Project Report: An Analysis of the Netflix Dataset

Sam Hawke

As the COVID-19 pandemic continues, Netflix increasingly becomes a part of modern life.

So what kind of content is Netflix producing? Do they have a bias in terms of which genres they make more available?

Additionally, certain shows on Netflix are more "successful" than other shows. In particular, the shows that get more seasons were renewed because Netflix executives thought they were worth it. Are there certain types of shows that get renewed more than others? Do the shows with many seasons have something in common?

This project will do descriptive statistics of the TV shows publicly available from the Netflix dataset found on Kaggle at this link: https://www.kaggle.com/shivamb/netflix-shows

We will also look at certain models to try to predict the nu mber of seasons for which a TV show runs based on other variables.

With the data on TV shows, we're interested in possibly predicting the number of seasons from the other variables. Some of these variables are highly unlikely to overfit. For example, title, director, cast, and description all likely uniquely identify the TV show. The useful predictor variables here appear to be country, release year, and listed in and/or rating. It might seem reasonable at first to fit a simple linear regression model from those 4 variables and see how well it predicts the number of seasons.

```
length(unique(tv$country)) * length(unique(tv$rating)) * length(unique(tv$listed_in))
```

```
## [1] 394320
```

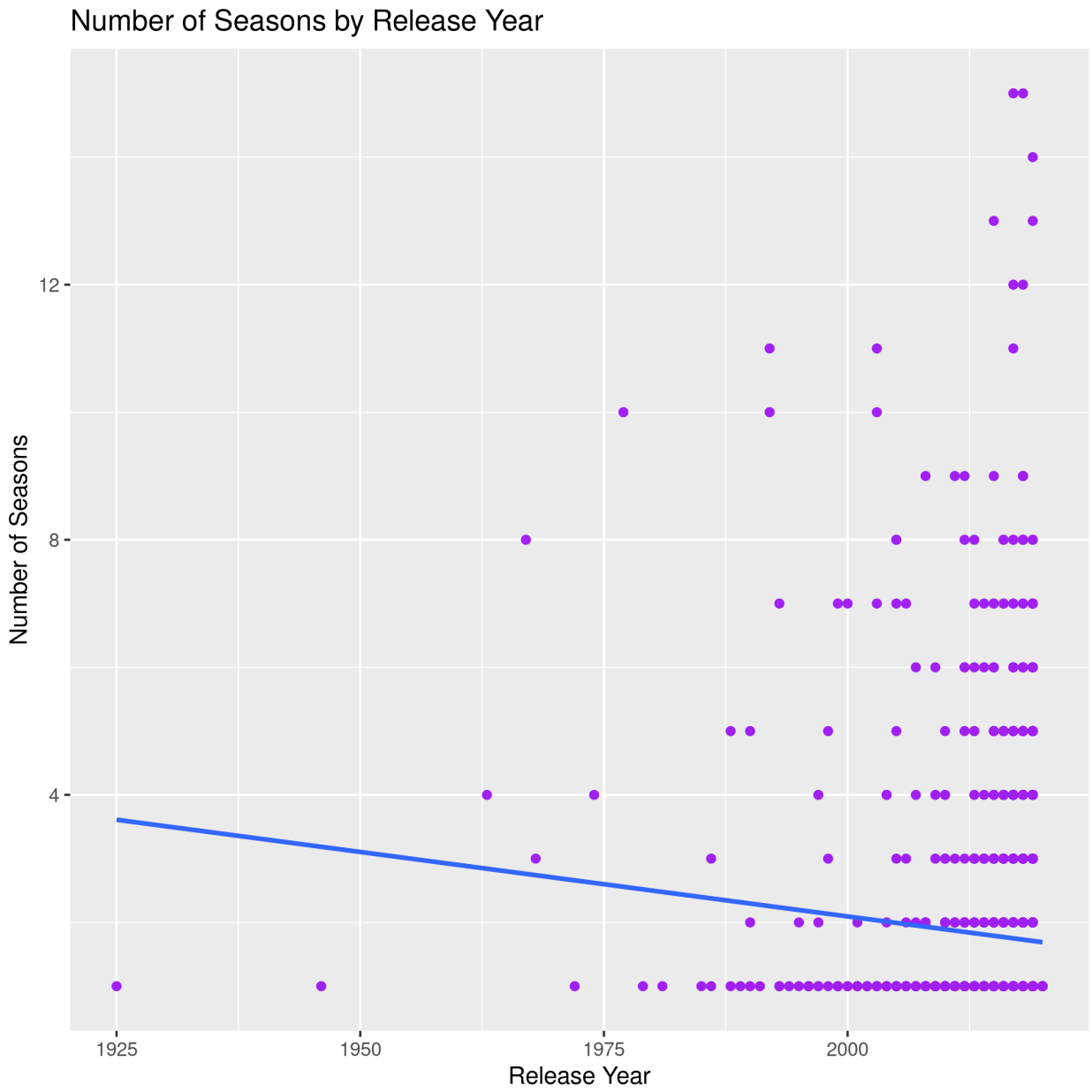
Unfortunately, this "simple" model estimates 394,320 parameters: one for each possible combination of country, rating, and category of TV show. So, let's try a much simpler model that takes only rating and release year into account:

```
##
## Call:
## lm(formula = num_seasons ~ release_year + rating, data = tv)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.2203 -0.8172 -0.6259  0.2314 13.2152
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    33.673085  13.133492   2.564   0.0104 *
## release_year   -0.016223   0.006496  -2.497   0.0126 *
## ratingG         0.032446   1.983136   0.016   0.9869
## ratingNR        1.830387   1.214523   1.507   0.1320
## ratingPG        1.064892   1.983264   0.537   0.5914
## ratingR         0.024334   1.619218   0.015   0.9880
## ratingTV-14     0.849672   1.146699   0.741   0.4588
## ratingTV-G      1.043008   1.161414   0.898   0.3693
## ratingTV-MA     0.674607   1.146811   0.588   0.5564
```

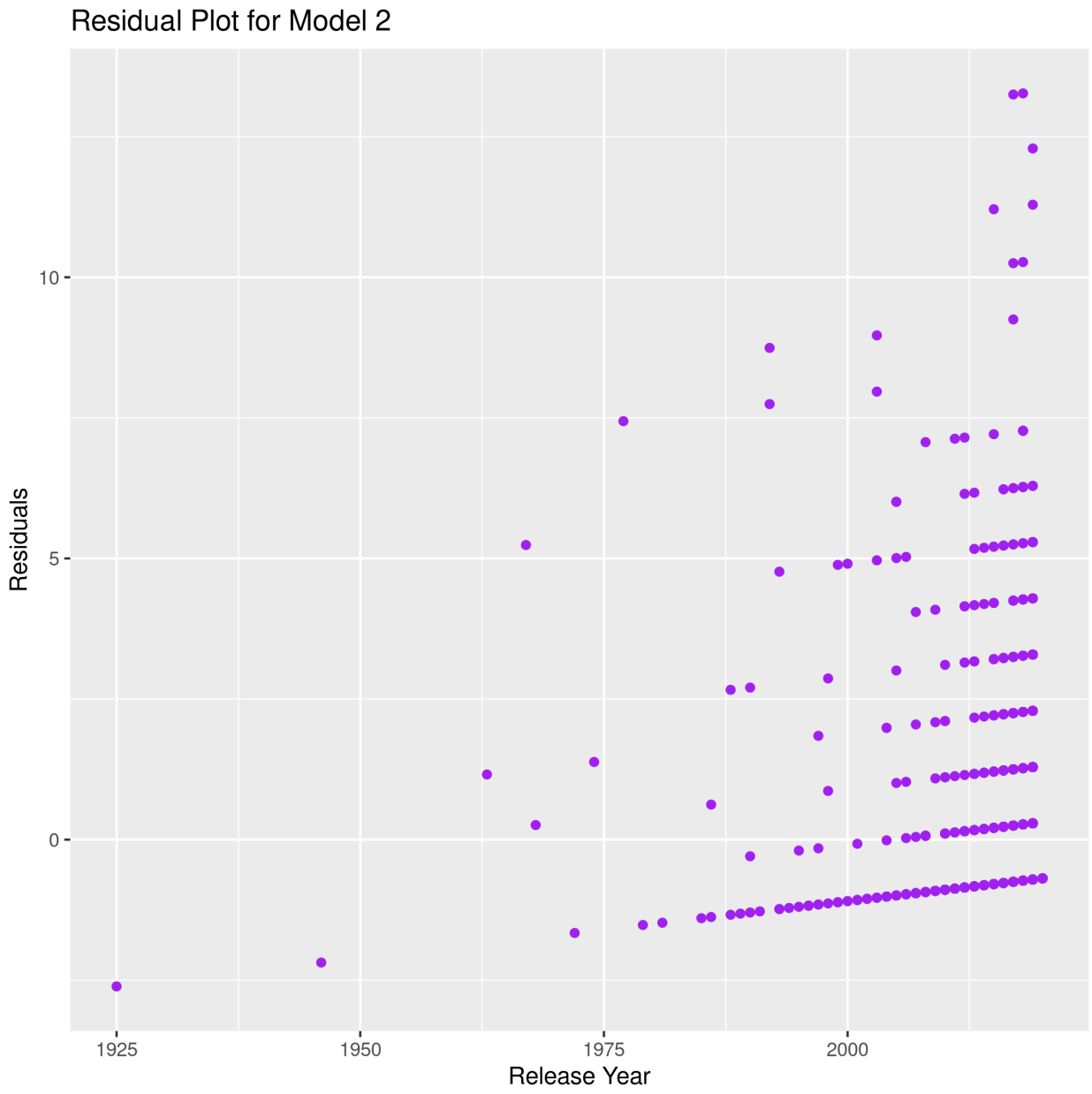
```
## ratingTV-PG     0.776426   1.149192   0.676   0.4994
## ratingTV-Y      0.910186   1.156164   0.787   0.4312
## ratingTV-Y7     0.946281   1.156380   0.818   0.4133
## ratingTV-Y7-FV  0.978051   1.161688   0.842   0.3999
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.619 on 1956 degrees of freedom
## Multiple R-squared:  0.01312,    Adjusted R-squared:  0.007061
## F-statistic: 2.166 on 12 and 1956 DF,  p-value: 0.01113
```

Interestingly, this model seems to suggest that rating is NOT a significant predictor for number of seasons. (It is important to note that, in this context, rating is taken to mean the specification of the appropriate audience, i.e. TV-14, TV-MA, etc.) Therefore, we will exclude it from the model, and instead examine by only release year.

```
##
## Call:
## lm(formula = num_seasons ~ release_year, data = tv)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6100 -0.7702 -0.7298  0.2702 13.2702
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  42.527776  12.699811   3.349 0.000827 ***
## release_year -0.020217   0.006301  -3.209 0.001355 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.621 on 1967 degrees of freedom
## Multiple R-squared:  0.005207,   Adjusted R-squared:  0.004701
## F-statistic: 10.29 on 1 and 1967 DF,  p-value: 0.001355
```



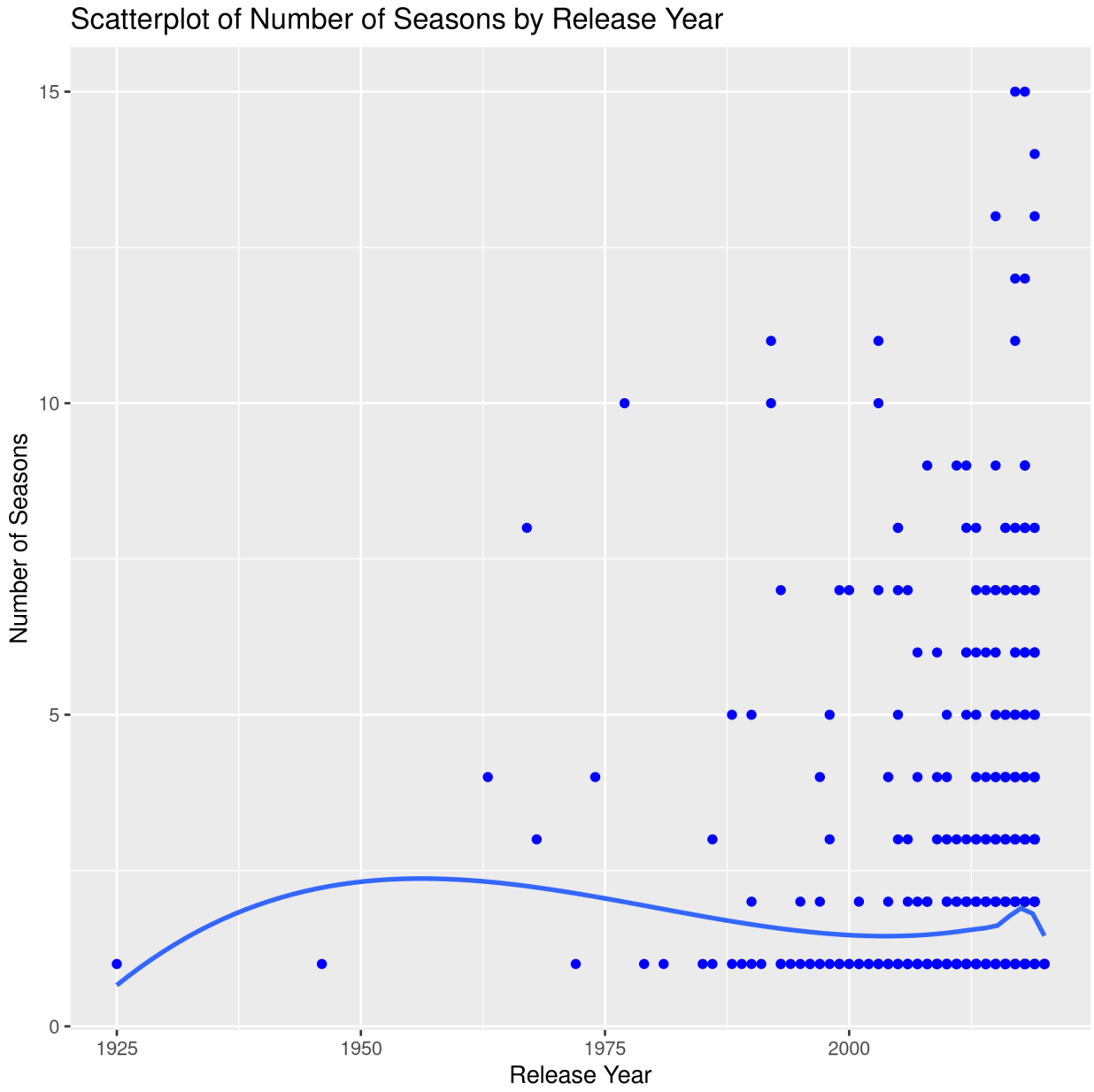
One first thing to notice is that most shows that have run for many seasons were released relatively recently. However, before interpreting any coefficients in the linear regression model, let's check to see whether the assumptions of the linear regression model are reasonable:



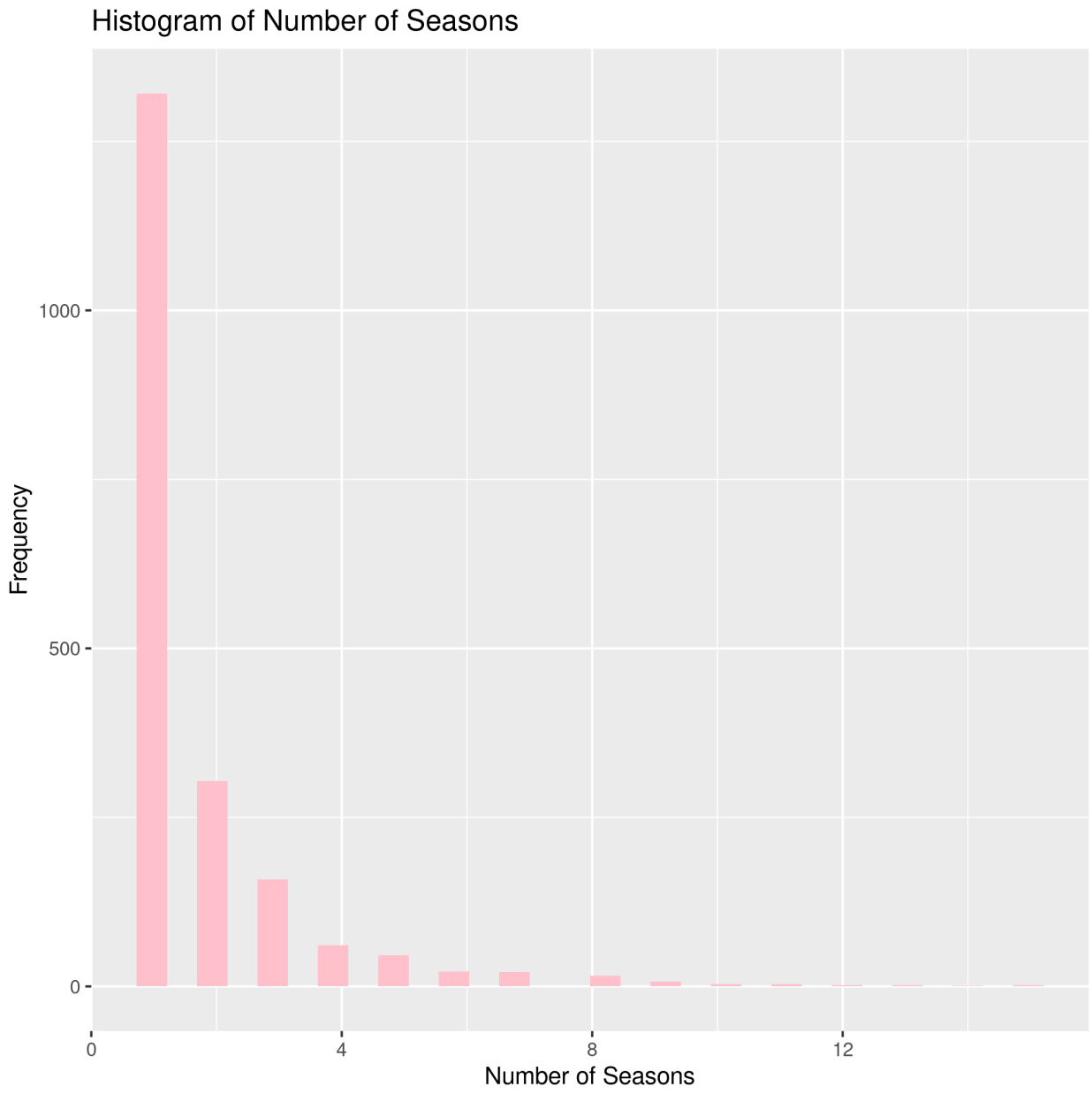
These data (clearly) do not satisfy the assumptions of linear regression that have to do with the error terms being normally distributed with mean zero and constant variance. Let's see if we can get a better fit by using a LOESS model.

```
## Call:
## loess(formula = num_seasons ~ release_year, data = tv)
##
## Number of Observations: 1969
## Equivalent Number of Parameters: 6.02
## Residual Standard Error: 1.627
## Trace of smoother matrix: 6.61  (exact)
##
## Control settings:
##   span     :  0.75
##   degree   :  2
```

```
##    family    :  gaussian
##    surface   :  interpolate      cell = 0.2
##    normalize:  TRUE
##  parametric:  FALSE
## drop.square:  FALSE
```

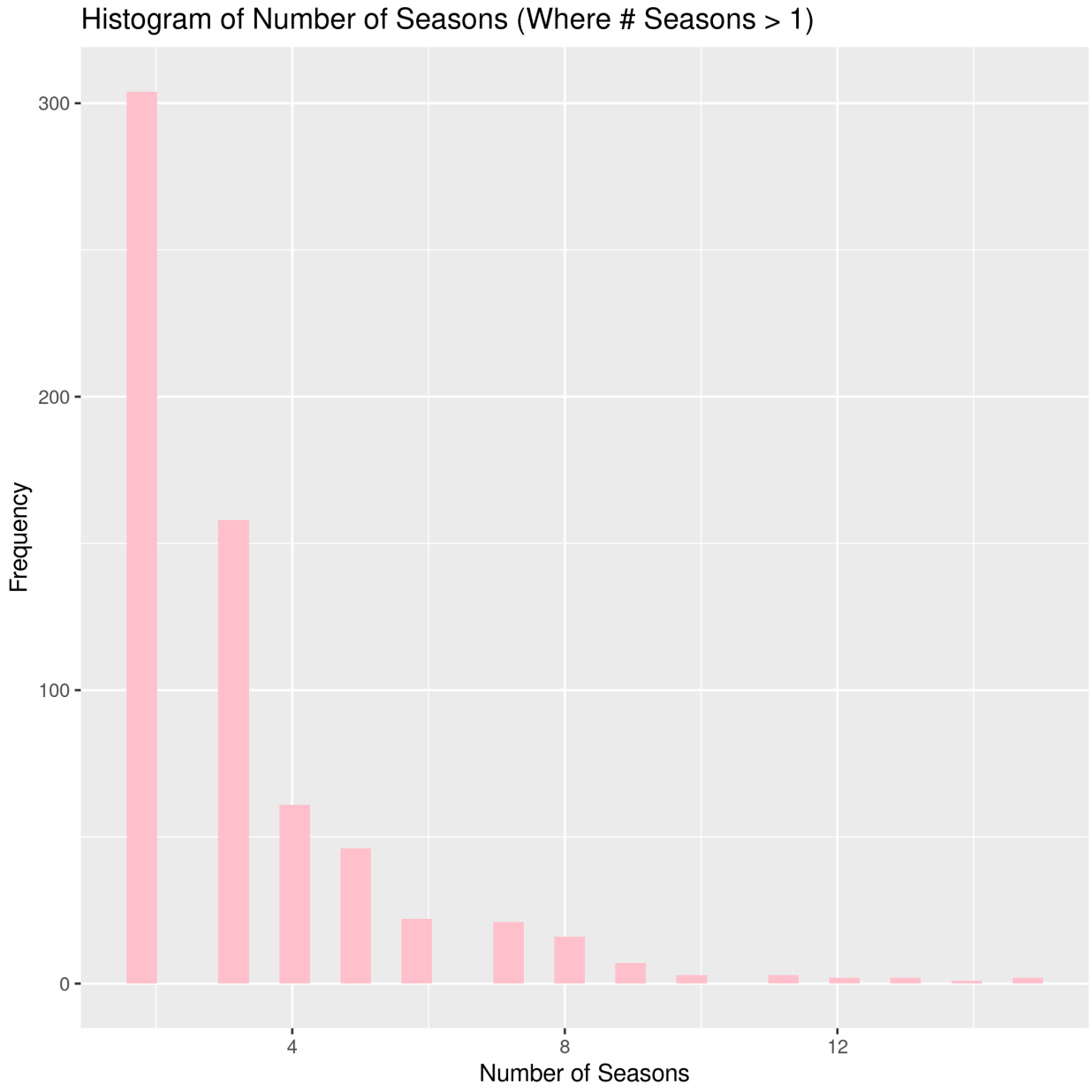


While this fit seems arguably somewhat more reasonable (but not much), the general trend appears to be completely dominated by a majority of TV shows with only 1 season. How many shows run for only 1 season?



It appears that the vast majority of shows on Netflix run for 1 season.

Further Explorations



While the conclusion found from these data was not super interesting, there are certain directions to explore which could be interesting. Firstly, we could turn it into a classification problem, where we try to predict a binary outcome: specifically whether or not a show runs for 1 season or more. Secondly, we could throw away the shows that ran for only 1 season and try to find a model to predict the number of seasons for which a show runs conditioned on the assumption that it runs for more than one season.

Thirdly, rather than simply using these given variables, we could try to find whether certain directors and/or actors account for more of the shows that run for longer, or whether those shows are more evenly distributed among different people. Finally, we could try to use some basic text mining to examine whether certain words appear in descriptions more than in others, and if certain words appear more often in longer-running shows than in others.