Categorical Data Analysis with rgates

Muhammad Ezzat

3/14/2020

Inroduction

This is a tutorial for categorical data analysis with rgates package, if you haven't checked it out yet here's a github link : https://github.com/MuhammadEzzatHBK/rgates . It's currently on Github only but I'm looking forward to a CRAN publish. Without any further to do let's get started. For this tutorial I'm using the Netflix shows dataset it has useful relatable easy to understand categorical & numerical variables that we can work with.

A link to the dataset : https://www.kaggle.com/shivamb/netflix-shows .

```r
summary(netflix)
```

```
##     show_id             type              title          
##  Min.   :  247747   Length:3774        Length:3774       
##  1st Qu.:70275815   Class :character   Class :character  
##  Median :80147322   Mode  :character   Mode  :character  
##  Mean   :75109075                                        
##  3rd Qu.:80240670                                        
##  Max.   :81235729                                        
##    director             cast             country         
##  Length:3774        Length:3774        Length:3774       
##  Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character  
##                                                          
##                                                          
##                                                          
##   date_added         release_year     rating               dur       
##  Length:3774        Min.   :1942   Length:3774        Min.   :  1.0  
##  Class :character   1st Qu.:2011   Class :character   1st Qu.: 87.0  
##  Mode  :character   Median :2016   Mode  :character   Median : 99.0  
##                     Mean   :2012                      Mean   : 99.9  
##                     3rd Qu.:2017                      3rd Qu.:117.0  
##                     Max.   :2020                      Max.   :228.0  
##      unit            listed_in         description       
##  Length:3774        Length:3774        Length:3774       
##  Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character  
##                                                          
##                                                          
##                                                          
```

Using rgates in filtering data

We all know the fliter function from the dplyr package, with rgates we can create more complex yet powerful logical conditions that we can filter our data based on in the simple dplyr framework.

```r
comedy_tv_shows <- filter(netflix, and(netflix$type=="TV Show",grepl('Comedies',netflix$listed_in)))
head(comedy_tv_shows,3)
```

```
## # A tibble: 3 x 13
##   show_id type  title director cast  country date_added release_year rating
##     <dbl> <chr> <chr> <chr>    <chr> <chr>   <chr>             <dbl> <chr>
## 1  8.02e7 TV S~ Come~ Jerry S~ Jerr~ United~ July 19, ~         2019 TV-14
## 2  8.02e7 TV S~ Dave~ Stan La~ Dave~ United~ March 21,~         2017 TV-MA
## 3  8.02e7 TV S~ Dave~ Stan La~ Dave~ United~ December ~         2017 TV-MA
## # ... with 4 more variables: dur <dbl>, unit <chr>, listed_in <chr>,
## #   description <chr>
```

Here we used the and() gate/function to filter the data for comdey tv shows.There are more advanced gates/filters, such as the inhibit() gate. It works by the term “X but not Y” so it returns TRUE only if X is so & Y isn’t. We can use it to extract non romantic movies like follows.

```
non_romantic_movies <- filter(netflix,inhibit(netflix$type =='Movie',grepl('Romantic',netflix$listed_in
head(non_romantic_movies,3)
```

```
## # A tibble: 3 x 13
##   show_id type  title director cast  country date_added release_year rating
##     <dbl> <chr> <chr> <chr>    <chr> <chr>   <chr>             <dbl> <chr>
## 1  8.01e7 Movie #rea~ Fernand~ Nest~ United~ September~         2017 TV-14
## 2  8.11e7 Movie #Sel~ Cristin~ Flav~ Romania June 1, 2~         2014 TV-MA
## 3  8.11e7 Movie #Sel~ Cristin~ Maia~ Romania June 1, 2~         2016 TV-MA
## # ... with 4 more variables: dur <dbl>, unit <chr>, listed_in <chr>,
## #   description <chr>
```

We can even chain gates inside the same filter function, in the next chunk what I’m showing you is basically an inhibit() gate running inside an and() gate, so the result coming from inhibit() is going inside the and() as one of it’s two inputs to extract non Drama movies produced in the United States.

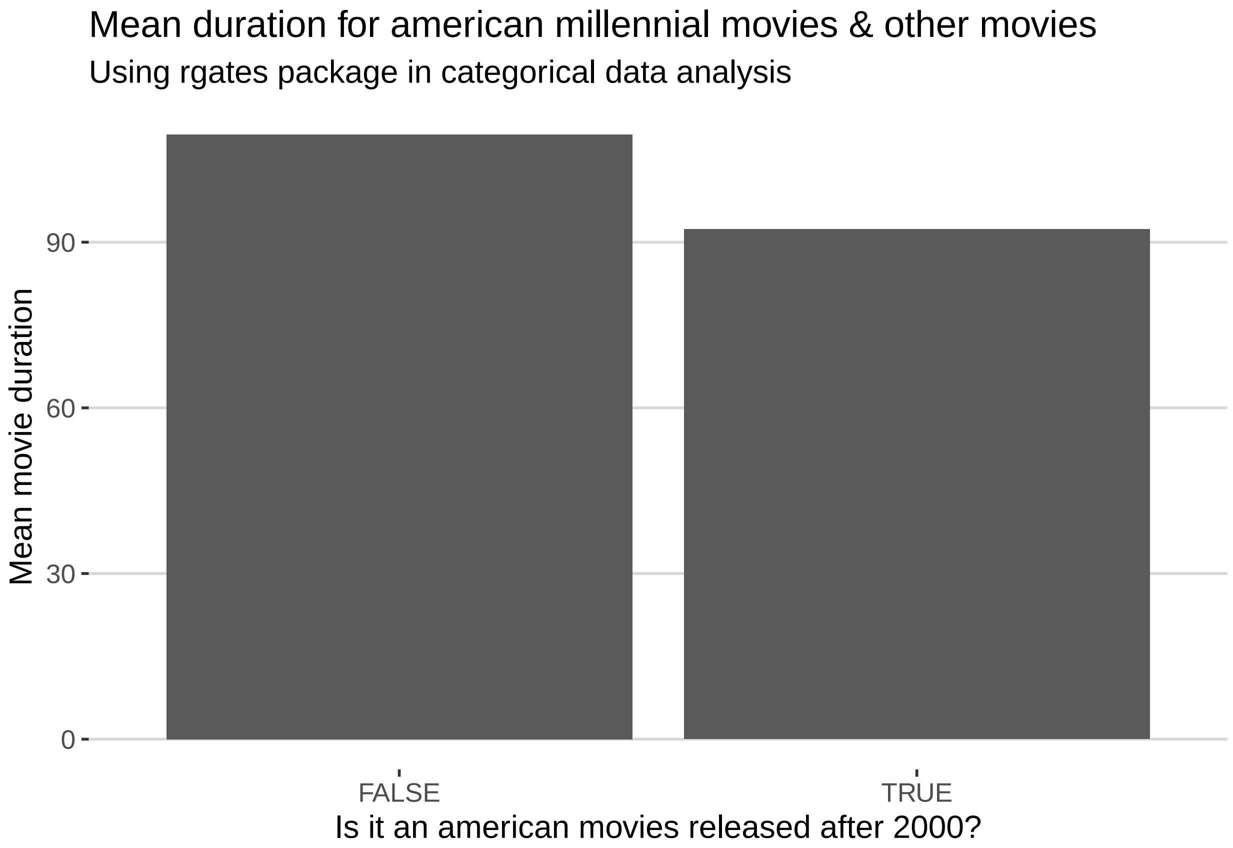
```
USA_nonDramaMovies <- filter(netflix,and(netflix$country=='United States',
                                         inhibit(netflix$type =='Movie',grepl('Drama',netflix$listed_in
head(USA_nonDramaMovies,3)
```

```
## # A tibble: 3 x 13
##   show_id type  title director cast  country date_added release_year rating
##     <dbl> <chr> <chr> <chr>    <chr> <chr>   <chr>             <dbl> <chr>
## 1  8.01e7 Movie #rea~ Fernand~ Nest~ United~ September~         2017 TV-14
## 2  8.01e7 Movie 13 C~ Victor ~ PJ M~ United~ August 13~         2015 NR
## 3  7.03e7 Movie 13 S~ Daniel ~ Mark~ United~ January 1~         2014 R
## # ... with 4 more variables: dur <dbl>, unit <chr>, listed_in <chr>,
## #   description <chr>
```

Using rgates in various analysis tasks

If you thought about it, this package is just a group of functions that produces logical/boolean vectors, surely it’s merge & integration in frameworks and pipelines is feasable.The next chunk shows a pipeline integrated with rgates functions, don’t be intimidated by it as I’ll break it down step by step.

```
filter(netflix,netflix$type=='Movie')%>%
        mutate(is_american_millennial =
            and(release_year>=2000,grepl('United States',country)))%>%
                group_by(is_american_millennial) %>% summarize(mean_duration = mean(dur))%>%
  ggplot(aes(x=is_american_millennial,y=mean_duration))+geom_col()+
                theme_hc()+theme(panel.grid.major.x = element_blank())+
  xlab('Is it an american movies released after 2000?')+
  ylab('Mean movie duration')+
  ggtitle('Mean duration for american millennial movies & other movies',
        subtitle = 'Using rgates package in categorical data analysis')
```



First of all we filter our data for movies with a single condition, although we could use the transefer() gate but we don't really need that.

Then we create a new column with the mutate function, that column is a logical column which basically an and() gate for two conditions regarding the place & time of movie release (USA & 2000's).

Then we can simply group by this column as it only has two values TRUE & FALSE. TRUE means it is an american millennial movie while FALSE means it isn't.

Then we summarize for the mean movie duration for both movie groups. And of course drawing a plot is better in conveying information.

So we drew a bar plot with the summarized data, rest of the chunk is just adding themes & titles. It can't win the ggplo2 beauty competition but it does the job.

Practice

If you are really intrested in such topic I suggest that you download the package & the dataset from links above **NOW**. Start practicing by recreating this pipelines or even creating your own pipelines using other real-world data. Until next time friends. See you again.