MA20277 2022 - Coursework 2

JAKE FORSDYKE 23281

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

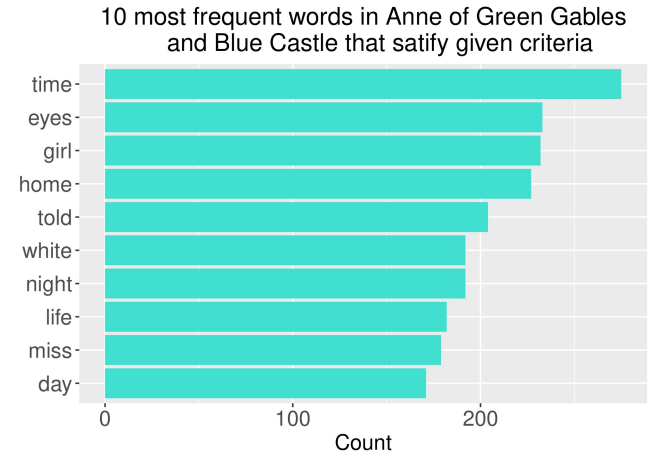
Question 1 [9 marks]

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]**

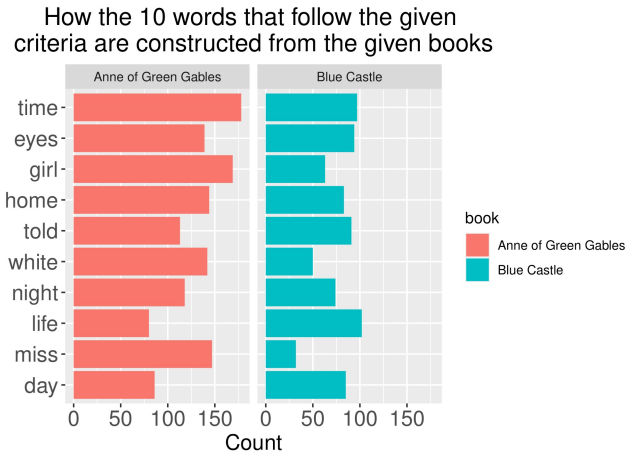
```
textRaw_Green <- readLines("Anne of Green Gables UTF8.txt")
textRaw_Blue <- readLines("Blue Castle UTF8.txt")
textRaw_Green <- data.frame(line=1:10721, text=textRaw_Green )
textRaw_Blue <- data.frame(line=1:8080, text=textRaw_Blue )

Common_words <-c("i'm", "don't", "it's", "didn't", "i've","i'll")

Green_count <- textRaw_Green %>%
  unnest_tokens( word, text )%>%
  mutate( word = gsub( "\\_", "", word ) )%>%
  anti_join( stop_words ) %>%
    count( word, sort=TRUE )
green_filter<-Green_count%>%
  filter(!word %in% Common_words)%>%
    mutate( 'term frequency'=n/sum(n), rank=row_number() )%>%
  filter( n>= 5 )
```

```
Blue_count <- textRaw_Blue %>%
  unnest_tokens( word, text )%>%
  mutate( word = gsub( "\\_", "", word ) )%>%
  anti_join( stop_words ) %>%
  count( word, sort=TRUE )
blue_filter<- Blue_count%>%
  filter(!word %in% Common_words)%>%
    mutate( 'term frequency'=n/sum(n), rank=row_number() )%>%
  filter( n>= 5 )

turquoise<-full_join( blue_filter, green_filter, by="word" ) %>%
  drop_na() %>%
 mutate("total number"=n.x+n.y)%>%
  arrange(desc(`total number`))%>%
  rename("number in AOGG" = n.x, "number in blue castle" = n.y)%>%
  rename("term frequency in AOGG"= `term frequency.x`, "term frequency in blue castle" =
           `term frequency.y`)%>%
  slice_head( n=10 )

turquoise%>%
  mutate(word=reorder(word,`total number`))%>%
     ggplot( aes( x=`total number`, y=word) ) + geom_col(fill='turquoise') +
     labs( x="Count", y="" ,title = "10 most frequent words in Anne of Green Gables
     and Blue Castle that satify given criteria") +
  theme( axis.text=element_text(size=15), axis.title=element_text(size=15)
         ,plot.title = element_text(hjust = 0.5,size=17) )
```




```
most_frequent_words <- data.frame(word=c ("time","eyes", "girl",
                                          "home","told","white","night", "life","miss","day"))
greenbook_count<-Green_count%>%
   mutate(book="Anne of Green Gables")
  greenbook<-inner_join(greenbook_count,most_frequent_words)

bluebook_count<-Blue_count%>%
    mutate(book="Blue Castle")
  bluebook<-inner_join(bluebook_count,most_frequent_words)
```

```
blue_green<-rbind(greenbook,bluebook)
blue_green%>%
  mutate(word=reorder(`word`,n))%>%
     ggplot( aes( x=n, y=word) ) + geom_col(aes(fill=book))+facet_wrap(~book)+
     labs( x="Count", y="" ,title = "How the 10 words that follow the given
criteria are constructed from the given books") +
  theme( axis.text=element_text(size=15), axis.title=element_text(size=15),
         plot.title = element_text(hjust = 0.5,size=17) )
```



This data set shows the most common 10 words that satisfy the criteria, they have been ordered from most frequent to the least. We find that there are no common character in each book and however they both have a female lead character explaining why miss and girl are so frequent. As you can see from the second graph these 10 words are influenced by Anne fo Green Gables much more than Blue Castle and neither book follows the same order as in graph 1.

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]**

```
textRaw_rebecca <- readLines("Rebecca of Sunnybrook Farm UTF8.txt")
textRaw_rebecca<-data.frame(line=1:7970,text=textRaw_rebecca)
text_Rebecca<- textRaw_rebecca %>%
    unnest_tokens( word, text ) %>%
  mutate( word = gsub( "\\_", "", word ) )%>%
  mutate(book="Rebecca of Sunnybrook Farm")

text_Greengables <- textRaw_Green %>%
  unnest_tokens( word, text )%>%
  mutate( word = gsub( "\\_", "", word ) )%>%
  mutate(book="Anne of Green Gables")

bothbooks<-rbind(text_Rebecca,text_Greengables)
bothbooks_Count <- bothbooks %>% count( book, word, sort=TRUE )
bothbooks_tf.idf <- bothbooks_Count %>%
  bind_tf_idf( word, book, n ) %>%
```

```
  arrange( desc(tf_idf) )
head( bothbooks_tf.idf )
```

```
##                          book    word    n          tf       idf      tf_idf
## 1       Anne of Green Gables    anne 1102 0.010670437 0.6931472 0.007396183
## 2 Rebecca of Sunnybrook Farm rebecca  572 0.007727850 0.6931472 0.005356537
## 3       Anne of Green Gables marilla  795 0.007697819 0.6931472 0.005335722
## 4       Anne of Green Gables   diana  385 0.003727875 0.6931472 0.002583966
## 5       Anne of Green Gables     i'm  362 0.003505171 0.6931472 0.002429599
## 6       Anne of Green Gables matthew  337 0.003263101 0.6931472 0.002261809
```

```
filter(bothbooks_tf.idf,book=="Anne of Green Gables")%>%
  slice_head(n=2)
```

```
##                   book    word    n          tf       idf      tf_idf
## 1 Anne of Green Gables    anne 1102 0.010670437 0.6931472 0.007396183
## 2 Anne of Green Gables marilla  795 0.007697819 0.6931472 0.005335722
```

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

```
##                          book    word   n          tf       idf      tf_idf
## 1 Rebecca of Sunnybrook Farm rebecca 572 0.007727850 0.6931472 0.005356537
## 2 Rebecca of Sunnybrook Farm   don't 147 0.001986003 0.6931472 0.001376593
```

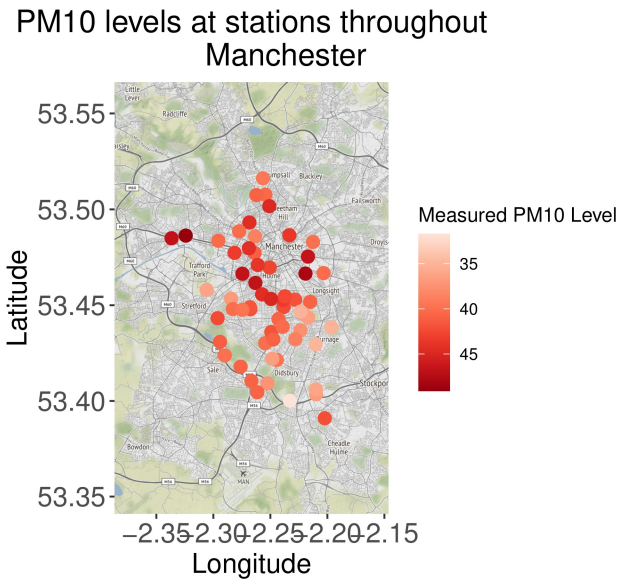
As you can see the words with the highest tf_idf value are the protagonists. this is to be expected as they are specific to the book and so will be important to that book and not the other. the unexpected result is "dont" having such a high value for rebecca of sunnybrook farm this is especially as there are other protagonists in the book like aunt marinda.

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]**

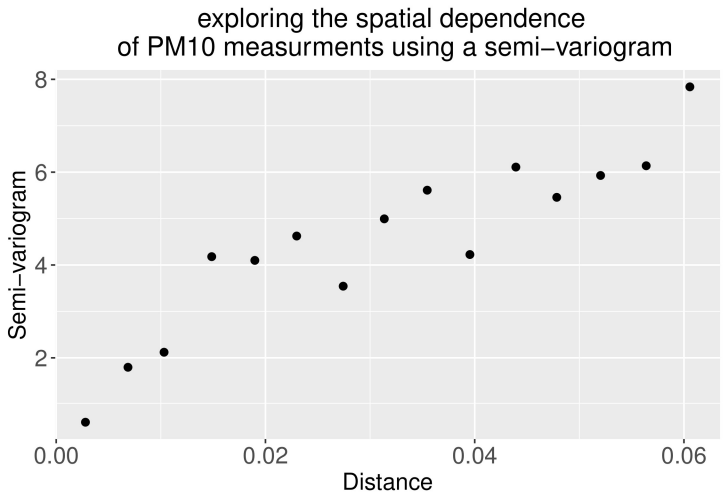
```
Manchester<- read.csv( "Manchester.csv" )
PlotDim <- c( left=min(Manchester$Lon)-0.05, right=max(Manchester$Lon)+0.05,
              top=max(Manchester$Lat)+0.05, bottom=min(Manchester$Lat)-0.05 )

ggmap( get_stamenmap(PlotDim, maptype="terrain", zoom=12) )+
 geom_point(data=Manchester, aes(x=Lon,y=Lat,color=Level), size=3 )+
  scale_color_distiller( palette="Reds", trans="reverse" ) +
  labs( x="Longitude", y="Latitude", color="Measured PM10 Level",
        title="PM10 levels at stations throughout
        Manchester" )+
theme( axis.text=element_text(size=15), axis.title=element_text(size=15),
       plot.title = element_text(hjust = 0.5,size=17) )
```



As you can see the highest levels of PM10 concentrate at the center of Manchester this is likely due to there being a denser population and it being a more built up area, however levels don't decrease too significantly as you go further out from the city center as you are still within the suburbs.

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

```
SST_gamma<-Manchester
coordinates( SST_gamma ) <- ~Lon+Lat
estim <- variogram( Level~1, SST_gamma )

ggplot( estim, aes( x=dist, y=gamma/2 ) ) + geom_point( size= 2 ) +
  labs( x="Distance", y="Semi-variogram" , title=" exploring the spatial dependence
  of PM10 measurments using a semi-variogram")+
theme( axis.text=element_text(size=15), axis.title=element_text(size=15),
       plot.title = element_text(hjust = 0.5,size=17) )
```



The nugget of this variogram is 0, the range is the entire area and there is no sill.The variogram indicates

positive correlation of spatially close sites, and that the degree of spatial dependence decrease with increasing spatial distance.

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] )
}
coord <- cbind( Manchester$Lon, Manchester$Lat )
s_star <- c( -2.275, 53.354 )
location1=IDW( X=Manchester$Level, S=coord, s_star, p=2 )

coord <- cbind( Manchester$Lon, Manchester$Lat )
s_star <- c( -2.250, 53.471 )
location2=IDW( X=Manchester$Level, S=coord, s_star, p=2 )
location1
```

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

```
location2
```

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

these measurements are unreliable as the power may not be optimal creating a larger mse then necessary, the pm10 levels are also dictated by external factors such as smoking and household products, this means that if there are no houses between two stations that read high levels of pm10 the estimate will also be high but this wont necessarily be the case making the readings unreliable.

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".

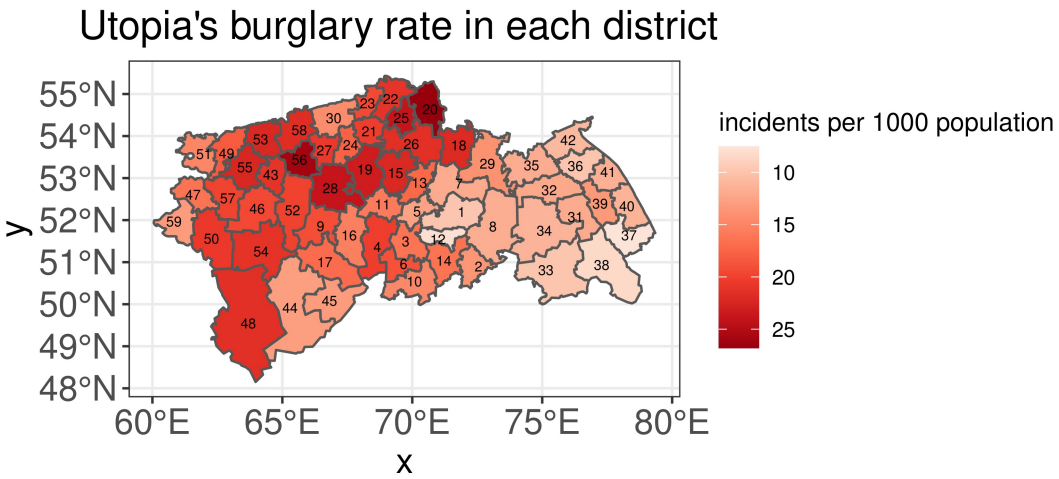
a) *What are the three most common crimes in Utopia? Create a map that visualizes the districts worst affected by the most common crime in terms of number of incidents per 1,000 population.* **[5 marks]**

```
Crimes<- read.csv( "UtopiaCrimes.csv" )
most_common_crimes<-Crimes%>%group_by(Category)%>%
  summarise(`most common crimes`=n())%>%
  arrange(desc(`most common crimes`))
  slice_head( most_common_crimes,n=3)
```

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

from the data we can see that the most common crimes are burglarys, drug possession and assault with burglary's being by far the most common crime.

```
population <-read.csv("UtopiaPopulation.csv")
UTO<-read_sf("UtopiaShapefile.shp")
UTO<-mutate(UTO,District_ID=as.numeric(gsub("District", "", NAME_1)))
most_common_crime<-Crimes%>%
  group_by(District_ID)%>%
  filter(Category=="Burglary")%>%
   summarise("Number"=n())
Burglary_crime<-inner_join(x=population,y=most_common_crime, by="District_ID")
Utopia_crime<-inner_join(x=UTO,y=Burglary_crime, by="District_ID")%>%
  mutate("incidents per 1000 population"= (Number/(Population)*1000))

ggplot(data=Utopia_crime,aes(fill=`incidents per 1000 population`))+ geom_sf() +
  theme_bw() +
 scale_fill_distiller( palette="Reds", trans="reverse" ) +
 theme( axis.title=element_text(size=15), axis.text=element_text(size=15),
        plot.title = element_text(hjust = 0.5,size=17) ) +
  labs(title = "Utopia's burglary rate in each district ")+
  geom_sf_text(aes(label = District_ID),size=2)
```

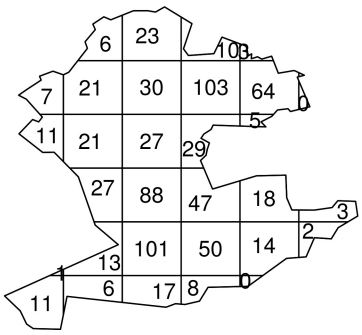


As you can see the district with the most burglarys per 100 people is district 20,We see a relatively strong spatial variation in the proportions, with values ranging from practically 7.5 per 1000 to 27 per 1000. The areas with the higher proportions appear to be in the western areas of the map, this may suggest that there are less policemen in the area or that these are richer neighbourhoods so that its more prolific to burgle the houses.

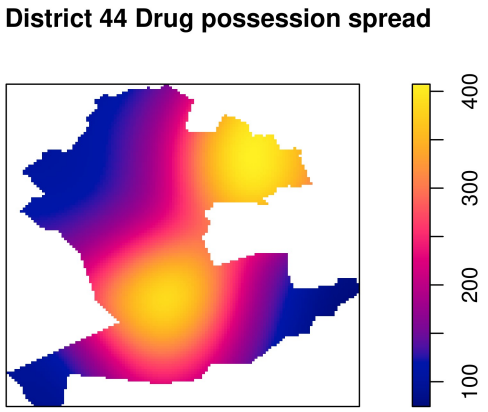
b) *You are told that District 44 is notorious for drug possession. The police is planning to conduct a raid to tackle the issue, but they are unsure on which area of the district they should focus on. Help them make the correct decision.* [**5 marks**]

```
UTO<-read_sf("UtopiaShapefile.shp")
District_44<-filter(UTO,NAME_1=="District 44")

Drugs_44<-Crimes%>%
group_by(District_ID)%>%
  filter(Category=="Drug Possession")%>%
  filter(District_ID=='44')

District44_sp <- as(District_44 , "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=Drugs_44$Longitude,y=Drugs_44$Latitude,
                       window = District44_win )
DrugsinDistrict44 <- quadratcount( District44_ppp, nx=6, ny=6 )

plot( DrugsinDistrict44, main= "District 44 Drug possession spread" )
```

District 44 Drug possession spread




```
lambdaC <- density.ppp( District44_ppp, edge=TRUE)
plot( lambdaC, main = "District 44 Drug possession spread")
```




```
K_Regular <- Kest( District44_ppp, correction = c("isotropic","translate") )
```

from the quadratcount graph we can see that the drug possession spread follows a more clusted spread, this is also shown by the estimated K-function (i havent plotted it as it exceeds the 3 graphs), following this information the police force should focus there raids on these cluster points and that there is a pattern to the spread. using the kernel intensity function we know that the focus points should be in the center of district 44 and the eastern point in the north.

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]**

```
burglarys<-Crimes%>%group_by(District_ID)%>%
  separate(col=Description, into=c('number of criminals',
                                   'victim type','entrance','stolen goods'),sep=';')%>%
  filter(Category=='Burglary')
burglarys%>%group_by(`victim type`)%>%summarise('number of victims'=n())%>%
  arrange(desc(`number of victims`))
```

```
## # A tibble: 6 x 2
##   `victim type`          `number of victims`
##   <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
```

```
burglarys%>%group_by(`number of criminals`)%>%summarise('frequency'=n())
```

```
## # A tibble: 4 x 2
##   `number of criminals`    frequency
##   <chr>                        <int>
## 1 "More than 3 criminals "      1874
## 2 "One criminal "               6972
## 3 "Three Criminals "            2156
## 4 "Two criminals "              5511
```

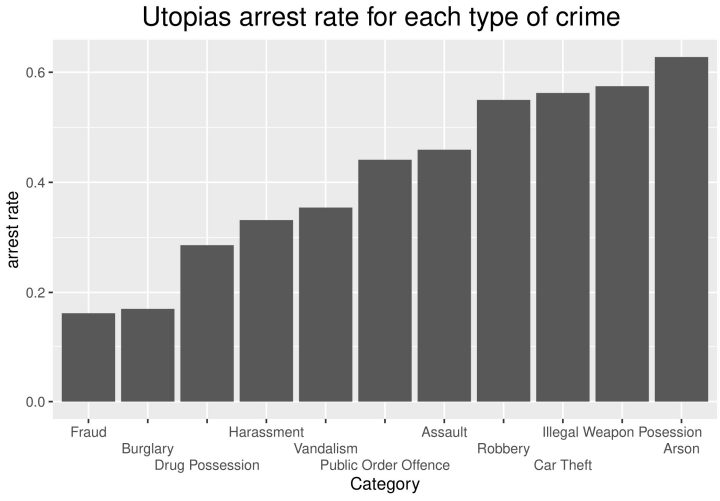
```
a=(1874+2156)/(1874+6972+2156+5511)
percent(a)
```

```
## [1] 24.41%
```

using the summary function we see that the group of people that are at most risk of being burgled are the elderly single. the percentage of berglaries that are commited by more that 2 criminals is 24.41%

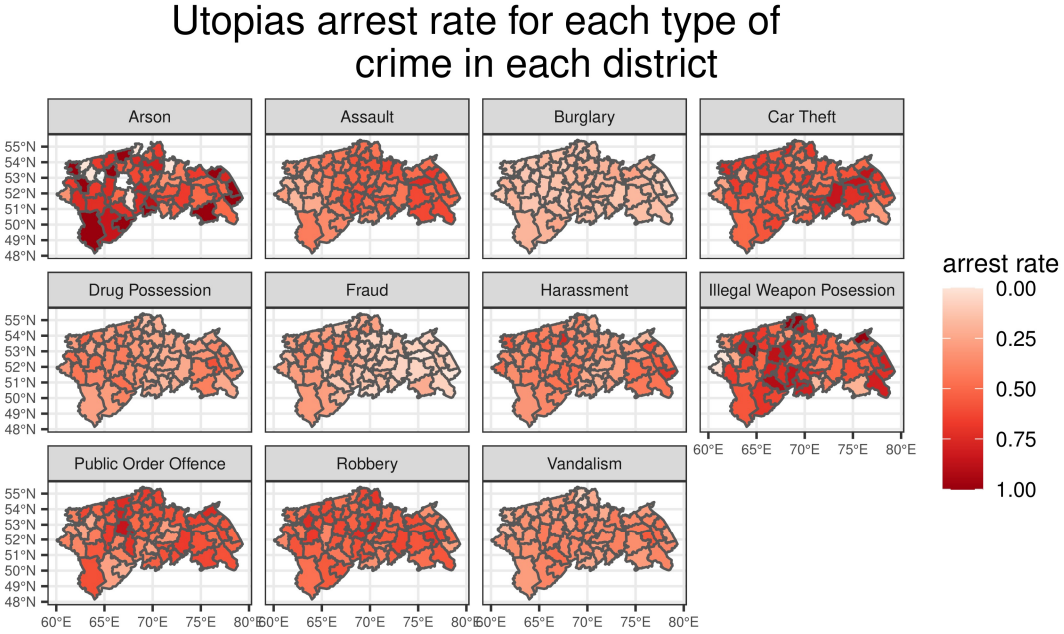
d) *Make up your own question and answer it. Your question should consider 1-2 aspects different to that in parts 2a)-2c). Originality will be rewarded.* [**7 marks**] What do the utopian police consider the most severe crime and which districts are the utopia police force most successful?

```
arrests<-Crimes%>%
   mutate("arrest type"=case_when( Arrest == "Yes" ~ 1,Arrest == "No" ~ 0  ))

arrest_rate<-arrests%>%
  group_by(Category)%>%
  summarise("arrest rate" =sum(`arrest type`==1)/(sum(`arrest type`==0) +
                                                    sum(`arrest type`==1)))%>%
  arrange(desc("arrest rate"))
arrest_rate%>%
  mutate(Category=reorder(Category,`arrest rate`))%>%
ggplot(aes(x=Category,y=`arrest rate`)) + geom_col()+
  scale_x_discrete(guide=guide_axis(n.dodge=3))+
  theme(plot.title = element_text(hjust = 0.5,size=17) ) +
  labs(title = "Utopias arrest rate for each type of crime ")
```



This data shows that the arrest rate for arson is the highest, this suggest that the utopian police department value this crime as more damaging to society than the other and so should increase police activity where there is higher rates. We also know from previous questions that this is the least occurring crime, this may confirm my finding as your less inclined to commit a crime that carries a worse a sentence and are more likely to be arrested for.There seems to be an inverse correlation between arrest rate and popularity of that crime.

```
UTO<-read_sf("UtopiaShapefile.shp")
UTO<-mutate(UTO,District_ID=as.numeric(gsub("District", "", NAME_1)))

arrests<-Crimes%>%
   mutate("arrest type"=case_when( Arrest == "Yes" ~ 1,Arrest == "No" ~ 0  ))
arrest_rate<-arrests%>%
  group_by(District_ID,Category)%>%
  summarise("arrest rate" =sum(`arrest type`==1)/(sum(`arrest type`==0) +
                                                    sum(`arrest type`==1)))%>%
  arrange(desc("arrest rate"))

Utopia_arrest_rate<-inner_join(x=UTO,y=arrest_rate, by="District_ID")

ggplot(data=Utopia_arrest_rate,aes(fill=`arrest rate`))+ geom_sf() +
  theme_bw() + facet_wrap(~Category,nrow=3)+
 scale_fill_distiller( palette="Reds", trans="reverse" ) +
 theme( axis.title=element_text(size=15), axis.text=element_text(size=6),
 plot.title = element_text(hjust = 0.5,size=20),strip.text=element_text(size=7) )+
 theme(plot.title = element_text(hjust = 0.5,size=17) ) +
 labs(title = "Utopias arrest rate for each type of
            crime in each district ")
```



From these graphs we can clearly see that the the crimes with the higher arrest rate have a much darker mapping however. we know from previous questions that there was a biased of the number of crimes occurring per 1000 being towards the west of utopia however we can see clearly that there seems to be a fairly even arrest rate throughout utopia. this suggests that were the policemen are sent out on patrol is fairly even throughout the map even though we have learnt there is an uneven crime rate. we can also see that the police forces are effective more effective in some areas of the map than other with the arrest rate of arsen being nearly one in some areas of the map. this information is useful as it could lead to a better overall police force as the more successful police department can help other departments.

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]**

In part a i found that burglarys were the most common crime committed in utopia. However,from part d this crime had one of the least arrests rate of all suggesting to me that the police department may need to change there crime tackling strategy and focus on the most popular crimes, this is because it will reduce the likelihood of a criminal not being convicted intern decreasing the crimes poplarity.I also found that that there was a pattern with where the burglary crimes were committed, they usually are targeted at districts in the west and north, however from part d we know that the arrest rates stay fairly constant throughout all the districts leading me to suggest that the police department should remove police officers from the eastern regions and place in the west to tackle the popularity of this crime.

from part b we can see that within district 44 there is a pattern as to where drug possession is spread with a concentrated amount in the middle and north east region of the district, this is where the police should commit the most officers to a raid. we also see from part a that drug possession is the second most popular crime however from part d its one of the least arrest rate crimes, this lead me to find that there is a relatively inverse relationship between arrest rate and the popularity of crimes. I would therefore recommend committing a large amount of police officers to this raid as it will reduce drug possession rates for fear of getting arrested. From part c i found that about a quarter of burglarys are committed by more that 2 people and that the majority of victims are single elders, this is to be expected as they are a vulnerable population therefore to counter this i would have a greater percentage of police in the districts with an older average population. lastly when a police officer files a police report i would record the ages of the criminals as, if theres a pattern such as teenagers commit the most burglarys, then you can prevent this by increasing awareness throughout schools.