CMSE381 final report: Analysis on Avengers Data

Xingyu Yang

4/18/2021

Project Proposal

In the history of Marvel, hundreds of Superheros and Legends are told. However, the group of avengers diminished over time since many of them died throughout the story. Wondering what have caused the first death ('Death1' in the data), we want to fit a few models to see if any features have an impact on the death as well as to predict whether the avenger will die based on those features including their time of appearance, year they join and so on.

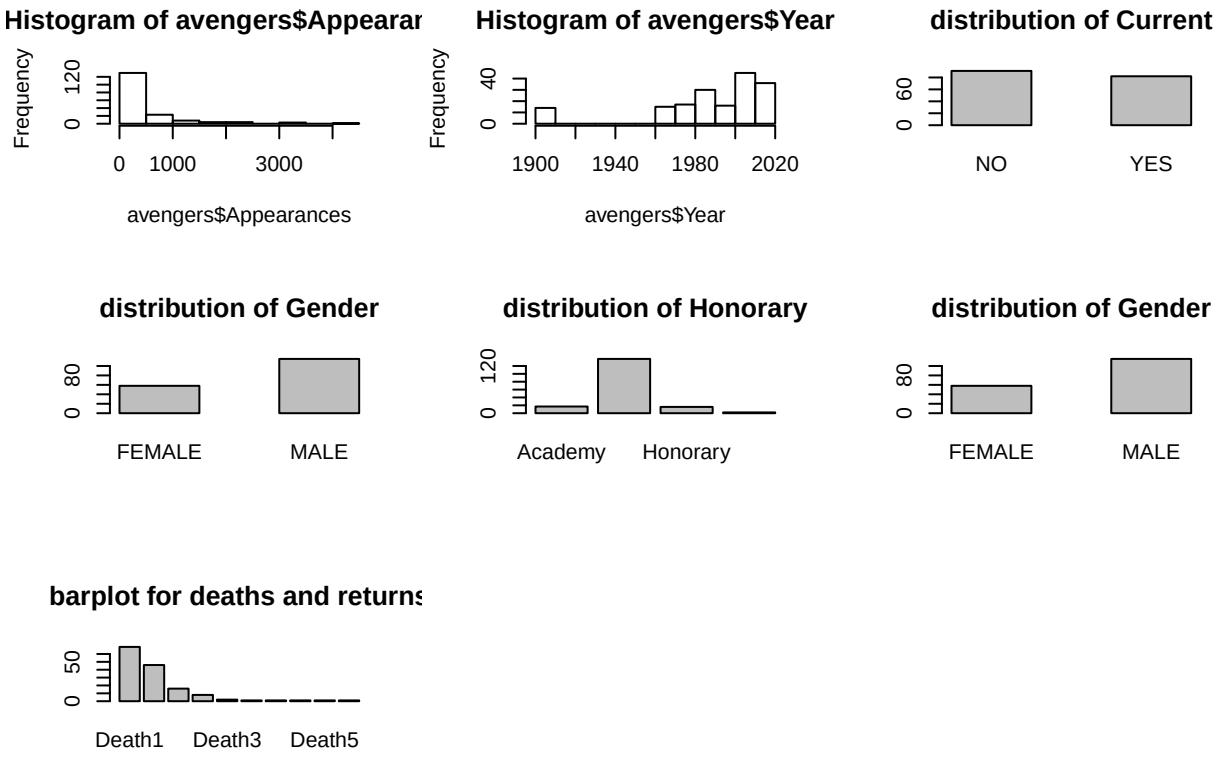
Take a look at the data

According to the structure of the data, it contains 173 rows and 21 variables. 18 of the variables are Factors and 3 of then are integers. According to the summary for the 3 integer variable, the statistics for them are shown in the below table.

```
##   Appearances          Year      Years.since.joining
##  Min.   :   2.0   Min.   :1900   Min.   :  0.00
##  1st Qu.:  58.0   1st Qu.:1979   1st Qu.:  5.00
##  Median : 132.0   Median :1996   Median : 19.00
##  Mean   : 414.1   Mean   :1988   Mean   : 26.55
##  3rd Qu.: 491.0   3rd Qu.:2010   3rd Qu.: 36.00
##  Max.   :4333.0   Max.   :2015   Max.   :115.00
```

Regenerate the Report

To compare with the results in the Article *Joining The Avengers Is As Deadly As Jumping Off A Four-Story Building*, I did some simple analysis on the Avenger Data. These histogram and boxplot below help to visualize general distributions for each features in the Data.



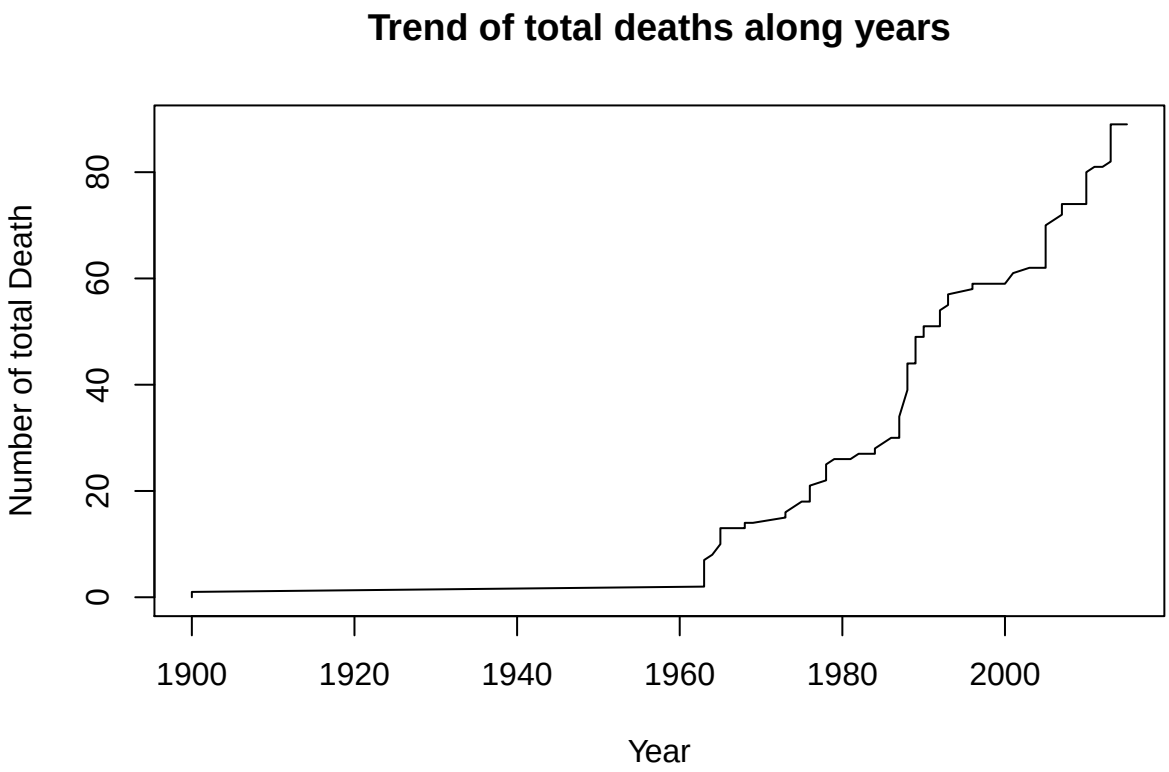
If we count the total number of death for all avengers, we can see that the percentage is around 40%.

If we do an analysis on the recovery rate, we get the statistic below: The total number of deaths is 89. The total number of returns is 57. The percentage of avengers recovered from a 2nd or 3rd death are both 50%.

Now let's promote our 'MVP of Death and Return' to the Avenger that Recovered five times: Jocasta!!

```
##                                          URL Name.Alias Appearances
## 33 http://marvel.wikia.com/Jocasta_(Earth-616)#    Jocasta         141
##    Current. Gender Probationary.Introl Full.Reserve.Avengers.Intro Year
## 33      YES FEMALE              Jul-80                      Nov-88 1988
##    Years.since.joining Honorary Death1 Return1 Death2 Return2 Death3
## 33                  27     Full    YES     YES    YES     YES    YES
##    Return3 Death4 Return4 Death5 Return5
## 33     YES    YES     YES    YES     YES
##
## 33 From her article: Death1: "Defeated Ultron and reversed the process leaving Jocasta a mindless hu
```

Finally, Along with Avengers' 53 years in operation, the number of total death haven't been changed much through year 1900 to 1960, however, it increases steeply from around 1960 all the way to the year 2015.



Designing new models

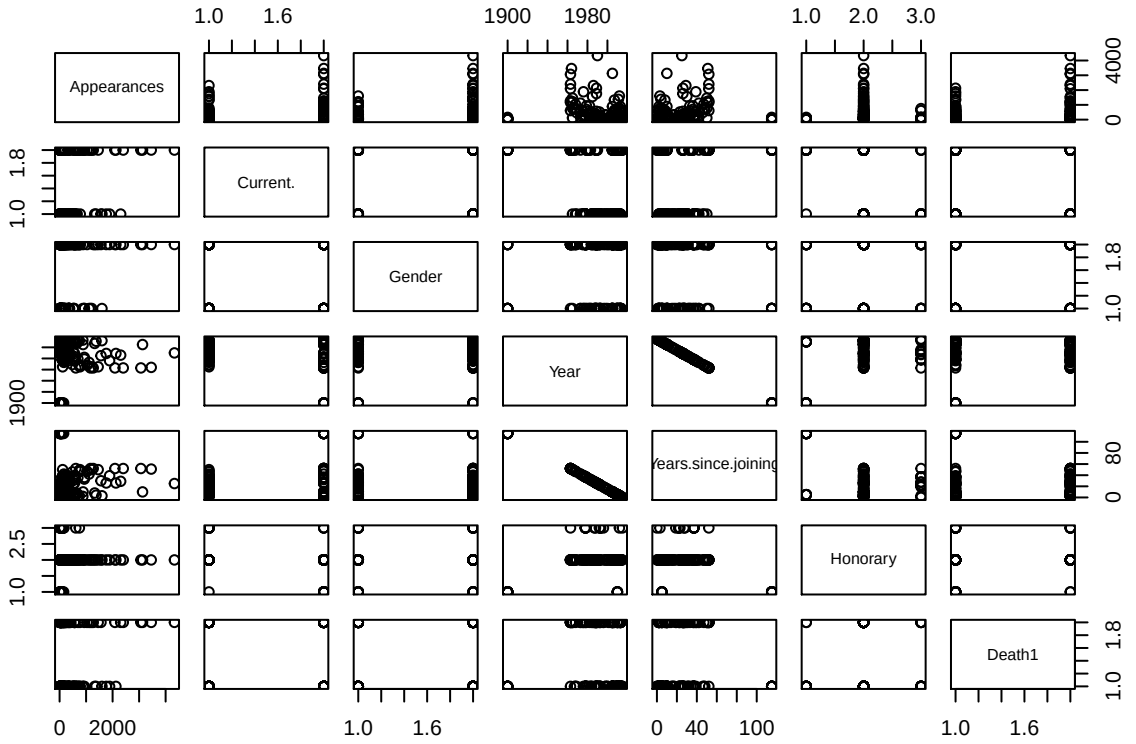
The paper only did some statistical analysis on the mortality rate as well as the recovery rate for Avengers, but did not state exactly what features are impacting the Death variable. Concerning whether a character will die whenever I watch the series of Avengers, instead of the theoretical reasons, how in fact the numeric numbers affect the death for each character? Thus, I would like to fit some models to analyze.

In order to fit a model for predicting Death1, we will use both logistic regression and K-nearest Neighbors to fit the model, applying cross-validation on both of them to tune the parameters, and calculate the test errors with randomly generated train/test datasets

Cleaning the data

Considering the difficulty for converting variables, I choose to exclude the variable 'URL', 'Name.Alias', 'Probationary.Introl', 'Full.Reserve.Avengers.Intro' and 'Notes'. Also, since there are many missing values for 'Return1', 'Death2', 'Return2', 'Death3','Return3', 'Death4', 'Return4', 'Death5' and 'Return5' and all these features are depended on our response variable, I choose to dropped these columns. Thus, we are using all other variables for modeling. I also droped row 76 and 77 since there are only to 'Probationary' in the Honorary and this might cause bias in our result.

Now let's make a pair plot to see the realtions between all variables.

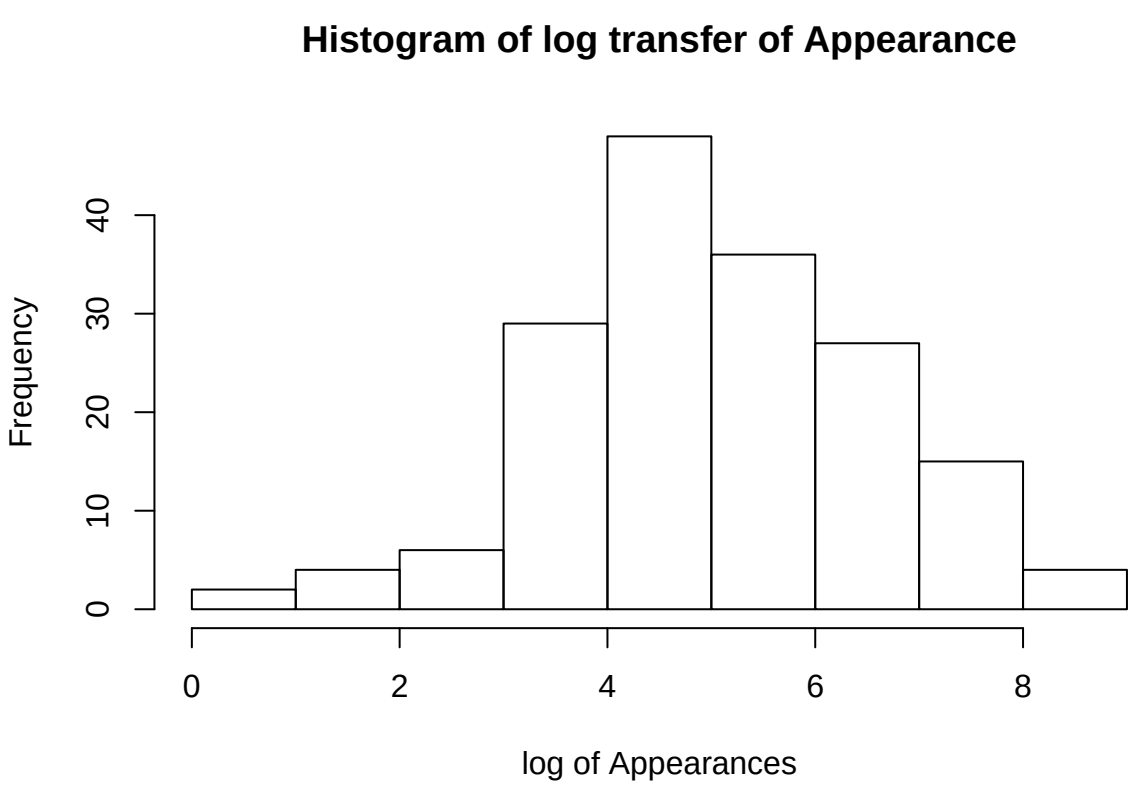


According to the paired scatterplot matrices, we can see that the variable 'Year' and 'Years.since.joining' are proportional to each other, therefore, I only keep 'Year' as a predicting feature to eliminate the repeat in predictor features. Then we split the data into a train set and a test set with drawing random samples.

Logistic regression

First, we want to fit logistic regressions with the formula $Death1 = \beta_0 + \beta_1 * Appearances + \beta_2 * Current. + \beta_3 * Gender + \beta_4 * Year + \beta_5 * Honorary$. We apply a LOOCV to the entire dataset to see the performance of this model. The error estimated is 0.4210526.

Then, noting the non-normal distribution for Appearances, I log transfer the Appearances into a more normaly distributed shape as shown in the graph below



However, the LOOCV error for the model $Death1 = \beta_0 + \beta_1 * log(Appearances) + \beta_2 * Current. + \beta_3 * Gender + \beta_4 * Year + \beta_5 * Honorary$ is 0.4327485 which is even slightly higher than the previous model.

Thus, I choose to use the first formula to fit a logistic regression on the training set, and got an Test Error of 0.3522727 based on the data. The table shows the first 5 rows of the predicted values along with the true results.

```
##   predicted.values true.values
## 1              YES         YES
## 2              YES         YES
## 5              YES         YES
## 6               NO          NO
## 8              YES         YES
```

```
## [1] 0.3522727
```

KNN method

Now we perform KNN method on model fitting with transfering each factor columns into type 'numeric'. Current: 0-NO, 1-YES, Gender: 0-MALE, 1-FEMALE, Honorary: 0-Academy, 1-Honorary, 2-Full, Death1: 0-NO, 1-YES.

To tune the best value of K, I perform a LOOCV again with KNN. Iterating through K=2,5,10,15,20,25,30. The result shows that K=15 has the lowest LOOCV error and therefore we will use this value to do train test fit.

```
##   k.values LOOCV.error
```

```
## 1        2   0.4698795
## 2        5   0.5060241
## 3       10   0.4216867
## 4       15   0.3975904
## 5       20   0.3975904
## 6       25   0.4096386
## 7       30   0.4096386
```

The Test Error for K-nearest neighbors method is 0.3636364, comparing to the Test Error of logistic regression, logistic regression performs better than KNN.

Discussion and conclusions

After fitting the three models on predicting 'Death1', the error rate for all the models are around 0.36. Obviously, these models do not perform perfectly since the errors are relavantly large comparing to rigorous data processing. And I came up with three reasons that would cause this:

1. Just like what is stated in the article, 'the only thing that can truly kill an Avenger is Marvel Studios not having the rights to the character or the performer portraying them'. So the Death of a character does not depend on any features at all.
2. The number of set of data is pretty small so we don't have enough information to fit a best model.
3. There are many more features for each character that can impact their death in a more significant way, but are not contained in this dataset.

In conclusion, although our models are not perfect, the test error 0.36 is still smaller than 50%. Fortunately, we still made some progress and are able to predict most of the 'First Death' with a logistic regression and prove that the variable 'Death1' is indeed someway impacted by the features.