

# CompleteChart.R

erikchen

2020-02-24

```
# Data Visualization - EChen & CDeLaney 2/23/2020  
# Complete Chart Assignment  
# Start Code
```

```
# Load packages ggplot, scales, dplyr, extrafont  
library(ggplot2)  
library(scales)  
library(dplyr)
```

```
##  
## Attaching package: 'dplyr'  
  
## The following objects are masked from 'package:stats':  
##  
##   filter, lag  
  
## The following objects are masked from 'package:base':  
##  
##   intersect, setdiff, setequal, union
```

```
# load 4 movie data files, all sourced from imdb.com  
# this file contains unique movie title ids, actual title names, year produced, genre  
main.movie.file <- read.csv("/Users/erikchen/Desktop/Data Viz Programs/Policy Brief/Data Files/title.ba  
# this file contains unique movie title ids, average ratings, and number of votes  
movies.ratings <- read.csv("/Users/erikchen/Desktop/Data Viz Programs/Policy Brief/Data Files/title.rat  
# this file contains unique actor ids and their associated role in a film  
movies.staractor <- read.csv("/Users/erikchen/Desktop/Data Viz Programs/Policy Brief/Data Files/title.p  
# this file contains the actual names of the individual actors  
actors.names <- read.csv("/Users/erikchen/Desktop/Data Viz Programs/Policy Brief/Data Files/name.basics  
  
# look at the column names of main file and movie ratings file  
str(main.movie.file)
```

```
## 'data.frame':   1048575 obs. of  9 variables:  
## $ tconst       : Factor w/ 1048575 levels "tt0000001","tt0000002",...: 1 2 3 4 5 6 7 8 9 10 ...  
## $ titleType    : Factor w/ 10 levels "movie","short",...: 2 2 2 2 2 2 2 2 1 2 ...  
## $ primaryTitle : Factor w/ 741621 levels "_grau","- Estrés + Chill",...: 103269 364906 471182 69332  
## $ originalTitle : Factor w/ 748043 levels "_grau","- Estrés + Chill",...: 102966 373426 482576 69819  
## $ isAdult      : int   0 0 0 0 0 0 0 0 0 0 ...  
## $ startYear    : Factor w/ 136 levels "\\N","1888","1889",...: 8 6 6 6 7 8 8 8 8 9 ...  
## $ endYear      : Factor w/ 80 levels "\\N","1933","1938",...: 1 1 1 1 1 1 1 1 1 1 ...  
## $ runtimeMinutes: Factor w/ 619 levels "\\N","0","1",...: 3 445 376 37 3 3 3 3 416 3 ...  
## $ genres       : Factor w/ 1728 levels "\\N","Action",...: 1192 673 576 673 966 1709 1710 1192 1681
```

```
str(movies.ratings)
```

```
## 'data.frame': 1028136 obs. of 3 variables:
## $ tconst : Factor w/ 1028136 levels "tt00000001","tt00000002",...: 1 2 3 4 5 6 7 8 9 10 ...
## $ averageRating: num 5.6 6.1 6.5 6.2 6.1 5.2 5.4 5.4 5.4 6.9 ...
## $ numVotes : int 1586 192 1249 119 2003 109 628 1721 88 5679 ...
```

```
# merge main movie file and movie ratings file by $tconst values ($tconst is the unique identifier for
movies <- merge(x = main.movie.file, y = movies.ratings, by = "tconst", all = TRUE)
dim(movies)
```

```
## [1] 1638909 11
```

```
summary(movies)
```

```
##          tconst          titleType          primaryTitle
## tt00000001:      1  tvEpisode:503610  Episode #1.1: 2622
## tt00000002:      1  movie      :232174  Episode #1.2: 2035
## tt00000003:      1  short      :140788  Episode #1.3: 1950
## tt00000004:      1  video      : 59053  Episode #1.4: 1855
## tt00000005:      1  tvSeries : 45429  Episode #1.5: 1749
## tt00000006:      1  (Other)  : 67521  (Other)      :1038364
## (Other) :1638903  NA's      :590334  NA's          : 590334
##          originalTitle          isAdult          startYear          endYear
## Episode #1.1: 2622  Min.      :0      2006      : 64083  \\N      :1032789
## Episode #1.2: 2035  1st Qu.:0      2005      : 61771  2004      : 687
## Episode #1.3: 1950  Median :0      2004      : 52966  2005      : 675
## Episode #1.4: 1855  Mean    :0      2003      : 43701  2003      : 599
## Episode #1.5: 1749  3rd Qu.:0      2002      : 36468  2001      : 574
## (Other)      :1038364  Max.    :1      (Other):789586  (Other): 13251
## NA's          : 590334  NA's      :590334  NA's          :590334  NA's          : 590334
## runtimeMinutes          genres          averageRating          numVotes
## \\N      :530468  Drama      :104260  Min.      : 1.0  Min.      : 5.0
## 30      : 49679  Comedy     : 87448  1st Qu.: 6.1  1st Qu.: 8.0
## 60      : 38020  \\N        : 83850  Median : 7.1  Median : 19.0
## 90      : 21583  Documentary: 50441  Mean    : 6.9  Mean    : 947.9
## 22      : 13263  Short      : 37516  3rd Qu.: 7.9  3rd Qu.: 75.0
## (Other):395562  (Other)    :685060  Max.    :10.0  Max.    :2192705.0
## NA's      :590334  NA's      :590334  NA's      :610773  NA's      :610773
```

```
# take out movies with no ratings
movies <- movies[which(!is.na(movies$averageRating) == TRUE),]
dim(movies)
```

```
## [1] 1028136 11
```

```
# look at the column names of star actors file and actor names file
str(movies.staractor)
```

```
## 'data.frame': 1048575 obs. of 6 variables:
## $ tconst : Factor w/ 119704 levels "tt00000001","tt00000002",...: 1 1 1 2 2 3 3 3 3 4 ...
## $ ordering : int 1 2 3 1 2 1 2 3 4 1 ...
## $ nconst : Factor w/ 250045 levels "nm00000001","nm00000002",...: 225023 3835 87115 159661 220249 1
## $ category : Factor w/ 12 levels "actor","actress",...: 11 7 5 7 6 7 9 6 8 7 ...
## $ job : Factor w/ 11964 levels "'It All Came True'",...: 3 3 2053 3 3 3 9344 3 3 3 ...
## $ characters: Factor w/ 307659 levels "[\\\"?\\\"]","[\\\". (Ep.: Ausflug in den Wald\\\"]",...: 122219 307
```

```
str(actors.names)
```

```
## 'data.frame': 1048575 obs. of 6 variables:
## $ nconst : Factor w/ 1048575 levels "nm0000001","nm0000002",...: 1 2 3 4 5 6 7 8 9 10 ...
## $ primaryName : Factor w/ 983789 levels ":Leif Sundberg",...: 300867 543781 118867 458184 38499...
## $ birthYear : Factor w/ 373 levels "\\N","1130","1150",...: 254 279 289 304 273 270 254 279 2...
## $ deathYear : Factor w/ 328 levels "\\N","1031","1191",...: 292 319 1 287 312 287 262 309 289...
## $ primaryProfession: Factor w/ 11618 levels "","actor","actor,actress",...: 9835 1289 1309 638 11362...
## $ knownForTitles : Factor w/ 751903 levels "\\N","tt0000005",...: 119056 297351 74532 142592 66449...
```

```
# merge star actors file and actor names file by $nconst values ($nconst is the unique id for individual)
actors.info <- merge(x = actors.names, y = movies.staractor, by = "nconst", all = TRUE)
dim(actors.info)
```

```
## [1] 1888030 11
```

```
summary(actors.info)
```

```
##          nconst          primaryName          birthYear
## nm0000305: 893 Mel Blanc : 893 \\N :905791
## nm0784407: 762 Mack Sennett : 762 1925 : 12413
## nm0000428: 580 D.W. Griffith : 580 1908 : 12392
## nm0005658: 527 G.W. Bitzer : 527 1930 : 12266
## nm0281487: 495 Dave Fleischer: 495 1892 : 12093
## nm0293989: 466 (Other) :1831632 (Other):879934
## (Other) :1884307 NA's : 53141 NA's : 53141
##          deathYear          primaryProfession
## \\N :1204272 actor : 314565
## 1979 : 9889 actress : 228790
## 1971 : 9500 writer : 55299
## 1994 : 9500 : 52037
## 1989 : 9295 miscellaneous: 39173
## (Other): 592433 (Other) :1145025
## NA's : 53141 NA's : 53141
##          knownForTitles          tconst
## \\N : 61405 tt0000376: 10
## tt0094939,tt0045708,tt0096438,tt0077278: 893 tt0000439: 10
## tt0023040,tt0012701,tt0229411,tt0018951: 762 tt0000488: 10
## tt0004972,tt0006864,tt0009559,tt0010484: 580 tt0000557: 10
## tt0249710,tt0322496,tt0431889,tt0004972: 527 tt0000574: 10
## (Other) :1770722 (Other) :1048525
## NA's : 53141 NA's : 839455
##          ordering          category          job
## Min. : 1.0 actor :318678 \\N :845917
## 1st Qu.: 3.0 actress :182400 producer : 62973
## Median : 5.0 writer :163828 screenplay: 25643
## Mean : 5.2 director :114275 story : 18580
## 3rd Qu.: 8.0 cinematographer: 75592 writer : 11445
## Max. :10.0 (Other) :193802 (Other) : 84017
## NA's :839455 NA's :839455 NA's :839455
##          characters
## \\N :588471
## ["Himself"] : 6782
## ["Herself"] : 2815
## ["Narrator"] : 1748
```

```
## ["Himself - Host"]: 567
## (Other) :448192
## NA's :839455

# make new dataset with only relevant columns in movie names file, eliminate movies prior to 1970
movies <- movies[c(1:3, 6, 9:11)]
movies$startYear <- as.numeric(as.character(movies$startYear)) # convert the stateYear variable to a numeric

## Warning: NAs introduced by coercion

str(movies) # look to see that conversion was successful

## 'data.frame': 1028136 obs. of 7 variables:
## $ tconst : Factor w/ 1638909 levels "tt0000001","tt0000002",...: 1 2 3 4 5 6 7 8 9 10 ...
## $ titleType : Factor w/ 10 levels "movie","short",...: 2 2 2 2 2 2 2 2 1 2 ...
## $ primaryTitle : Factor w/ 741621 levels "_grau","- Estrés + Chill",...: 103269 364906 471182 693320 ...
## $ startYear : num 1894 1892 1892 1892 1893 ...
## $ genres : Factor w/ 1728 levels "\\N","Action",...: 1192 673 576 673 966 1709 1710 1192 1681 ...
## $ averageRating: num 5.6 6.1 6.5 6.2 6.1 5.2 5.4 5.4 5.4 6.9 ...
## $ numVotes : int 1586 192 1249 119 2003 109 628 1721 88 5679 ...

movies.sub <- subset(movies, startYear >= 1970)
dim(movies.sub)

## [1] 338628 7

# make another new dataset with only relevant columns in actor info file, keep only relevant columns, p
actors.info <- actors.info[c(1:4, 7, 9, 11)]
actors.info$birthYear <- as.numeric(as.character(actors.info$birthYear))

## Warning: NAs introduced by coercion

actors.info.sub <- subset(actors.info, birthYear >= 1930 & (category == "actor" | category == "actress"))
dim(actors.info.sub) # check to see if it worked

## [1] 176871 7

# merge sub-data sets
movies.by.actors <- merge(x = actors.info.sub, y = movies.sub, by = "tconst", all = TRUE)
movies.by.actors <- movies.by.actors[which(!is.na(movies.by.actors$primaryTitle) == TRUE),]
movies.by.actors <- movies.by.actors[which(!is.na(movies.by.actors$nconst) == TRUE),]
dim(movies.by.actors) # check to see if it worked

## [1] 140084 13

# filter out only Robert De Niro's filmography
robert.deniro.only <- filter(.data = movies.by.actors, primaryName == "Robert De Niro")
str(robert.deniro.only) #check to see if it's accurate

## 'data.frame': 41 obs. of 13 variables:
## $ tconst : Factor w/ 1638910 levels "tt0000001","tt0000002",...: 63966 67797 68387 69534 71989 ...
## $ nconst : Factor w/ 1089500 levels "nm0000001","nm0000002",...: 134 134 134 134 134 134 134 134 134 134 ...
## $ primaryName : Factor w/ 983789 levels ":Leif Sundberg",...: 793815 793815 793815 793815 793815 793815 793815 793815 793815 793815 ...
## $ birthYear : num 1943 1943 1943 1943 1943 ...
## $ deathYear : Factor w/ 328 levels "\\N","1031","1191",...: 1 1 1 1 1 1 1 1 1 1 ...
## $ category : Factor w/ 12 levels "actor","actress",...: 1 1 1 1 1 1 1 1 1 1 ...
## $ characters : Factor w/ 307659 levels "[\\\"?\\\"]","[\\\". (Ep.: Ausflug in den Wald)\\\"]",...: 149829 149829 149829 149829 149829 149829 149829 149829 149829 149829 ...
## $ titleType : Factor w/ 10 levels "movie","short",...: 1 1 1 1 1 1 1 1 1 1 ...
## $ primaryTitle : Factor w/ 741621 levels "_grau","- Estrés + Chill",...: 284038 65983 404286 614375 614375 614375 614375 614375 614375 614375 ...
## $ startYear : num 1970 1973 1973 1974 1976 ...
```

```
## $ genres      : Factor w/ 1728 levels "\\N","Action",...: 830 1324 1010 994 1247 1306 994 1272 1332
## $ averageRating: num  6.2 6.9 7.3 9 7.7 6.3 8.3 6.7 8.1 8.2 ...
## $ numVotes     : int  4001 5143 91957 1058265 21149 7544 675223 17197 294832 302581 ...

# single out the 1st listed genre category
robert.deniro.only$Genre <- gsub("^(.*?),.*", "\\1", robert.deniro.only$genres)
str(robert.deniro.only) # check to see if it worked

## 'data.frame': 41 obs. of 14 variables:
## $ tconst      : Factor w/ 1638910 levels "tt0000001","tt0000002",...: 63966 67797 68387 69534 71989
## $ nconst      : Factor w/ 1089500 levels "nm0000001","nm0000002",...: 134 134 134 134 134 134 134 134 134 134
## $ primaryName  : Factor w/ 983789 levels ":Leif Sundberg",...: 793815 793815 793815 793815 793815 793815 793815 793815 793815 793815
## $ birthYear    : num  1943 1943 1943 1943 1943 ...
## $ deathYear    : Factor w/ 328 levels "\\N","1031","1191",...: 1 1 1 1 1 1 1 1 1 1
## $ category     : Factor w/ 12 levels "actor","actress",...: 1 1 1 1 1 1 1 1 1 1
## $ characters   : Factor w/ 307659 levels "[\\?\\"]","[\\. (Ep.: Ausflug in den Wald)\\"]",...: 149829 149829 149829 149829 149829 149829 149829 149829 149829 149829
## $ titleType    : Factor w/ 10 levels "movie","short",...: 1 1 1 1 1 1 1 1 1 1
## $ primaryTitle : Factor w/ 741621 levels "_grau","- Estrés + Chill",...: 284038 65983 404286 614375 614375 614375 614375 614375 614375 614375
## $ startYear    : num  1970 1973 1973 1974 1976 ...
## $ genres       : Factor w/ 1728 levels "\\N","Action",...: 830 1324 1010 994 1247 1306 994 1272 1332
## $ averageRating: num  6.2 6.9 7.3 9 7.7 6.3 8.3 6.7 8.1 8.2 ...
## $ numVotes     : int  4001 5143 91957 1058265 21149 7544 675223 17197 294832 302581 ...
## $ Genre        : chr  "Comedy" "Drama" "Crime" "Crime" ...

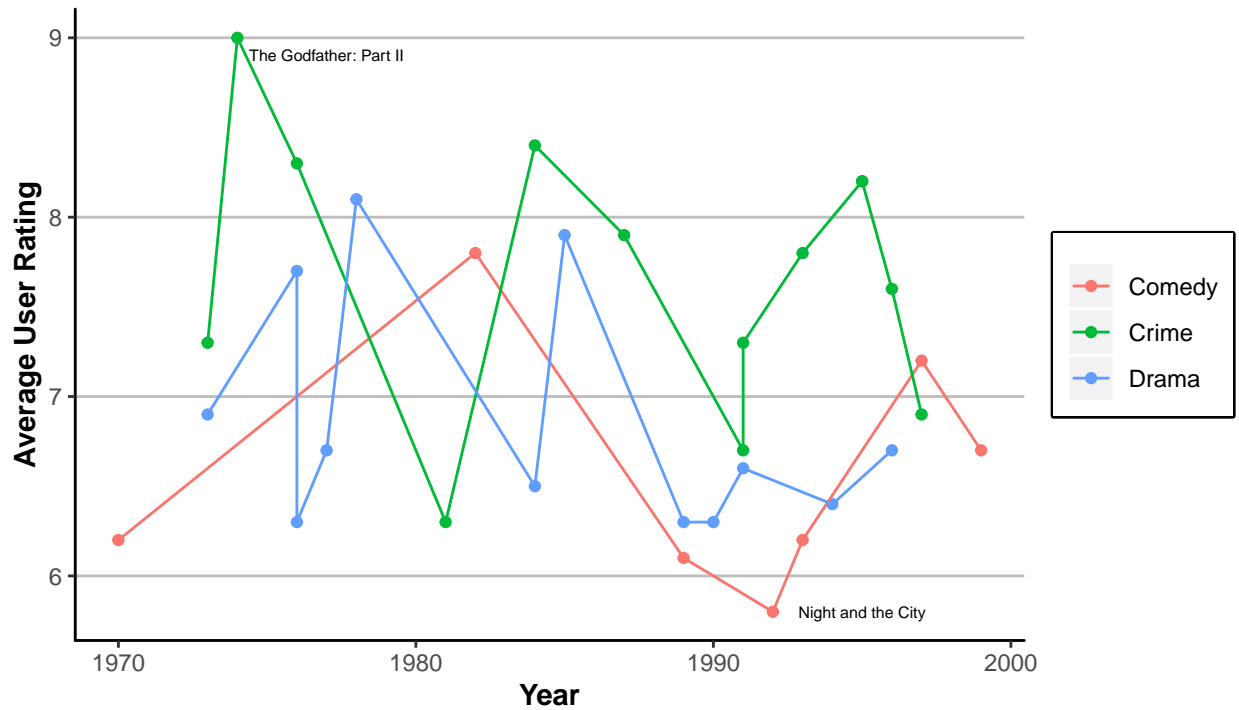
# subset filmography to take out genres where De Niro has not consistently acted in throughout the 30-year period
deniro.sub <- subset(robert.deniro.only, Genre != "Action" & Genre != "Adventure" & Genre != "Biography")

# plot graph, points and trace those points
deniro.chart <- ggplot(deniro.sub, aes(x = startYear, y = averageRating, color = Genre)) + geom_point() +
  geom_path() +
  labs(title = "IMDB User Ratings of Robert De Niro's Films 1970 - 2000",
        subtitle = "Ratings of De Niro's diverse portfolio of films in a 30-year period",
        x = "Year", y = "Average User Rating", caption = "Source: IMDB") +
  # add annotations to highest rated and lowest rated film
  annotate("text", x = 1977, y = 8.9, label = "The Godfather: Part II", size = 2) +
  annotate("text", x = 1995, y = 5.8, label = "Night and the City", size = 2) +
  # add themes to graph to make it aesthetically easy to look at, edit the gridlines & legend
  theme(panel.background = element_rect(fill = "white"),
        panel.grid.major.y = element_line(color = "grey"),
        panel.grid.major.x = element_blank(),
        legend.box.background = element_rect(color = "black", fill = NA, size = 1),
        legend.title = element_blank()) +
  # add titles, stylize typeface
  theme(axis.line.x = element_line(color = "black", size = 0.5),
        axis.line.y = element_line(color = "black", size = 0.5),
        axis.title.x = element_text(face = "bold"),
        axis.title.y = element_text(face = "bold"),
        plot.title = element_text(hjust = 0.5, face = "bold", size = 10),
        plot.subtitle = element_text(hjust = 0.5, face = "bold", size = 7.5),
        plot.caption = element_text(hjust = 0))

deniro.chart
```

### IMDB User Ratings of Robert De Niro's Films 1970 – 2000

Ratings of De Niro's diverse portfolio of films in a 30-year period are scattered across genres, suggesting star power is not a major influence on ratings



Source: IMDB