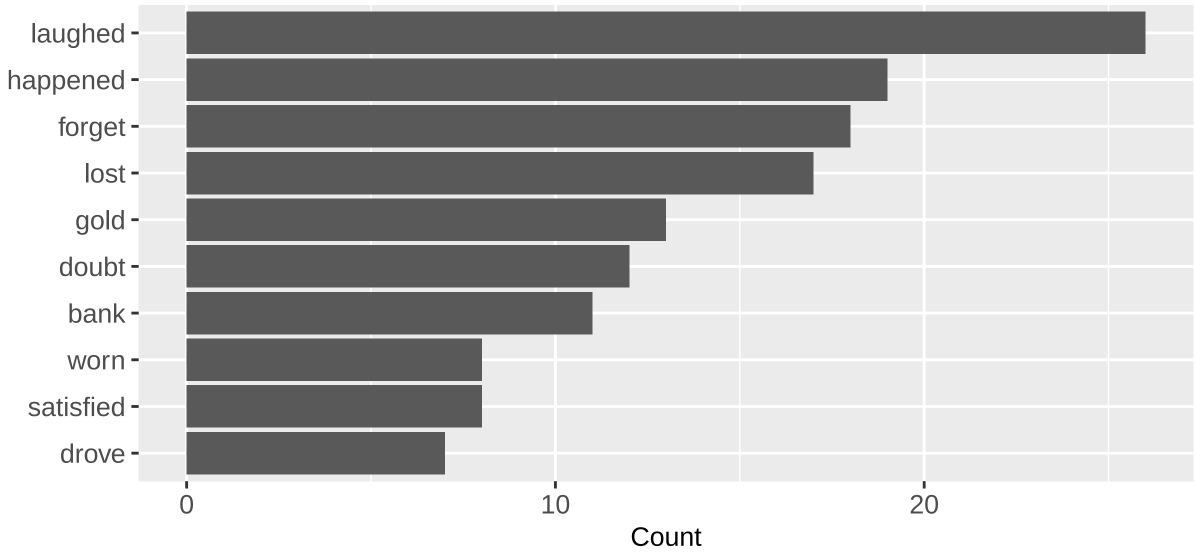
```
Anne_count <- Anne %>% count( word, sort=TRUE )
Anne_count <- filter(Anne_count, n >= 5)
Blue_count <- Blue %>% count( word, sort=TRUE )
Blue_count <- filter(Blue_count, n >= 5)
```

Then we want to join them by "inner_join()" because we only want words in both books. Then we will create a bar chart using "geom_col()" :

```
LMMontgomery <- inner_join(Anne_count, Blue_count)
```

```
## Joining, by = c("word", "n")
```

```
LMMontgomery %>%
  slice_head( n=10 ) %>%
  mutate( word = reorder(word,n) ) %>%
  ggplot( aes( x=n, y=word) ) + geom_col() + labs(
    x="Count", y="" ) +
  theme( axis.text=element_text(size=10),
         axis.title=element_text(size=10) )
```



We see the ten most frequent words fitting this criteria above.

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**]

First we need to turn the book "Rebecca of Sunnybrook Farm" into tokens, remove the stop words, and I am also going to remove the words from the list given in (a). I will then extract the word frequency into "Rebecca_count":

```
Rebecca_raw <- readLines("Rebecca of Sunnybrook Farm.txt")

Rebecca <- tibble(text = Rebecca_raw)
Rebecca <- Rebecca %>% unnest_tokens(word, text)
Rebecca$word <- gsub("\\_", "", Rebecca$word)
Rebecca <- Rebecca %>% anti_join(stop_words, copy = TRUE) %>%
  anti_join(contractions, copy = TRUE)
```

```
## Joining, by = "word"
## Joining, by = "word"
```

```
Rebecca_count <- Rebecca %>% count( word, sort=TRUE )
```

Next we need to combine the two books into one data frame and state the books' titles so that they are the only books in the corpus. We then need to count the number of appearances of each word separately for each of the books using the "count()" function:

```
Anne_count <- Anne_count %>% mutate(Title = "Anne of Green Gables")
Rebecca_count <- Rebecca_count %>% mutate(Title = "Rebecca of Sunnybrook Farm")
Corpus <- full_join(Anne_count, Rebecca_count, by = c("Title", "word", "n"))
```

Now we need to use "bind_tf_idf()" and extract the top two words with the highest term frequency-inverse document frequency for each of the two books:

```
Corpus_tf.idf <- Corpus %>%
  bind_tf_idf( word, Title, n ) %>%
  arrange( desc(tf_idf) )

Corpus_tf.idf %>% filter(Title == "Anne of Green Gables") %>% slice_head(n=2)
```

```
## # A tibble: 2 x 6
##   word         n Title                    tf   idf tf_idf
##   <chr>    <int> <chr>                 <dbl> <dbl>  <dbl>
## 1 anne      1202 Anne of Green Gables 0.0452 0.693 0.0313
## 2 marilla    849 Anne of Green Gables 0.0319 0.693 0.0221
```

```
Corpus_tf.idf %>% filter(Title == "Rebecca of Sunnybrook Farm") %>% slice_head(n=2)
```

```
## # A tibble: 2 x 6
##   word         n Title                          tf   idf  tf_idf
##   <chr>    <int> <chr>                       <dbl> <dbl>   <dbl>
## 1 rebecca    574 Rebecca of Sunnybrook Farm 0.0217  0.693 0.0150
## 2 emma       156 Rebecca of Sunnybrook Farm 0.00589 0.693 0.00408
```

We see that the top two words for both books are names of characters and that the words in Anne of Green Gables have a higher term frequency-inverse distance frequency. This could suggest that the scholars that say that "Anne of Green Gables" is patterned after the book "Rebecca of Sunnybrook Farm", are correct.

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**]

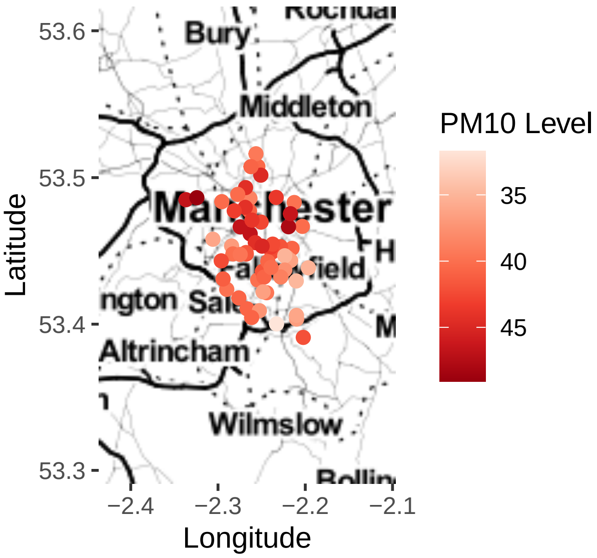
To do this, I want to represent the data on a toner map since it clearly shows Manchester graphically:

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

PlotDim <- c(left = min(Manchester$Lon)-0.1, right = max(Manchester$Lon)+0.1,
             top = max(Manchester$Lat)+0.1, bottom = min(Manchester$Lat)-0.1)

ggmap( get_stamenmap(PlotDim, maptype="toner", zoom=9) ) +
  geom_point( data=Manchester, aes(x=Lon, y=Lat, color= Level), size=2 ) +
  scale_color_distiller( palette="Reds", trans="reverse" ) +
  labs( x="Longitude", y="Latitude", color = "PM10 Level")
```

```
## i Map tiles by Stamen Design, under CC BY 3.0. Data by OpenStreetMap, under ODbL.
```

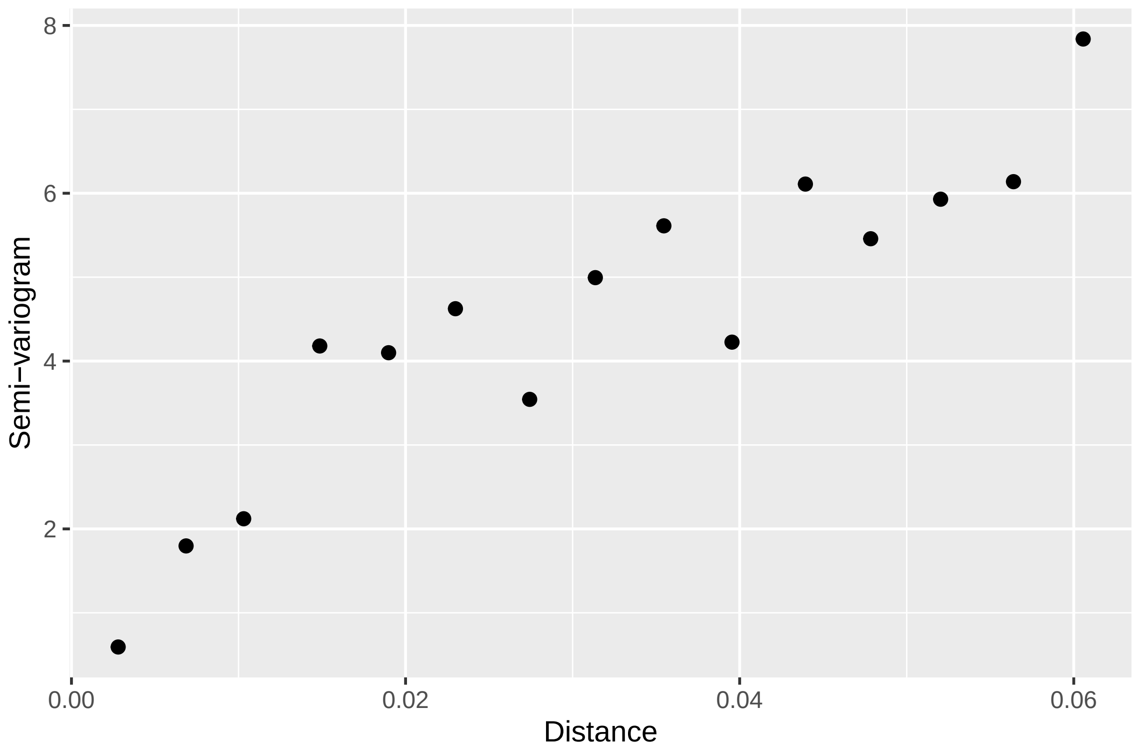


We see a concentration of higher PM10 levels in the centre of Manchester, and in general, lower levels as we move away from the centre. This holds with the fact that there is generally worse air quality (indicated by higher PM10 levels) in the more built up areas like the centre of Manchester.

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

To explore spatial dependence, I am going to use "variogram()" to create an estimated semi-variogram, which will display the spatial dependence on a graph:

```
coordinates( Manchester ) <- ~Lon+Lat
estim <- variogram( Level~1, Manchester,)
ggplot( estim, aes( x=dist, y=gamma/2 ) ) + geom_point( size=2 ) +
  labs( x="Distance", y="Semi-variogram" )
```



The semi-variogram suggests a positive correlation of spatially close sites, and that the degree of spatial dependence decreases with increasing spatial distance. This means that points that are close in distance, you would expect them to have similar values

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**]

To do this we need to IDW() function we made in lectures:

```
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] )
}
```

Then to estimate location (1) we use this function. I have chosen to use a p of 0.5 since the range of PM10 levels in Manchester is quite small, there is no need for big variety in prediction:

```
coord <- cbind( Manchester$Lon, Manchester$Lat )
s_star <- c(53.354, -2.275)
IDW( X=Manchester$Level, S=coord, s_star, p=0.5 )
```

```
## [1] 40.87385
```

Estimation for location (2):

```r
coord <- cbind( Manchester$Lon, Manchester$Lat )
s_star <- c(53.471, -2.250)
IDW( X=Manchester$Level, S=coord, s_star, p=0.5 )
```

```
## [1] 40.87385
```

These Locations have exactly the same prediction therefore, I do not feel that my estimates are reliable, while they are close, I do not feel that they are close enough to have exactly the same PM10 level, to 5 decimal places.

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".

Before I answer the questions, I am going to load the files provided:

```r
U_Crimes <- read.csv("UtopiaCrimes.csv")
U_Population <- read.csv("UtopiaPopulation.csv")
Utopia <- read_sf("UtopiaShapefile.shp")
```

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**]

To find the 3 most common crimes we need to group by crime, summarise, arrange and display the top 3:

```r
U_Crimes %>% group_by(Category) %>% summarise(n= n()) %>%
  arrange(desc(n)) %>% slice_head(n = 3)
```

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

We see that Burglary, Drug possession and Assault are the 3 most common crimes.

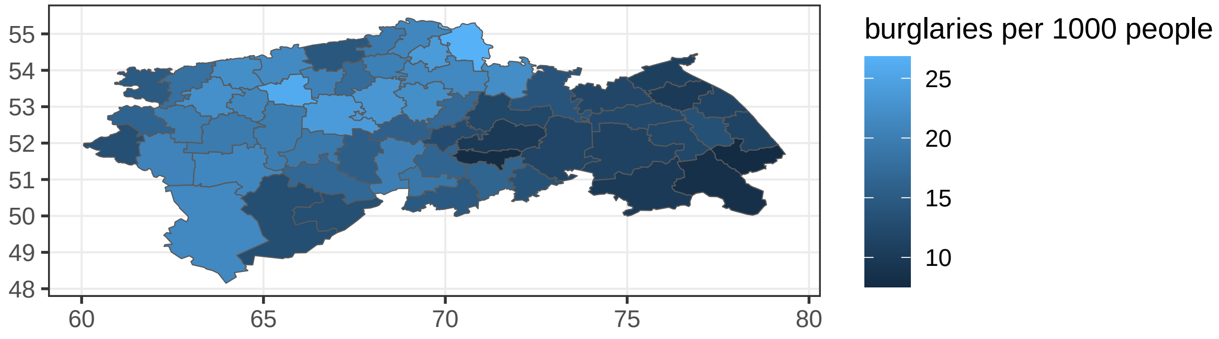
To visualise the districts worst affected by Burglary (the most common crime), first we need to filter the dataframe and group by district. Then, to find it in terms of incidents per 1,000 population, we need to join the data Burglaries with Population and create a new column:

```r
BurglariesByDistrict <- U_Crimes %>% filter(Category == "Burglary") %>%
  group_by(District_ID) %>% summarise(n=n())
BurglariesByDistrict <- full_join(BurglariesByDistrict, U_Population, by = "District_ID") %>%
  mutate(per_1000 = n/(Population/1000))
```

Now we need to plot this data on the shapefile by joining the two by the district. To join them we need to change the format of the variable "District_ID" in Burglaries:

```
BurglariesByDistrict$District_ID <- sub("^","District ",BurglariesByDistrict$District_ID)
U_Burglaries <- inner_join(BurglariesByDistrict, Utopia, by = c("District_ID"="NAME_1"))

ggplot(U_Burglaries, aes(fill= per_1000, geometry = geometry)) +
  geom_sf() + theme_bw() + labs(fill = "burglaries per 1000 people")
```



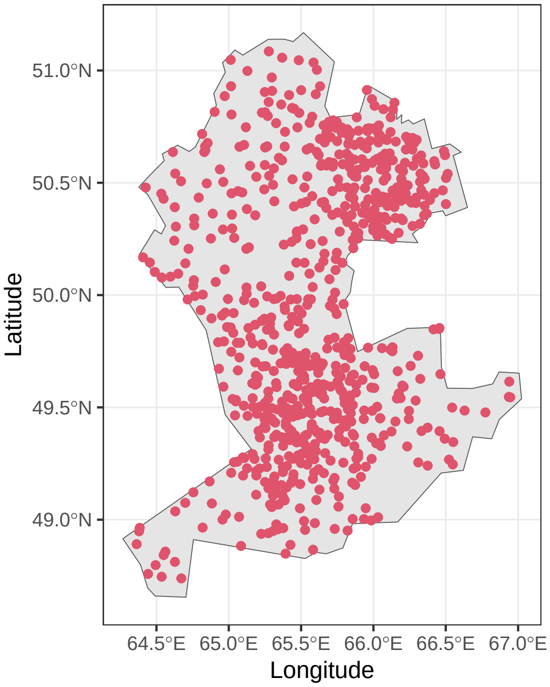
We see that there is a higher number of burglaries per 1000 people in a strip from the North to the South-West, where East Utopia has much less of a burglary problem.

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**]

To do this we need to extract District 44 from the shapefile and filter the crimes file so that we only have drug possessions in District 44:

```
District44 <- Utopia %>% filter(NAME_1 == "District 44")
DrugPoss <- U_Crimes %>% filter(Category == "Drug Possession") %>% filter(District_ID == "44")

ggplot(District44) + geom_sf() + theme_bw() +
  geom_point(data = DrugPoss, aes(x = Longitude, y = Latitude), colour=2 )
```



We see two main points with high concentration of drug possession offences, the North-East and in the middle to the south. Therefore, I would suggest to focus of these two areas when conducting the raid.

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**]

To do this I need to filter to only get burglaries, then extract the description of the victims from the variable "Description":

```
Burglaries <- U_Crimes %>% filter(Category == "Burglary")
Burglaries <- Burglaries %>%
  separate(Description, into = c("Criminals", "Victim", "Means", "Other"),
           sep=";", extra="drop")
Burglaries %>% group_by(Victim) %>% summarise(n = n()) %>%
  arrange(desc(n)) %>% slice_head(n=1)
```

```
## # A tibble: 1 x 2
##   Victim                 n
##   <chr>              <int>
## 1 " elderly single "  4410
```

We see that elderly single people are most at risk of burglary.

To find what proportion of burglaries that involve more than two criminals we need to look at when it was committed by "Three Criminals" or "More than 3 criminals"

```
mean(Burglaries$Criminals == "Three Criminals " |
       Burglaries$Criminals == "More than 3 criminals ")
```

```
## [1] 0.2440501
```

We see that 24.4% of burglaries are committed by more than two criminals.

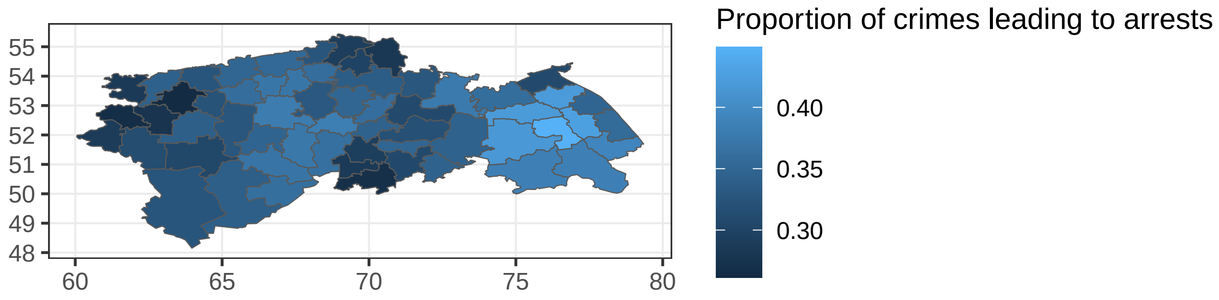
d) *Make up your own question and answer it. Your question should consider 1-2 aspects different to that in parts 2a)-2c). Originality will be rewarded.* [**7 marks**]

I want to investigate how effective each district's police department is at investigating and finding the perpetrators of the crimes.
To do this, I am going to look at the proportion of crimes that lead to arrests,

```
Crime_types <- U_Crimes %>% group_by(Category) %>% summarise(n= n())
Arrests <- U_Crimes %>% group_by(District_ID) %>% summarise(prop_arrest = mean(Arrest == "Yes"))

Arrests$District_ID <- sub("^","District ",Arrests$District_ID)
U_Arrests <- inner_join(Arrests, Utopia, by = c("District_ID"="NAME_1"))

ggplot(U_Arrests, aes(fill= prop_arrest, geometry = geometry)) +
  geom_sf() + theme_bw() + labs(fill = "Proportion of crimes leading to arrests")
```



I am going to extract the three districts with the lowest proportions of arrests made:

```
Arrests %>% arrange(prop_arrest) %>% slice_head(n=3)
```

```
## # A tibble: 3 x 2
##   District_ID prop_arrest
##   <chr>             <dbl>
## 1 District 55       0.262
## 2 District 47       0.267
## 3 District 10       0.270
```

We see that Districts 55, 47, and 10 have the lowest proportion of arrests made. Assuming that the reason crimes didn't lead to arrests is that the criminals weren't found by the police department, these districts are the least effective at investigating crimes.

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**]

Dear Utopia Police Department,
I have taken a look at the data you sent me and I have some recommendations resulting from my analysis. We see that the 3 most common crimes in Utopia are burglary, drug possession, and assault; to combat these I would recommend getting residents to install security systems, do an anti-drug campaign and increase police presence on the streets,to combat these crimes respectively. We see that East Utopia is affected a lot less by burglaries than the rest of the country, so I would suggest that you investigate what the factors affecting this are and accordingly implement changes to the rest of the districts. You plan on conducting a drug raid in district 44, the map I created shows two points to focus on. Therefore, I would recommend sending in two teams, one in the North-East (centred around latitude 50.5 and longitude 66.0)and the other in the middle to the south (centred around latitude 49.5 and longitude 65.5).

We see that single elderly people are most at risk of being burgled so I would recommend providing all single elderly residents with security systems to deter burglars. We also see that only 24.4% of burglaries are committed by more than two criminals, this good because it is much easier to deter one or two people than more than two. I investigated how effective each district's police department is at investigating and finding the perpetrators of crimes, and found that districts 55, 47 and 10 are the least effective at this. Therefore, assuming that the reason crimes didn't lead to arrests is that the criminals weren't found by the police department, I would recommend considering giving them more funding and hiring more staff. Alternatively, you could give them training sessions run by the higher performing districts like district 31 to improve their investigation skills and increase the proportion of arrests made.