R Notebook

Code

Hide

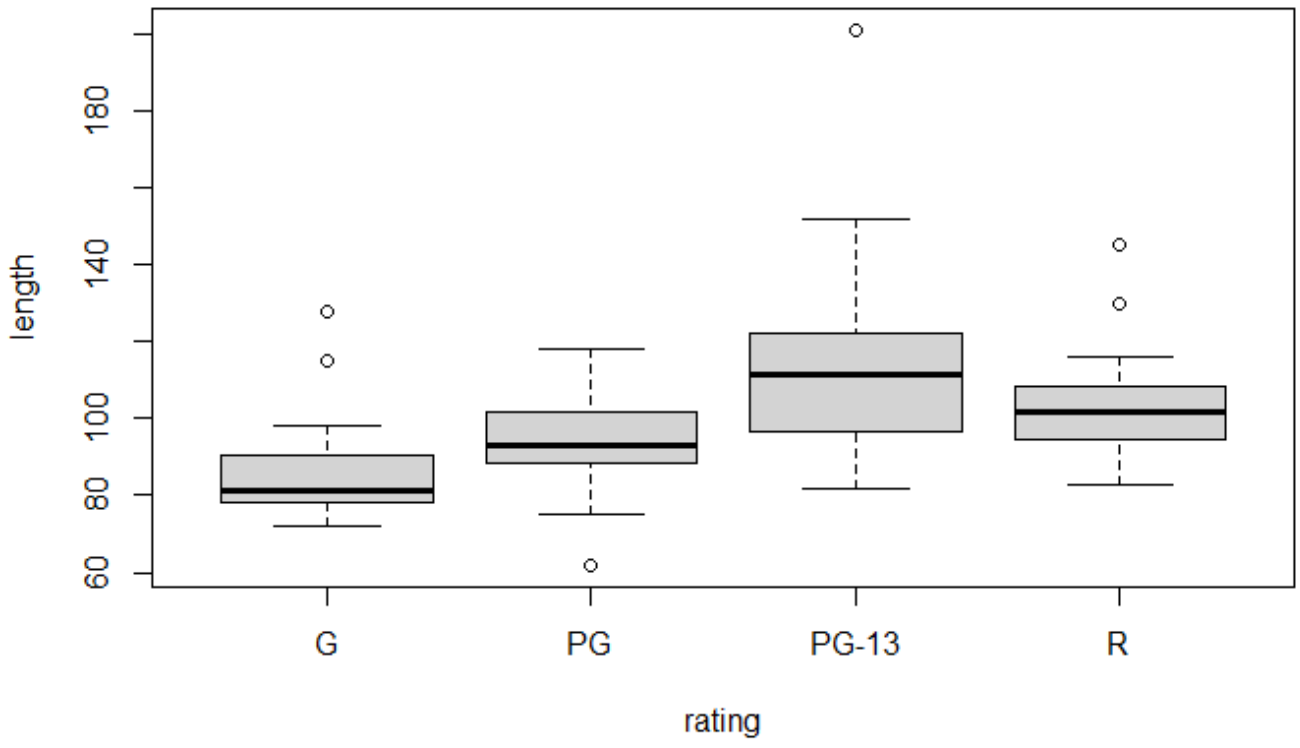
```
library(s20x)
Movies.df = read.table("movies.txt", header = TRUE)

Movies.df$rating = as.factor(Movies.df$rating)

Movies.df = Movies.df[-c(1)]
head(Movies.df,4)
```

	rating <fctr>	length <int>
1	G	115
2	G	72
3	G	78
4	G	89

4 rows

Hide

```
boxplot(length ~ rating, data = Movies.df)
```



All movie ratings seem to be uniformly distributed

PG - 13 movies have a much larger variance than the others, this may be due to the 1 extremely long movie

There is no obvious trend to be seen apart from maybe a slight increase in length as ratings increase. Though it is difficult to tell with PG-13 seemingly having the longer run time on average.

Hide

```
Movies.fit1 = lm(length ~ rating, data = Movies.df)
summary(Movies.fit1)
```

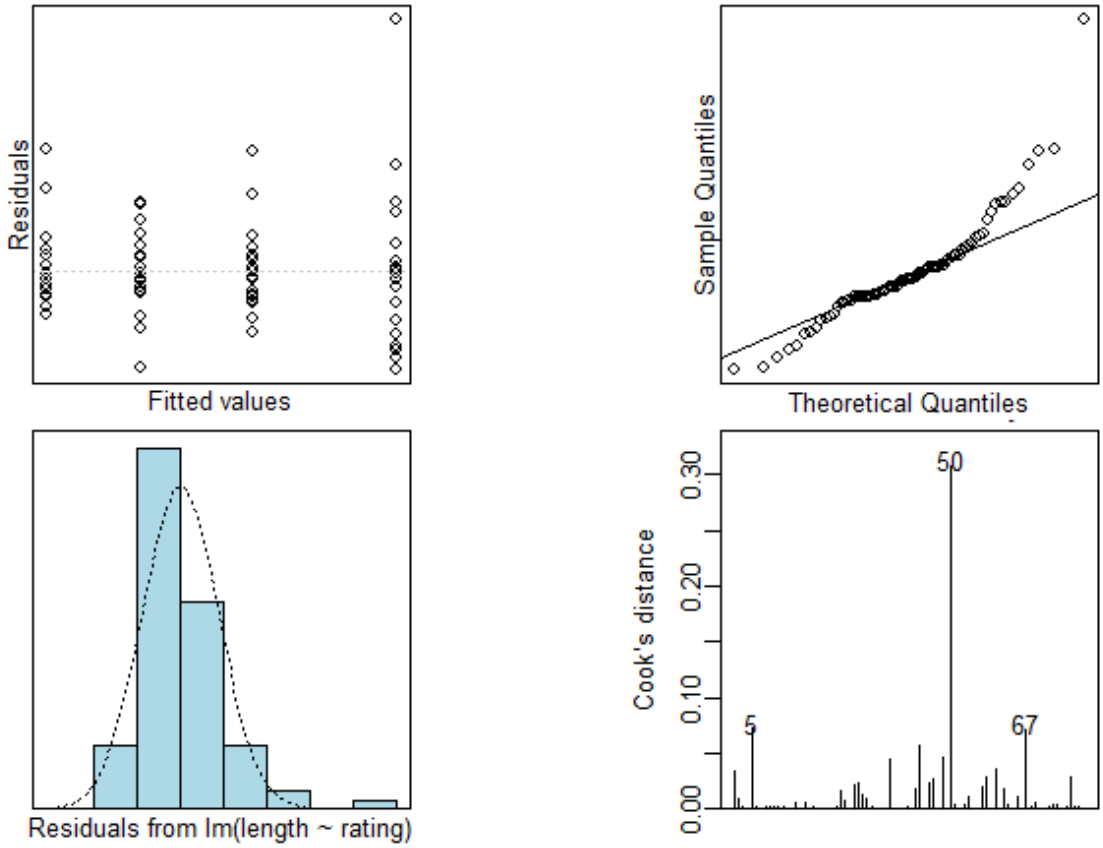
```

Call:
lm(formula = length ~ rating, data = Movies.df)

Residuals:
    Min      1Q  Median      3Q     Max 
-33.400  -8.600  -2.625   5.362  85.600 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   86.600      4.060  21.331  < 2e-16 ***
ratingPG       7.750      5.741   1.350  0.18108    
ratingPG-13   28.800      5.741   5.016 3.37e-06 ***
ratingR       17.050      5.741   2.970  0.00399 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18.16 on 76 degrees of freedom
Multiple R-squared:  0.2694,    Adjusted R-squared:  0.2406 
F-statistic: 9.343 on 3 and 76 DF,  p-value: 2.485e-05
```

With the model we fitted, we can estimate the means of each movie rating to be: G rated = 86.6 PG rated = 86.6 + 7.75 = 94.35 PG-13 rated = 86.6 + 28.8 = 115.4 R rated = 86.6 + 17.05 = 103.65

This only allows us to show us the difference between G rated and other rated groups

Hide

```
modcheck(Movies.fit1)
```

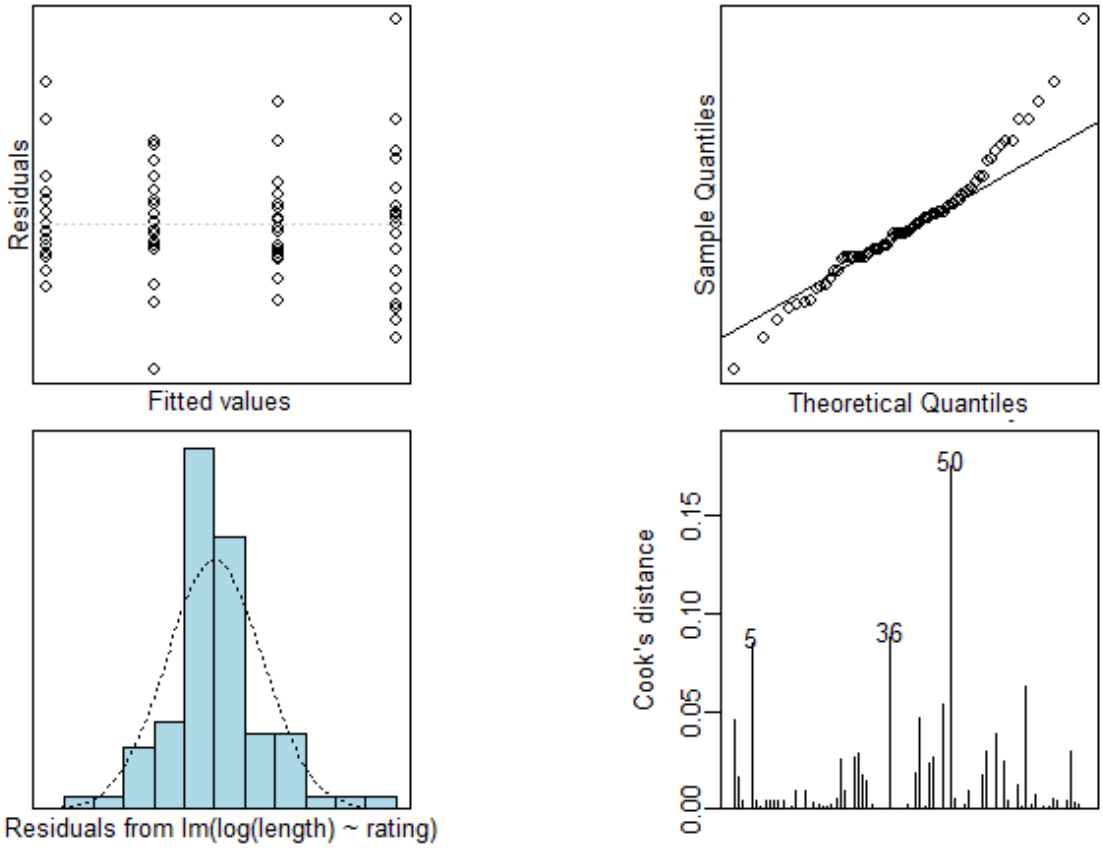


The scatter of points is not very consistent although most points do seem to crowd the zero residual mark. You can also see on the top right graph, the middle data quantiles match the normal quantiles but towards the edges, they deviate quite a lot.This is not good for a model. The histogram plot shows a slightly right skewed plot. The outlier in the Cooks plot is quite substantial but not large enough to alter the model in a substantial way.

As these plots are not up to quality for a model, we will log the data.

Hide

```
Movies.fit2 = lm(log(length) ~ rating, data = Movies.df)
modcheck(Movies.fit2)
```



The residual scatter spread with the logged value is much more even, with the outlier on the right having much less significance.

The normal and data quantiles appear to have been more tightly pushed towards the normal line, making for better data

The logged data also appears to be that of a normal distribution.Compared to the right skew before.

The cooks distance graph shows that the outlier has a lower value then before, and thus will have much less of an affect on the final model.

Hide

```
summary(Movies.fit2)
```

```
Call:
lm(formula = log(length) ~ rating, data = Movies.df)

Residuals:
     Min       1Q   Median       3Q      Max
-0.40944 -0.09059 -0.01877  0.06110  0.57765

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.45072    0.03630 122.603  < 2e-16 ***
ratingPG     0.08585    0.05134   1.672 0.098585 .
ratingPG-13  0.27493    0.05134   5.355 8.79e-07 ***
ratingR      0.18202    0.05134   3.545 0.000675 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1623 on 76 degrees of freedom
Multiple R-squared:  0.2976,    Adjusted R-squared:  0.2699
F-statistic: 10.73 on 3 and 76 DF,  p-value: 5.843e-06
```

Hide

```
anova(Movies.fit2)
```

```
Analysis of Variance Table

Response: log(length)
          Df  Sum Sq  Mean Sq F value    Pr(>F)
rating     3 0.84862 0.282875  10.732 5.843e-06 ***
Residuals 76 2.00312 0.026357
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Hide

```
multipleComp(Movies.fit2)
```

```
               Estimate Tukey.L Tukey.U Tukey.p
G  -  PG     -0.08585208 -0.2207  0.0490  0.3454
G  -  PG-13  -0.27493454 -0.4098 -0.1401  0.0000
G  -  R      -0.18202093 -0.3169 -0.0472  0.0037
PG  -  PG-13 -0.18908245 -0.3239 -0.0542  0.0024
PG  -  R     -0.09616884 -0.2310  0.0387  0.2482
PG-13  -  R   0.09291361 -0.0419  0.2278  0.2768
```

Hide

```
exp(multipleComp(Movies.fit2))
```

```
              Estimate   Tukey.L   Tukey.U  Tukey.p
G  -  PG     0.9177300 0.8019572 1.0502204 1.412555
G  -  PG-13  0.7596219 0.6637830 0.8692713 1.000000
G  -  R      0.8335839 0.7284036 0.9538966 1.003707
PG  -  PG-13 0.8277183 0.7233226 0.9472426 1.002403
PG  -  R     0.9083106 0.7937395 1.0394586 1.281716
PG-13  -  R  1.0973669 0.9589657 1.2558341 1.318903
```

Methods and Assumptions test:
We have movie run times for 4 different age ratings so we should use one-way ANOVA

We can assume that the movies are independent of each other. The equality of variance on the first model was not satisfied mainly due to the right skew of data. Due to this we transform the data using log. After the data has been logged, there is no longer any skew as it is uniformly distributed. There are no obvious outliers which will effect that data.

The model fitted is $log(Movies)_i$ = β_1 + $\beta_2 x_{2,i}$ + $\beta_3 x_{3,i}$ + $\beta_4 x_{4,i}$ + ε_i where $x_{2,i}$, $x_{3,i}$ and $x_{4,i}$ = 1 if the observation is from PG, PG-13 and R movies respecitvely and $\varepsilon_i \approx^{iid}$ N(0,σ^2), and 0 otherwise. The baseline has rating G.

The model can also be described as $log(Movies)_{ij}$ = μ + a_i + ε_{ij} where μ is the mean log length of movies, a_i is the effect of being at each age rating, and $\varepsilon_{ij} \approx^{iid}$ N(0,σ^2)

Executive Summary:
20 Movies with different age ratings had their run times recorded. We would like to determine if there is and differences in lengths of movies with relation to their age rating (G, PG, PG-13, R).

To perform the analysis we had to transform the data. This results in our movie run times with different age ratings relating to medians and being in multiplicative terms.

There is no evidence of a different run time for G and PG movies, PG and R movies, and PG-13 and R movies.

We discovered that the median run times of the PG-13 and R movies were longer then the G rated movies. We also found that the median run time of PG-13 movies were longer then PG movies.

We can estimate that the median run time of movies that are G rated are between 0.66 and 0.95 times the median run times of PG-13 and R rated movies.

Our model explains 30% of the variation in the run time of movies for each age rating.