Data Wrangling Steps

Rebecca Lee

4/11/2018

Anime Data

There are two data sets that will be used in this project, one called anime that includes information on the anime itself, and one called rating, which includes information on the user and their rating of a particular anime.

No change was needed to the rating data, since the rating is automatically set at -1 if the user did not rate the anime (but did watch it) instead of an NA value. However there were quite a few changes required for the anime data, which I will outline below.

Step 1: Checking Column Names

I decided to change the column in anime from rating to avg_rating, since I think this is more descriptive of what the value actually is. All other column names did not need to be changed.

Step 2a: Checking for missing values

The first step I took was to check if there were any NA values in any of the columns for the anime data set. I did this by doing the following:

```
colSums(is.na(anime))
```

```
##   anime_id       name      genre       type   episodes avg_rating
##          0          0          0          0          0        230
##    members
##          0
```

This shows that there are 230 values in avg_rating that are NA. In order to keep this constant with the rating data set, I decided to change all the NA values in avg_rating to -1. Double checking this by using colSums showed that I succesfully changed those values.

Next I decided to check for empty values, since these are also values that do not make sense in any of the categories.

```
colSums(anime == '')
```

```
##   anime_id       name      genre       type   episodes avg_rating
##          0          0         62         25          0          0
##    members
##          0
```

This showed me that there were 25 empty values in the type column, and 62 in the genre column. I decided to tackle the type first.

After taking a look at just the anime that had an empty value for type it also shows that these anime were not rated (all had a value of -1). I then assumed that this meant the reason the type was unknown was because the anime was unreleased at the time, meaning it had some information on the website but it was not complete. Because of this, I decided to delete these 25 anime from the data set altogether, since it would not affect the future analysis.

```
anime <- subset(anime, type!='')
```

ADD WHAT TO DO ABOUT GENRE LATER

Step 2b: Checking that the values make sense

I took a look at the structure of each column for the anime data set, and found out that the episodes type was set as factor, instead of int. Upon further examination of the data, I realized there were some values of "Unknown" instead of a number. I decided to first change the data type to int, and then to figure out what to do with the resulting NA values.

The next step was more difficult, because I then had to decide what exactly to replace the values of NA with. Since there are different types and I assumed number of episodes would be affected by these types (e.g. a movie is usually only 1 episode, not more), I decided to check the mean, median, and mode of each of the types.

```
sapply(split(anime$episodes, anime$type), mean, na.rm = TRUE)
```

```
##               Movie     Music       ONA       OVA   Special        TV
##       NaN  1.102389  1.131417  6.877651  2.417663  2.561341 35.915595
```

```
sapply(split(anime$episodes, anime$type), median, na.rm = TRUE)
```

```
##           Movie   Music     ONA     OVA Special      TV
##      NA       1       1       2       2       1      24
```

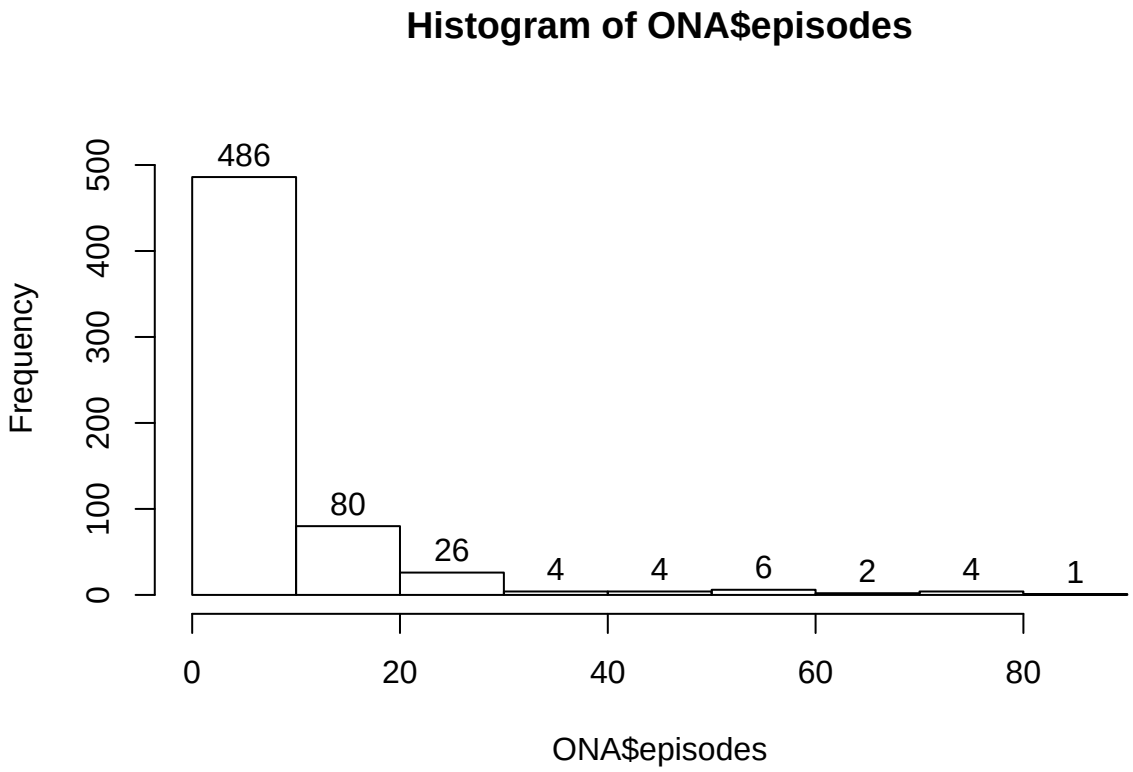
```
sapply(split(anime$episodes, anime$type), Mode)
```

```
##           Movie   Music     ONA     OVA Special      TV
##      NA       1       1       1       1       1      12
```

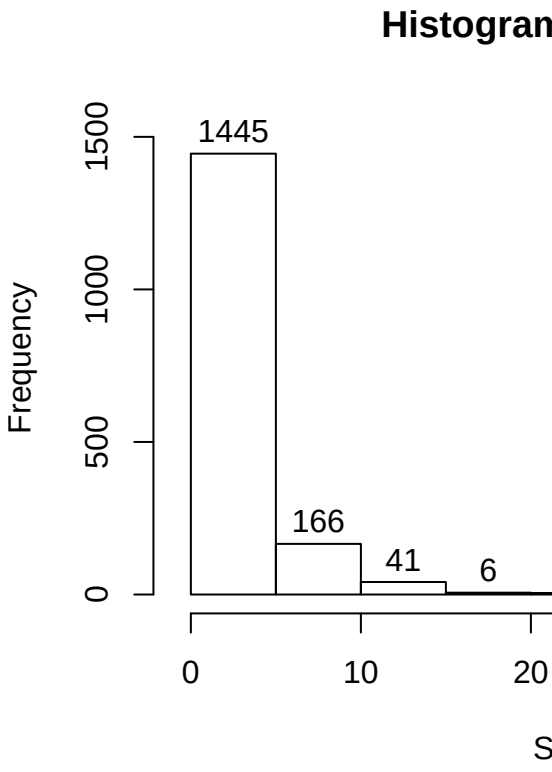
I decided that mode was not a good value to replace the NA with since they were almost all the same value, and varied more differently to the mean than the median does.

For both movie and music anime, I decided to change the value of NA to 1 since both mean and median were about 1. I did the same with OVA, except I used the value of 2.

To decide what to do with the ONA type anime, I decided to look at a histogram graph to help me decide between mean and median.



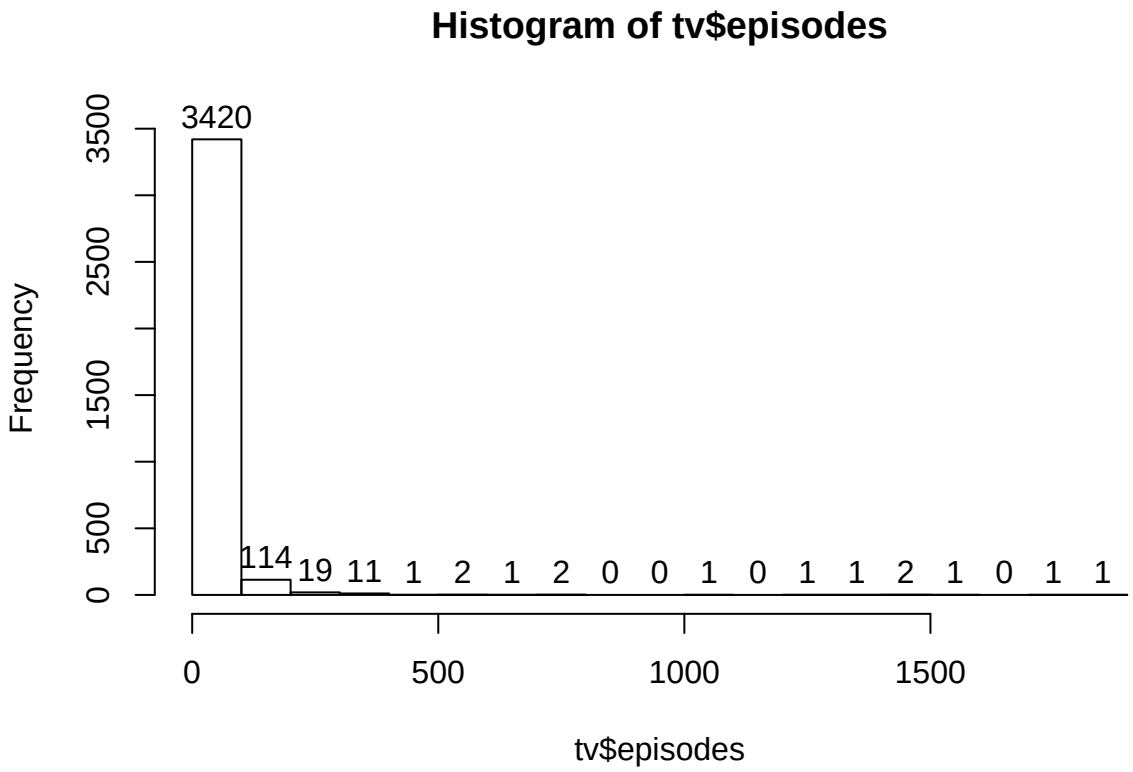
I decided to use the mean since there were no outliers, and rounded this up to 7.



Again to decide which value to use for Specials I used a histogram graph to help.

It looks like there are a couple of outliers, so I decided to use the median to replace the NA values, which is 2.

The same was done with TV, and the histogram showed some outliers as well so I chose to replace the NA with the median, a value of 24.


Histogram of tv$episodes
Frequency
0
500
1500
2500
3500
3420
114
19
11
1
2
1
2
0
0
1
0
1
1
2
1
0
1
1
0
500
1000
1500
tv$episodes