MA20277 2022 - Coursework 2

Noel Howley

```
#LIBRARIES
suppressPackageStartupMessages(library(tidyr, warn.conflicts = F, quietly = T))
suppressPackageStartupMessages(library( tidytext , warn.conflicts = F, quietly = T))
suppressPackageStartupMessages(library( widyr , warn.conflicts = F, quietly = T))
suppressPackageStartupMessages(library( dplyr, warn.conflicts = F, quietly = T ))
suppressPackageStartupMessages(library( ggplot2 , warn.conflicts = F, quietly = T))
suppressPackageStartupMessages(library( patchwork , warn.conflicts = F, quietly = T))
suppressPackageStartupMessages(library( ggmap , warn.conflicts = F, quietly = T))
```

```
## i Google's Terms of Service: <https://mapsplatform.google.com>

## i Please cite ggmap if you use it! Use `citation("ggmap")` for details.
```

```
suppressPackageStartupMessages(library( sf , warn.conflicts = F, quietly = T))
suppressPackageStartupMessages(library( spatstat , warn.conflicts = F, quietly = T))
suppressPackageStartupMessages(library( maptools , warn.conflicts = F, quietly = T))
suppressPackageStartupMessages(library( gstat , warn.conflicts = F, quietly = T))
options(dplyr.summarise.inform = FALSE)
```

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]**

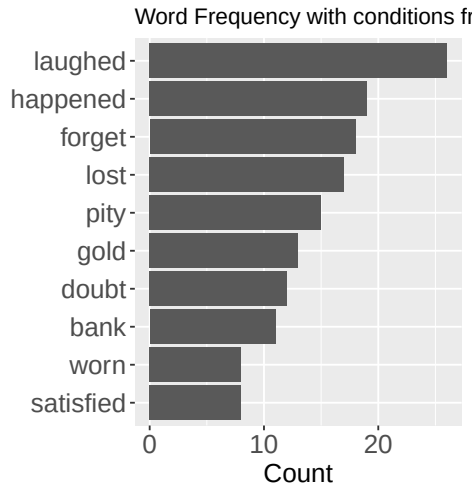
```
AoGG_raw = readLines( "Anne of Green Gables.txt" );BC_raw = readLines( "Blue Castle.txt" )
data( "stop_words" );AoGG_raw <- data.frame( text=AoGG_raw );BC_raw <- data.frame( text=BC_raw )
#Combine the two DataFrames and Clean data, and remove stop words, condition(2).
AoGG <- AoGG_raw %>%  unnest_tokens( word, text ) %>%  mutate( word = gsub( "\\_", "", word ) )%>%
  anti_join( stop_words );BC <- BC_raw %>%  unnest_tokens( word, text ) %>%
  mutate( word = gsub( "\\_", "", word ) )%>%   anti_join( stop_words )
```

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

```
# Modify the dataframes to not include the words as specified, condition(3)
AoGG_Cond_2_3 <- AoGG%>%
  mutate( word = gsub("I'm", "", word))%>%mutate( word = gsub("don't", "", word))%>%
  mutate( word = gsub("didn't", "", word))%>%mutate( word = gsub("I've", "", word))%>%
  mutate( word = gsub("I'll", "", word));BC_Cond_2_3 <- BC%>%
  mutate( word = gsub("I'm", "", word))%>%mutate( word = gsub("don't", "", word))%>%
  mutate( word = gsub("didn't", "", word))%>%  mutate( word = gsub("I've", "", word))%>%
  mutate( word = gsub("I'll", "", word));#Apply the count and other functions
#to the dataframe and filter the words to have n > = 5, #condition (1)
AoGG_Cond_2_3_Count = AoGG_Cond_2_3 %>% count( word, sort=TRUE )%>%
  filter(n >=5);BC_Cond_2_3_Count = BC_Cond_2_3 %>%  count( word, sort=TRUE )%>%
  filter(n >=5);#Finally, Combine the DataFrames.
BothBooks_all_conditions_count = AoGG_Cond_2_3_Count %>%inner_join(BC_Cond_2_3_Count)
```

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

```
Question1 = BothBooks_all_conditions_count%>%  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=15),
                                               axis.title=element_text(size=15) ) +
  labs(title = "Word Frequency with conditions from Question");Question1
```



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]**

```
RoSF_raw = readLines( "Rebecca of Sunnybrook Farm.txt" )
RoSF_raw <- data.frame( text=RoSF_raw );RoSF <- RoSF_raw %>%
  unnest_tokens( word, text ) %>%mutate( word = gsub( "\\_", "", word ) )%>%
  anti_join( stop_words );#define the count
```

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

```
RoSF_Count = RoSF %>%  count( word, sort=TRUE )%>%
  mutate(Story = "Rebecca of Sunnybrook Farm");AoGG_Count = AoGG %>%
  count( word, sort=TRUE )%>%mutate(Story = "Anne of Green Gables")
corpus = RoSF_Count%>%full_join(AoGG_Count);Q2_tf.idf = corpus %>%
  bind_tf_idf(word,Story, n);Q2_tf.idf%>%group_by(Story)%>%
  arrange(desc(tf_idf))%>%slice_head(n = 2)%>%  select(-tf,-idf)
```

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

```
## # A tibble: 4 x 4
## # Groups:   Story [2]
##   word          n Story                       tf_idf
##   <chr>     <int> <chr>                        <dbl>
## 1 anne       1102 Anne of Green Gables       0.0207
## 2 marilla     795 Anne of Green Gables       0.0149
## 3 rebecca     572 Rebecca of Sunnybrook Farm 0.0150
## 4 rebecca's   105 Rebecca of Sunnybrook Farm 0.00275
```

The Top two words with the highest term frequecy-inverse document frequency for each of the two books are "anne", "marilla" for Anne of Green Gables, and "rebecca", "rebecca's" for Rebecca of Sunnybrook Farm .

Question 2 [9 marks]

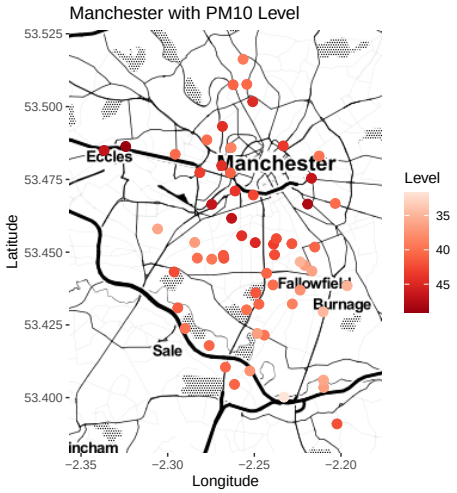
We were given PM10 measurements from 60 measurement stations in the Greater Manchester area, including the locations of the stations. The data can be found in the file "Manchester.csv". A detailed description of the variables is provided in the file "DataDescriptions.pdf".

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

```
a = read.csv("Manchester.csv")
PlotDim <- c("left"=min(a$Lon)-0.02, "right"=max(a$Lon)+0.02,
             "top"= max(a$Lat)+0.01, "bottom"=min(a$Lat)-0.01 )
man <- ggmap( get_stamenmap(PlotDim, maptype="toner", zoom=11) )
```

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

```
manchester = man +geom_point( data=a, size=3,aes( x=Lon, y=Lat, color=Level ) ) +
scale_color_distiller( palette="Reds", trans="reverse" ) +
labs( color="Level", x="Longitude", y="Latitude" ,
      title = "Manchester with PM10 Level");suppressWarnings(manchester)
```

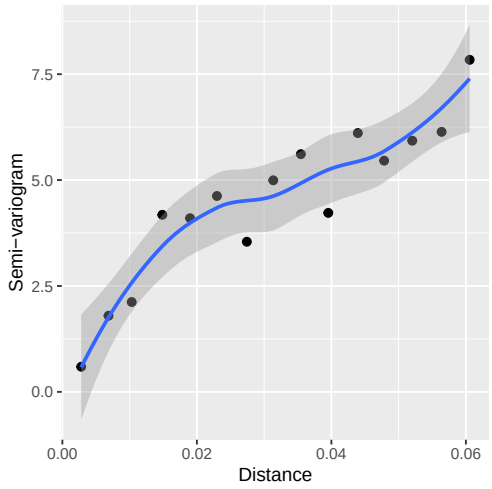


It seems that the data has higher values of PM10 "Level" happen to be arpound the city of Manchester and further from the fields and surrounding countryside.

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

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

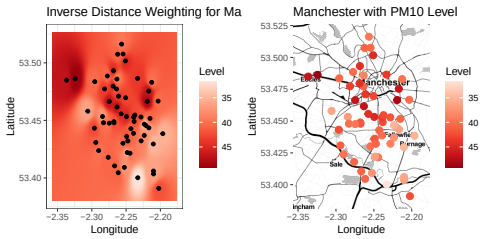
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



The variogram indicates positive correlation of spatially close sites, and that the spatial dependence is increasing with increasing spatial distance between sites.

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
#define 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] )
};a = read.csv("Manchester.csv")
#define the grid via longitde and latitude points from csv
points_Lat <- seq( min(a$Lat)-0.01, max(a$Lat)+0.01, by=0.001 )
points_Lon <- seq( min(a$Lon)-0.02, max(a$Lon)+0.02, by=0.001 )
pixels <- as.matrix( expand.grid( points_Lon, points_Lat ) )
#predict
Predict <- c();coord <- cbind( a$Lon, a$Lat )
for( j in 1:length(pixels[,1]) )
  Predict[j] <- IDW( a$Level, coord, pixels[j,], p=4 )
IDW_predict <- data.frame( "Lon"=pixels[,1], "Lat"=pixels[,2], "Pred"=Predict )
#plot
ggplot() + theme_bw() +  geom_raster( data=IDW_predict, aes(x=Lon, y=Lat, fill=Pred)) +
  scale_fill_distiller( palette="Reds", trans="reverse" ) +
  geom_point( data=a, aes(x=Lon, y=Lat) )+labs( x="Longitude", y="Latitude",fill="Level" ,
        title = "Inverse Distance Weighting for Manchester") + manchester
```




```r
coord <- cbind( a$Lon, a$Lat ) #PREDICTING THE TWO POINTS
A = c(-2.275,53.354) #POINT A;
Prediction1 = IDW(a$Level, S =coord, A, p=4)
B = c(-2.250,53.471) #POINT B
Prediction2 = IDW(a$Level, coord, B, p=4)
print(c(Prediction1,Prediction2)) # PRINT
```

```
## [1] 39.56061 42.72760
```

For point A the x value isnt in the plot range, hence this point is far from other points implying lower spacial dependence via question b). This means we may have low precision on this value depending on p in Inverse Distance Weighting function. However we can say with some certainty that point B, is more precise due to it being closer to the other points and seems reasonable being near the the city of Manchester.

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");places = read.csv("UtopiaPopulation.csv") #Reading the csv
u_shape <- st_read( "UtopiaShapefile.dbf" ) #Reading the Utopia Shapefile
```

```
## Reading layer 'UtopiaShapefile' from data source
##   'C:\Users\noelh\OneDrive\Documents\MSDS Yr2\DS\cw2\UtopiaShapefile.dbf'
##   using driver 'ESRI Shapefile'
## Simple feature collection with 59 features and 1 field
## Geometry type: MULTIPOLYGON
## Dimension:     XY
## Bounding box:  xmin: 60.05291 ymin: 48.15122 xmax: 79.35101 ymax: 55.43181
## Geodetic CRS:  WGS 84
```

```
a = crimes %>%  group_by(Category)%>%summarise(Count = n())%>%  arrange(desc(Count))
ThreeMostFrequent = a%>% slice_head(n = 3);print(ThreeMostFrequent)
```

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

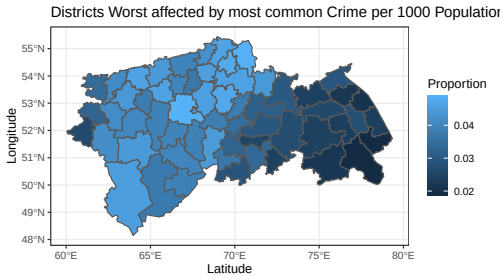
We see that the Three most common types of crime in Utopia is Burglary, Drug Possession and Assault.

```
worstAffected = crimes %>%
  #filter the catagory to only be the top three most common types of crime
  filter(Category == "Burglary"|Category == "Drug Possession"| Category == "Assault")%>%
  group_by(District_ID)%>%arrange(desc(District_ID))%>%summarise(Count = n())
#join the data with the places dataframe to gain the population collumn
b = worstAffected%>% inner_join(places) %>%mutate(Proportion = Count/Population)#define the proportion
```

```
## Joining, by = "District_ID"
```

```
#change the data type of the District ID to match the shape file
b = b%>% mutate(District_ID = as.character(District_ID))
#set the col equal to eachother so they have something in common to inner join
b$District_ID = u_shape$NAME_1
#define a shape file to be the last shapefule but now with the data we want
```

```
u_Shape2 = inner_join( u_shape, b, by=c("NAME_1"="District_ID") )
#plot the result
ggplot( u_Shape2, aes(fill=Proportion) ) + geom_sf() + theme_bw() +
  labs(x = "Latitude",y = "Longitude",
       title = "Districts Worst affected by most common Crime per 1000 Population")
```



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

```
#getting the data of the district 44
b2 = worstAffected%>% inner_join(places) # same as before for part a)
```

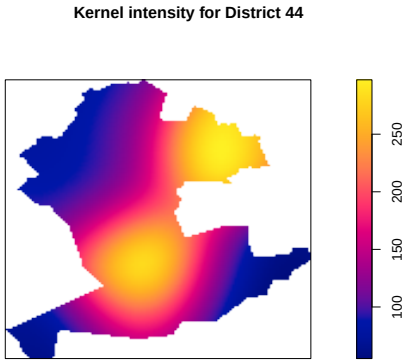
```
## Joining, by = "District_ID"
```

```
b244 = b2 %>% filter(District_ID == 44) # select the district 44 data
crimes44 = crimes%>% filter(District_ID == 44) # same as line above for crimes data frame
b244plus = b244 %>% inner_join(crimes44) # combine the two data frames
```

```
## Joining, by = "District_ID"
```

```
b244plusDrugsOnly = b244plus %>% filter(Category == "Drug Possession")
# select only the drug possession crimes
b244plusDrugsOnly = b244plus %>% filter(Arrest == "No")
# select only the crimes where the criminals got away
b244plusDrugsOnly = b244plusDrugsOnly%>%
  mutate(District_ID = as.character(District_ID))
# convert the district id to character
b244plusDrugsOnly = b244plusDrugsOnly%>% select(Longitude,Latitude,District_ID)
b244plusDrugsOnlyNoNAN = na.omit(b244plusDrugsOnly)
#SELECTING ONLY THE DISTRICT 44 SHAPE
ushape44 = u_Shape2%>% filter(NAME_1 == "District 44")
# make the districtID column the same
b244plusDrugsOnlyNoNAN$District_ID = ushape44$NAME_1
```

```
# merge the shapefile and the data
ShapeAndDataB = inner_join(ushape44,b244plusDrugsOnlyNoNAN,
                           by = c("NAME_1"="District_ID"))
#handle the shapefile and data variabel to use Intensity weighting
us44_sp <- as( ushape44, "Spatial" )
us44_sp <- slot( us44_sp, "polygons" )
FourtyFour_win <- lapply( us44_sp, function(z) { SpatialPolygons(list(z)) } )
FourtyFour_win <- lapply( FourtyFour_win, as.owin )[[1]]
crimes_ppp <- ppp( x=ShapeAndDataB$Longitude, y=ShapeAndDataB$Latitude,
                       window = FourtyFour_win )
#plot result with edge correction
plot( density.ppp(crimes_ppp,edge = TRUE),
      main="Kernel intensity for District 44" )
```



after reviewing the plot I recommend the police department targeting the North-East region of District 44 as this is the area where the most amount of Drug Possession crimes are taking place. Should the Police department decide they needed another region within District 44, just South of Central area of district 44 is the second most amount of drug possession crimes are taking place, hence I would suggest there as a second best position to target their efforts.

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

```
#dataframe only containing the burgalaries
Burgal = crimes %>% filter(Category == "Burglary")
#description data without NA
descNoNans = na.omit(Burgal$Description)
#list of variables to search for
x = c("young single","young couple","middle-aged single","middle-aged couple",
      "elderly single","elderly couple")
df = data.frame(line = 1:length(descNoNans),text =descNoNans)
df_count = df %>% unnest_tokens( word,text, token = "ngrams", n = 2) %>%
  count(word,sort = TRUE)
```

```
df_count_part1 = df_count %>% filter(word %in% x) %>% arrange(desc(n))
sufferingMostFromBurg = df_count_part1 %>% slice_head(n = 1)
print(sufferingMostFromBurg)
```

```
##             word    n
## 1 elderly single 4410
```

We see that the demographic most at risk from burglary is "elderly single"

```
#SECOND PART
x2 = c("two criminals", "three criminals","more than")
# apply the criteria
df_count_part2 = df_count %>% filter(word %in% x2) %>% arrange(desc(n))
#sum the values as we are interested in more than 2
TotalBurglariesMoreThanTwo = sum(df_count_part2$n)
# total burglaries is the length of the column
TotalBurglaries = length(Burgal$Category)
#define the proportion
proportionOfBurgalariesMoreThanTwo = TotalBurglariesMoreThanTwo/TotalBurglaries
print(proportionOfBurgalariesMoreThanTwo)# print result
```

```
## [1] 0.5777872
```

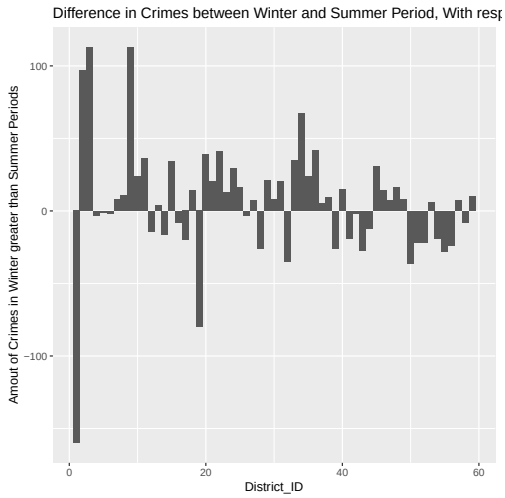
We see that 0.5777872 of all burglaries with description are committed by more than two criminals at the scene of the crime.

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**]

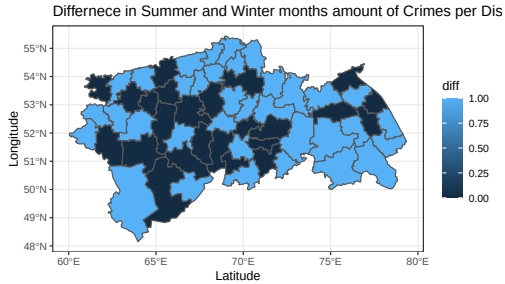
d)QUESTION: ANALISE WETHER THE WARMER MONTHS OF SPRING AND SUMMER HAVE ANY EFFECT ON THE CRIME RATE OF UTOPIA. FURTHER SHOW WHICH DISTRICTS ARE MORE SUSEPTIBLE TO CRIME IN THE DIFFERENT TIMES OF YEAR IN A PLOT.

```
#define new dataframe to be the old crime dataframe but with a binary column, 1 is it is Summer/Spring
SummerOrSpringDataFrame = crimes%>%
  mutate(SummerOrSpring = as.numeric(Month>= 3 & Month <= 8))
# get the names of the district_ID's form crimes
namedf = crimes %>% group_by(District_ID) %>%
  summarise(count = n()) %>% select(- count)
u_shape2 = u_shape
#alter the names in the shapefile to match the dataframe for later
u_shape2$NAME_1 = namedf$District_ID
#define a dataframe to count the amount of crimes per district and per binary value
d = SummerOrSpringDataFrame %>% group_by(SummerOrSpring,District_ID)%>%
  summarise(count = n())
#merge the shapefile to dataframe
ShapeAndDataD = inner_join(u_shape2,d,by =c("NAME_1" = "District_ID"))
#pivot wider the dataframe  for another column, one for eacH value of Summmer
df = d%>% pivot_wider(names_from = SummerOrSpring, values_from = count )
#define the difference in crimes between Autumn/Winter months and Summer months
df= df%>% mutate(diff = `0` - `1`)
```

```r
# minus values imply more crimes in Summer/Spring than Winter/Autumn.
#plot result
ggplot(df, aes(x =District_ID, y = diff)) + geom_col() +
  labs(x = "District_ID",
       y = "Amout of Crimes in Winter greater than Summer Periods",
       title = "Difference in Crimes between Winter and Summer Period, With respect to distrxict")
```




```r
#define the cumulative sum of the differences between Winter/Autumn months and
#Summer/Spring months as a column
df2 = df %>% mutate( absoluteDiff = cumsum( diff ) )
# defeine a binary operator to show the districts that have more crimes in summer or winter period
df2  = df2 %>% mutate(diff = as.numeric(diff>0))
#merge the shape file and the data
ShapeAndDataD = inner_join(u_shape2,df2,by =c("NAME_1" = "District_ID"))
#plot the shapefile and color according the the winter crime rate vs summer crime rate
ggplot( ShapeAndDataD, aes(fill=diff) ) + geom_sf() + theme_bw() +
  labs(x = "Latitude", y = "Longitude",
       title = "Differnece in Summer and Winter months amount of Crimes per District")
```



```r
#Number Of Districts That Suffer From Winter Period Increased Crimes
n = sum(df2$diff)
print(c("Number Of Districts That Suffer From Winter Period Increased Crimes =",n))
```

```
## [1] "Number Of Districts That Suffer From Winter Period Increased Crimes ="
## [2] "35"
```

```r
# Difference in number of Crimes from Winter Period to Summer Period over all Districts
cumsum = df2$absoluteDiff[length(df2$absoluteDiff)]
print(c("Difference in number of Crimes from Winter Period to Summer Period over all Districts ="
        ,cumsum))
```

```
## [1] "Difference in number of Crimes from Winter Period to Summer Period over all Districts ="
## [2] "351"
```

From the plot above we see that Utopia as a country suffers from more crimes over the winter months as the cumulative difference in Summer and Winter months amount of Crimes per District is +351, implying there are 351 MORE crimes in the Autumn/Winter period than in the Summer/Spring period over all Districts. On a District by District case, we see that majority of districts, 35 out of 59, suffer from more crimes over the winter period than summer period. As such i would reckoned the Utopia police department to focus it's efforts more on the districts highlighted in light blue in the second plot over winter months and districts colored in deep blue over the summer months.

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]** REPORT: In conclusion from our report, we have shown that Utopia is at most risk from Burglary crimes and second to that is Drug Possession crimes. We first discuss the former. We have shown that the demographic most at risk from burglary is the Elderly Single demographic, and that just over half of all burglary crimes are committed by more than two criminals. This can be of great help for the Police department as they may wish to raise awareness to the elderly demographic to help reduce the burglary crime rate. In relation to the Drug Possession crimes, we have shown that District 44- a district that is notorious for drug possession- should have efforts targeted on the central region and North-East region in particular. This is due to where District 44 contains the majority of Drug Possession crimes that have taken place over the years 2015- 2022. Reviewing Utopia as a whole, we have identified that there is an increase in crimes committed in Utopia over the Autumn and Winter months, September till February, compared to the warmer Spring and Summer months, March till August. Here in lies another area for the Police department to become more vigilant and focus their efforts. Should the Police department decide to review the details of this report, they may so choose to increase their vigilance over the Spring and Summer Months for certain districts of utopia and upon the Autumn and Winter months "swap" their increased efforts for the districts found to have more crime over the colder Months, all in effort to reduce focus police department effort when it is needed most.