**Accidental Deaths**

**NAME: ADITYA PATIL**

**CIN: 304746057**

Fall 2016

CIS 5270 – 01 Business Intelligence

Prof. Shilpa Balan

Aditya Patil

**A. DATA SETS**

<https://catalog.data.gov/dataset/accidental-drug-related-deaths-january-2012-sept-2015>

Data set primarily consists of thirty-one columns as well as over two thousand five hundred rows. Information set incorporates date of the occurrence happened, information set additionally expresses the sex, race, age of the demised. It likewise states point by point data about the living arrangement city, living arrangement state, habitation region and additionally passing city, demise state, demise area of the demised. Moreover, information set not just categorizes fourteen distinctive sorts of medications which may have assumed essential part in unplanned demise additionally gives Immediate reason for death of the individual.

A posting of every unplanned demise connected with medication overdose in Connecticut from 2012 to 2015. A "Y" esteem under the distinctive substance sections demonstrates that specific substance was identified. Information is derived from an examination by the Office of the Chief Medical Examiner which incorporates the toxicity report, death certificate, as well as a scene investigation.

In addition, in the information set in some heroin passing’s, the harmfulness results won't demonstrate whether the morphine is from heroin or remedy morphine. In these cases, the Medical Examiner might have the capacity to decide the cause in view of the scene examination, (for example, discovering heroin needles). On the off chance that they discover remedy morphine at the scene it is confirmed as "Morphine (not heroin)." Therefore, the Cause of Death may show Morphine, however the Heroin or Morphine (Not Heroin) may not be demonstrated.

**B. DATA CLEANING**

More often than not the information includes pointless data. That should be erased or adjusted and the information sets should be refined to make sense. The accompanying pictures will clarify what was the unrequired data in the information sets and how it was cleaned.

**1.Deleting Column:**

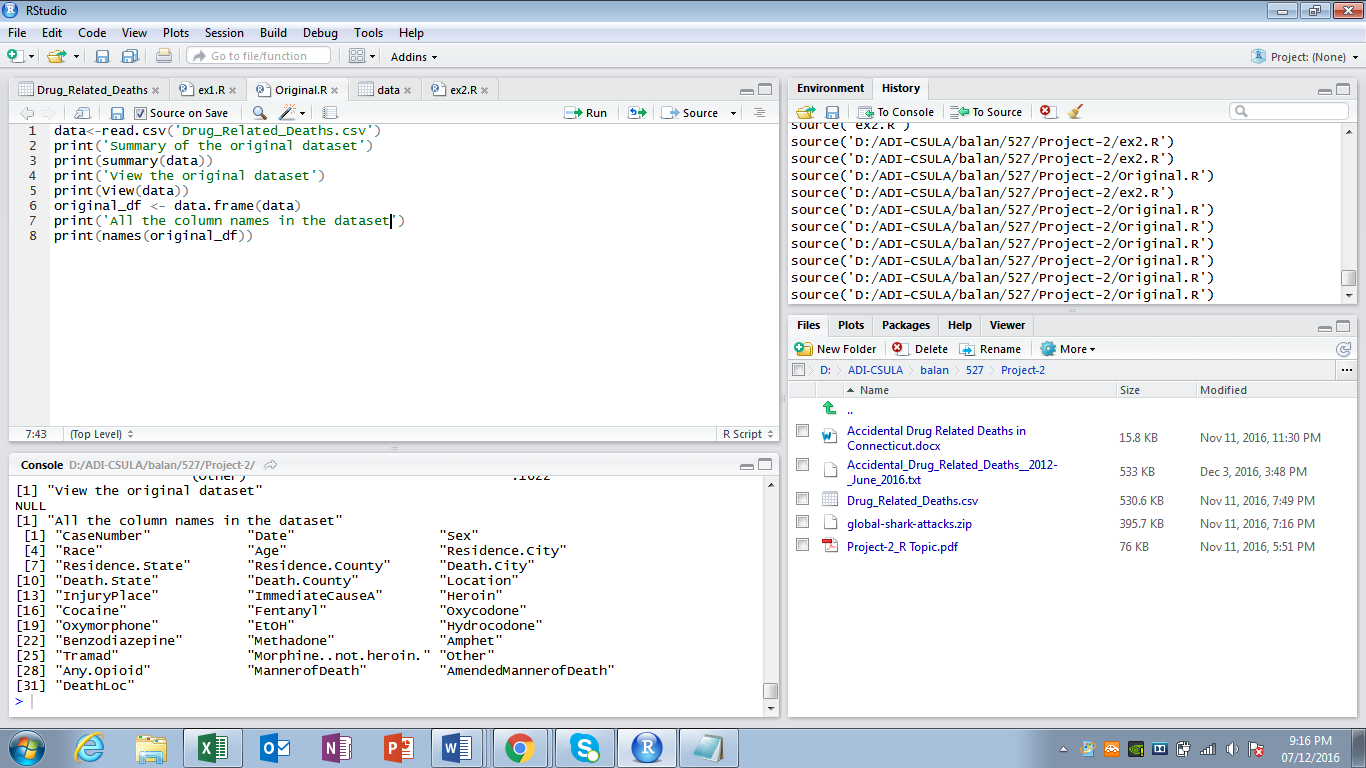
In the dataset I had an empty column, ‘AmendedMannerofDeath’. Since I had no use of the empty column I had to clean the dataset of that particular column.

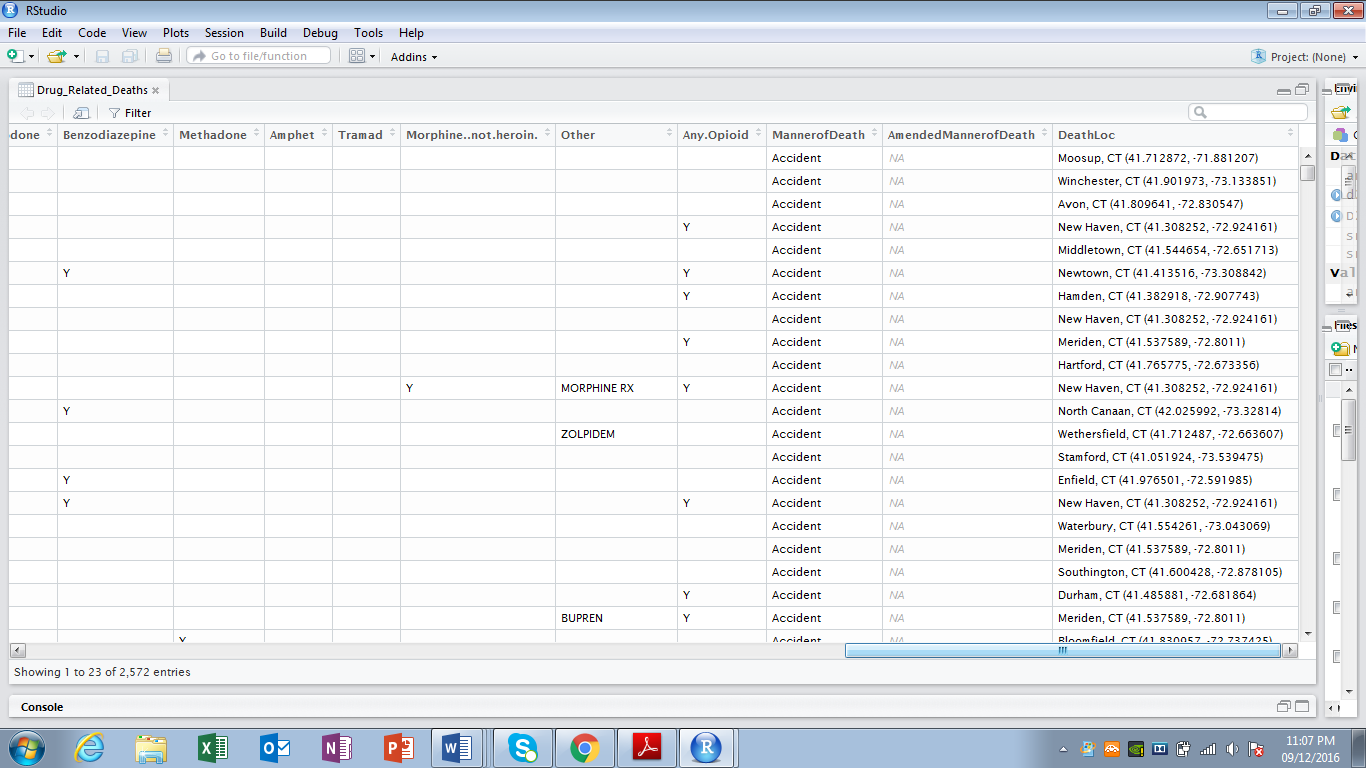
**2.Empty Values:**

Furthermore, I had number of columns and rows which had empty values. Such values cannot be considered in the analysis of the data. Considering such empty values will mess up the analysis hence I had to fill up such blanks with ‘NA’.

**Before Shot:**

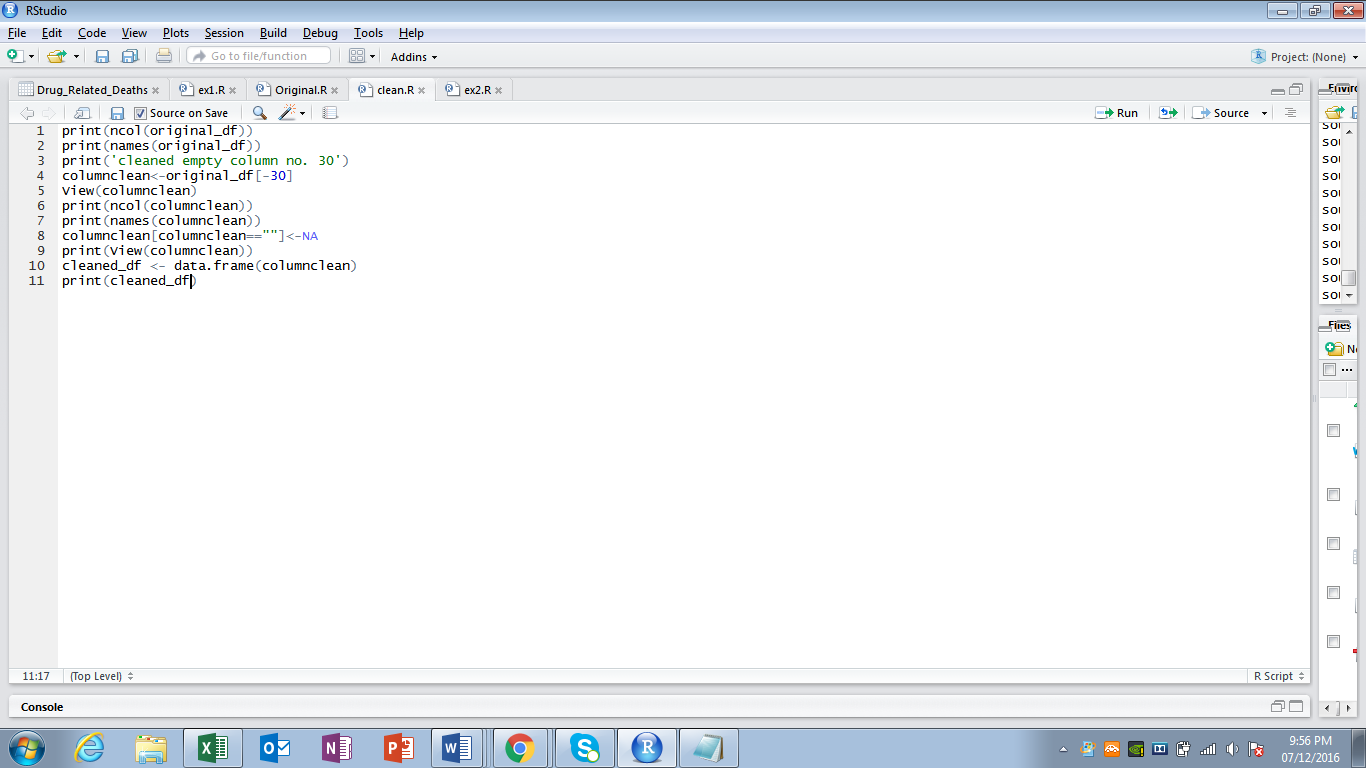
Here are the before shot of the uncleaned dataset.

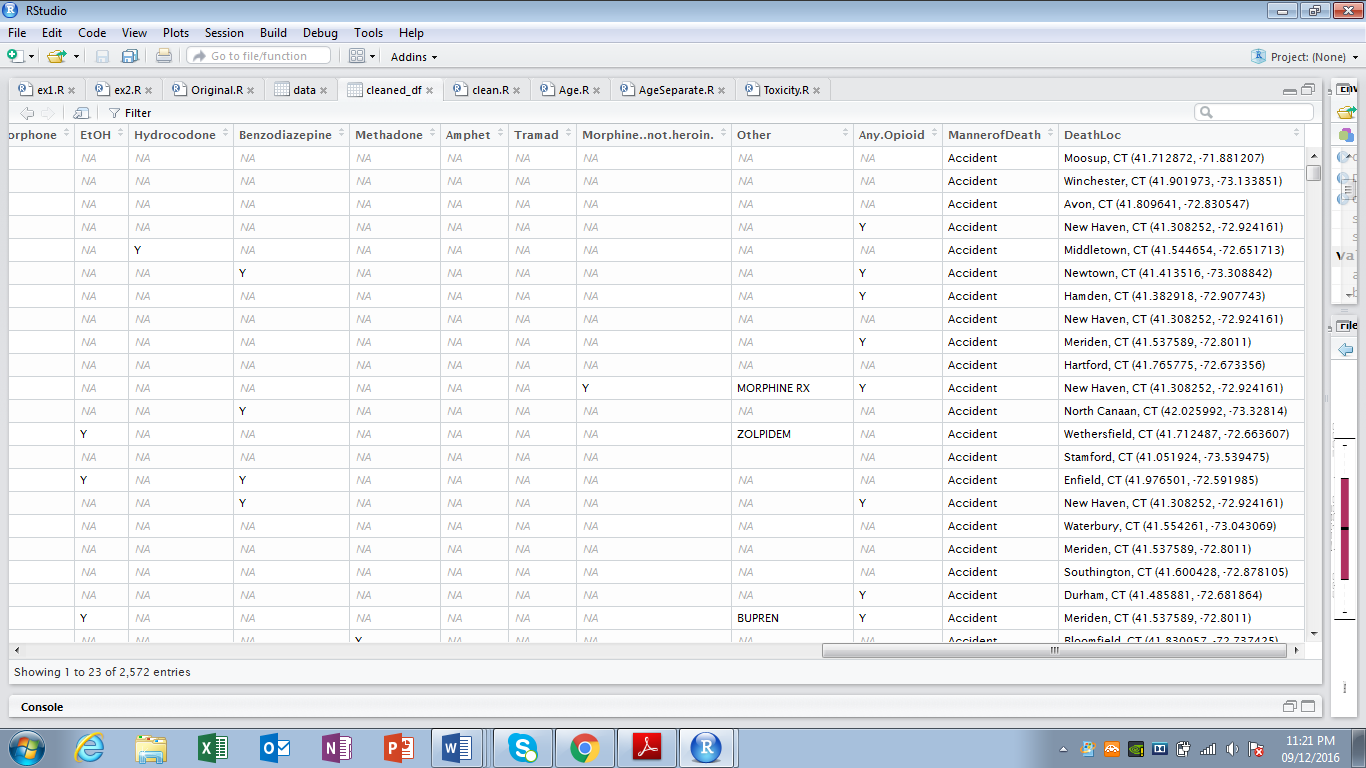




**After Shot:**

Here are the screenshots of the dataset after cleaning.

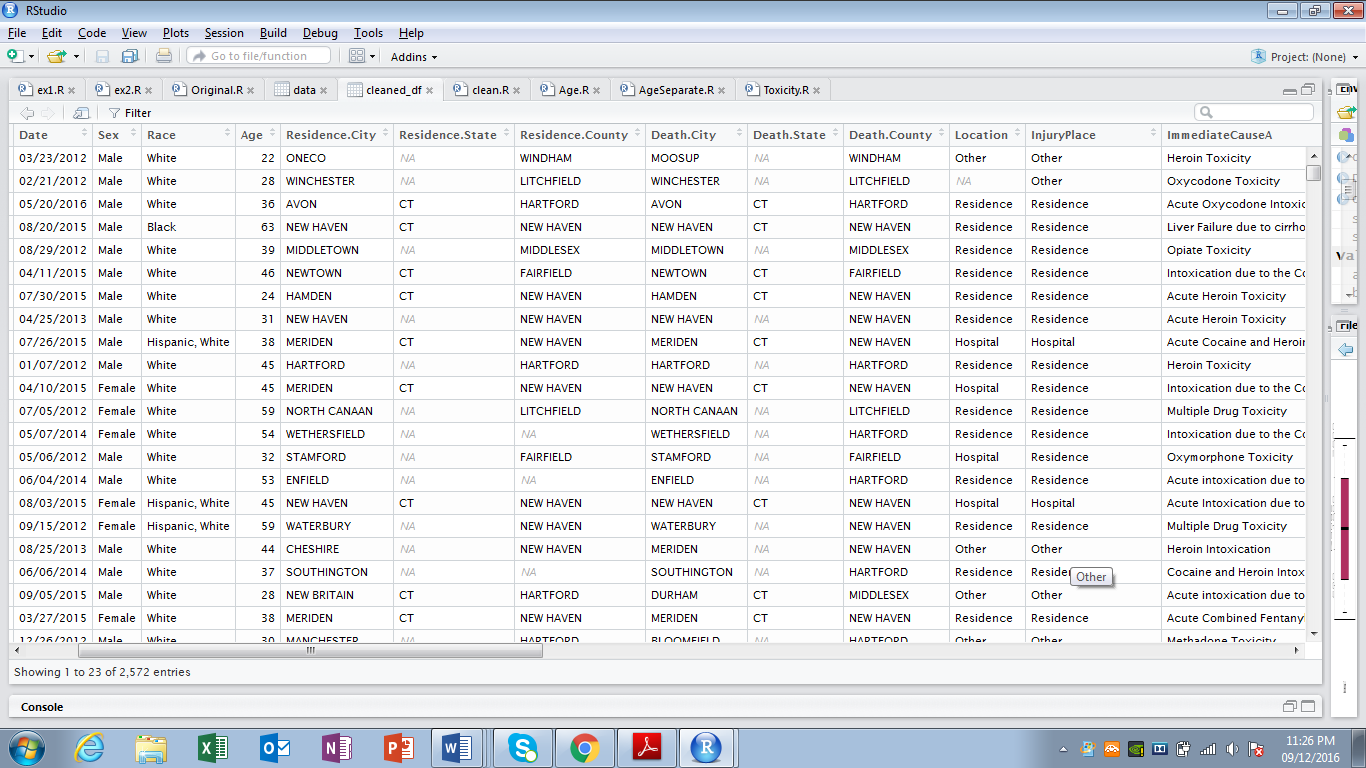




As you can see the empty spaces are filled with ‘NA’ values as well as the unnecessary column has been deleted.

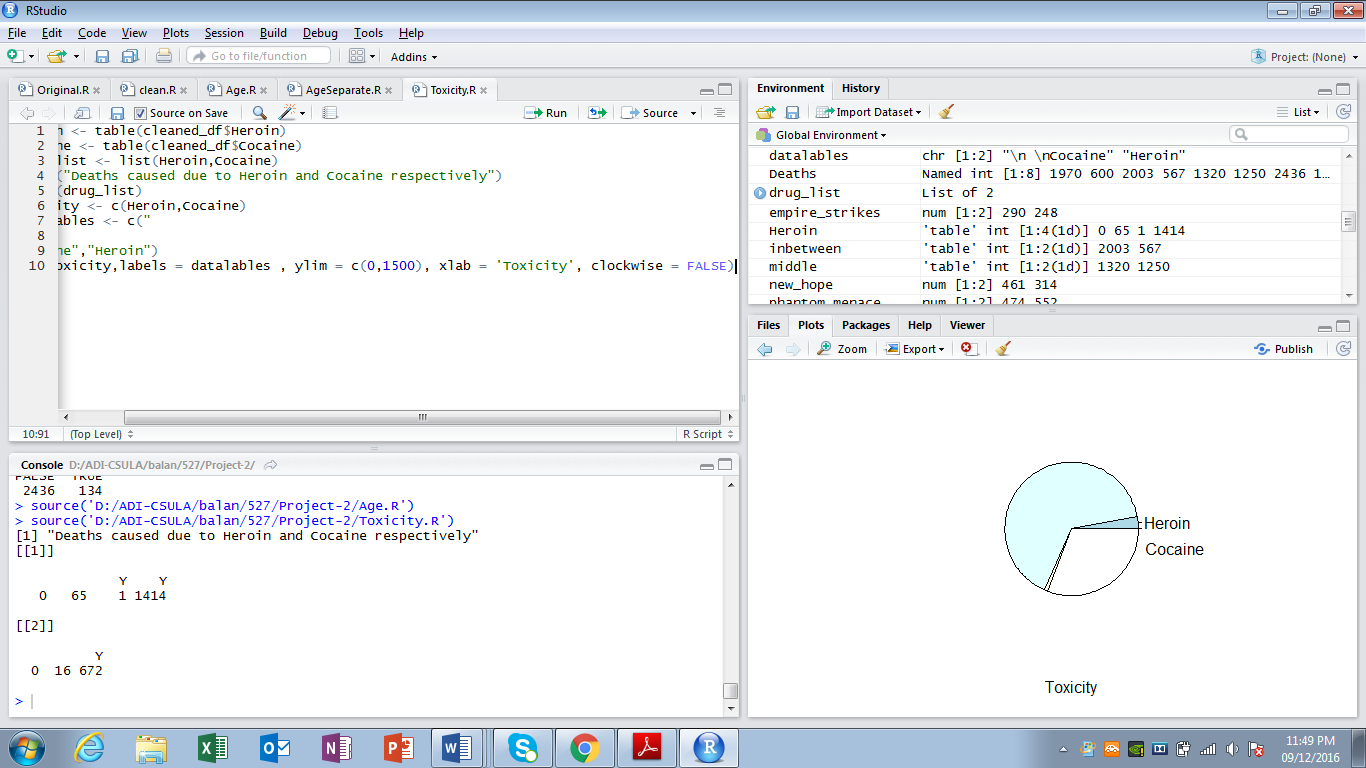
**3. Irrelevant Value:**

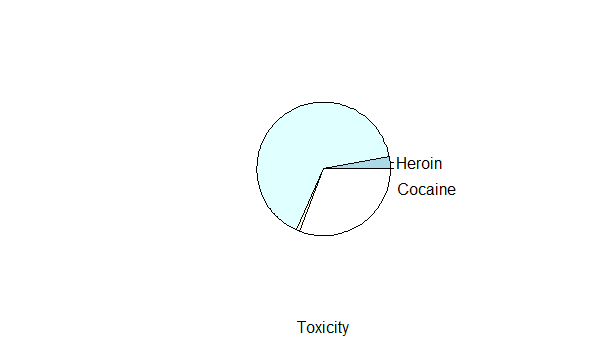
Except the above data cleaning necessities there were no irrelevant values in the dataset or missing data present. Here is the screenshot of the dataset.



**3.Analysis & Visualization:**

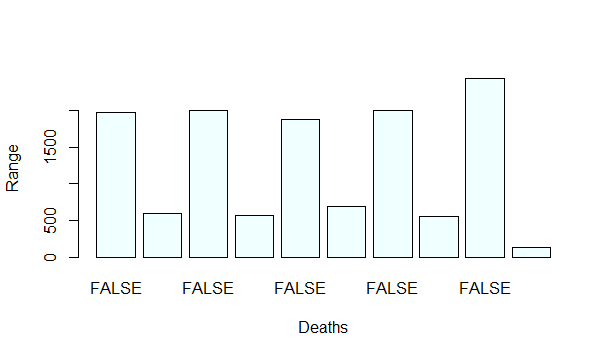
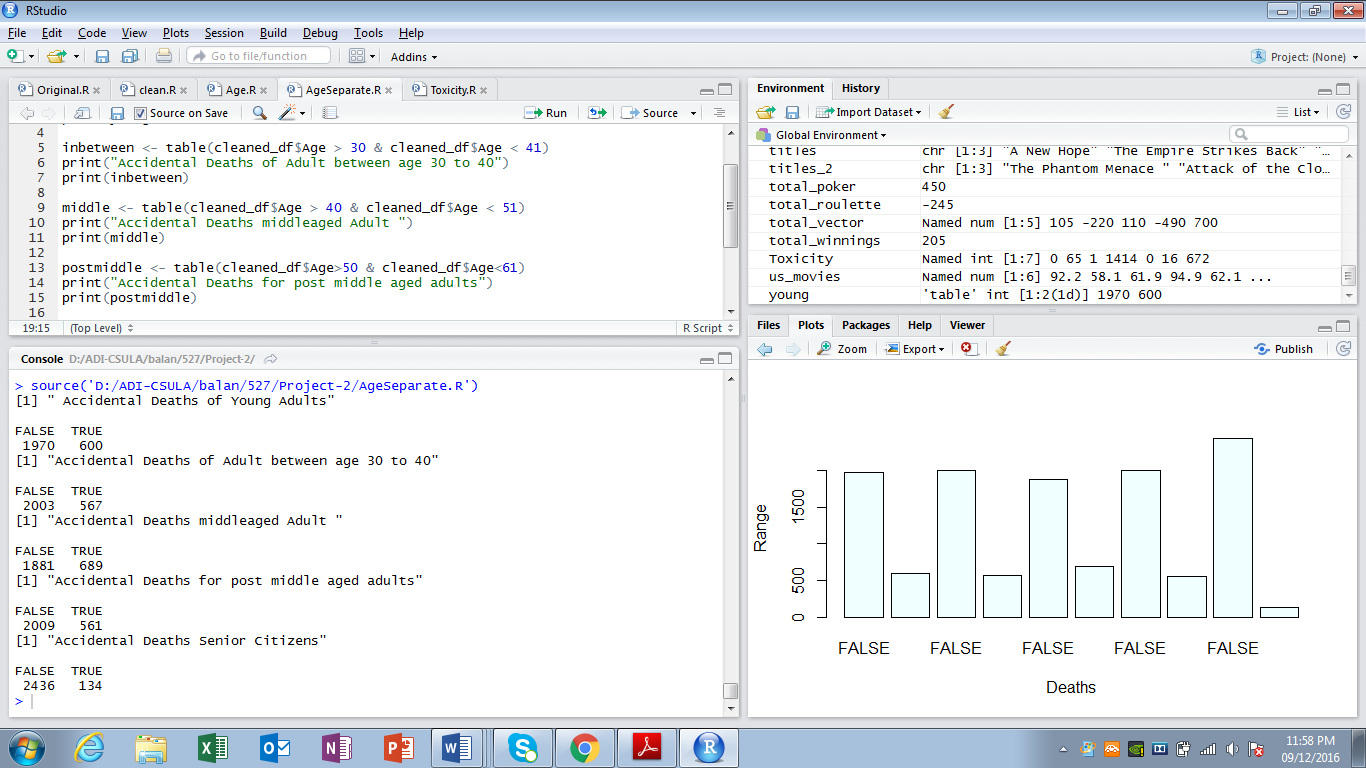
**1. Among the toxicity due to Heroin and Cocaine, which one caused the most accidental deaths in adults?**





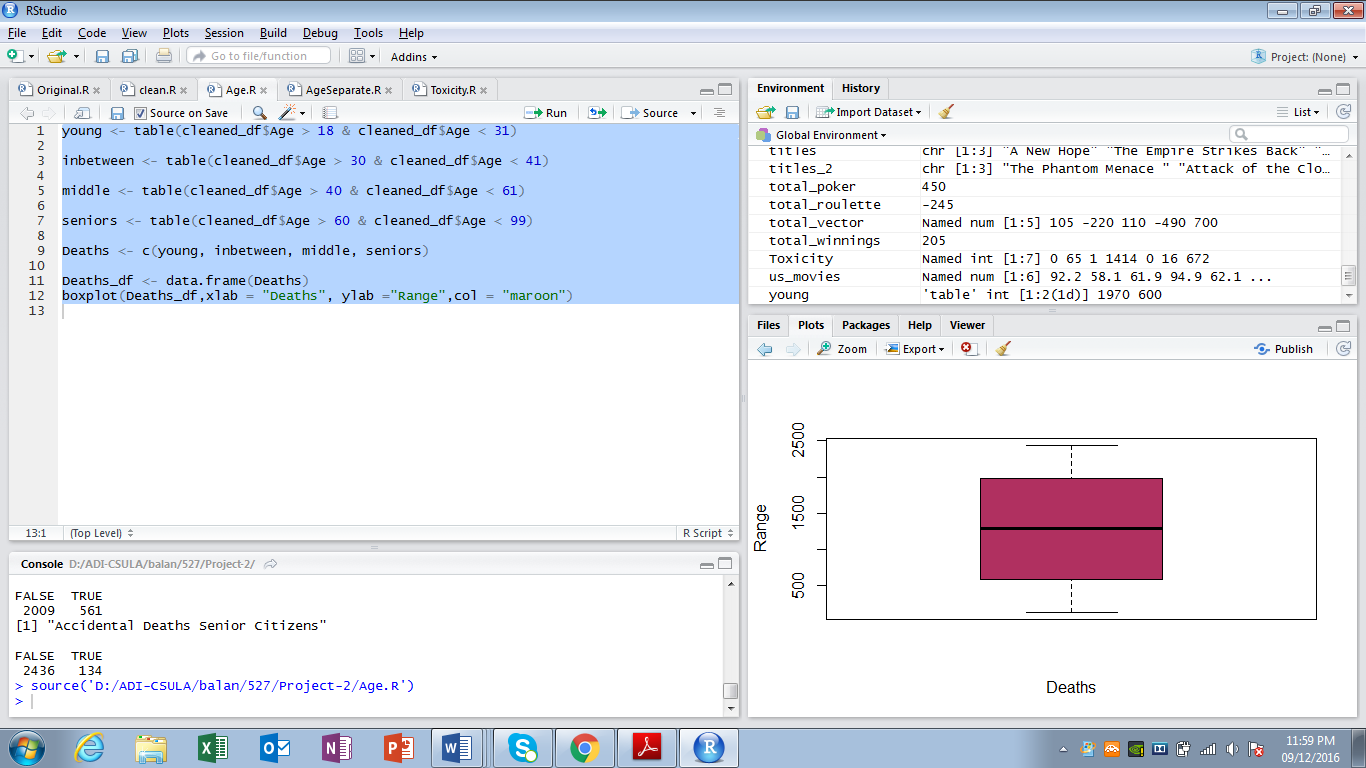
Here, in the question it is asked to analyze the Which of the Drug among Heroin and Cocaine has led to maximum number of accidental deaths. From the above visualization it is found that Heroin took more lives than the Cocaine. In above visualization both blue and lighter blue color show the accidental deaths caused due to Heroin. The total number of deaths occurred due to Heroin is 1480 throughout the given years. Furthermore, Cocaine took less lives compared to heroin, shows the analysis. In the above visualization the white as well as off-white color shows the deaths due to cocaine. The total number of deaths occurred due to cocaine is 688 which is significantly late than heroine.

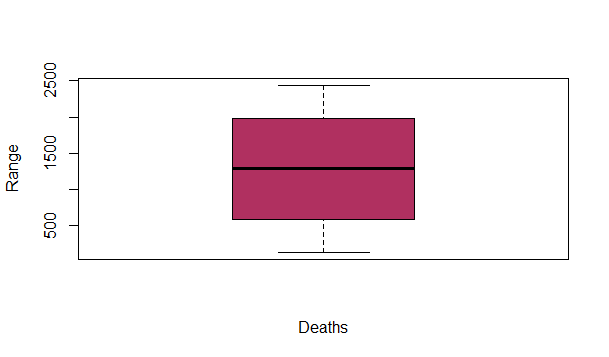
**2. How many persons among the demised were in their 20’s, 30’s, 40’s, 50’s and how many were the senior citizens?**



In the above chart, the second, fourth, sixth, eighth and tenth chart represents the Accidental Deaths of Young Adults representing age group 18 to 30, Accidental Deaths of Adult between age 30 to 40, Accidental Deaths middle aged Adult age group 40 to 50, Accidental Deaths for post middle aged adults age group 50 to 60 and Accidental Deaths Senior Citizens that is 60 above respectively! The larger bars show the total number of deaths counted altogether. From the analysis it can be clearly seen that number of Accidental Deaths of Young Adults due to drugs are 600. While number of accidental deaths caused due to drugs in people between age 30 to 40, middle aged Adult, post middle aged adults and Senior Citizens are 567, 689, 561 and 134 respectively. From the analysis and visualization, it can be seen clearly that the most accidental drug related deaths occurred in the age group of 40 to 50.

**3. What is the range of number deaths occurred in different years altogether?**





From the above analysis and the visualization, it can be seen that the range of accidental deaths occurred in the years of 2012 to 2015 is from 500 to 2000. The boxplot gathers the data from four different years and gives the range.

**R-Scripts Code:**

data<-read.csv('Drug\_Related\_Deaths.csv')

print('Summary of the original dataset')

print(summary(data))

print('View the original dataset')

print(View(data))

original\_df <- data.frame(data)

print('All the column names in the dataset')

print(names(original\_df))

View(cleaned\_df)

print(cleaned\_df)

print(ncol(original\_df))

print(names(original\_df))

print('cleaned empty column no. 30')

columnclean<-original\_df[-30]

print(ncol(columnclean))

print(names(columnclean))

columnclean[columnclean==""]<-NA

print(View(columnclean))

#columnclean[columnclean$Heroin=='y']<-'Y'

#replace(columnclean,"y","Y")

print(columnclean)

View(columnclean)

cleaned\_df <- data.frame(columnclean)

print(cleaned\_df)

young <- table(cleaned\_df$Age > 18 & cleaned\_df$Age < 31)

inbetween <- table(cleaned\_df$Age > 30 & cleaned\_df$Age < 41)

middle <- table(cleaned\_df$Age > 40 & cleaned\_df$Age < 61)

seniors <- table(cleaned\_df$Age > 60 & cleaned\_df$Age < 99)

Deaths <- c(young, inbetween, middle, seniors)

Deaths\_df <- data.frame(Deaths)

boxplot(Deaths\_df,xlab = "Deaths", ylab ="Range",col = "maroon")

young <- table(cleaned\_df$Age > 18 & cleaned\_df$Age < 31)

print(" Accidental Deaths of Young Adults")

print(young)

inbetween <- table(cleaned\_df$Age > 30 & cleaned\_df$Age < 41)

print("Accidental Deaths of Adult between age 30 to 40")

print(inbetween)

middle <- table(cleaned\_df$Age > 40 & cleaned\_df$Age < 51)

print("Accidental Deaths middleaged Adult ")

print(middle)

postmiddle <- table(cleaned\_df$Age>50 & cleaned\_df$Age<61)

print("Accidental Deaths for post middle aged adults")

print(postmiddle)

seniors <- table(cleaned\_df$Age > 60 & cleaned\_df$Age < 99)

print("Accidental Deaths Senior Citizens")

print(seniors)

Deaths <- c(young, inbetween, middle, postmiddle, seniors)

Deaths\_df <- data.frame(Deaths)

barplot(Deaths, xlab = "Deaths", ylab = "Range", col = "azure")

Heroin <- table(cleaned\_df$Heroin)

Cocaine <- table(cleaned\_df$Cocaine)

drug\_list <- list(Heroin,Cocaine)

print("Deaths caused due to Heroin and Cocaine respectively")

print(drug\_list)

Toxicity <- c(Heroin,Cocaine)

datalables <- c("Cocaine","Heroin")

pie(Toxicity,labels = datalables , ylim = c(0,1500), xlab = 'Toxicity', clockwise = FALSE)