EDA1

Vishaka Jhunjhunwala

October 26, 2019

## 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:

## Exploratory Data Analysis of Drug Review Dataset.

## Review of variables,attributes and their Data types

#install.packages("tidyverse")  
library(tidyverse)

## Warning: package 'tidyverse' was built under R version 3.5.3

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

## v ggplot2 3.1.0 v purrr 0.3.1  
## v tibble 2.0.1 v dplyr 0.8.3  
## v tidyr 0.8.3 v stringr 1.3.1  
## v readr 1.3.1 v forcats 0.4.0

## Warning: package 'ggplot2' was built under R version 3.5.3

## Warning: package 'tibble' was built under R version 3.5.3

## Warning: package 'tidyr' was built under R version 3.5.3

## Warning: package 'readr' was built under R version 3.5.3

## Warning: package 'purrr' was built under R version 3.5.3

## Warning: package 'dplyr' was built under R version 3.5.3

## Warning: package 'forcats' was built under R version 3.5.3

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

train\_data <- read.delim("C:/Users/saraz/Downloads/drugsCom\_raw/drugsComTrain\_raw.tsv", sep = '\t', header = TRUE,col.names = c("ID","Name","Condition","review","rating","Date","Hit\_count"))  
View(train\_data)  
summary(train\_data)

## ID Name   
## Min. : 2 Levonorgestrel : 3657   
## 1st Qu.: 58063 Etonogestrel : 3336   
## Median :115744 Ethinyl estradiol / norethindrone : 2850   
## Mean :115924 Nexplanon : 2156   
## 3rd Qu.:173776 Ethinyl estradiol / norgestimate : 2117   
## Max. :232291 Ethinyl estradiol / levonorgestrel: 1888   
## (Other) :145293   
## Condition review rating   
## Birth Control : 28788 "Good" : 33 Min. : 1.000   
## Depression : 9069 "Great" : 20 1st Qu.: 5.000   
## Pain : 6145 "Good." : 18 Median : 8.000   
## Anxiety : 5904 "Very good" : 16 Mean : 6.994   
## Acne : 5588 "Great." : 14 3rd Qu.:10.000   
## Bipolar Disorde: 4224 "Works well.": 12 Max. :10.000   
## (Other) :101579 (Other) :161184   
## Date Hit\_count   
## March 1, 2016 : 146 Min. : 0   
## September 23, 2015: 141 1st Qu.: 6   
## January 13, 2016 : 138 Median : 16   
## January 7, 2017 : 136 Mean : 28   
## April 25, 2017 : 135 3rd Qu.: 36   
## August 3, 2016 : 135 Max. :1291   
## (Other) :160466

## Checking missing Data

sum(is.na(train\_data))

## [1] 0

## Grouping Data by different attributes and visualizing plots

length(unique(train\_data$Condition))

## [1] 885

# There are 885 unique medical conditions that are reviewed by patients in this dataset.

## Number of drugs reviewed per condition

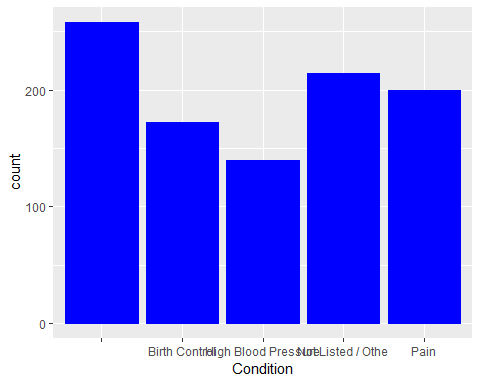
Data\_by\_condition<-train\_data%>%group\_by(Condition)%>%summarise(count = n\_distinct(Name))%>%arrange(desc(count))  
Data\_by\_condition %>% top\_n(10)

## Selecting by count

## # A tibble: 10 x 2  
## Condition count  
## <fct> <int>  
## 1 "" 258  
## 2 Not Listed / Othe 214  
## 3 Pain 200  
## 4 Birth Control 172  
## 5 High Blood Pressure 140  
## 6 Acne 117  
## 7 Depression 105  
## 8 Rheumatoid Arthritis 98  
## 9 Diabetes, Type 2 89  
## 10 Allergic Rhinitis 88

Data\_by\_condition %>% top\_n(5) %>% ggplot(aes(Condition,count))+geom\_col(show.legend=FALSE, color = "blue", fill = "blue")

## Selecting by count

 ## Number of times each drug is reviewed

length(unique(train\_data$Name))

## [1] 3436

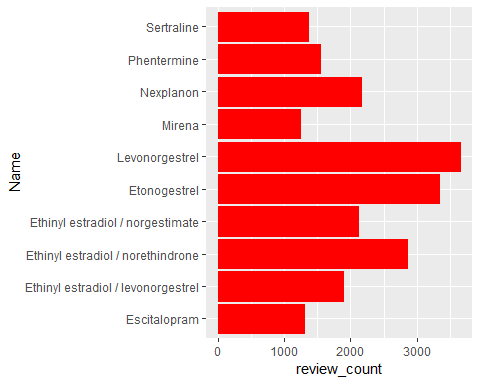
# There are 3436 unique drug names in this dataset   
  
# Plot of top 10 Drugs along with their review count  
  
Data\_By\_drug= train\_data %>% group\_by(Name) %>% summarise(review\_count= n\_distinct(ID))%>%arrange(desc(review\_count))  
Data\_By\_drug %>% top\_n(10)

## Selecting by review\_count

## # A tibble: 10 x 2  
## Name review\_count  
## <fct> <int>  
## 1 Levonorgestrel 3657  
## 2 Etonogestrel 3336  
## 3 Ethinyl estradiol / norethindrone 2850  
## 4 Nexplanon 2156  
## 5 Ethinyl estradiol / norgestimate 2117  
## 6 Ethinyl estradiol / levonorgestrel 1888  
## 7 Phentermine 1543  
## 8 Sertraline 1360  
## 9 Escitalopram 1292  
## 10 Mirena 1242

Data\_By\_drug %>% top\_n(10) %>% ggplot(aes(Name,review\_count))+geom\_col(show.legend=FALSE, color = "red", fill = "red")+coord\_flip()

## Selecting by review\_count

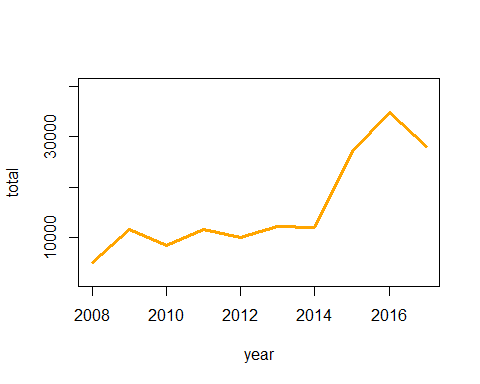


## Plotting reviews based on time

train\_data$Date= as.Date(train\_data$Date,format = "%B%d,%Y")  
  
#Review count per year  
  
rc\_year<-train\_data %>%mutate(year = format(train\_data$Date, "%Y")) %>% group\_by(year)%>% summarise(total = n())  
  
rc\_year

## # A tibble: 10 x 2  
## year total  
## <chr> <int>  
## 1 2008 5137  
## 2 2009 11636  
## 3 2010 8413  
## 4 2011 11682  
## 5 2012 10007  
## 6 2013 12278  
## 7 2014 12051  
## 8 2015 27164  
## 9 2016 34842  
## 10 2017 28087

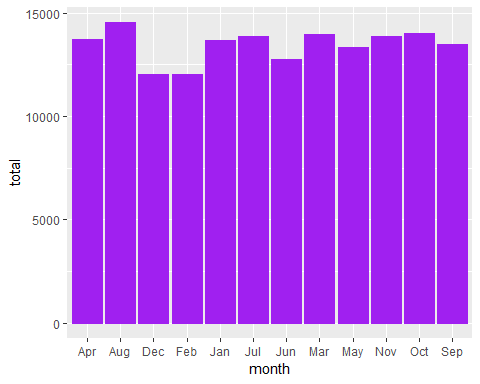
plot(rc\_year,type = "l",lwd=3,col ="Orange",xlab ="year",ylab="total",ylim=c(2000, 40000))



# Review Count per Month  
  
  
rc\_month<-train\_data %>%mutate(month = format(train\_data$Date, "%b")) %>% group\_by(month)%>% summarise(total = n())  
  
rc\_month

## # A tibble: 12 x 2  
## month total  
## <chr> <int>  
## 1 Apr 13736  
## 2 Aug 14554  
## 3 Dec 12019  
## 4 Feb 12041  
## 5 Jan 13657  
## 6 Jul 13852  
## 7 Jun 12749  
## 8 Mar 13956  
## 9 May 13345  
## 10 Nov 13875  
## 11 Oct 14038  
## 12 Sep 13475

rc\_month%>%ggplot(aes(month,total))+geom\_col(show.legend=FALSE, color = "purple", fill = "purple")



Data\_By\_hitcount<- train\_data %>% group\_by(Name) %>% summarise(count= n\_distinct(train\_data$Hit\_count))%>%arrange(desc(count))

## Plotting top 10 drugs based on hitcount

Data\_By\_hitcount<- train\_data %>% group\_by(Name,Hit\_count)  
Data\_By\_hitcount\_top10<- aggregate(Hit\_count~Name,Data\_By\_hitcount,sum)%>%arrange(desc(Hit\_count))%>%top\_n(10)

## Selecting by Hit\_count

Data\_By\_hitcount\_top10%>%ggplot(aes(Name,Hit\_count))+geom\_col(show.legend=FALSE, color = "Brown", fill ="Brown")+coord\_flip()

