First Analysis

Jingwen Xu

1/30/2017

```
oscar <- read.csv("oscar.csv")
df <- subset(oscar, select = -c(X, X.1, X.2) )

# histogram
summary(df)
```

```
##                  Film           Year          Awards     Nominations
##  A Star Is Born    :   2   Min.   :1927   1      :898   1      :507
##  Cleopatra         :   2   1st Qu.:1950   2      :138   2      :129
##  Cyrano de Bergerac:   2   Median :1971   3      : 69   4      :104
##  Henry V           :   2   Mean   :1972   4      : 42   5      : 94
##  King Kong         :   2   3rd Qu.:1994   0 (1)  : 31   3      : 92
##  Little Women      :   2   Max.   :2015   5      : 25   7      : 71
##  (Other)           :1244                  (Other): 53   (Other):259
```

```
# cannot directly use change to numeric, b/c the "()" the resulting number would be weird
# df$Awards <-as.numeric(df$Awards)
# df$Nominations <- as.numeric(df$Nominations)
# View(df)
# hist(df$Awards, df$Nominations)

# regex tryout
# 1. remove the " ( )"
# a <- c("0 (1)", "11 (4)")
#a <- gsub("[] [!#$%()*,.:;<=>@^_`|~.{}].*[!#$%()*,.:;<=>@^_`|~.{}]", "", a)
# 2. remove the "[ ]"
# b <- c("10", "10[11]")
# b <- gsub("\\[.*\\]", "", b)


# delete the honoured awards, keep only the competitive awards
# honoured awards are in brackets
df$Awards <- gsub("[] [!#$%()*,.:;<=>@^_`|~.{}].*[!#$%()*,.:;<=>@^_`|~.{}]", "", df$Awards)
df$Nominations <- gsub("[] [!#$%()*,.:;<=>@^_`|~.{}].*[!#$%()*,.:;<=>@^_`|~.{}]", "", df$Nominations )
df$Nominations <- gsub("\\[.*\\]", "", df$Nominations)

# change to numeric
df$Awards <-as.numeric(df$Awards)
df$Nominations <- as.numeric(df$Nominations)

# remove the head of the long tail
df_nozero <- df[df$Awards != 0,]
# check plots again
table(df_nozero$Awards, df_nozero$Nominations)
```

```
##
```

```
##        1   2   3   4   5   6   7   8   9  10  11  12  13  14
##   1  504 108  77  75  59  30  28  10   6   4   3   2   0   0
##   2    0  15  13  19  25  21  18  12   6   5   4   1   0   0
##   3    0   0   3  10   6   7  15  16   3   5   2   2   1   0
##   4    0   0   0   1   3   5   8  10   7   4   1   3   1   0
##   5    0   0   0   0   1   1   2   3   4   7   5   1   2   0
##   6    0   0   0   0   0   0   0   2   2   2   1   1   2   1
##   7    0   0   0   0   0   0   0   2   0   5   1   2   1   0
##   8    0   0   0   0   0   0   0   0   0   2   2   2   2   0
##   9    0   0   0   0   0   0   0   0   2   0   0   1   0   0
##   10   0   0   0   0   0   0   0   0   0   0   1   0   0   0
##   11   0   0   0   0   0   0   0   0   0   0   1   1   0   1
```

```
# remove the head of the long tail
df_noone <- df_nozero[df_nozero$Awards != 1,]
# check plots again
table(df_noone$Awards, df_noone$Nominations)
```

```
##
##       2  3  4  5  6  7  8  9 10 11 12 13 14
##   2  15 13 19 25 21 18 12  6  5  4  1  0  0
##   3   0  3 10  6  7 15 16  3  5  2  2  1  0
##   4   0  0  1  3  5  8 10  7  4  1  3  1  0
##   5   0  0  0  1  1  2  3  4  7  5  1  2  0
##   6   0  0  0  0  0  0  2  2  2  1  1  2  1
##   7   0  0  0  0  0  0  2  0  5  1  2  1  0
##   8   0  0  0  0  0  0  0  0  2  2  2  2  0
##   9   0  0  0  0  0  0  0  2  0  0  1  0  0
##   10  0  0  0  0  0  0  0  0  0  1  0  0  0
##   11  0  0  0  0  0  0  0  0  0  1  1  0  1
```

```
fit1 <- glm(df_noone$Awards ~df_noone$Nominations, family = gaussian, data = df_noone)
awards_with_14_nomination  = 0.52549+14*0.398
# After my first analysis with linear regression, the predicted awards for La La Land (2016)
# who has 14 nominations, is predicted to have around 6 awards:)
```