

Smoking Trends Among the Youth of USA



Submitted to: **Dr. Shilpa Balan**

DATASET

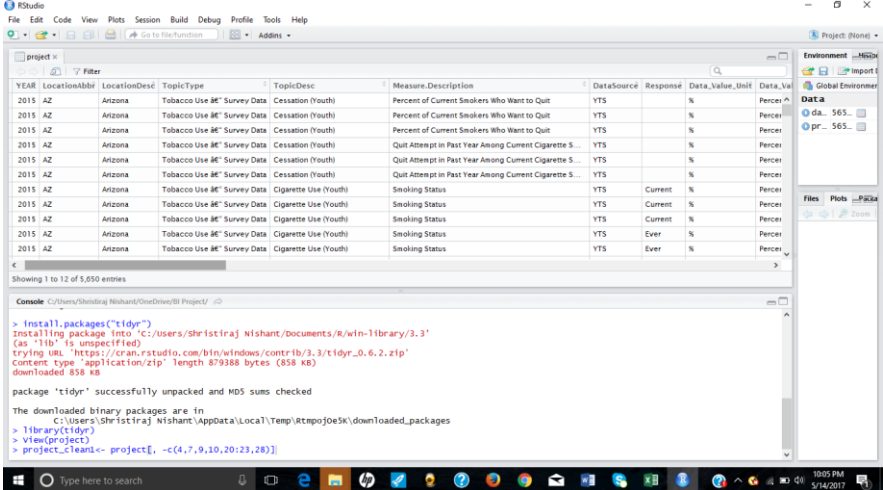
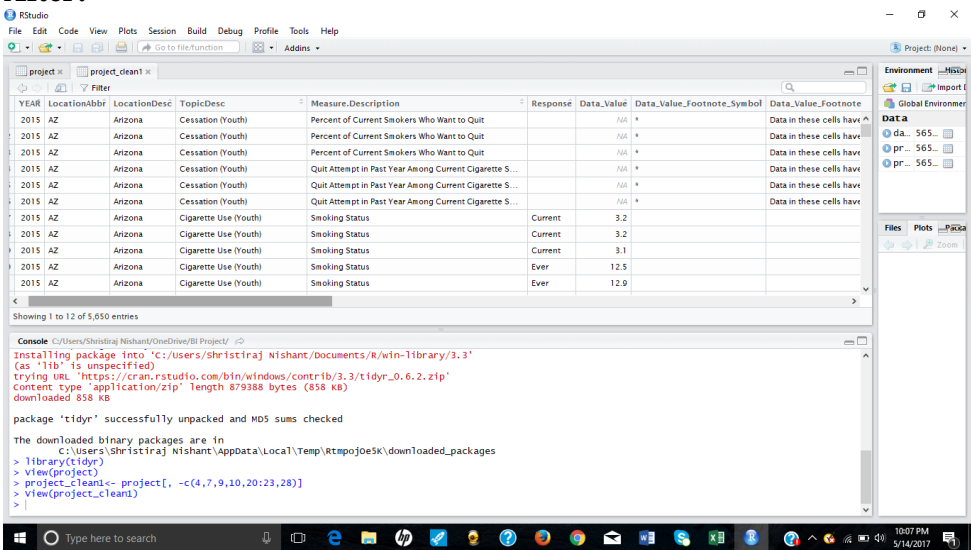
URL: <https://chronicdata.cdc.gov/api/views/4juzx2tp/rows.csv?accessType=DOWNLOAD>

The Youth Tobacco Survey under State Tobacco Activities Tracking System (STATE), provides dataset for this project. The dataset is extensive and contains more than 9000 different records. It has measures such as smoking status, user status and cessation which opens wide scope for analysis of data. The records in the dataset has entries from the year 1999 till 2015. It also has different stratification measures which opens the scope for critical analysis.

The dataset contains Locations (abbreviated and described), Years (only 2005-2015 has been used), data values in percentages to carry out different analysis. There are different Stratification measures, such as Generation, Age, Race, and Education which provides an array of different combinations for our analysis. For this project, StratificationID1, which contains parameters: 1 GEN, 2 GEN, and 3 GEN, can be used to identify the number of smokers who tried to quit or to calculate the whole cessation rate for any generation. Gender column helps us to identify and carry out analysis on male and female smokers. Education column can be handy in terms of analyzing the smokers from either high school or middle school.

The available dataset was listed from 1999- 2015 but, for analysis purpose, the data from 2005-2015 is considered to make finding more recent and not been affected by historical data. This analysis can be used to analyze smoking trends among the youth of the USA and identify the gender, generation and education parameters which contribute to smoking on a whole.

DATA CLEANING

Categories	Un-Cleaned and Cleaned Data
Deleting/ Removing Unnecessar y Columns	<p>Before:</p>  <p>After:</p>  <p>R Code:</p> <pre> Source Console C:/Users/Shristiraj Nishant/OneDrive/BI Project/ > setwd("C:/Users/Shristiraj Nishant/OneDrive/BI Project") > getwd() [1] "C:/Users/Shristiraj Nishant/OneDrive/BI Project" > data1<- read.csv("R Proj Data.csv", header = T, sep = ",") > project<- data.frame(data1) > view(project) > install.packages("tidyr") Error in install.packages : Updating loaded packages Restarting R session... > install.packages("tidyr") Installing package into 'C:/Users/Shristiraj Nishant/Documents/R/win-library/3.3' (as 'lib' is unspecified) trying URL 'https://cran.rstudio.com/bin/windows/contrib/3.3/tidyr_0.6.2.zip' Content type 'application/zip' length 879539 bytes (858 KB) downloaded 858 KB package 'tidyr' successfully unpacked and MD5 sums checked The downloaded binary packages are in C:/Users/Shristiraj Nishant/AppData/Local/Temp/RtmpQVn6xz/downloaded_packages > library(tidyr) > project_clean1<- project[, -c(4,7,9,10,20:23,28)] > view(project_clean1) </pre>

Splitting Columns

Before:

```

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to File Function Addins
project -> project_clean1 -> data1
Filter
Data_Value_Std_Err Low_Confidence_Limit High_Confidence_Limit Sample_Size Gender Race.Age.Education StratificationID1 StratificationID2 StratificationID3 StratificationID4
1.8 0.7 16.8 566 Male All Races-All Ages-High School 2GEN BAGE BRAC 2EDU
1.7 12.9 19.7 653 Female All Races-All Ages-High School 3GEN BAGE BRAC 2EDU
1.4 9.9 15.2 1226 Overall All Races-All Ages-High School 1GEN BAGE BRAC 2EDU
2.1 17.9 26.1 559 Male All Races-All Ages-High School 2GEN BAGE BRAC 2EDU
0.5 1.6 3.6 653 Female All Races-All Ages-High School 3GEN BAGE BRAC 2EDU
1.7 24.0 30.7 1221 Overall All Races-All Ages-High School 1GEN BAGE BRAC 2EDU
2.8 14.3 48.2 158 Male All Races-All Ages-High School 2GEN BAGE BRAC 2EDU
1.4 11.8 17.1 691 Female All Races-All Ages-High School 3GEN BAGE BRAC 2EDU
1.0 3.8 7.9 1226 Overall All Races-All Ages-High School 1GEN BAGE BRAC 2EDU
1.6 7.5 13.9 559 Male All Races-All Ages-High School 2GEN BAGE BRAC 2EDU
0.5 0.0 2.0 653 Female All Races-All Ages-High School 3GEN BAGE BRAC 2EDU
Showing 5,639 to 5,650 of 5,650 entries

Console C:/Users/Shristiraj Nishant/OneDrive/BI Project/
warning message:
Too many values at 1 location: 1
> project_clean1<- project_clean1[,]
> project_clean2<- separate(project_clean1, Race.Age.Education, c("Race", "Age", "Education"), sep = "-")
warning message:
Too many values at 1 location: 1
> project_clean2<- separate(project_clean1, Race.Age.Education, c("Race", "Age", "Education"), sep = "-")
warning message:
Too many values at 1 location: 1
> project_clean2<- project[, c(4,7,9,10,20:23,28)]
> view(project_clean2)
> project_clean2<- data.frame(project_clean2)
> clean2<- separate(project_clean2, )
  
```

After:

```

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to File Function Addins
project -> project_clean1 -> clean2 -> data1
Filter
Data_Value_Std_Err Low_Confidence_Limit High_Confidence_Limit Sample_Size Gender Race Age Education StratificationID1 StratificationID2 StratificationID3 StratificationID4
because ... NA NA NA NA NA Overall All Races All Ages Middle School 1GEN BAGE BRAC 1E
because ... NA NA NA NA NA Male All Races All Ages Middle School 2GEN BAGE BRAC 1E
because ... NA NA NA NA NA Female All Races All Ages Middle School 3GEN BAGE BRAC 1E
because ... NA NA NA NA NA Overall All Races All Ages Middle School 1GEN BAGE BRAC 1E
because ... NA NA NA NA NA Male All Races All Ages Middle School 2GEN BAGE BRAC 1E
because ... NA NA NA NA NA Female All Races All Ages Middle School 3GEN BAGE BRAC 1E
1.5 0.3 6.1 1377 Overall All Races All Ages Middle School 1GEN BAGE BRAC 1E
1.5 0.3 6.2 700 Male All Races All Ages Middle School 2GEN BAGE BRAC 1E
1.8 0.1 6.1 671 Female All Races All Ages Middle School 3GEN BAGE BRAC 1E
2.7 7.2 17.9 1331 Overall All Races All Ages Middle School 1GEN BAGE BRAC 1E
2.6 7.7 18.1 669 Male All Races All Ages Middle School 2GEN BAGE BRAC 1E
Showing 1 to 12 of 5,650 entries

Console C:/Users/Shristiraj Nishant/OneDrive/BI Project/
> project_clean1<- data.frame(project_clean1)
> clean2<- separate(project_clean1, Race.Age.Education, c("Race", "Age", "Education"), sep = "-")
> view(clean2)
  
```

R Code:

```

Console C:/Users/Shristiraj Nishant/OneDrive/BI Project/
> project_clean1<- data.frame(project_clean1)
> clean2<- separate(project_clean1, Race.Age.Education, c("Race", "Age", "Education"), sep = "-")
> view(clean2)
  
```

Updating Rows/Columns with Empty Values

Before:

RStudio interface showing the 'Before' state of a dataset. The data table has columns: Data_Value, Data_Value_Footnote_Symbol, Data_Value_Footnote, Data_Value_Std_Err, Low_Confidence_Limit, High_Confidence_Limit, Sample_Size, Gender, and Race. The console shows the following R code:

```

> clean2 <- clean1[, ~c(23:26)]
> View(clean2)
> mean(clean2$Data_Value)
[1] NA
> mean(clean2$Data_Value, na.rm = T)
[1] 18.32166
> mean(clean2$Data_Value_Std_Err, na.rm = T)
[1] 1.849671
> mean(clean2$Low_Confidence_Limit, na.rm = T)
[1] 14.71259
> mean(clean2$High_Confidence_Limit, na.rm = T)
[1] 21.94098
> mean(clean2$Sample_Size, na.rm = T)
[1] 1138.773

```

After:

RStudio interface showing the 'After' state of the dataset. The data table now has numerical values instead of NA in the previously empty columns. The console shows the following R code and an error message:

```

> script.R
Error: object 'script.R' not found
> source('C:/Users/Shristiraj Nishant/OneDrive/BI Project/script.R')
> 
> source('C:/Users/Shristiraj Nishant/OneDrive/BI Project/script.R')
[1] "H1"
> clean2$Data_Value[is.na(clean2$Data_Value)] = mean(clean2$Data_Value, na.rm = T)
> clean2$Data_Value_Std_Err[is.na(clean2$Data_Value_Std_Err)] = mean(clean2$Data_Value_Std_Err, na.rm = T)
> clean2$Low_Confidence_Limit[is.na(clean2$Low_Confidence_Limit)] = mean(clean2$Low_Confidence_Limit, na.rm = T)
> clean2$High_Confidence_Limit[is.na(clean2$High_Confidence_Limit)] = mean(clean2$High_Confidence_Limit, na.rm = T)
> clean2$Sample_Size[is.na(clean2$Sample_Size)] = mean(clean2$Sample_Size, na.rm = T)
> View(clean2)

```

R code:

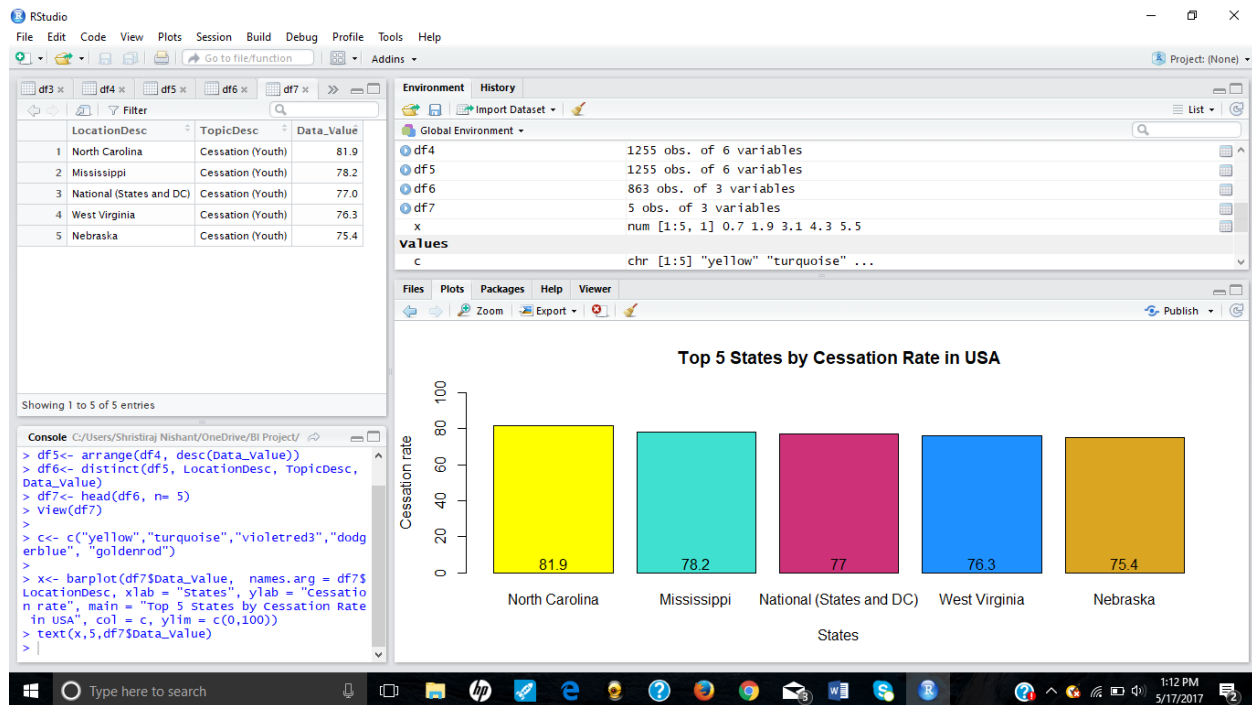
```

> clean2 <- separate(project_clean1, Race.Age.Education, c("Race", "Age", "Education"), sep = "-")
> View(clean2)
> clean2 <- data.frame(clean2)
> clean2$Data_Value[is.na(clean2$Data_Value)] = mean(clean2$Data_Value, na.rm = T)
> clean2 <- data.frame(clean2)
> View(clean2)
> clean2$Data_Value_Std_Err[is.na(clean2$Data_Value_Std_Err)] = mean(clean2$Data_Value_Std_Err, na.rm = T)
> 
> clean2$Data_Value_Std_Err[is.na(clean2$Data_Value_Std_Err)] = mean(clean2$Data_Value_Std_Err, na.rm = T)
> clean2$Sample_Size[is.na(clean2$Sample_Size)] = mean(clean2$Sample_Size, na.rm = T)
> clean2$High_Confidence_Limit[is.na(clean2$High_Confidence_Limit)] = mean(clean2$High_Confidence_Limit, na.rm = T)
> clean2$Data_Value_Std_Err[is.na(clean2$Data_Value_Std_Err)] = mean(clean2$Data_Value_Std_Err, na.rm = T)
> View(clean2)
> clean2$Low_Confidence_Limit[is.na(clean2$High_Confidence_Limit)] = mean(clean2$Low_Confidence_Limit, na.rm = T)
> View(clean2)
> clean2$Low_Confidence_Limit[is.na(clean2$Low_Confidence_Limit)] = mean(clean2$Low_Confidence_Limit, na.rm = T)
> View(clean2)
> |

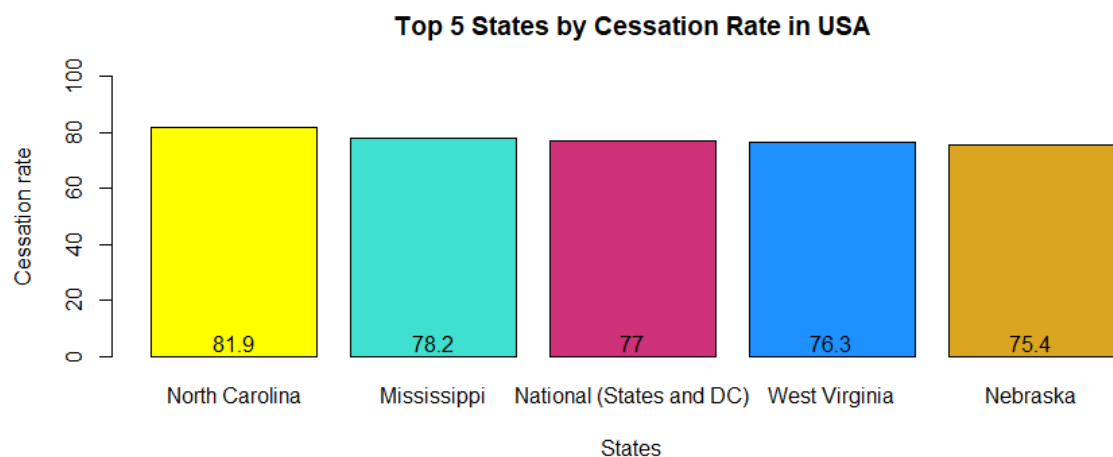
```

DATA VISUALIZATION

1. What are the top 5 states with maximum cessation rate?



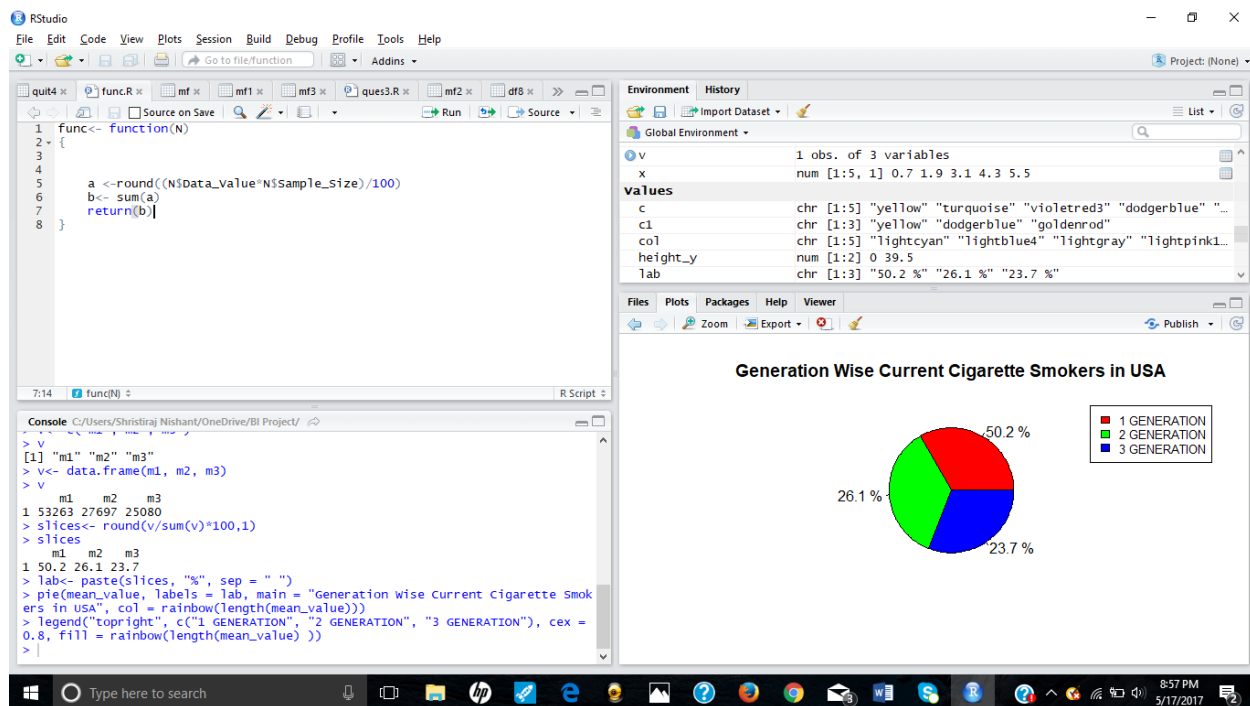
(Highlights: dplyr package, barplot)



The above visualization shows the different rates with the maximum cessation rate. Cessation rate had two different parameters in the dataset: “Smokers who wanted to quit” and “Quit attempts in

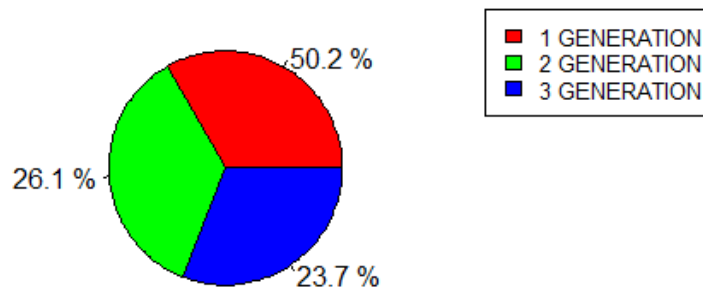
past year”. Both these parameters are considered while carrying out the visuals. The people in state of North Carolina have the highest motivation to leave smoking as evident from the 81.9% cessation rate. The state of Mississippi stands second with 78.2%. Dplyr package used in this visualization provided with functions like filter(), arrange(), and distinct(), which were helpful in segregating the huge dataset to a very small data frame and plot the graph.

2. Which generation (1st GEN, 2nd GEN or 3rd GEN) of smokers has highest current cigarette smokers?



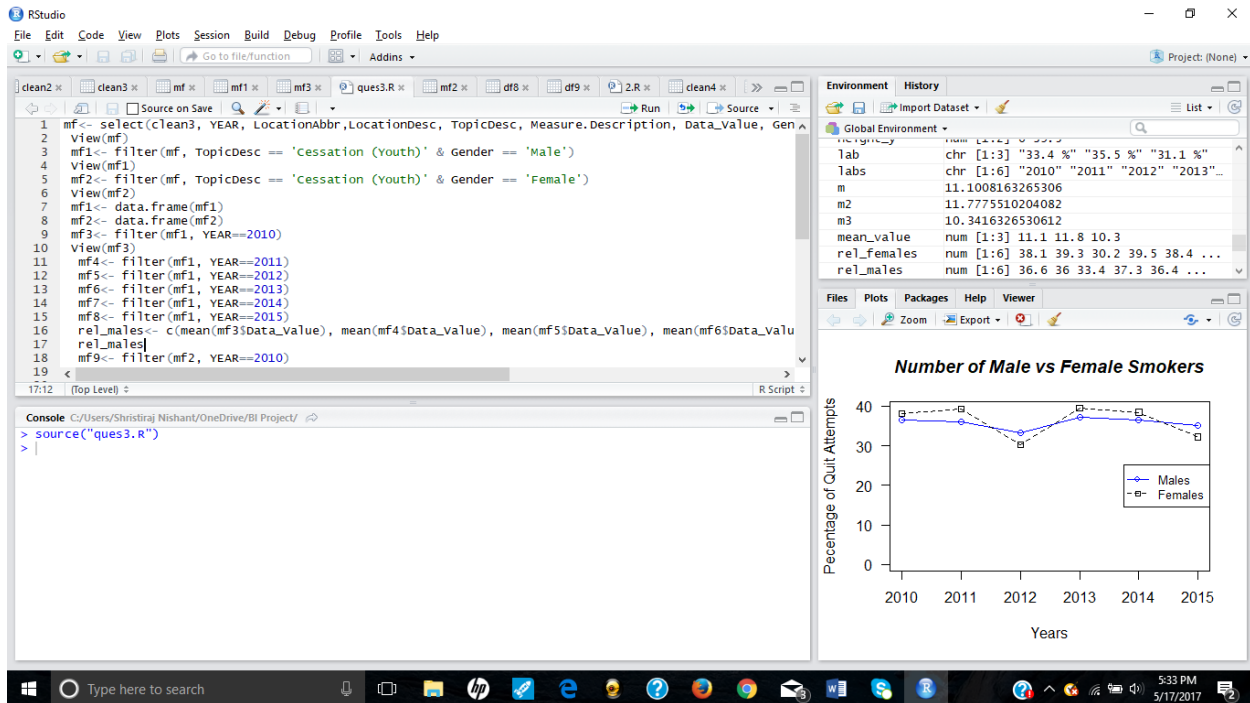
(Highlights: function, pie chart)

Generation Wise Current Cigarette Smokers in USA

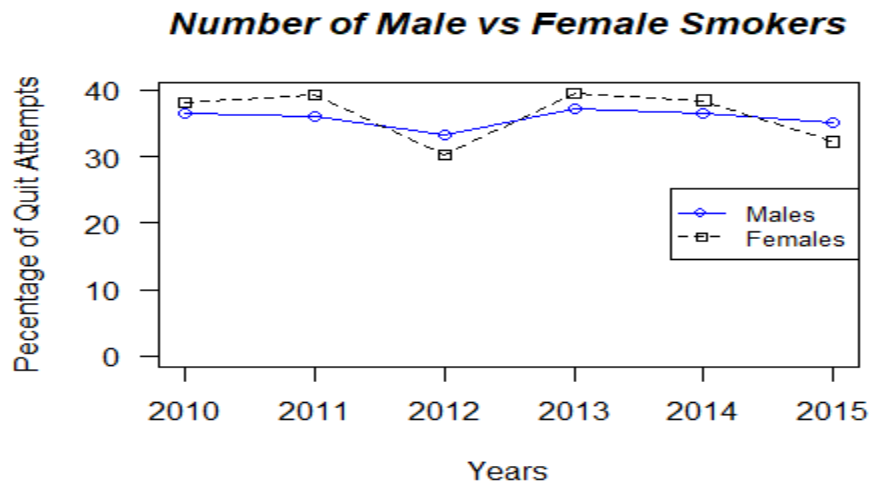


The following pie chart represents the generation wise current cigarette smokers in the whole of USA. This chart uses a user defined function called **func()**. This function takes data frame as an input from the user and calculates the number of smokers from the given percentage of data values for each row. Then it sums up the value for all the rows and returns that sum, i.e., total number of users for that data frame. 1st Generation smokers refers to the people who were the first in their family to start smoking during their middle or high school. Similarly, 2nd generation means there were people in the family(parents) who smoked before them. The chart clearly indicates that from between 2005- 2015, 1st Generation smokers were the highest current cigarette smokers in the whole USA.

3. Which year (from 2010- 2015) has the highest number of smokers (male and female) who tried to quit smoking or had past attempt in the last year?

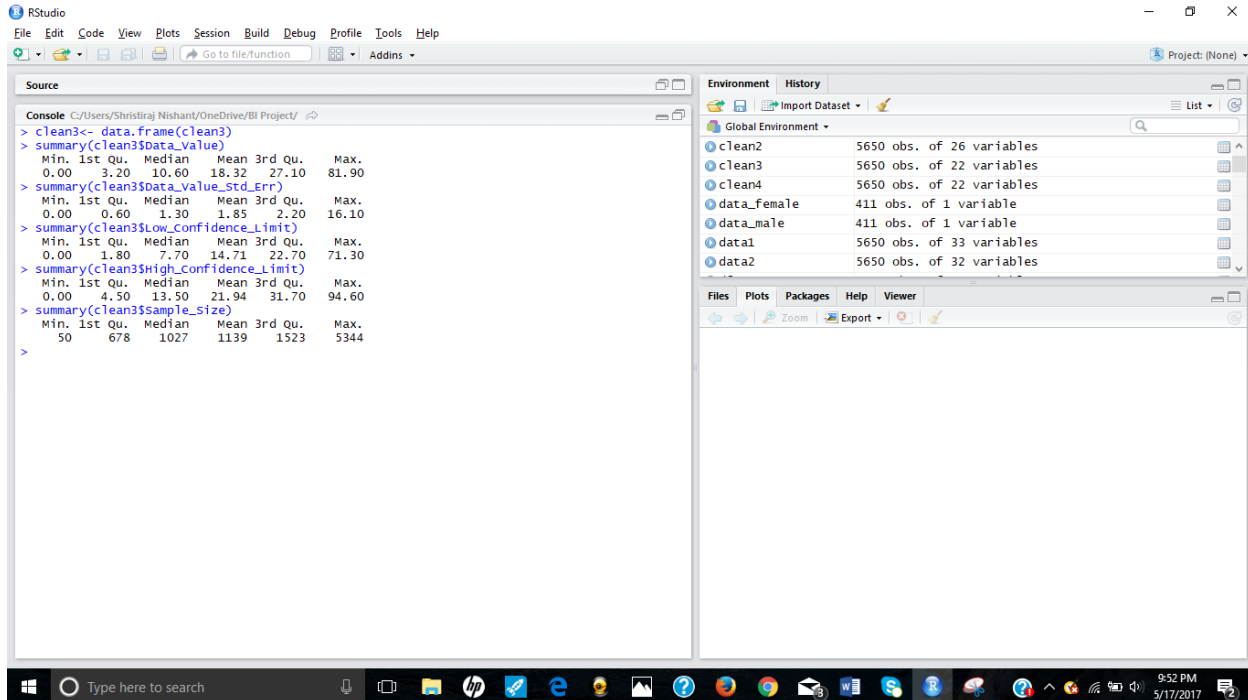


(Highlights: R-script, dplyr package, line chart)



This visualization is executed using a R-script. The code has been written in the script and not in the console of R-studio, and run by using `source("script_name.R")`. The visualization utilizes data frames from different years starting from 2010 till 2015, filtered on male or female and line charts are plotted for the same. It is indicative that Female smokers have high cessation rate than Male smokers in the USA. The highest value was 39.5% for females and 37.3% for males in the year 2013, which indicates that year 2013 saw a high cessation in smoking in USA. The only exceptions are year 2012 and 2015, where male smokers have high cessation rate. The `filter()` function is used to extract the necessary elements from the dataset to make a new and meaningful data frame suitable for visualization.

Statistical Averages



The command `summary(dataset$column_name)` shows different statistical averages of our dataset in value columns. These averages include minimum, maximum, mean, median, 1st Quartile and 3rd Quartile values. The dataset contains 5 numeric columns namely: Data_Value, Data_Value_Std_Err, High_Confidence_Limit, Low_Confidence_Limit and Sample_Size. The statistical averages for all the 5 columns have been shown in the screenshot above.

R CODE FOR ANALYSIS AND VISUALS:

- **Data Cleaning**

```
setwd("C:/Users/anus/OneDrive/BI Project")
```

```
getwd()
```

```
data1<- read.csv("R Proj Data.csv", header = T, sep = ",")
```

```
project<- data.frame(data1)
```

```
View(project)
```

```
install.packages("tidyr")
```

```
library(tidyr)
```

```
project_clean1<- project[, -c(4,7,9,10,20:23,28)]
```

```
View(project_clean1)
```

```
project_clean1<- data.frame(project_clean1)
```

```
clean2<- separate(project_clean1, Race.Age.Education, c("Race", "Age", "Education"), sep = "-"  
)
```

```
View(clean2)
```

```
clean2$Data_Value[is.na(clean2$Data_Value)]= mean(clean2$Data_Value, na.rm = T)
```

```
clean2$Data_Value_Std_Err[is.na(clean2$Data_Value_Std_Err)]= mean(clean2$Data_Value_Std_Err, na.rm = T)
```

```
clean2$Data_Value_Std_Err[is.na(clean2$Data_Value_Std_Err)]= mean(clean2$Data_Value_Std_Err, na.rm = T)
```

```
clean2$High_Confidence_Limit[is.na(clean2$High_Confidence_Limit)]= mean(clean2$High_Confidence_Limit, na.rm = T)
```

```
clean2$Sample_Size[is.na(clean2$Sample_Size)]= mean(clean2$Sample_Size, na.rm = T)
```

```
View(clean2)
```

```
clean3<- data.frame(clean2)
```

```
View(clean3)
```

- **Barplot for Top 5 States in terms of Cessation Rate**

```
install.packages(dplyr)

library(dplyr)

df3<- select(clean3, YEAR, LocationDesc, TopicDesc, Measure.Description, Data_Value)

View(df3)

df4<- filter(df3, TopicDesc == 'Cessation (Youth)')

df4<- data.frame(df4)

View(df4)

df5<- arrange(df4, desc(Data_Value))

View(df5)

df6<- distinct(df5, LocationDesc, TopicDesc, Data_Value)

View(df6)

df7<- head(df6, n= 5)

View(df7)

df6<- distinct(df5, LocationAbbr,LocationDesc, TopicDesc, Data_Value)

df7<- head(df6, n= 5)

View(df7)

c<- c("yellow","turquoise","violetred3","dodgerblue", "goldenrod")

x<- barplot(df7$Data_Value, names.arg = df7$LocationAbbr, xlab = "States", ylab = "Cessation rate", main = "Top 5 States in Terms of Cessation Rate in USA", col = c, ylim = c(0,100))

text(x,5,df7$Data_Value)
```

- **Pie-Chart for Generation Wise Current Cigarette Smokers**

```
View(clean3)

df8<- subset(clean3, Response == 'Current')

View(df8)

df8<- subset(clean3, Response == 'Current' & TopicDesc == 'Cigarette Use (Youth)' & StratificationID1 == '1GEN')

df9<- subset(clean3, Response == 'Current' & TopicDesc == 'Cigarette Use (Youth)' & StratificationID1 == '2GEN')

df10<- subset(clean3, Response == 'Current' & TopicDesc == 'Cigarette Use (Youth)' & StratificationID1 == '3GEN')

m1<- func(df8)

m2<- func(df9)

m2<- func(df10)

v<- data.frame(m1,m2,m3)

slices<- round(v/sum(v)*100,1)

lab<- paste(slices, "%", sep = " ")

pie(v, labels = lab, main = "Generation Wise Current Cigarette Smokers in USA", col = rainbow(length(v)))

legend("topright", c("1 GENERATION", "2 GENERATION", "3 GENERATION"), cex = 0.8, fill = rainbow(length(v)))
```

User Defined Function: func()

```
func<- function(N)

{

a <-round((N$Data_Value*N$Sample_Size)/100)

b<- sum(a)

return(b)

}
```

- **R- Script for Line chart to find number of Male and Female Smokers who tried to quit smoking**

```
library(dplyr)

mf<- select(clean3, YEAR, LocationAbbr,LocationDesc, TopicDesc, Measure.Description,
Data_Value, Gender)

View(mf)

mf1<- filter(mf, TopicDesc == 'Cessation (Youth)' & Gender == 'Male')

View(mf1)

mf2<- filter(mf, TopicDesc == 'Cessation (Youth)' & Gender == 'Female')

View(mf2)

mf1<- data.frame(mf1)

mf2<- data.frame(mf2)

mf3<- filter(mf1, YEAR==2010)

View(mf3)

mf4<- filter(mf1, YEAR==2011)

mf5<- filter(mf1, YEAR==2012)

mf6<- filter(mf1, YEAR==2013)

mf7<- filter(mf1, YEAR==2014)

mf8<- filter(mf1, YEAR==2015)

rel_males<- c(mean(mf3$Data_Value), mean(mf4$Data_Value), mean(mf5$Data_Value),
mean(mf6$Data_Value),mean(mf7$Data_Value), mean(mf8$Data_Value))

rel_males

mf9<- filter(mf2, YEAR==2010)

mf10<- filter(mf2, YEAR==2011)

mf11<- filter(mf2, YEAR==2012)

mf12<- filter(mf2, YEAR==2013)

mf13<- filter(mf2, YEAR==2014)

mf14<- filter(mf2, YEAR==2015)
```

```

rel_females<- c(mean(mf9$Data_Value), mean(mf10$Data_Value), mean(mf11$Data_Value),
mean(mf12$Data_Value),mean(mf13$Data_Value), mean(mf14$Data_Value))

rel_females

height_y<- range(0,rel_males, rel_females)

labs<- c("2010", "2011", "2012","2013","2014","2015")

plot(rel_males, type = "o", col= "blue", ylim = height_y, axes = F, ann = F)

axis(1, at=1:6, labels = labs)

axis(2, las=1, at=10*0:height_y[2])

lines(rel_females, type="o", pch=22, lty=2, col="black")

title(main="Number of Male vs Female Smokers", col.main="Black", font.main=4, xlab =
"Years", ylab = "Percentage of Quit Attempts")

box()

legend("right", c("Males","Females"), cex=0.8, col=c("blue", "black"), pch=21:22, lty=1:2)

```

Execution of script:

```
source("program.R")
```

REFERENCES

- [1] "Function in R, passing a dataframe and a column name." *Function in R, passing a dataframe and a column name - Stack Overflow*. N.p., n.d. Web. 18 May 2017.
- [2] "Top Questions." *Stack Overflow*. N.p., n.d. Web. 18 May 2017.
- [3] *Introduction to dplyr*. N.p., 23 June 2016. Web. 18 May 2017.