WORK SAMPLE

By

Suhasini Singh (1811068)

Department of Commerce

CHRIST (Deemed to be University)

Bangalore

2019-2020

Business Stats (COH233)

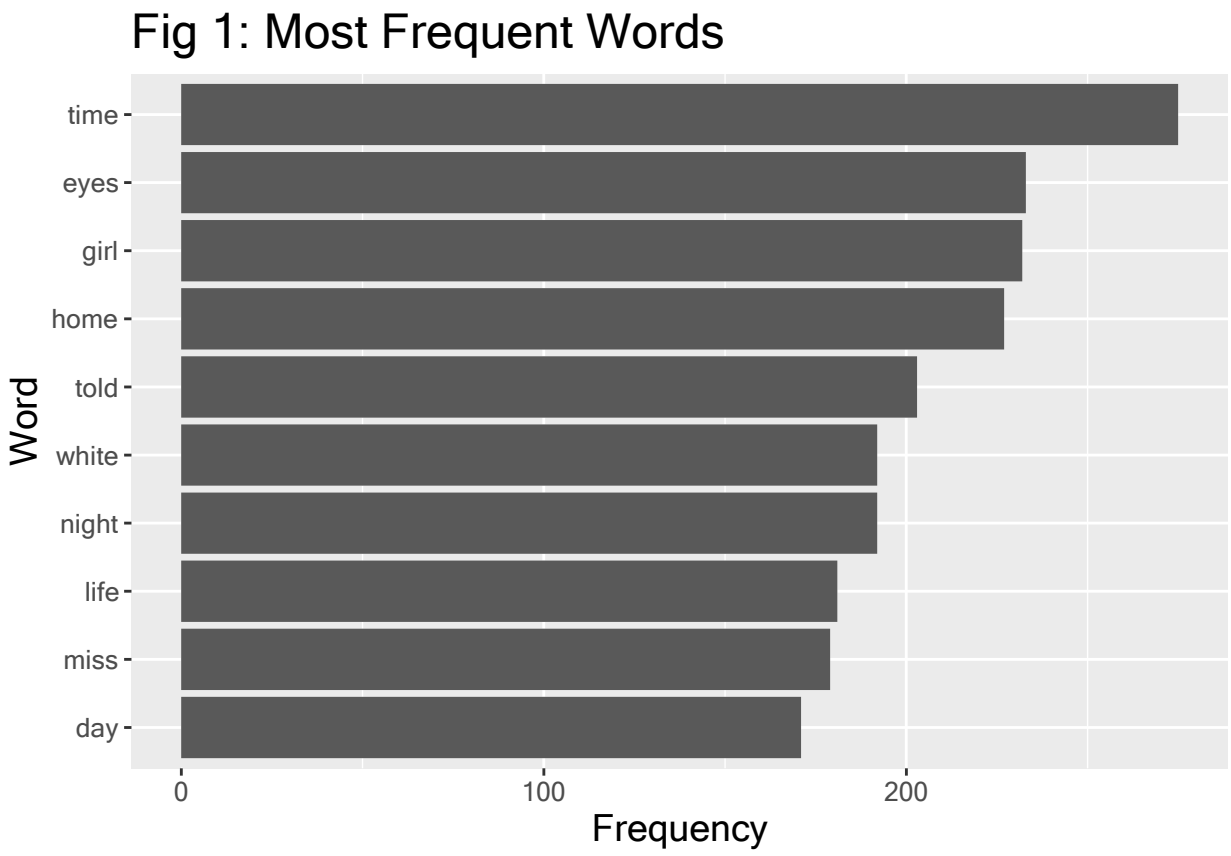
SUMMARY
The document outlines a series of statistical analysis tasks using R programming, focusing on three main areas: literary analysis of specific words in novels, spatial analysis of environmental data in Manchester, and crime data analysis in a fictional country, Utopia. Each task involves data manipulation, visualization, and interpretation, highlighting the application of R in diverse contexts like literature, environmental science, and criminology. The exercises aim to demonstrate the practical use of statistical techniques and data analysis using R.

Task 1 [9 marks]

We want to analyze the books "Anne of Green Gables" and "Blue Castle" by Lucy Maud Montgomery.The two books are provided in the files "Anne of Green Gables.txt" and "Blue Castle.txt". a) *Visualize the frequency of the 10 most frequent words that satisfy the following three criteria: (1) The word occurs at least five times in each book, (2) The word is not a stop word according to the usual stop list considered in the lectures, (3) The word is not "I'm", "don't", "it's", "didn't", "I've" or "I'll".* **[6 marks]**

```
#Loading the text of the books
text_Anne = readLines( "Anne of Green Gables
UTF8.txt" )text_Blue = readLines( "Blue
Castle UTF8.txt")
#Creating the data frame of the booksquote_Anne = data.frame (text_Anne) quote_Blue =
data.frame (text_Blue) Anne_tidy = quote_Anne %>% unnest_tokens (word,text_Anne)
Blue_tidy = quote_Blue %>% unnest_tokens (word,text_Blue) Anne_Count = Anne_tidy %>%
count (word,sort=TRUE) %>%
mutate ("term frequency"= n/sum(n),rank =
row_number()) Blue_Count = Blue_tidy %>%
count (word,sort=TRUE) %>%
mutate ("term frequency"= n/sum(n),rank = row_number())
#Join books
Books_Count = full_join (Anne_Count,Blue_Count,by = "word")
#Replace na entries by 0
Books_Count [is.na(Books_Count)] = 0
#Get total freq of words
Books_Count$n = Books_Count$n.x + Books_Count$n.y
#Removing stop words
Books_Count = Books_Count %>% anti_join (stop_words)
#Removing the words specified
Books_Count = Books_Count[Books_Count$word
!= "it' s",]#Word should occur more than 5 times
in each book and#Visualising top 10 frequent words
Books_Count %>% filter (Books_Count$n.x >= 5 & Books_Count$n.y >= 5) %>% select
(c("word","n")) %>% arrange (desc(n)) %>% slice_head (n=10) %>%  mutate (word =
reorder(word,n)) %>%
rename (c("Word"="word","Frequency"="n")) %>% ggplot (aes( x=Frequency, y=Word )) +
```

```
geom_col() + labs (title = "Fig 1: Most Frequent Words", x="Frequency", y="Word") +
theme (title = element_text(size = 15), axis.title=element_text(size = 14),
axis.text=element_text(size = 10))
```



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]**

```
#For Anne  textbook
#Add title of book in front of words
title = replicate (nrow(Anne_Count),"Anne of Green Gables")
Anne_Count$Title = title
#For Rebecca textbook, same process
text_Rebecca = readLines( "Rebecca of Sunnybrook Farm UTF8.txt" )
quote_Rebecca = data.frame (text_Rebecca)
Rebecca_tidy = quote_Rebecca %>% unnest_tokens (word,text_Rebecca)
Rebecca_Count = Rebecca_tidy %>%
count (word,sort=TRUE) %>%
mutate ("term frequency" = n/sum(n),rank=row_number())
title_1 = replicate (nrow(Rebecca_Count),"Rebecca of Sunnybrook Farm")
Rebecca_Count$Title = title_1
#Combine books
Books = full_join (Anne_Count,Rebecca_Count, by = c( "word", "Title"))
```

```
Books[is.na(Books)] = 0
Books$n = Books$n.x + Books$n.y
Books$"termfrequency" = Books$"term frequency.x" + Books$"term frequency.y"
#For Anne
Books_1 = Books %>% bind_tf_idf (word,Title,n) %>%
select (c("word","Title","n","tf_idf")) %>% filter (Title == "Anne of Green Gables")  %>%
arrange (desc(tf_idf)) %>% rename (c("Word"="word","Frequency"="n"))
slice_head(Books_1, n=2)
```

```
##      Word                Title Frequency      tf_idf
## 1    anne Anne of Green Gables      1102 0.007396183
## 2 marilla Anne of Green Gables       795 0.005335722
```

```
#For Rebecca
Books_2 = Books %>% bind_tf_idf(word,Title,n) %>%
select (c("word","Title","n","tf_idf")) %>% filter (Title == "Rebecca of Sunnybrook Farm")  %>%
arrange (desc(tf_idf)) %>% rename (c("Word"="word","Frequency"="n"))
slice_head(Books_2, n=5)
```

```
##        Word                       Title Frequency       tf_idf
## 1   rebecca Rebecca of Sunnybrook Farm       572 0.0053565374
## 2     don't Rebecca of Sunnybrook Farm       147 0.0013765927
## 3      it's Rebecca of Sunnybrook Farm       105 0.0009832805
## 4 rebecca's Rebecca of Sunnybrook Farm       105 0.0009832805
## 5      cobb Rebecca of Sunnybrook Farm        90 0.0008428118
```

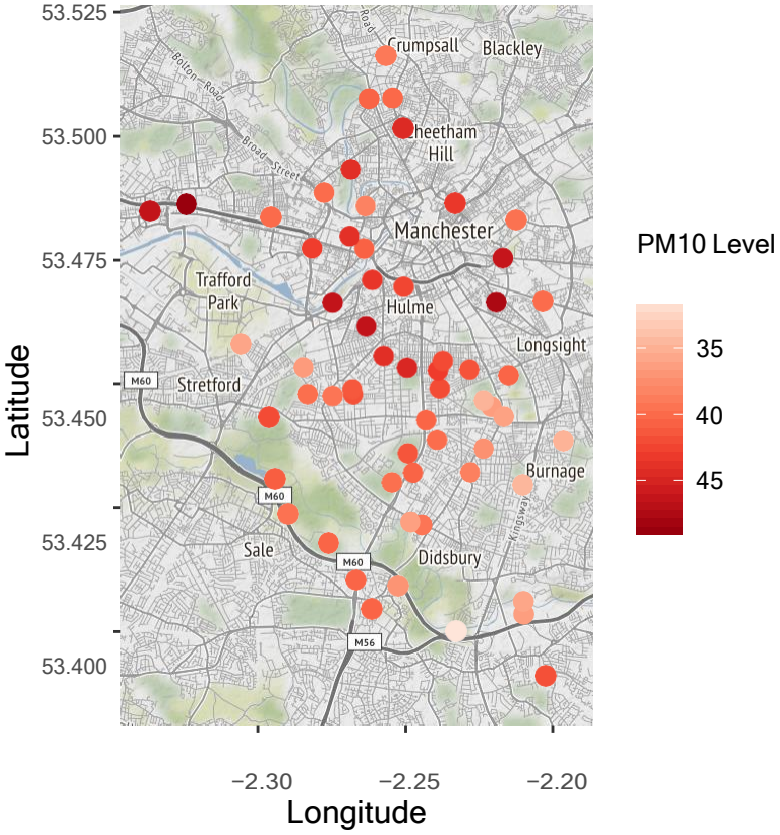
For the textbook **Anne of Green Gables**, the words with the highest term frequency-inverse distance frequency are **Anne and Marilla**. While for **Rebecca of Sunnybrook Farm**, the required words are **Rebecca and Cobb**. We ignore the other words as *don't* and *it's* are stop words; while *Rebecca's* is the same thing as *Rebecca*.

Task 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]**

```
#Reading the CSV file
Manchester = read.csv("Manchester.csv")
#Plotting
PlotDim = c(left = min(Manchester$Lon) - 0.01,right = max(Manchester$Lon) + 0.01,
top = max(Manchester$Lat) + 0.01, bottom = min(Manchester$Lat) - 0.01)
ggmap (get_stamenmap (PlotDim,maptype = "terrain",zoom = 12)) +
geom_point (data=Manchester,aes(x = Lon,y = Lat,color = Level),size = 3) +
scale_color_distiller (palette = "Reds",trans = "reverse") +
labs (title = "Fig 2: PM10 Levels Recorded by Stations",
x = "Longitude",y = "Latitude",color = "PM10 Level") +
theme (title = element_text(size = 10), axis.title=element_text(size = 12),
axis.text=element_text(size = 8))
```

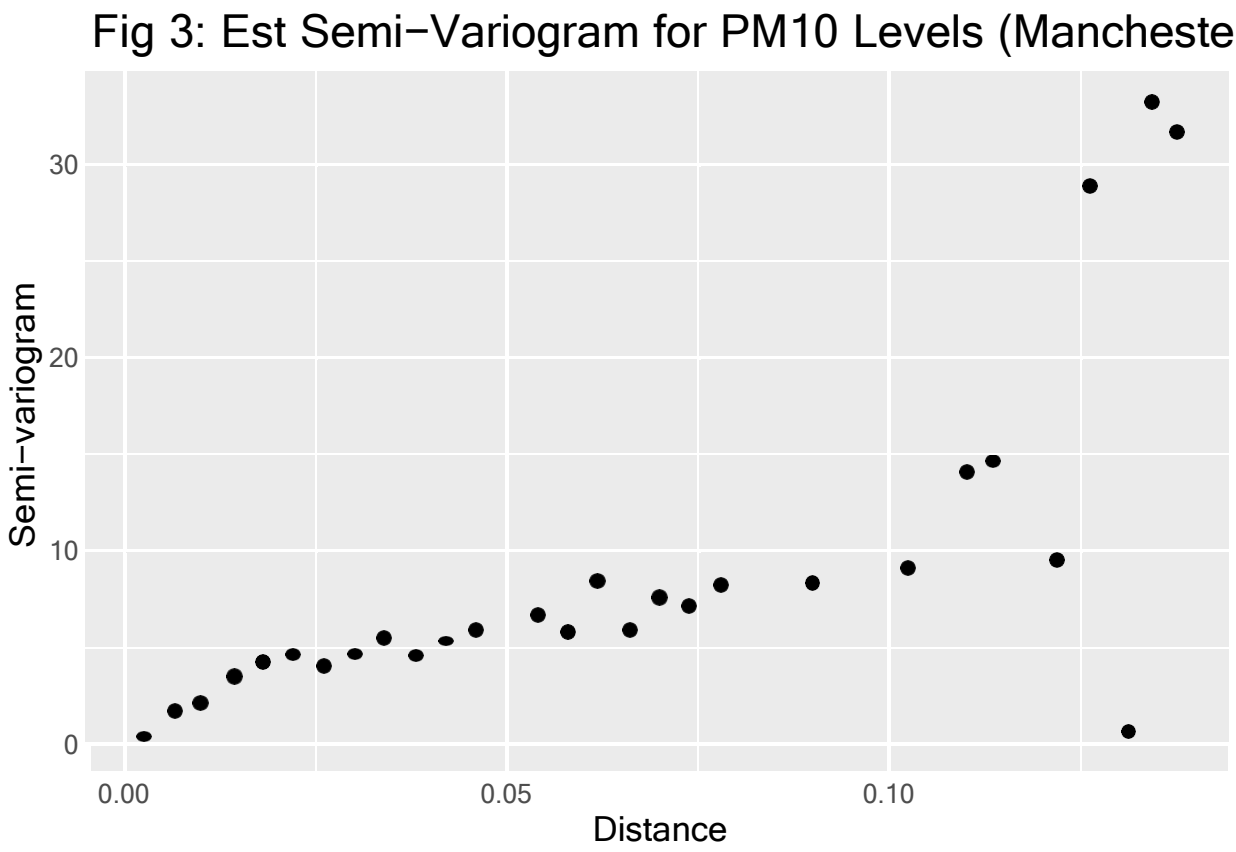
Fig 2: PM10 Levels Recorded by Stations



We observe that the station which recorded the least PM10 level is located south of Didsbury and lies in greener area. The stations with higher measurements are located near Manchester, where there is greater density of population and roads. This leads to more traffic and homes, and hence higher pollution levels due to vehicular emissions, cooking and smoking. The two stations towards east and northwards of Trafford Park, are near the city airport of Barton. Hence, higher levels of PM10 were observed. The stations near the M56 and M60 highways also record higher PM10 levels due to increased vehicular emissions.

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

```
Man_gamma = drop_na (Manchester,Level)
coordinates (Man_gamma)  = ~ Lon + Lat
gamma_hat = variogram (Level~1,Man_gamma,cutoff = 0.14,width = 0.004)
ggplot (gamma_hat,aes (x = dist,y = gamma/2)) + geom_point (size=2) +
labs (title = "Fig 3: Est Semi-Variogram for PM10 Levels (Manchester)",
x = "Distance",y = "Semi-variogram") + theme (title = element_text(size = 14),
axis.title=element_text(size = 13), axis.text=element_text(size = 10))
```



We infer that as distance increases, spatial dependence decreases gradually. The stations that are spatially close exhibit spatial dependence, but at a distance of about 0.025 degrees in latitude or longitude, the PM10 level measurements are close to independent. This seems logical as the stations near to each other would not experience vastly different air qualities. While the stations that lie far away can have a different environment, leading to different pollution levels. The range of the variogram is 0.025. Interestingly, stations that are more than 0.13 degrees far also exhibit spatial dependence. This indicates that our analysis is not ideal. The PM10 levels may depend on other factors, such as population. So, we should incorporate those factorsin our model, and then explore dependence on spatial residuals.

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]**

```
#Defining the 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])}
coord = cbind (Manchester$Lon,Manchester$Lat)
#The 2 locations
s_star_1 = c(-2.275,53.354)
s_star_2 = c(-2.250,53.471)
#Predicting the PM10 level
IDW (Manchester$Level,coord,s_star_1,1.5 )
```

```
## [1] 40.32368
```

```
IDW (Manchester$Level,coord,s_star_2,1.5 )
```

```
## [1] 42.46621
```

Hence, the predicted PM10 level at **(-2.275,53.354)** is 40.32 $\mu g/m^3$, and 42.46 $\mu g/m^3$ at **(-2.250,53.471)**. Our estimates are not completely reliable because majority of the recorded stations lie far away from the stations for which we have predicted the PM10 levels. Because our weighing system is inversely proportional to the distance; our estimates are not accurate due to lack of nearby stations. Also, as mentioned previously, basing our predictions completely on distance is wrong, as over large distances, population density changes as well as the natural environment. We need to include these extra variables to get a more relaible and accurate prediction.

Task 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".

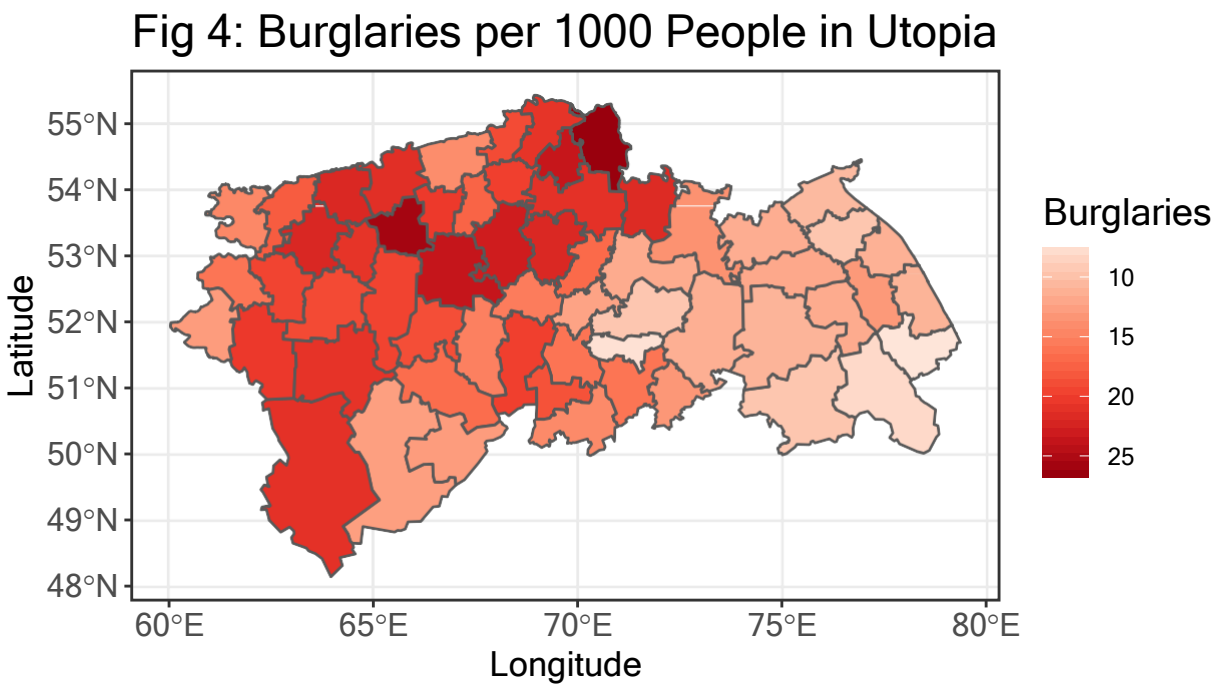
a) *What are the three most common crimes in Utopia? Create a map that visualizes the districts worst affected by the most common crime in terms of number of incidents per 1,000 population.* **[5 marks]**

```
#Read the data files
Crime = read.csv("UtopiaCrimes.csv")
Pop = read.csv("UtopiaPopulation.csv")
Crime %>% count (Category,sort=TRUE) %>% slice_head (n=3) %>% rename ("Crime" = "Category","Count" = n)
```

```
##             Crime Count
## 1        Burglary 16513
## 2 Drug Possession 10551
## 3         Assault 10169
```

Hence, the three most common crimes are **burglary, drug possession and assault**.

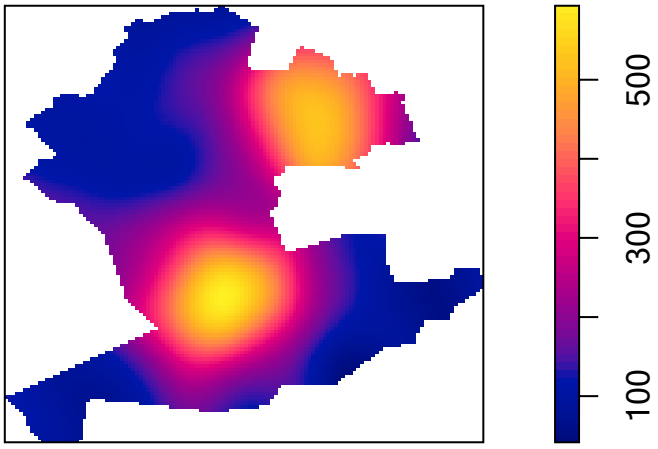
```
#Filtering out instances of burglary, grouping by dist
Common_Crime = Crime %>% filter (Category=="Burglary") %>%
group_by (District_ID) %>% summarise ("Number_of_Burglaries" = n())
#Join with population data
Data = full_join (Common_Crime,Pop,by = "District_ID")
#Crime per 1000
Data$Burg_per_1000 = (Data$Number_of_Burglaries/Data$Population)*1000
#Mapping
Utopia = read_sf ("UtopiaShapefile.shp")
ggplot (data = Utopia,aes (fill = Data$Burg_per_1000)) + geom_sf () + theme_bw () +
scale_fill_distiller ( palette = "Reds", trans = "reverse" ) +
theme (title = element_text(size = 14),axis.title=element_text(size = 13),
axis.text=element_text(size = 12) ) + labs (title = "Fig 4: Burglaries per 1000 People in Utopia",
x = "Longitude",y = "Latitude",fill = "Burglaries" )
```



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]**

```
#Filtering out drug possessioj instances in dist 44
Crime_44 = Crime %>% filter (District_ID == 44,Category == "Drug Possession")
#Mapping
Dist_44 = filter (Utopia,NAME_1=="District 44")
Dist_44_sp = as( Dist_44, "Spatial" )
Dist_44_sp = slot( Dist_44_sp, "polygons" )
Dist_44_win = lapply( Dist_44_sp, function(z) { SpatialPolygons(list(z)) } )
Dist_44_win = lapply( Dist_44_win, as.owin )[[1]]
Crime_44_ppp = ppp (x = Crime_44$Longitude,y = Crime_44$Latitude,window = Dist_44_win)
par (mfrow=c(1,2),mai=c(0.01,0.01,0.5,0.01))
lambdaC = density.ppp (Crime_44_ppp,edge = TRUE,sigma = 0.15)
plot (lambdaC,main = "Fig 5: Drug Possession - District 44")
```

Fig 5: Drug Possession − District 44



We observe from the smoothed kernel intensity estimates that more crimes of drug possession occur towards the northeast corner of the district 44 and and the middle southwestern part. Hence, the police should conduct the raids in these areas.

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]**

```
#Filter out burglaries
#Split the description into diff rows using ; as separator
#Count instances for each description
Crime_des = Crime %>% filter (Category=="Burglary") %>%
separate_rows (Description,sep=";") %>% group_by (Description) %>%
summarise ("No_of_Instances" = n())
#Get all kinds of age groups
#Then get most vulnerable group
Crime_des %>% filter (Description == " elderly single "|
Description == " elderly couple " |
Description == " middle-aged couple "|
Description == " young single "|
Description == " young couple " |
Description == " middle-aged single " ) %>%
arrange (desc(No_of_Instances)) %>% slice_head()
```

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

Hence, **elderly single** citizens have the highest risk of being a victim of burglaries.

```
#Filter out instances of 1/2 criminals
Crime_des_crim = Crime_des %>% filter (Description == "One criminal "|
Description == "Two criminals ")
#Required prop
Required_Prop = 1 - sum(Crime_des_crim$No_of_Instances) / nrow(Crime %>% filter(Category=="Burglary"))
print(Required_Prop)
```

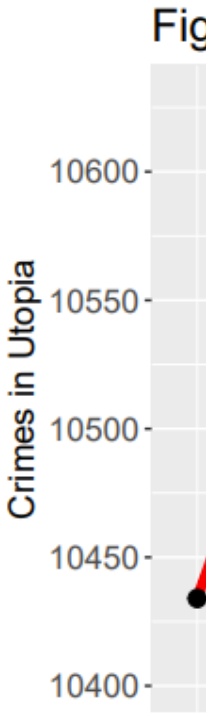
```
## [1] 0.2440501
```

Hence, 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]**

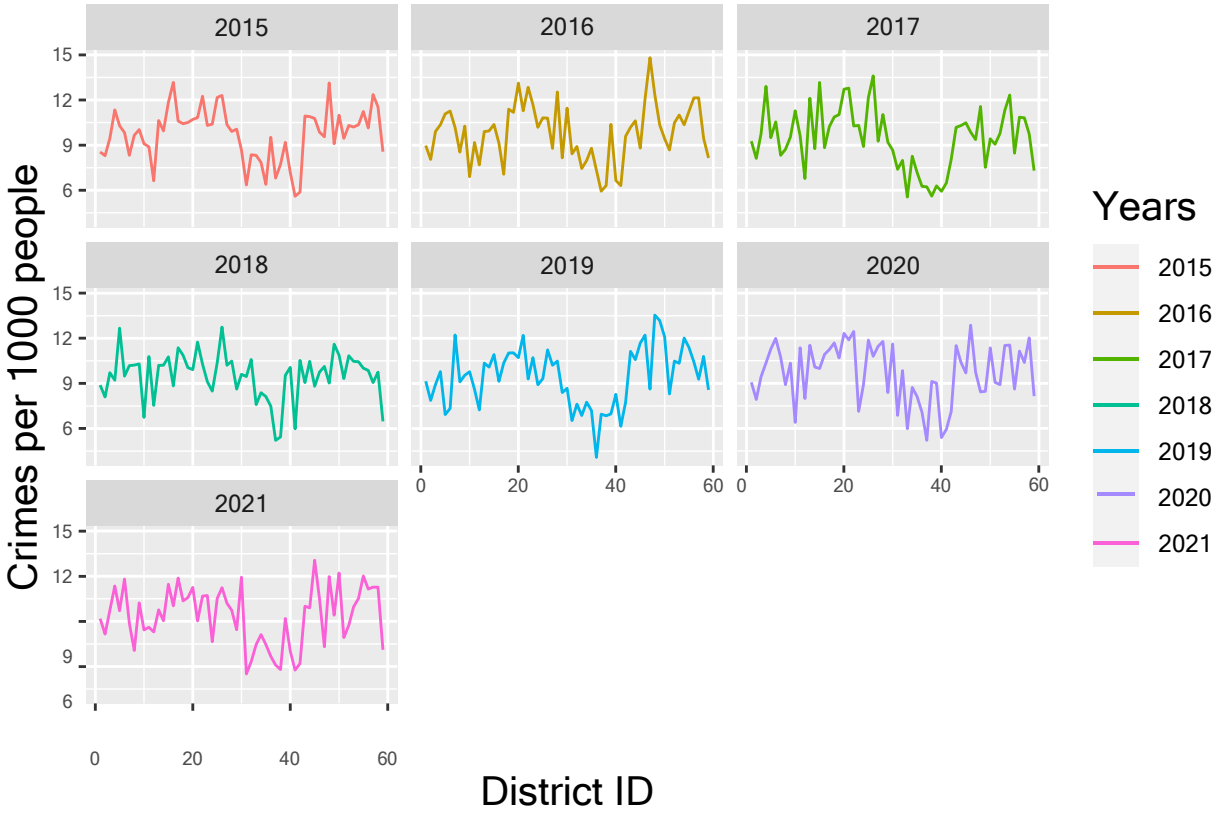
MY QUESTION: *Analyse the trends in crime rates in Utopia over the past 7 years.*

```
#Plotting crime data by years
Crime %>% group_by (Year) %>% summarise ("Crimes" = n()) %>%
ggplot (aes(x = Year,y = Crimes)) + geom_line (color = "Red",size = 2) +
geom_point(size = 3) + labs(x = "Year", y = "Crimes in Utopia",
title = "Fig 6: Crimes in Utopia (2015-2021)") + theme (title = element_text(size = 15),
axis.title=element_text(size = 14),axis.text=element_text(size = 13))
```



```r
#Filtering out entries by years, grouping by districts
C_15 = Crime %>% filter (Year == 2015) %>% select (c(District_ID,Year)) %>%
group_by (District_ID) %>% summarise ("noc_15" = n())
C_16 = Crime %>% filter (Year == 2016) %>% select (c(District_ID,Year)) %>%
group_by (District_ID) %>% summarise ("noc_16" = n())
C_17 = Crime %>% filter (Year == 2017) %>% select (c(District_ID,Year)) %>%
group_by (District_ID) %>% summarise ("noc_17" = n())
C_18 = Crime %>% filter (Year == 2018) %>% select (c(District_ID,Year)) %>%
group_by (District_ID) %>% summarise ("noc_18" = n())
C_19 = Crime %>% filter (Year == 2019) %>% select (c(District_ID,Year)) %>%
group_by (District_ID) %>% summarise ("noc_19" = n())
C_20 = Crime %>% filter (Year == 2020) %>% select (c(District_ID,Year)) %>%
group_by (District_ID) %>% summarise ("noc_20" = n())
C_21 = Crime %>% filter (Year == 2021) %>% select (c(District_ID,Year)) %>%
group_by (District_ID) %>% summarise ("noc_21" = n())
#Joining all these data sets
Crime_yrs = list (C_15,C_16,C_17,C_18,C_19,C_20,C_21) %>% reduce (full_join,by = "District_ID")
#Joining with pop dataset
Crime_yrs = full_join (Crime_yrs,Pop,by="District_ID")
#Get crimes per 1000 people
Crime_yrs$"2015" = (Crime_yrs$noc_15/Crime_yrs$Population)*1000
Crime_yrs$"2016" = (Crime_yrs$noc_16/Crime_yrs$Population)*1000
Crime_yrs$"2017" = (Crime_yrs$noc_17/Crime_yrs$Population)*1000
Crime_yrs$"2018" = (Crime_yrs$noc_18/Crime_yrs$Population)*1000
Crime_yrs$"2019" = (Crime_yrs$noc_19/Crime_yrs$Population)*1000
Crime_yrs$"2020" = (Crime_yrs$noc_20/Crime_yrs$Population)*1000
Crime_yrs$"2021" = (Crime_yrs$noc_21/Crime_yrs$Population)*1000
#Get required columns
Crime_yrs_ = Crime_yrs %>% select (c(District_ID,"2015","2016","2017","2018","2019","2020","2021"))
#Convert to tidy format and plot
Crime_yrs_ %>% pivot_longer (cols = c("2015","2016","2017","2018","2019","2020","2021"),
names_to = "Years") %>% ggplot (aes(x = District_ID, y = value)) +
geom_line (aes(color = Years)) + facet_wrap(~Years, ncol = 3) +
labs(x = "District ID",y = "Crimes per 1000 people",
title = "Fig 7: Crime Numbers Across Districts (2015-2021)") +
theme (title = element_text(size = 15),axis.title=element_text(size = 15),
axis.text=element_text(size = 7))
```

Fig 7: Crime Numbers Across Districts (2015–2021)



Over the past seven years, the number of crimes in Utopia range from 10,400 to 10,650. The lowest number of crimes were recorded in 2019, while the highest in 2020. Crimes increased from 2015 to 2017, fell in 2018 and 2019 before increasing again in 2020. Notice that there has been a marginal increase in number of crimes from 2015 to 2021. If we analyse the crime rates in Utopia across all districts, we see that none of the districts witnessed more than 15 crimes per 1000 people during any year. Interestingly, each year a different district had the highest crime rate. For the years 2015-17 we infer that districts 10-30 and 40-50 were performing comparatively poorly. In the next three years, districts 30-40 performed exceptionally well in bringing down crime rates. However, there was a gradual increase in districts 40-60. Finally, in 2021, the crime rates were almost similar to 2015, although lower than 2020.