MA20277 2022 - Coursework 2

Callum Gregory

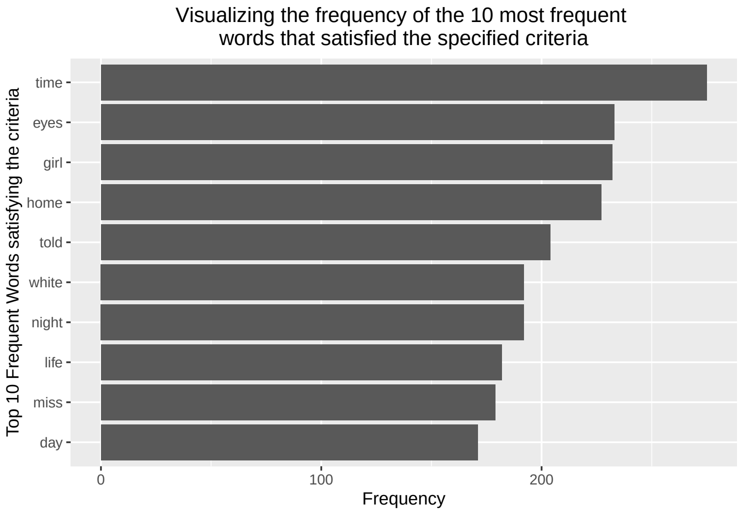
```
library(dplyr, warn.conflicts = FALSE)
library(lubridate, warn.conflicts = FALSE)
library(tidyr)
library(ggplot2)
library(patchwork)
library(ggmap)
library(sf)
library(sp)
library(spatstat)
library(maptools)
library(spdep)
library(tidytext)
library(widyr)
library(gstat)
```

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”.

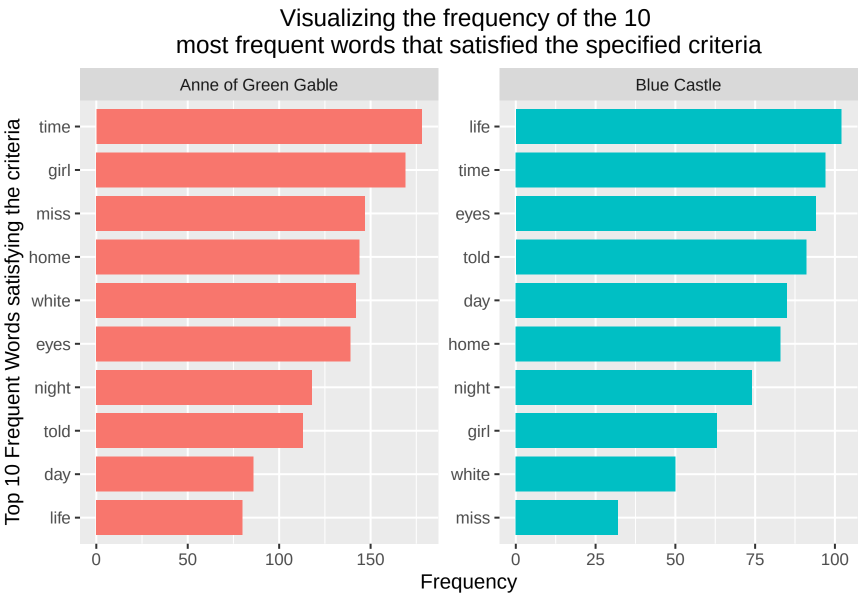
a) *Visualize the frequency of the 10 most frequent words that satisfy the following three criteria: (1) The word occurs at least five times in each book, (2) The word is not a stop word according to the usual stop list considered in the lectures, (3) The word is not “I’m”, “don’t”, “it’s”, “didn’t”,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

```
AOGG_raw <- readLines("Anne of Green Gables UTF8.txt", encoding = "UTF-8")
AOGG_raw <- data.frame(line=1:10721, text=AOGG_raw)
AOGG <- AOGG_raw %>% unnest_tokens(word, text) %>%
mutate(word = gsub("\\_", "", word)) %>% count(word, sort=TRUE) %>%
anti_join(stop_words)
Blue_Castle_raw <- readLines( "Blue Castle UTF8.txt", encoding = "UTF-8")
Blue_Castle_raw<- data.frame(line=1:8080, text=Blue_Castle_raw)
Blue_Castle <- Blue_Castle_raw %>% unnest_tokens(word, text) %>%
mutate(word = gsub("\\_", "",word)) %>% count(word, sort=TRUE) %>%
anti_join(stop_words)
six_words <- c("i’m", "don’t", "it’s", "didn’t", "i’ve", "i’ll")
words_proportion <- full_join(AOGG, Blue_Castle, by="word") %>%
rename(AOGG_n='n.x', Blue_Castle_n='n.y') %>% drop_na() %>%
filter(AOGG_n > 4, Blue_Castle_n > 4, !word %in% six_words) %>%
mutate(Total_n=AOGG_n+Blue_Castle_n) %>% arrange(desc(Total_n)) %>% slice_head(n=10)
ggplot(words_proportion%>%mutate(word=reorder(word,Total_n)), aes(x=Total_n, y=word)) +
geom_col() + labs(x="Frequency", y="Top 10 Frequent Words satisfying the criteria",
title="Visualizing the frequency of the 10 most frequent
words that satisfied the specified criteria") + theme(plot.title=element_text(hjust = 0.5))
```



The data graphic above demonstrates the total frequency of each of the 10 most frequent words satisfying the specified criteria, but it might be more interesting to see how each of the different books contributed to the frequencies for each specific words. This is highlighted below:

```
words_prop_tidy <- words_proportion %>%
pivot_longer(cols=c(AOGG_n, Blue_Castle_n), names_to="Book") %>% rename(Number=value) %>%
arrange(desc(Number)) %>% mutate(word=reorder(word,Number))
words_prop_tidy%>% group_by(Book) %>% slice_max(order_by=Number, n=10) %>% ungroup() %>%
ggplot(aes(x=Number, y=reorder_within(word,Number,Book), fill=Book)) +
facet_wrap(~Book, scales="free", labeller = as_labeller(c(`AOGG_n`="Anne of Green Gable",
`Blue_Castle_n`="Blue Castle"))) + geom_col(show.legend = FALSE, width=0.8) +
scale_y_reordered() + labs(x="Frequency", y="Top 10 Frequent Words satisfying the criteria",
title="Visualizing the frequency of the 10
most frequent words that satisfied the specified criteria") + theme(plot.title=
element_text(hjust = 0.5))
```



From this data graphic, it is noticeable that certain words like "girl" and "miss" have a much higher frequency in the book Anne of Green Gables compared to Blue Castle. This isn't really that surprising though, given that Anne of Green Gables is a bigger book than Blue Castle.

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 distance 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**]

```
ROSF_raw <- readLines("Rebecca of Sunnybrook Farm UTF8.txt", encoding="UTF-8")
ROSF_raw<- data.frame(line=1:7970, text=ROSF_raw)
both_books=bind_rows(list(`Anne of Green Gables`=AOGG_raw,
`Rebecca of Sunnybrook Farm`=ROSF_raw), .id = 'book')
Books=both_books%>% unnest_tokens(word, text) %>% mutate(word = gsub("\\_", "", word)) %>%
count(book, word, sort=TRUE) %>% anti_join(stop_words)
Books_tf.idf <- Books %>% bind_tf_idf(word, book, n) %>% arrange(desc(tf_idf))
Books_tf.idf %>% filter(book=="Anne of Green Gables") %>% slice_head(n=2)
```

```
                 book     word    n         tf       idf     tf_idf
1 Anne of Green Gables     anne 1102 0.02987745 0.6931472 0.02070947
2 Anne of Green Gables marilla  795 0.02155406 0.6931472 0.01494014
```

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

```
                       book      word   n          tf       idf     tf_idf
1 Rebecca of Sunnybrook Farm   rebecca 572 0.021612635 0.6931472 0.01498074
2 Rebecca of Sunnybrook Farm rebecca's 105 0.003967354 0.6931472 0.00274996
```

- It is found that that the top two words with the highest term frequency-inverse distance frequency for the book Anne of Green Gables are "anne" and "marilla"

- It is found that that the top two words with the highest term frequency-inverse distance frequency for the book Rebecca of Sunnybrook Farm are "rebecca" and "rebecca's". However, these two words don't really tell us two differing pieces of information, so to get two distinct, differing words for this book we take the word with the next highest term frequency-inverse distance frequency - which is "cobb".

```
Books_tf.idf %>% filter(book=="Rebecca of Sunnybrook Farm", word!="rebecca's") %>%
  slice_head(n=2)
```

```
                        book    word   n          tf       idf      tf_idf
1 Rebecca of Sunnybrook Farm rebecca 572 0.021612635 0.6931472 0.014980737
2 Rebecca of Sunnybrook Farm    cobb  90 0.003400589 0.6931472 0.002357109
```

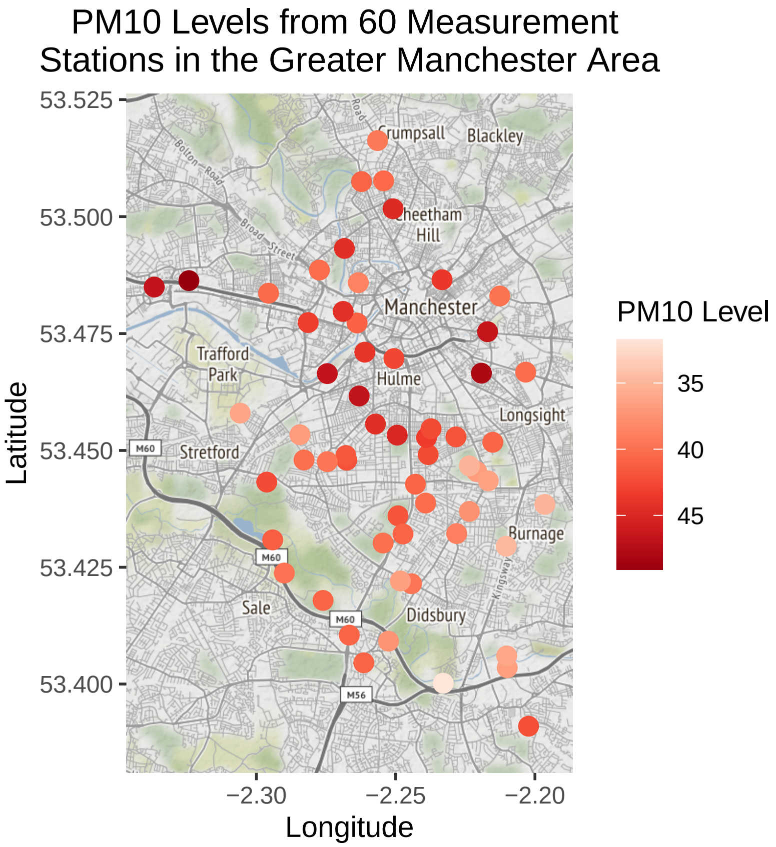
It is unsurprising that character names have the highest term frequency-inverse distance frequency, since the named characters only appear in the correspondingly-related book and not the other.

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")
Manchester <- Manchester %>% rename(Longitude=Lon, Latitude=Lat, PM10_Level=Level)
PlotDim <- c(left=min(Manchester$Longitude)-0.01, right=max(Manchester$Longitude)+0.01,
top=max(Manchester$Latitude)+0.01, bottom=min(Manchester$Latitude)-0.01)
ggmap(get_stamenmap(PlotDim, maptype="terrain", zoom=12)) +
geom_point( data=Manchester, aes(x=Longitude, y=Latitude, color=PM10_Level), size=3) +
scale_color_distiller(palette="Reds", trans="reverse") +
labs(x="Longitude", y="Latitude", color="PM10 Level",
     title="PM10 Levels from 60 Measurement
Stations in the Greater Manchester Area")  + theme(plot.title = element_text(hjust = 0.5))
```

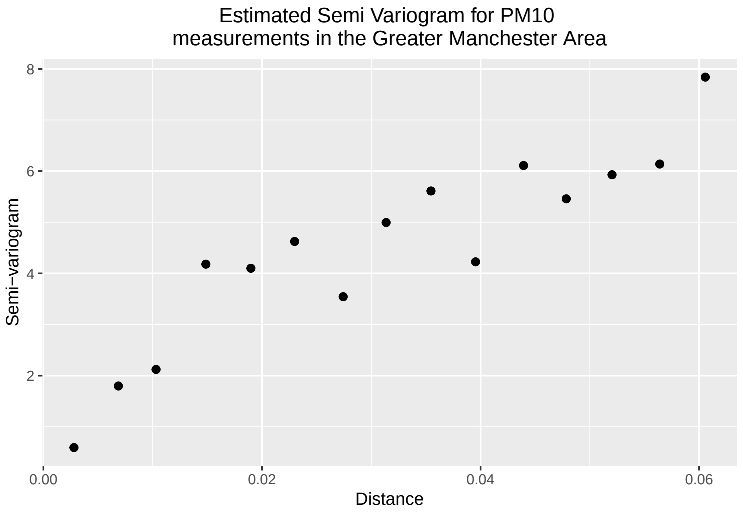


From the data graphic, there is generally a higher PM10 Level recorded around Manchester city centre - for example around Hulme. This intuitively makes sense as high levels of air pollution are likely to occur here, since there will be a lot of traffic and congestion. Towards the South of Manchester - near Didsbury and Burnage - there is generally lower PM10 Levels presumably since there is less traffic and congestion.

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

Since we have point-referenced data, to explore the spatial dependence of the PM10 measurements a semi variogram is used. We also make the assumption of isotropy. That is, the exact locations of each measurement station are not relevant for analyzing the spatial dependence as long as we know the distance between every pair of measurement stations.

```r
SST_gamma <- Manchester
coordinates(SST_gamma) <- ~Longitude+Latitude
gamma_hat <- variogram(PM10_Level~1, SST_gamma)
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 the Greater Manchester Area") + theme(plot.title=element_text(hjust=0.5))
```



The variogram indicates positive correlation of spatially close measurement sites, and that the degree of spatial dependence decreases with increasing spatial distance between measurement sites in the Greater Manchester Area.

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**]

```r
# Inverse-distance weighting function
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] )
}

# Estimate for Location 1
coord <- cbind( Manchester$Longitude, Manchester$Latitude )
s_star <- c(-2.275, 53.354)
IDW( X=Manchester$PM10_Level, S=coord, s_star, p=2)
```

```
[1] 40.14383
```

```r
# Estimate for Location 2
s_star <- c(-2.250, 53.471)
IDW( X=Manchester$PM10_Level, S=coord, s_star, p=2)
```

```
[1] 42.68654
```

```r
# Extremities of the data
c(`Min. Longitude`=min(Manchester$Longitude), `Max. Longitude`=max(Manchester$Longitude),
`Max. Latitude`=max(Manchester$Latitude), `Min. Latitude`=min(Manchester$Latitude))
```

```
Min. Longitude Max. Longitude  Max. Latitude  Min. Latitude
     -2.336692      -2.196436      53.516253      53.390959
```

- The estimate for Location 1 is not very reliable - it is out of the range of the data we currently possess (extrapolation) since the lowest latitude value in the data is 53.39096.
- The estimate for Location 2 is more reliable since it is within the cluster of 60 stations.
- It is also worth noting that the p-value has also been chosen by "eye". It would be worth using the concept of cross-validation to help find a p-value that performs best for predictions to get a more reliable estimate for Location 2.

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**]

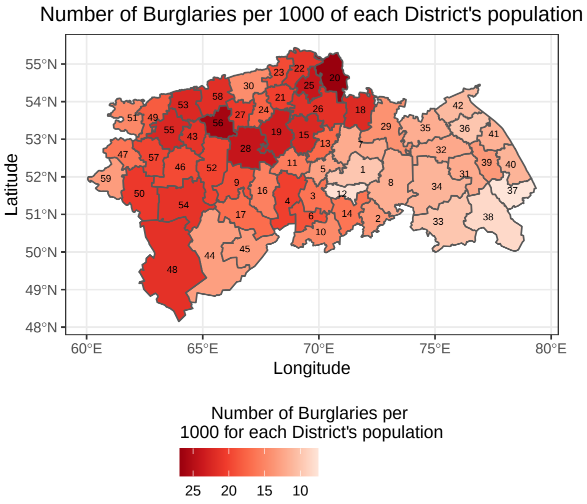
```
Utopia_Crimes <- read.csv("UtopiaCrimes.csv")
Utopia_Population <- read.csv("UtopiaPopulation.csv")
# What are the 3 most common crimes in Utopia
Utopia_Crimes %>% group_by(Category) %>% summarise("Number of Incidents"=n())%>%
arrange(desc(`Number of Incidents`)) %>% slice_head(n=3)
```

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

We see that the 3 most common crimes are Burglaries, Drug Possession and Assaults.

```
# Create a map that visualizes the districts worst affected by the most common crime
# in terms of number of incidents per 1,000 population
Most_Common_Crime <- Utopia_Crimes %>% group_by(District_ID) %>%
filter(Category=="Burglary") %>% summarise("Number of Incidents"= n())
Utopia_Burglaries <- inner_join(Utopia_Population, Most_Common_Crime, by="District_ID")%>%
mutate(`No. Incidents per 1000`=`Number of Incidents`/(Population/1000))
UTO <- read_sf( "UtopiaShapefile.shp" ) %>%
mutate(District_ID=as.numeric(gsub("District", "", NAME_1)))
Utopia <- inner_join( x=UTO, y=Utopia_Burglaries, by=c("District_ID"="District_ID"))
ggplot(data=Utopia, aes(fill=`No. Incidents per 1000`)) + geom_sf() + theme_bw() +
geom_text(label=UTO$District_ID, size=2, aes(geometry=geometry), stat="sf_coordinates") +
```

```r
scale_fill_distiller(palette="Reds", trans="reverse") +
labs(title="Number of Burglaries per 1000 of each District's population", x ="Longitude",
y="Latitude", fill="Number of Burglaries per
1000 for each District's population") + theme(plot.title = element_text(hjust = 0.5),
legend.position="bottom") + guides(fill = guide_colorbar(title.position="top",
title.hjust=0.5, label.hjust=0.5))
```

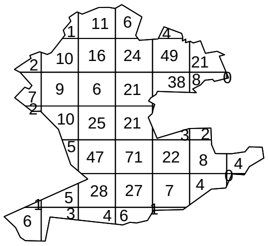


From the data graphic, there appears to be a higher number of burglaries per 1000 of each District's population on the left hand side of Utopia compared to the right hand side.

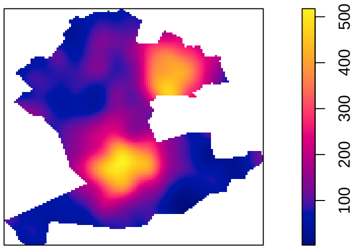
b) *You are told that District 44 is notorious for drug possession. The police is planning to conduct a raid to tackle the issue, but they are unsure on which area of the district they should focus on. Help them make the correct decision.* [**5 marks**]

Intuitively, it makes sense to assume that criminals who are found in possession of drugs won't tend to get found in the same places again. Thus, if the police want to conduct a raid it seems reasonable to only explore the locations where criminals have not been found. It is also worth discovering how the criminals are distributed across District 44, so the police can target an area where there a lot of criminals to make the raid a good use of time.

```r
District_44_data <- Utopia_Crimes %>% filter(District_ID==44, Category=="Drug Possession",
Arrest=="No")
District_44 <- filter(UTO, NAME_1=="District 44")
District_44_sp <- as(District_44, "Spatial")
District_44_sp <- slot(District_44_sp, "polygons")
District_44_win <- lapply(District_44_sp, function(z) { SpatialPolygons(list(z)) })
District_44_win <- lapply(District_44_win, as.owin)[[1]]
District_44_ppp <- ppp(x=District_44_data$Longitude,y=District_44_data$Latitude,
window=District_44_win)
District44 <- quadratcount(District_44_ppp, nx=7, ny=7)
plot(District44, main="Quadrat Count for Number of
Criminals in possession of drugs in District 44", cex.main=0.5)
lambdaC <- density.ppp(District_44_ppp, edge=TRUE, sigma=0.1)
plot(lambdaC, main="Modelling the intensity of Drug Possession
using the uniformly corrected smoothed kernel intensity function", cex.main=0.5)
```



Quadrat Count for Number of
Criminals in possession of drugs in District 44



Modelling the intensity of Drug Possession
using the uniformly corrected smoothed kernel intensity function

From the quadrat count, it appears that the distribution of criminals in possession of drugs seems to be clustered and thus the point process is non-homogeneous. This is unsurprising as criminals in possession of drugs may tend to work in "drug gangs" . It seems advisable to suggest that the police should conduct a raid just below the centre of District 44 as this is where there is highest intensity.

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**]

```
Utopia_Burglary <- Utopia_Crimes %>% filter(Category=="Burglary") %>%
mutate(separate(data.frame(text=Description), text, into=c("Number of Criminals",
"Victim", "Crime Location", "Crime Committed"), sep=";", fill="right", extra="drop"))
Utopia_Burglary %>% group_by(Victim) %>% summarise("Number of Burglaries"=n()) %>%
arrange(desc(`Number of Burglaries`)) %>% slice_head(n=1)
```

```
# A tibble: 1 x 2
  Victim             ‘Number of Burglaries‘
  <chr>                               <int>
1 " elderly single "                   4410
```

```
Criminals <- Utopia_Burglary %>% group_by(`Number of Criminals`) %>%
summarise("Number of Burglaries"=n())
(1874+2156)/(1874+6972+2156+5511)
```

```
[1] 0.2440501
```

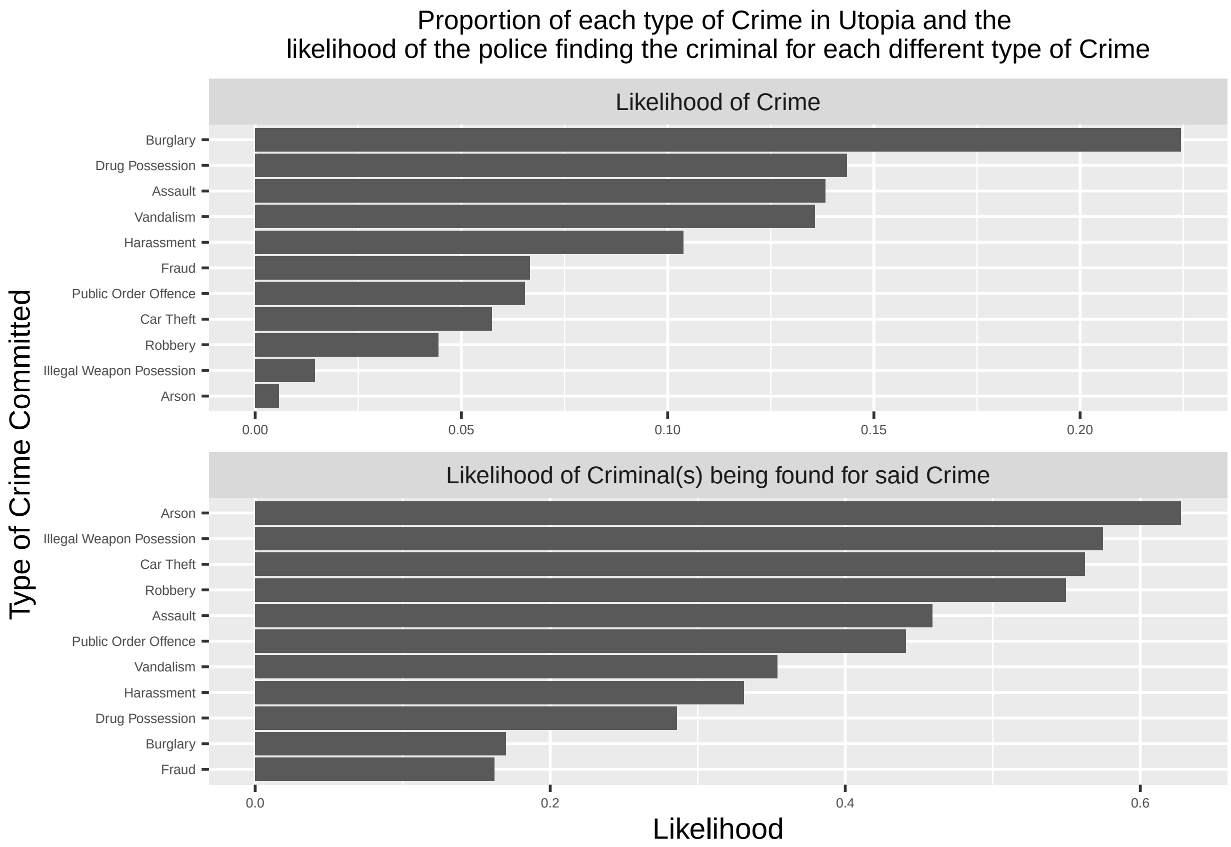
- It is found that single elderly people suffer the most number of burglaries out of all the victim groups.
- The proportion of burglaries that involved more than two criminals is 0.244 (24.4%)

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**]

Does the type of Crime and the District in which the Crime took place affect how likely the criminals are to be found?

We begin by observing broadly over the whole of Utopia.

```
data1 <- Utopia_Crimes %>% group_by(Category) %>% summarise("n"=n()) %>%
mutate(`Likelihood of Crime`=n/sum(n)) %>% select(-n)
data2 <- Utopia_Crimes %>% group_by(Category, Arrest) %>% summarise("n"=n())
a <- rowsum(data2$n,rep(1:11,each=2))
dim(a) <- c(1,11)
b <- c(rbind(a, a))
dim(b) <- c(22,1)
data2 <- data2 %>% ungroup() %>% mutate(`No. of Crimes`=b,
`Likelihood of Criminal(s) being found for said Crime`=n/`No. of Crimes`) %>%
filter(Arrest=="Yes")
data1$`Likelihood of Criminal(s) being found for said Crime` <-
data2$`Likelihood of Criminal(s) being found for said Crime`
data1 <- data1 %>%
pivot_longer(cols=
`Likelihood of Crime`:`Likelihood of Criminal(s) being found for said Crime`,
names_to = "Analysis") %>% mutate(Category = reorder_within(Category, value, Analysis))
ggplot(data1, aes(y=Category, x=value)) + geom_col() + scale_y_reordered() +
facet_wrap(~Analysis, scales="free", nrow=2) + labs( x="Likelihood",
y="Type of Crime Committed", title = "Proportion of each type of Crime in Utopia and the
likelihood of the police finding the criminal for each different type of Crime") +
theme(axis.text=element_text(size = 5), plot.title=element_text(hjust = 0.5, size=10))
```



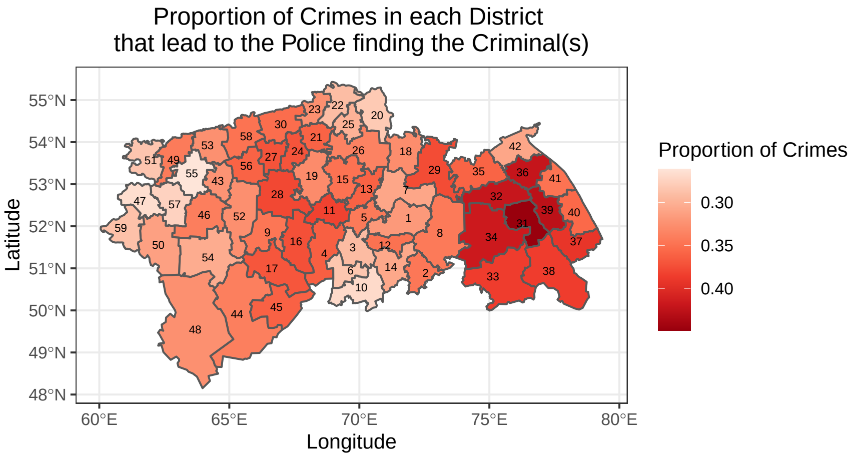
From the data graphic above, unsurprisingly we can observe that the two least likely crimes that occur in Utopia have the two highest likelihoods that the criminal will be found. This is likely because both Arson and being in possession of Illegal Weapons are extreme cases, so a lot of effort is put in by the police to find the criminals. We also see that there is a low proportion of criminals found who either commit burglaries or are in possession of drugs and this could be due to the sheer number of cases of these two crimes so the police force is overwhelmed.

Given we have noticed that there are certain crimes that generally result in the police finding the criminals more often than in some other crimes, it becomes clear to see that the number and types of crimes that happen in each district are likely to play a role in the proportion of criminals the police find in each district.

Since Utopia is fairly large it also seems reasonable to assume that each District has their own set of officers and collectively they all form Utopia's Police Department. This means we can see how well the police in each district do with regards to finding criminals.

```
Utopia1<-Utopia_Crimes%>%group_by(District_ID, Arrest)%>%summarise("Number"=n())
Utopia2<-inner_join(x=UTO, y=Utopia1, by=c("District_ID"="District_ID"))
Utopia3 <- inner_join(x=Utopia2,y=Utopia_Population,by=c("District_ID"="District_ID"))
a <- rowsum(Utopia3$Number,rep(1:59,each=2))
dim(a) <- c(1,59)
b<-c(rbind(a, a))
dim(b) <- c(118,1)
Utopia3<-Utopia3%>%mutate(Total_Criminals=b, Proportion_Caught=Number/Total_Criminals)

ggplot(data=Utopia3%>%filter(Arrest=="Yes"), aes(fill=Proportion_Caught)) + geom_sf() +
theme_bw() + geom_text(label=UTO$District_ID, size=2, aes(geometry = geometry), stat =
"sf_coordinates") + scale_fill_distiller(palette="Reds", trans="reverse") + guides(fill =
guide_colourbar(title.position = "top"))+labs(title="Proportion of Crimes in each District
that lead to the Police finding the Criminal(s)", fill="Proportion of Crimes",
x ="Longitude", y="Latitude") + theme(plot.title=element_text(hjust = 0.5)) +
guides(fill = guide_colorbar(title.position="top", title.hjust=0.5, label.hjust=0.5))
```



```
Utopia4<-Utopia3%>%filter(Arrest=="Yes")
neighbours_Utopia <- poly2nb(Utopia4)
neighbours_Utopia <- nb2listw( neighbours_Utopia, style="B")
MoranLocal <- localmoran( x=as.vector(Utopia4$`Proportion_Caught`),
listw=neighbours_Utopia)
print(c(Utopia4$NAME_1[31], MoranLocal[[31]]))
```

```
[1] "District 31"      "23.0382717055803"
```

This data graphic highlights the proportion of criminals found depending on the district. On observation, we see that on the right hand side of Utopia there appears to be a higher proportion of criminals being found by the police - most specifically in the districts 31, 32, 34, 36 and 39. The Local Moran's I is also very high for District 31 implying that there is positive spatial dependence for the proportion of criminals being found. This intuitively makes sense, as if there is a high proportion of criminals found in District 31, it and it's neighbouring districts would equally see an increase in the number of criminals found, as criminals don't tend to return to places where they have been caught and they will inform their criminal buddies to avoid the location. Hence, the number of cases go down which in turn gives the police more time to find the criminals as there are less of them.

So overall, we can see that the Police do the best under the given crime instances in District 31, but it is hard to compare how well the police have performed in each district as each district has differing numbers and types of crimes being committed. Further, since we have assumed that each district has it's own set of police officers it would perhaps be advisable for the other districts in Utopia to match the ratio of police officers to population that there is in District 31, in the hopes that the number of criminals found begins to increase throughout Utopia. Obviously though, it is not quite as simple as that due to the factors noted above and the fact that criminals are always adapting to remain uncaught.

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 analysing the crime data from Utopia's Police Department, it becomes evident that burglaries are a huge problem in the upper middle to left side of Utopia - most notably in District 20 and 56 - whereas on the right hand side of Utopia there is a significantly lower number of burglaries per 1000 of each District's respective population. Given that there is such a problem with burglaries, it might make sense for Utopia's Police Department to give crime prevention and home security tips to try and reduce the number of burglaries that happen. It was also uncovered that single elderly people are most likely to fall victim of burglaries and this could be because criminals see single elderly people as easy targets. Obviously, it is hard for the Police to protect the single elderly people from becoming victims, but one other thing Utopia's Police Department could do (apart from the ideas stated above) is ensure there is good emotional support for single elderly victims since being burgled can have a big impact on them and they may not have many people to confide in to make them feel safe again.

Another major crime in Utopia is the possession of drugs and upon investigating the whereabouts of drug possession in District 44 it is recommended that the Utopia Police Department should raid just below the centre of District 44. This is because, this area is where there is likely to be a lot of criminals in possession of drugs, so the Police stand a good chance of finding a lot of criminals and not wasting their time and resources on a pointless raid. Furthermore, it is also uncovered that the proportion of criminals found on the right hand side of Utopia is significantly higher than elsewhere in Utopia. It would therefore be advisable for Utopia's Police Department to look at the ratio of Police Officers to population size in those Districts where there are high proportions of criminals being found, and try and mimic that across the rest of Utopia in the hopes of finding more criminals. However, it is worth noting that this will only help the problem to a certain extent as certain crimes result in the criminals being harder to find than others and the crimes that happen in each District may not be similar.