R Notebook

Load the libraries

```
library(tidyverse)
```

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

```
## v ggplot2 3.1.0     v purrr   0.2.5
## v tibble  2.0.1     v dplyr   0.7.8
## v tidyr   0.8.2     v stringr 1.3.1
## v readr   1.3.1     v forcats 0.3.0
```

```
## -- Conflicts --------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Read the data

```
train<- read.csv("drugsComTrain_raw.csv", stringsAsFactors = FALSE)
glimpse(train)
```

```
## Observations: 161,297
## Variables: 7
## $ uniqueID    <int> 206461, 95260, 92703, 138000, 35696, 155963, 16590...
## $ drugName    <chr> "Valsartan", "Guanfacine", "Lybrel", "Ortho Evra",...
## $ condition   <chr> "Left Ventricular Dysfunction", "ADHD", "Birth Con...
## $ review      <chr> "\"It has no side effect, I take it in combination...
## $ rating      <int> 9, 8, 5, 8, 9, 2, 1, 10, 1, 8, 9, 10, 4, 4, 3, 9, ...
## $ date        <chr> "20-May-12", "27-Apr-10", "14-Dec-09", "3-Nov-15",...
## $ usefulCount <int> 27, 192, 17, 10, 37, 43, 5, 32, 11, 1, 19, 54, 8, ...
```

Group Data by conditions and removing absurd and missing values of condition

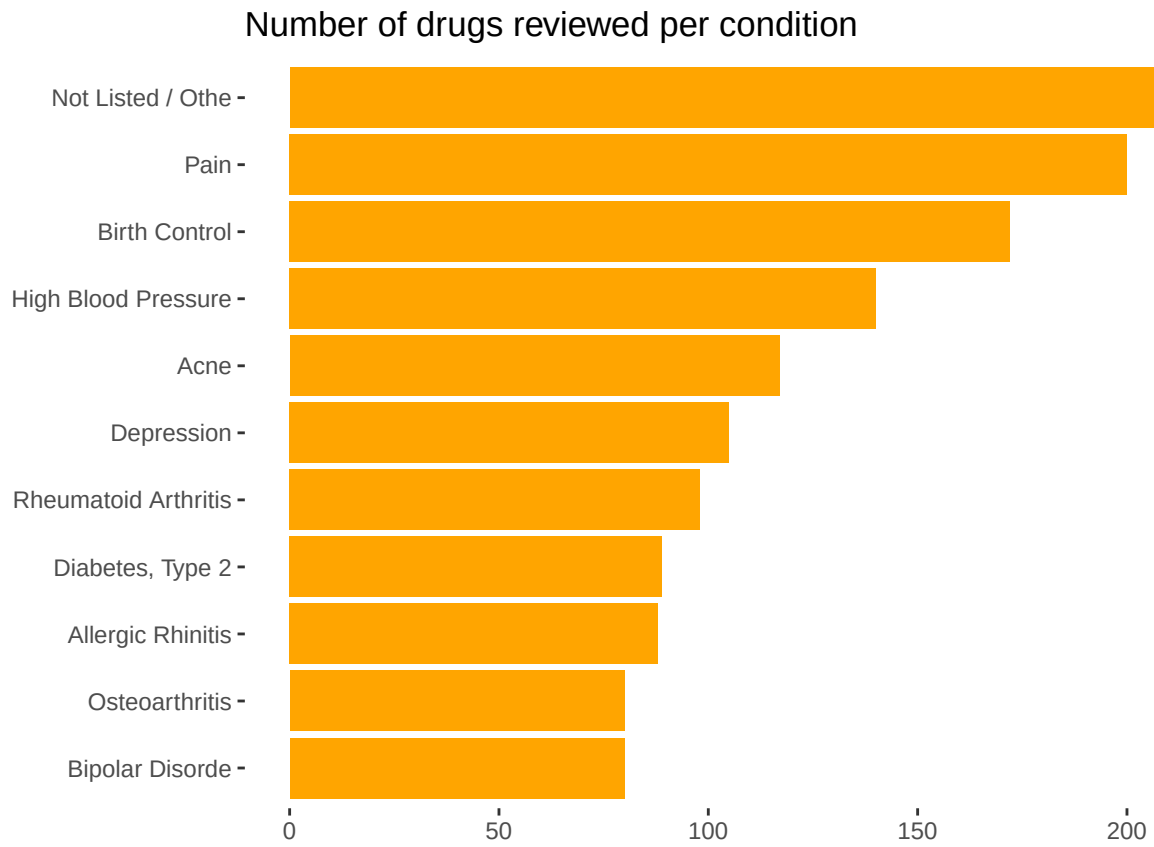
```
length(unique(train$condition))
```

```
## [1] 885
```

```
Bycondition= train %>% group_by(condition) %>% filter(!grepl("^[0-9]", condition)) %>% filter(!condition
Bycondition %>% arrange(desc(number_of_drugs))
```

```
## # A tibble: 811 x 2
##    condition            number_of_drugs
##    <chr>                          <int>
##  1 Not Listed / Othe                214
##  2 Pain                             200
##  3 Birth Control                    172
##  4 High Blood Pressure              140
##  5 Acne                             117
##  6 Depression                       105
##  7 Rheumatoid Arthritis              98
##  8 Diabetes, Type 2                  89
##  9 Allergic Rhinitis                 88
## 10 Bipolar Disorde                   80
## # ... with 801 more rows
```

```
Bycondition %>% top_n(10, number_of_drugs) %>%
  ggplot()+geom_bar(aes(x = reorder(condition, number_of_drugs,sum), y= number_of_drugs ), stat= "identi
panel.background = element_blank(), axis.line = element_blank())+coord_flip()
```

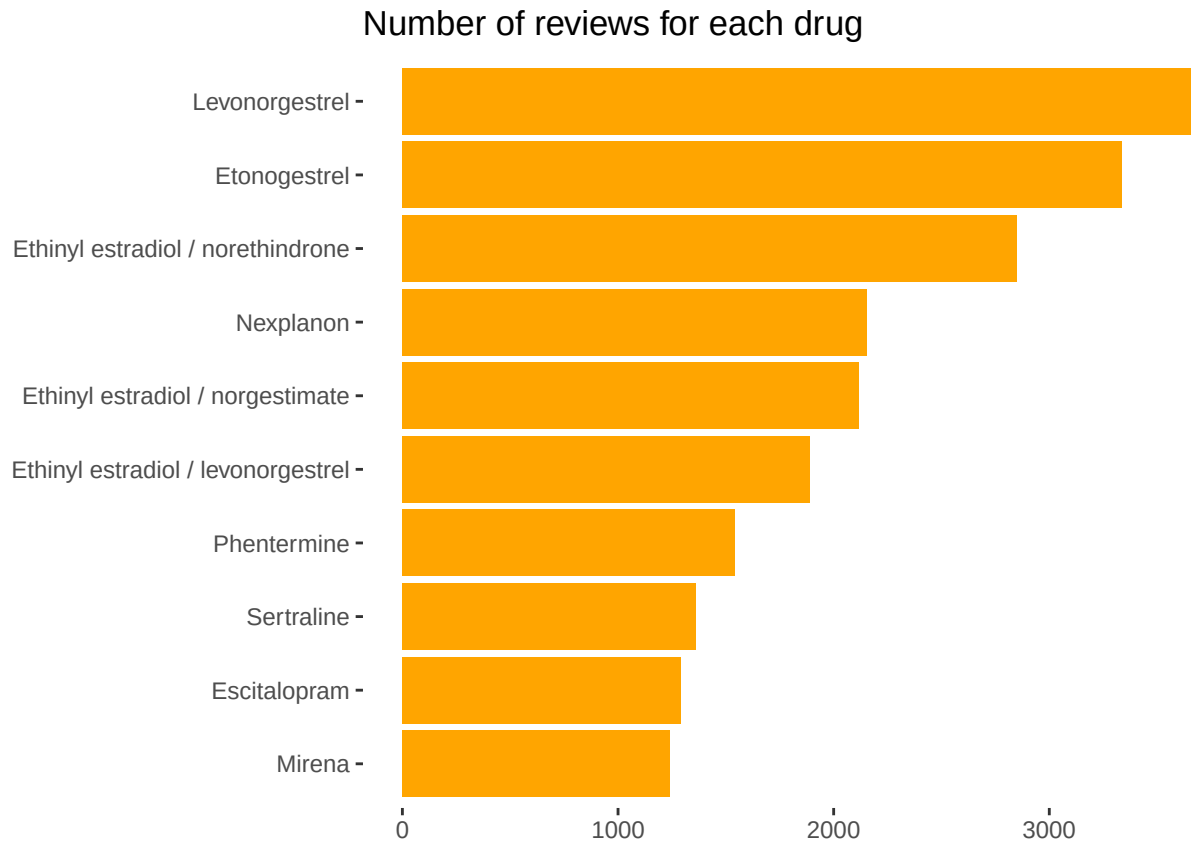


Number of times each drug is reviewed and its average rating.

```
length(unique(train$drugName))
```

```
## [1] 3436
```

```
Bydrug= train %>% group_by(drugName) %>% summarise(number_of_reviews= n_distinct(uniqueID), average_rati
# plot of top reviewed drugs
Bydrug %>% top_n(10, number_of_reviews) %>% ggplot()+geom_bar(aes(x = reorder(drugName, number_of_reviev
```



Top rated drugs.(Overall average)

```
Bydrug %>% top_n(10, average_rating)
```

```
## # A tibble: 490 x 3
##    drugName                                number_of_revie~ average_rating
##    <chr>                                              <int>          <dbl>
##  1 A / B Otic                                             1             10
##  2 A + D Cracked Skin Relief                              1             10
##  3 Absorbine Jr.                                          1             10
##  4 Accolate                                               2             10
##  5 Acetaminophen / caffeine / magnesium sa~               1             10
##  6 Acetaminophen / dextromethorphan / doxy~               1             10
##  7 Acetaminophen / phenylephrine                          2             10
##  8 Acetaminophen / pseudoephedrine                        7             10
##  9 Acetic acid / antipyrine / benzocaine /~               1             10
## 10 Acrivastine / pseudoephedrine                          1             10
## # ... with 480 more rows
```

Top rated drugs as per condition

```
Bycondition_drug= train %>% group_by(condition, drugName) %>% filter(!grepl("^[0-9]", condition)) %>% f
```

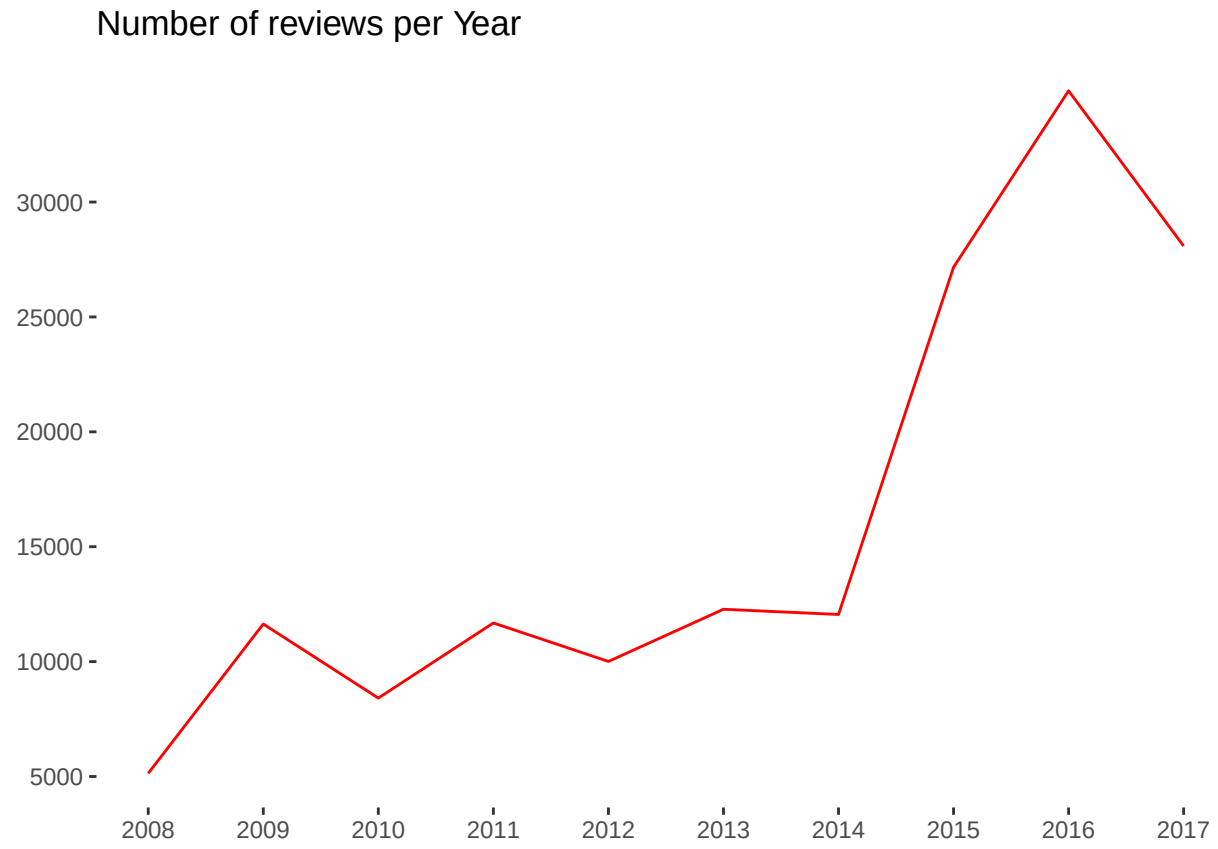
Great!!! Now we have a table that gives highest rated (average) drug according to the condition.

Now lets discuss reviews by time. First format the date and then extract year month and day fromit.

```
Bydate = train%>%mutate(date = as.Date(date,format="%d-%B-%y"))%>% mutate(month=as.numeric(format(date,
```

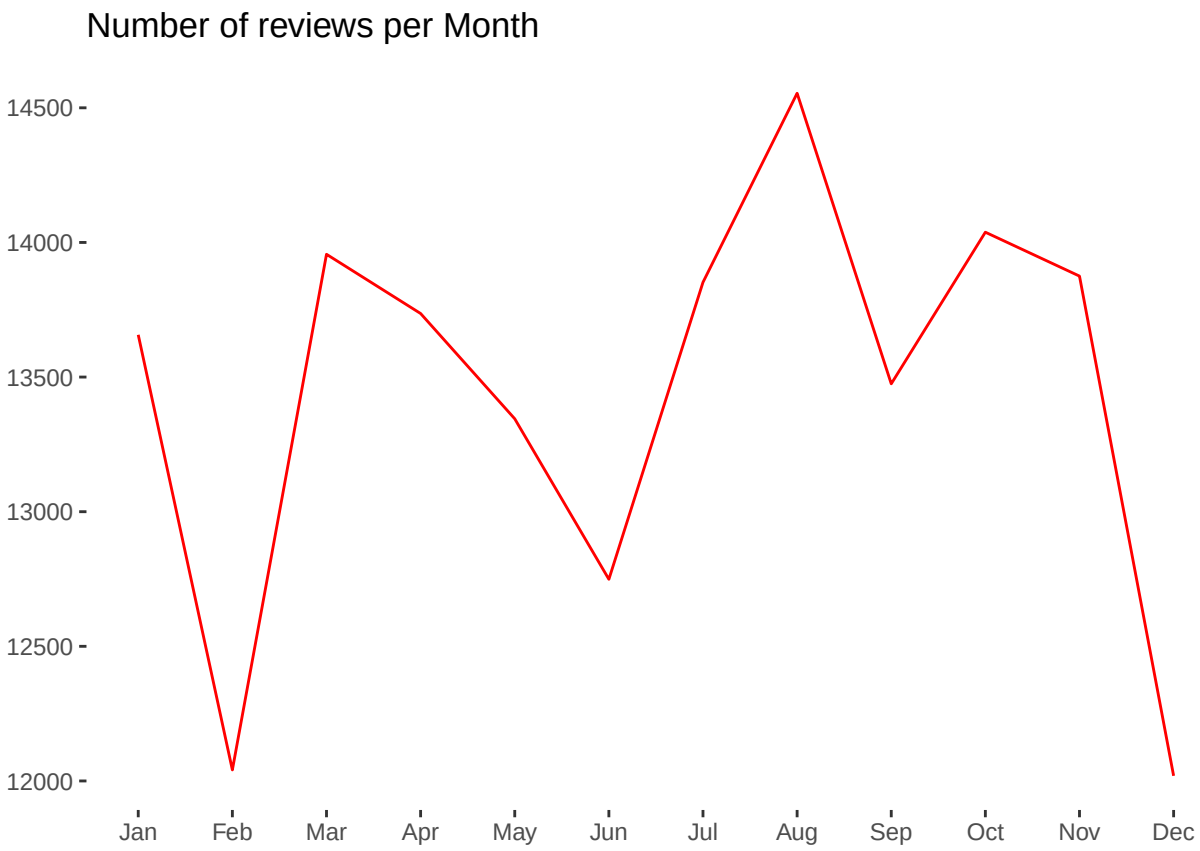
Number of Reviews as per year

```
Bydate%>% group_by(year)%>%summarise(no_of_reviews= n())%>% ggplot()+geom_line(aes(x= year, y= no_of_rev
```



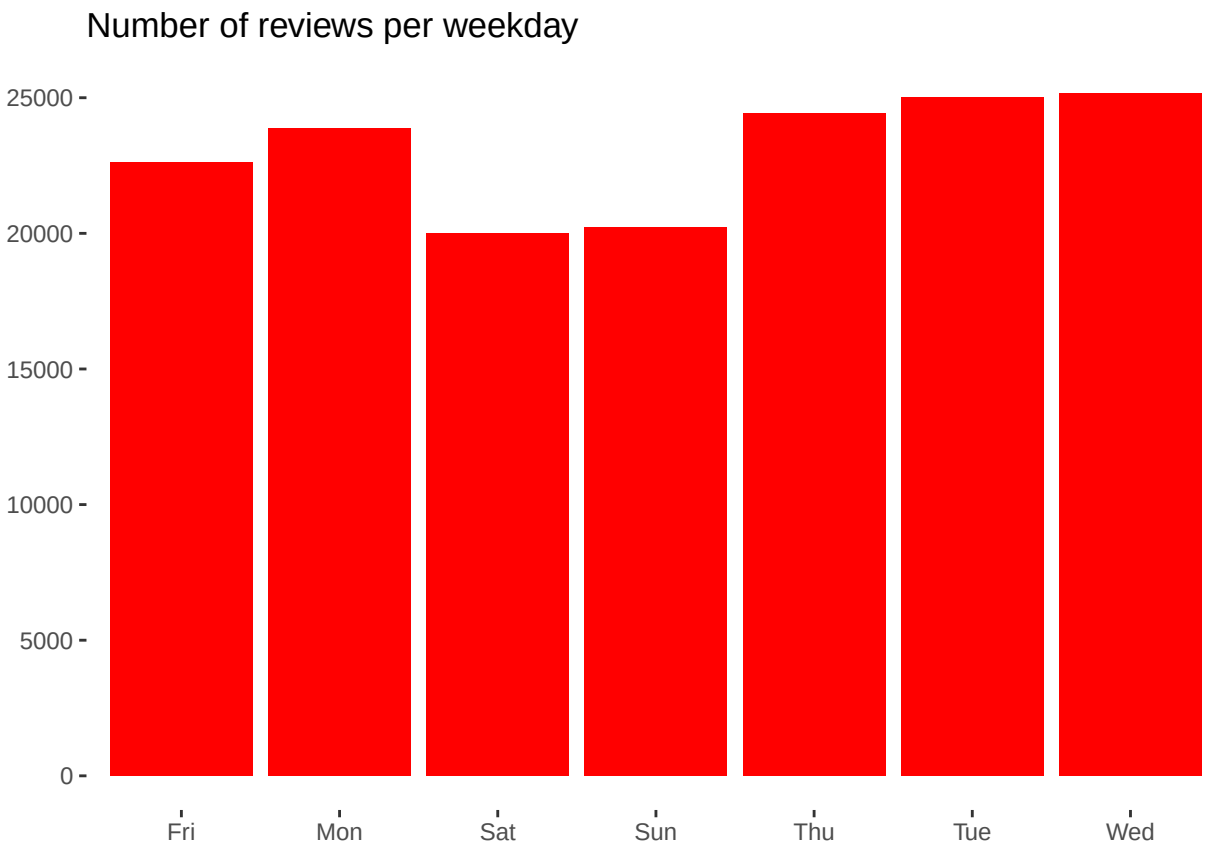
Number of Reviews as per month

```
Bydate %>% group_by(month) %>% summarise(no_of_reviews= n())%>% ggplot() + geom_line(aes(x= month, y= no
```



Number of Reviews as per weekday

```
Bydate%>% group_by(weekday)%>%summarise(no_of_reviews= n())%>% arrange() %>% ggplot()+geom_bar(aes(x= w
```


Number of reviews per weekday
25000
20000
15000
10000
5000
0
Fri
Mon
Sat
Sun
Thu
Tue
Wed