Final\_Project

## R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button, a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(ggplot2)  
library(gridExtra)

##   
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':  
##   
## combine

library(tidyverse)

## -- Attaching packages ------------------------------------------ tidyverse 1.2.1 --

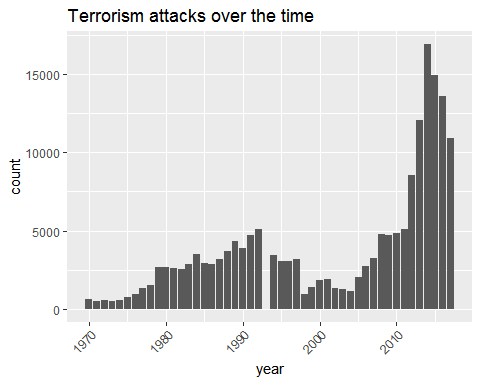
## v tibble 1.4.2 v purrr 0.2.5  
## v tidyr 0.8.2 v stringr 1.3.1  
## v readr 1.1.1 v forcats 0.3.0

## -- Conflicts --------------------------------------------- tidyverse\_conflicts() --  
## x gridExtra::combine() masks dplyr::combine()  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

MyData <- read.csv(file="D:/Master's/1st sem/Statistical and predictive modelling/globalterrorismdb\_0718dist.csv", header=TRUE, sep=",")  
glimpse(MyData)

## Observations: 181,691  
## Variables: 135  
##I have chosen the Global terrorism dataset for my final project. It is the terrorism data for all the countries from 1970 - 2016 with more than 170,000 cases. These data were collected from different country populations, and they were acquired and stored in different modalities. The database is maintained and updated periodically by researchers at the National Consortium for the Study of Terrorism and Responses to Terrorism (START) at the university of Maryland. It should be noted that the terrorism data has 1 year missing, we decided to just ignore this as it doesn't impact the overall analysis. The incidents of terrorism from 1993 are not present in the GTD because they were lost prior to START's compilation of the GTD from multiple data collection efforts. Several efforts were made to recollect these incidents from original news sources. The number of 1993 cases for which sources were identified is only 15% of estimated attacks. As a consequence, we exclude all 1993 attacks from the GTD data to prevent users from misinterpreting the low frequency in 1993 as an actual count.  
##Research Goals  
##The main objective of this project is to generate models using various analytical tools, to predict the possible terrorist attack of 2017 from the given set of attacks using a generated model. This will be helpful in taking the necessary precautions before any Incident can occur by checking on it. Therefore, it would be useful if a model can analyze patterns in these data and predict the attacks. Effective clustering methods would aid in categorizing data into different attack types. A properly generated regression model can predict the type of attack present in any region or country. This can be helpful in treating or identifying attack.  
##Attributes  
##The Global terrorism dataset has 182k rows, and 48 columns. Each row represents a terrorist attack. The first column of the dataset includes the eventid which is useful for identifying the type of attack present in the data. This column will be used for supervised learning as the response or target variable for the analyses.  
##Analysis  
##For this project, both supervised and unsupervised learning methods will be employed on the dataset. For supervised learning, various methods such as linear regression, multiple linear regression, ridge regression, lasso regression, neural networks etc., will be utilized to generate a model that can predict the response variable, which is the event id. Associated accuracies will be recorded and the best model will be presented. For unsupervised learning, various methods for clustering analysis and association will be employed. After generating optimal clusters, they will be cross validated with the event id. Useful graphics representing different regression and clustering methods will be displayed to get a better understanding about the analysis.  
  
  
  
PopulationData <- read.csv(file="D:/Master's/1st sem/Statistical and predictive modelling/UNpopfile.csv", header=TRUE, sep=",")  
  
pop\_1 <- PopulationData %>%  
 select(-MidPeriod, -PopMale, -PopFemale, -VarID)  
  
pop\_1 <- pop\_1 %>%   
 filter(Time > 1969 & Variant == 'Medium' & Time < 2017) %>%  
 select(-Variant, -LocID)  
  
futurepopulation <- PopulationData %>%  
 filter(Time >2016 & Variant == 'Medium') %>%  
 select(-Variant, -LocID, -MidPeriod, -PopMale, -PopFemale, -VarID)  
df <- MyData %>%  
 select(iyear, imonth, iday, country\_txt, region\_txt, city, latitude, longitude, summary, multiple, attacktype1\_txt, targtype1\_txt, targsubtype1\_txt, gname, weaptype1\_txt, nkill, nwound, nkillter)   
  
df <- df %>%  
 rename(year = iyear, month = imonth, day = iday, country = country\_txt, region = region\_txt, multiple\_attack = multiple, attacktype = attacktype1\_txt, target\_type = targtype1\_txt, target\_sub\_type = targsubtype1\_txt, group\_name = gname, weapon\_type = weaptype1\_txt)  
  
df <- df %>% mutate(decades =   
 if\_else(year< 1980, '70s',   
 if\_else(year < 1990, '80s',   
 if\_else(year < 2000, '90s',   
 if\_else( year < 2010, '2000s', '2010s')))))  
df$decades <- factor(df$decades, level=c("70s", "80s", "90s", "2000s", "2010s"))  
##Number of Terrorist Attacks   
plot1 <- print(ggplot(data=df, aes(x=year)) +  
 geom\_histogram(stat='count') +  
 theme(axis.text.x= element\_text(angle=45, hjust=1)) +  
 labs(title='Terrorism attacks over the time'))

## Warning: Ignoring unknown parameters: binwidth, bins, pad

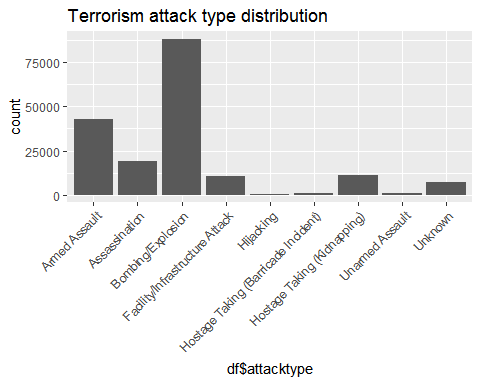


df %>%  
 summarise(number\_of\_attacks = n())

## number\_of\_attacks  
## 1 181691

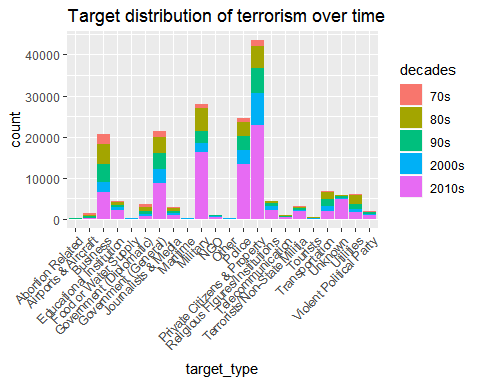
##Over 170000 attacks happening, and they seem to have inreased  
  
##Attack type Distribution  
plot2 <- print(ggplot(data = df, aes(x = df$attacktype)) +   
 theme(axis.text.x = element\_text(angle = 45, hjust = 1)) +   
 geom\_histogram(stat = "count") +  
 labs(title='Terrorism attack type distribution'))

## Warning: Ignoring unknown parameters: binwidth, bins, pad



#hist(MyData$attacktype1)  
#visual  
ggplot(data=df, aes(x=target\_type, fill=decades)) +  
 geom\_histogram(stat='count') +  
 theme(axis.text.x= element\_text(angle=45, hjust=1)) +  
 labs(title='Target distribution of terrorism over time')

## Warning: Ignoring unknown parameters: binwidth, bins, pad



## More than 80,000 Bombings, second biggest grouping is Armed assault ~40,000 attacks.  
  
##Target Distribution  
  
df %>%  
 group\_by(target\_type) %>%  
 summarise(number\_of\_attacks = n()) %>%  
 arrange(desc(number\_of\_attacks)) %>%  
 head(n=10)

## # A tibble: 10 x 2  
## target\_type number\_of\_attacks  
## <fct> <int>  
## 1 Private Citizens & Property 43511  
## 2 Military 27984  
## 3 Police 24506  
## 4 Government (General) 21283  
## 5 Business 20669  
## 6 Transportation 6799  
## 7 Utilities 6023  
## 8 Unknown 5898  
## 9 Religious Figures/Institutions 4440  
## 10 Educational Institution 4322

##the bigger target is for private citizens  
  
  
##top 10 locations of terrorist attacks by region, country and city.  
  
df %>%  
 group\_by(region) %>%  
 summarise(number\_of\_attacks = n()) %>%  
 arrange(desc(number\_of\_attacks)) %>%  
 head(n=10)

## # A tibble: 10 x 2  
## region number\_of\_attacks  
## <fct> <int>  
## 1 Middle East & North Africa 50474  
## 2 South Asia 44974  
## 3 South America 18978  
## 4 Sub-Saharan Africa 17550  
## 5 Western Europe 16639  
## 6 Southeast Asia 12485  
## 7 Central America & Caribbean 10344  
## 8 Eastern Europe 5144  
## 9 North America 3456  
## 10 East Asia 802

df %>%  
 group\_by(country) %>%  
 summarise( number\_of\_attacks = n()) %>%  
 arrange(desc(number\_of\_attacks)) %>%  
 head(n=10)

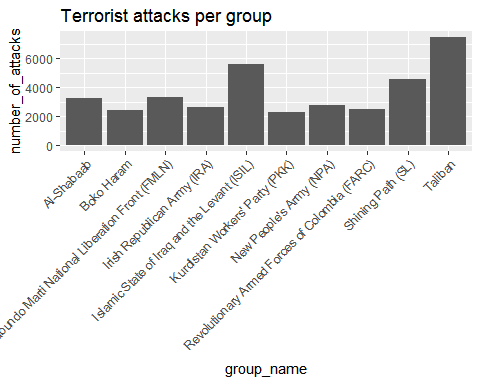
## # A tibble: 10 x 2  
## country number\_of\_attacks  
## <fct> <int>  
## 1 Iraq 24636  
## 2 Pakistan 14368  
## 3 Afghanistan 12731  
## 4 India 11960  
## 5 Colombia 8306  
## 6 Philippines 6908  
## 7 Peru 6096  
## 8 El Salvador 5320  
## 9 United Kingdom 5235  
## 10 Turkey 4292

df %>%  
 filter(city != 'Unknown') %>%  
 group\_by(city) %>%  
 summarise( number\_of\_attacks = n()) %>%  
 arrange(desc(number\_of\_attacks)) %>%  
 head(n=10)

## # A tibble: 10 x 2  
## city number\_of\_attacks  
## <fct> <int>  
## 1 Baghdad 7589  
## 2 Karachi 2652  
## 3 Lima 2359  
## 4 Mosul 2265  
## 5 Belfast 2171  
## 6 Santiago 1621  
## 7 Mogadishu 1581  
## 8 San Salvador 1558  
## 9 Istanbul 1048  
## 10 Athens 1019

##which group of terrorist is doing these attacks  
top10groups <- df %>%  
 filter(group\_name != "Unknown") %>%  
 group\_by(group\_name) %>%  
 summarise(number\_of\_attacks = n()) %>%  
 arrange(desc(number\_of\_attacks)) %>%  
 head(n=10)  
  
#visual  
ggplot(data=top10groups) +  
 stat\_summary(aes(x=group\_name, y=number\_of\_attacks), geom="bar") +  
 theme(axis.text.x= element\_text(angle=45, hjust=1)) +  
 labs(title='Terrorist attacks per group')

## No summary function supplied, defaulting to `mean\_se()

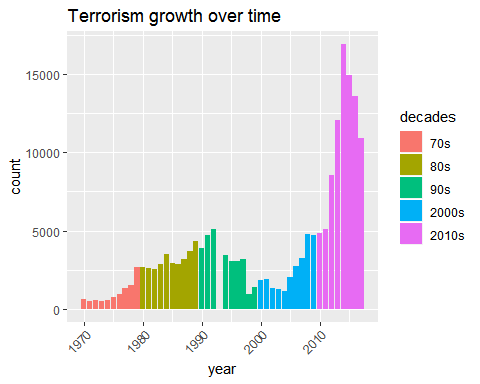


##Terorism Growth rate over time  
df %>%  
 group\_by(decades) %>%  
 summarise(number\_of\_attacks = n()) %>%  
 arrange(desc(number\_of\_attacks)) %>%head(n=10)

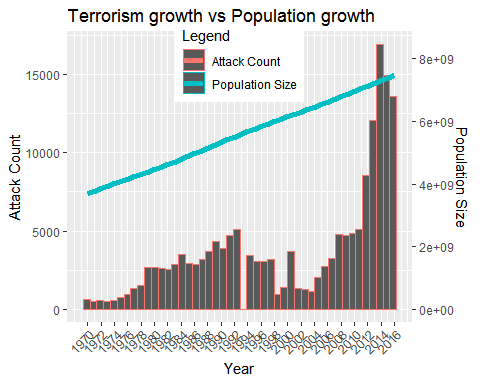
## # A tibble: 5 x 2  
## decades number\_of\_attacks  
## <fct> <int>  
## 1 2010s 86815  
## 2 80s 31160  
## 3 90s 28762  
## 4 2000s 25040  
## 5 70s 9914

#visual  
ggplot(data=df, aes(x=year, fill=decades)) +  
 geom\_histogram(stat='count') +   
 theme(axis.text.x= element\_text(angle=45, hjust=1)) +  
 labs(title='Terrorism growth over time')

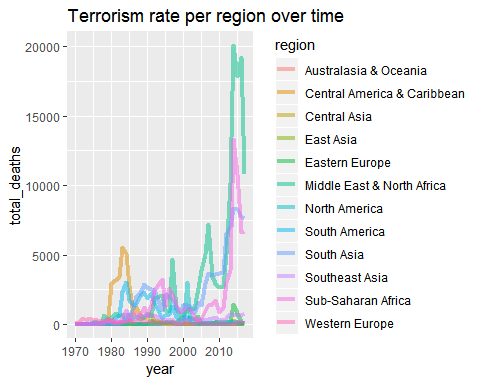
## Warning: Ignoring unknown parameters: binwidth, bins, pad



##considering the world population for the comparison of terrorism rate  
  
  
populationworld <- pop\_1 %>%  
 filter(Location == "World") %>%  
 select(-Location)  
   
df2 <- inner\_join(df, populationworld, by= c("year" = "Time"))  
  
pp1 <- ggplot(data=df2, aes(x=year)) +  
 geom\_histogram(aes(col='Attack Count'), bins=46) +   
 theme(axis.text.x= element\_text(angle=45, hjust=1)) +  
 scale\_x\_continuous(breaks=seq(1970, 2016, 2))  
  
pp1 +   
 geom\_line(aes(y=PopTotal/ 500, col='Population Size'), size=2) +   
 scale\_y\_continuous(sec.axis = sec\_axis(~ . \* 500000, name = "Population Size")) +  
 labs(y = "Attack Count", x = "Year", colour = "Legend") +  
 theme(legend.position = c(0.5, 0.9)) +  
 labs(title='Terrorism growth vs Population growth')



##terrorism growth has been grown after 2010. from the data Middle East and N Africa (~27% of total), South Asia (~24%) and S America (11%) are the top three regions in terms of number of attacks. Iraq (~12.9% of total), Pakistan,(~8%), Afghanistan (~6.6%), India(~6.4%) and Colombia (4.7%) are the top five countries in terms of number of attacks.   
  
#most of the deaths by terrorist attack have occurred during bomb attacks.   
  
  
  
weaponlethality <- df %>%  
 filter(weapon\_type != "Unknown") %>%  
 select(decades, weapon\_type, nkill)%>%  
 group\_by(decades,weapon\_type)%>%  
 summarise(number\_of\_deaths = n())%>%  
 top\_n(n=5, wt=number\_of\_deaths) %>%  
 mutate(percent\_deaths = (number\_of\_deaths/sum(number\_of\_deaths)\*100))  
   
##Terrorist group activity over time  
top10groups <- df %>%  
 filter(group\_name != "Unknown") %>%  
 group\_by(group\_name) %>%  
 summarise(number\_of\_deaths = n()) %>%  
 arrange(desc(number\_of\_deaths)) %>%  
 head(n=10)  
   
top10groupactivity <- df %>%  
filter(df$group\_name %in% c("Taliban", "Shining Path (SL)", "Islamic State of Iraq and the Levant (ISIL)", "Farabundo Marti National Liberation Front (FMLN)", "Al-Shabaab", "Irish Republican Army (IRA)", "Revolutionary Armed Forces of Colombia (FARC)", "New People's Army (NPA)", "Kurdistan Workers' Party (PKK)", "Boko Haram"))%>%   
select(year, group\_name)%>%  
group\_by(year, group\_name) %>%  
 summarise(number\_of\_deaths = n())%>%  
 arrange(desc(number\_of\_deaths))%>%  
 top\_n(n=10, wt=number\_of\_deaths)  
  
##The spike in Terrorist Activity has been maintained by 4 Main Groups- Taliban, Boko Haram,NPA,ISIL.  
  
  
#death rate by Region  
regionmort <- df %>%  
 filter(nkill != 'Unknown') %>%  
 select(region, nkill, nwound, year, group\_name, decades) %>%  
 group\_by(region, year) %>%  
 summarise(total\_deaths = sum(nkill))  
  
ggplot(data=regionmort, aes(x=year, y=total\_deaths, col=region, group= region)) +  
 geom\_line(size=1.5, alpha=0.5) +  
 labs(title='Terrorism rate per region over time')



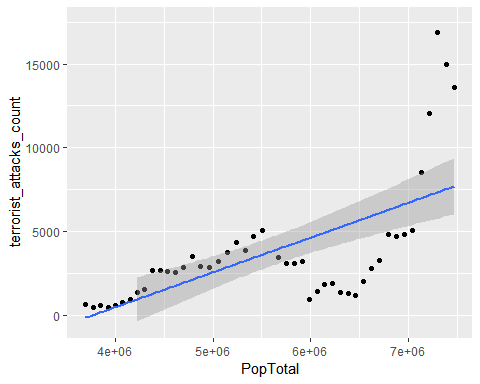
##from the data it can be visualized that The Middle East and North Africa have had an immense increase of deaths from 2003   
  
  
##Correlation analysis ofthe data over the population data   
##grouping the data -data reframing  
df3 <- df2 %>%  
 group\_by(year) %>%  
 summarise(terrorist\_attacks\_count = n())  
df3 <- inner\_join(df3, populationworld, by = c("year" = "Time"))  
  
df3 <- df3 %>%  
 mutate(decades =   
 ifelse(year<1980, '70s',   
 ifelse(year < 1990, '80s',   
 ifelse(year < 2000, '90s',   
 ifelse( year < 2010, '2000s', '2010s')))))  
  
  
df3$decades <- factor(df3$decades, levels=c("70s", "80s", "90s", "2000s", "2010s"))  
  
cor.test(df3$PopTotal, df3$terrorist\_attacks\_count, method="pearson")

##   
## Pearson's product-moment correlation  
##   
## data: df3$PopTotal and df3$terrorist\_attacks\_count  
## t = 5.5274, df = 44, p-value = 1.662e-06  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.4297245 0.7846445  
## sample estimates:  
## cor   
## 0.6401633

##Linear Model-with single variable regression  
  
  
mm1 <- lm(data=df3, terrorist\_attacks\_count ~ PopTotal)  
summary(mm1)

##   
## Call:  
## lm(formula = terrorist\_attacks\_count ~ PopTotal, data = df3)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4426.2 -1731.5 390.0 994.3 9563.6   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -7.894e+03 2.143e+03 -3.684 0.000626 \*\*\*  
## PopTotal 2.087e-03 3.776e-04 5.527 1.66e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2899 on 44 degrees of freedom  
## Multiple R-squared: 0.4098, Adjusted R-squared: 0.3964   
## F-statistic: 30.55 on 1 and 44 DF, p-value: 1.662e-06

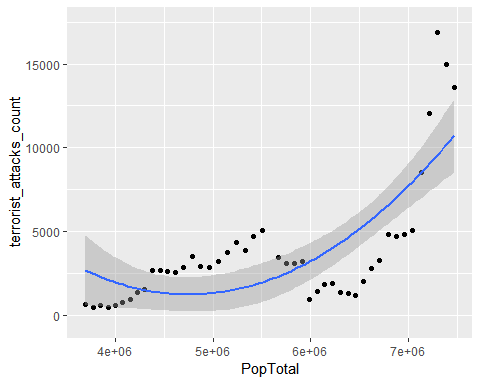
#Adjusted R-squared: 0.397 with p<0.05.  
  
ggplot(data=df3, aes(x=PopTotal, y=terrorist\_attacks\_count)) +  
 geom\_point() +  
 geom\_smooth(method="lm",formula= y ~ x)+  
 scale\_y\_continuous(limits = c(-500, 17500))



##since after sometime there was no linear relationship anymore as most of the points ~95% fell out of the fitted line hence we are considering the multiple regression  
  
mm2 <- lm(data=df3, terrorist\_attacks\_count ~ PopTotal  
+ I(PopTotal^2))  
summary(mm2)

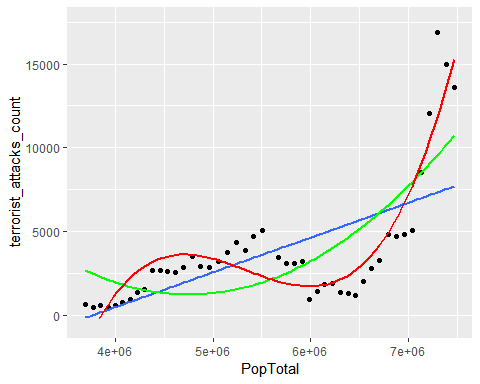
##   
## Call:  
## lm(formula = terrorist\_attacks\_count ~ PopTotal + I(PopTotal^2),   
## data = df3)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3811.5 -2036.3 38.3 1667.7 7335.4   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.040e+04 1.027e+04 2.960 0.004993 \*\*   
## PopTotal -1.227e-02 3.802e-03 -3.229 0.002385 \*\*   
## I(PopTotal^2) 1.291e-09 3.403e-10 3.792 0.000462 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2539 on 43 degrees of freedom  
## Multiple R-squared: 0.5577, Adjusted R-squared: 0.5371   
## F-statistic: 27.11 on 2 and 43 DF, p-value: 2.414e-08

#Adjusted R-squared: 0.536 with p<0.05.  
  
ggplot(data=df3, aes(x=PopTotal, y=terrorist\_attacks\_count)) +  
 geom\_point() +  
 geom\_smooth(method="lm",formula= y ~ x + I(x^2)) +  
 scale\_y\_continuous(limits = c(-500, 17500))



##the variance in terrorist attack count much better with an adjusted R-squared of 0.536 (p<0.05)  
ggplot(data=df3, aes(x=PopTotal, y=terrorist\_attacks\_count)) +  
 geom\_point() +  
 geom\_smooth(method="lm",formula= y ~ x, se=F) +  
 geom\_smooth(method="lm",formula= y ~ x + I(x^2), se=F, col='green') +  
 geom\_smooth(method="lm",formula= y ~ x + I(x^2) + I(x^3), se=F, col='red') +  
 scale\_y\_continuous(limits = c(-500, 17500))

## Warning: Removed 3 rows containing missing values (geom\_smooth).



m3 <- lm(data=df3, terrorist\_attacks\_count ~ PopTotal + I(PopTotal^2) + I(PopTotal^3))  
summary(m3)

##   
## Call:  
## lm(formula = terrorist\_attacks\_count ~ PopTotal + I(PopTotal^2) +   
## I(PopTotal^3), data = df3)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2692.1 -1004.4 -485.5 1195.6 5048.5   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -2.649e+05 3.366e+04 -7.870 8.52e-10 \*\*\*  
## PopTotal 1.545e-01 1.884e-02 8.205 2.90e-10 \*\*\*  
## I(PopTotal^2) -2.931e-08 3.437e-09 -8.530 1.03e-10 \*\*\*  
## I(PopTotal^3) 1.827e-15 2.048e-16 8.921 3.02e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1510 on 42 degrees of freedom  
## Multiple R-squared: 0.8472, Adjusted R-squared: 0.8363   
## F-statistic: 77.63 on 3 and 42 DF, p-value: < 2.2e-16

##we compare model 1 and model 2.  
anova(mm1, mm2, test="F")

## Analysis of Variance Table  
##   
## Model 1: terrorist\_attacks\_count ~ PopTotal  
## Model 2: terrorist\_attacks\_count ~ PopTotal + I(PopTotal^2)  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 44 369850350   
## 2 43 277168082 1 92682268 14.379 0.0004617 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

#p<0.05 for model 2, meaning it's the better model  
  
anova(mm2, m3, test="F")

## Analysis of Variance Table  
##   
## Model 1: terrorist\_attacks\_count ~ PopTotal + I(PopTotal^2)  
## Model 2: terrorist\_attacks\_count ~ PopTotal + I(PopTotal^2) + I(PopTotal^3)  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 43 277168082   
## 2 42 95745809 1 181422273 79.583 3.024e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

#p<0.05 for model 3, meaning it's the best model

## Next we do Regression Diagnostics by performing the Outlier Tests

## the main reasons is population growth has decoupled as a linear predictor for terrorism attacks due the data from past 10 years.

## Model validation will be done using k-fold cross-validation and predict the amount of future terrorism attacks based on future total population given by the predicted population dataset.

## Then we apply classification methods and Data preparation for first association analysis by using the Association rules and finally produce Interactive overview of Rule To determine the terrorism rate.